

Project**:** H2020-ICT-2014-2 5G NORMA

Project Name:

5G Novel Radio Multiservice adaptive network Architecture (5G NORMA)

Deliverable D3.2
5G NORMA network architecture – Intermediate report

Date of delivery:	31/01/2017	Version:	1.0
Start date of Project:	01/07/2015	Duration:	30 months

Document Properties

Document Number:	H2020-ICT-2014-2 5G NORMA/D3.2
Document Title:	5G NORMA network architecture – intermediate report
Authors:	Ignacio Labrador Pavon, Jorge Rivas Sánchez (Atos) , Alessandro Colazzo, Riccardo Ferrari (Azcom), Paul Arnold, Markus Breitbach, Heinz Droste, Dirk von Hugo (Deutsche Telekom), Vasilis Friderikos, Stan Wong (King's College London), Vincenzo Sciancalepore, Zarrar Yousaf (NEC), Marie Line Alberi-Morel, Mark Doll, Borislava Gajic, Sylvaine Kerboeuf, Christian Mannweiler, Diomidis Michalopoulos, Peter Rost, Bessem Sayadi, Peter Schneider, Dimitrios Schinianakis, Yu Ling (Nokia), Charles Chambers, Hassan Osman (Real Wireless), Ignacio Berberana, Rafael Cantó (Telefonica), Marcos Rates Crippa, Bin Han, Shreya Tayade (TU Kaiserslautern), Dario Bega, Albert Banchs (Universidad Carlos III de Madrid)
Editor(s):	Markus Breitbach, Heinz Droste (DT)
Contractual Date of Delivery:	31/01/2017
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	5G NORMA D3.2_v1.0.docx

Revision History

Revision	Date	Issued by	Description
1.0	31.01.17	5G NORMA WP3	Final version

Abstract

This second deliverable of WP3 describes the status of the 5G NORMA architecture after the second design iteration. It integrates the control and data layer functions developed in WPs 4 and 5 into a harmonized mobile network architecture, applying the paradigms of adaptive (de-)composition and allocation of network functions, programmable network control, and end-to-end network slicing. The deliverable depicts the design of a multi-service management & orchestration layer with dedicated interfaces for mobile network tenants. A 5G ecosystem analysis reveals the technical interaction between important stakeholders, and, using so-called offer types, the required infrastructure assets. The security threats arising in virtualized multi-tenant networks are discussed and novel security solutions are presented. The architecture verification applies a methodology with three evaluation cases and the generic 5G services to analyse the fulfilment of different requirement categories as defined in D2.1.

Keywords

Mobile network architecture & design principles; stakeholder roles & business relationships; network slicing; control and data layer split; SDM-based network management, orchestration and control; security; architecture verification

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

Executive Summary

This document reports on the intermediate results of the 5G NORMA mobile network architecture design after having completed the second design iteration. Latest results of RAN architecture components (WP 4, D4.1) and connectivity and QoE/QoS management mechanisms (WP 5, D5.1) are integrated into the overall architecture concept of WP3, considering the general design principles of the project. Thus, this document presents the first harmonised and integrated 5G NORMA mobile network architecture, including the control and data layer functions and interfaces of radio access and core network. Furthermore, the complete functional outline of the Management & Orchestration layer and the interaction with third party systems, e.g., from vertical sectors, is described.

As a starting point, this deliverable elaborates on the drivers, enablers, and challenges in designing the 5G mobile network architecture. Industry expectations regarding multi-tenant and multi-service networks motivate the need to design a network architecture based on the principles of network programmability, adaptive (de)composition and allocation of network functions, as well as network slicing. Resulting challenges include, among others, service-aware orchestration and network control, efficient resource sharing approaches, and the management and control of slicing in multi-technology, heterogeneous radio access networks.

In order to get an understanding of the overall 5G technical ecosystem and the security implications, this deliverable provides a thorough analysis of the stakeholders that are involved in 5G networks deployment and operations. Three fundamental players are identified: the infrastructure provider, the mobile service provider (MSP), and the mobile network tenant. Based on an inventory of the available and expected 5G network infrastructure, three typical offer types of MSPs for tenants are derived. The technical and security implications of these offer types are outlined and evaluated, particularly focusing on the varying involvement of the tenant in operating a network slice.

A major part of this document covers the end-to-end architecture design. Introducing a top-down approach, the high-level functional perspective and three-non-functional views illustrate how the design goals of 5G NORMA are achieved. By means of combining virtualisation, multiplexing, and multitasking as baseline technologies for resource sharing, network slicing with both physical and virtualized network functions slicing is realised in an end-to-end fashion. The novel control and data layer design enables both software-defined network control, in particular programmable mobility management and QoS/QoE control. Moreover, the Management & Orchestration layer extends the ETSI NFV MANO framework, accommodating network slice lifecycle management in a multi-tenant environment.

Security aspects are a major concern in shared network infrastructures. 5G NORMA tackles the specific security threads arising in NFV-based multi-service and software-programmable networks. Virtualised Authentication, Authorisation, and Accounting (V-AAA) functions, role-based VM introspection, a new access stratum security architecture and local Trust Zones are among the innovative concepts integrated into the overall architecture.

The deliverable finally includes the results of the architecture design verification after completion of the second iteration. The verification methodology describes the major objectives, verification tools, as well as the three selected evaluation cases: the baseline (e)MBB case, a multi-tenant case, and a multi-service case. Intermediate evaluations indicate that the current architecture design already fulfils a significant number of performance, functional, operational, and security requirements that have been defined at the beginning of the project. Remaining gaps will have to be closed in the third design iteration.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3GPP	Third Generation Partnership Project
5G NORMA	5G Novel Radio Multiservice Network Adaptive Architecture
AAA	Authentication, Authorization and Accounting
A-CPI	Application-Controller Plane Interface
AI	Air Interface
AKA	Authentication and Key Agreement
ANN	Artificial Neural Network
AP	Access Point
API	Application Programming Interface
AS	Access Stratum
ASIC	Application-specific integrated circuit
BB	Building Block
BS	Base Station
BSS	Business Support System
CAPEX	Capital Expenditure
CC	Central Cloud
CCCM	Central Cloud Connection Monitoring
CME	Central Management Entity
cMTC	Critical Machine Type Communication
CMTS (HFC)	Cable Modem Termination System (Hybrid Fiber Coax)
CNE	Core Network Element
CoMP	Cooperative Multipoint
CPU	Central Processing Unit
CQI	Channel Quality Indicator
C-RAN	Centralised RAN
D2D	Device-to-Device
D-CPI	Data-Controller Plane Interface
DoS	Denial of Service
DPI	Deep Packet Inspection
DPI	Deep Packet Inspection
DSP	Digital Signal Processor
E2E	End to End
EC	Edge Cloud
EC	European Commission
EDA	Event-Driven software Architecture
EJB	Enterprise JavaBeans
EM	Element Management (Element Manager)
eMBB	Evolved Mobile Broadband
eNB	Enhanced Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
ES	Emergency Services
ETSI	European Telecommunications Standards Institute
FBMC	Filter-Bank MultiCarrier
FCAPS	Fault, Configuration, Accounting, Performance, Security
FE	Functional Element
FFT	Fast Fourier Transform
FPGA	Field Programmable Gate Array
G.FAST	Fast Access to Subscriber Terminals
GDB	Geo-location data base
GPP	General Purpose Processing
GUI	Graphical User Interface
GWCN	Gateway Core Network
H2020	Horizon 2020
H2020	Horizon 2020
HSS	Home Subscriber Server

HT	Horizontal Topic
HW	Hardware
IaaS	Infrastructure as a Service
ICT	Information and Communication Technologies
IE	Information Element
IFFT	Inverse Fast Fourier Transform
IMS	IP-based Multimedia Services
iNC	iJOIN Network Controller
InP	Infrastructure Provider
IoT	Internet of Things
iRPU	iJOIN Radio Processing Unit
iSC	iJOIN Small Cell
KPI	Key Performance Indicator
LAA	Local Access Assistant
LDAP	Lightweight Directory Access Protocol
LSS	Local Subscriber Server
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution-Advanced
M2M	Machine-to-Machine
MAC	Medium Access Control
MANO	Management and Orchestration
MANO-F	Management and Orchestration Function
MBB	Massive Broadband
MCS	Modulation and Coding Scheme
MDD	Multi-Dimensional Descriptor
MM	Mobility Management
MME	Mobility Management Entity
mMTC	Massive Machine-Type Communication
MNO	Mobile Network Operator
MOCN	Multi Operator Core Network
MOS	Mean Opinion Score
MSC	Message Sequence Chart
MTC	Machine Type Communication
NaaS	Network as a Service
NAS	Non-Access Stratum
NEM	Network Element Manager
NF	Network Function
NF-FG	Network Functions Forwarding Graph
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
NFVO	Network Function Virtualization Orchestrator
NGMN	Next Generation Mobile Networks Alliance
NRT	Neighbor Relation Table
NSSF	Network Slice Selection Function
OAM	Operation And Maintenance
ONF	Open Network Foundation
OPEX	Operational Expenditure
OS	Operating System
OSS	Operation Support System
OTT	Over The Top
P Router	Provider Router
PA	Power Amplifier
PCRF	Policy and Charging Rules Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Packet Data Unit
PE Router	Provider Edge Router
P-GW	Packet Gateway
PHY	Physical Layer

PLMN	Public Land Mobile Network
PNF	Physical Network Function
PoC	Proof of Concept
PoP	Point-of-Presence
QoE	Quality of Experience
QoS	Quality of Service
RAM	Random Access Memory
RAN	Radio Access Network
RANaaS	RAN as a Service
RAT	Radio Access Technology
RBAC	Role-based Access Control
RF	radio frequency
RG	Requirement Group
RISC	Reduced Instruction Set Computer
RLC	Radio Link Control
RLC-AM	Radio Link Control, Acknowledged Mode
RLC-UM	Radio Link Control, Unacknowledged Mode
RNE	Radio Network Element
RNM	Radio Node Management
ROADM	Reconfigurable Optical Add-Drop Multiplexer
RRC	Radio Resource Control
RRH	Remote Radio Head
RTT	Round Trip Time
SA	Security Auditing
SBI	Southbound Interface
SCTP	Stream-Control Transport Protocol
SDMC	Software Defined Mobile Network Control
SDM-C	Software-Defined Mobile network Controller
SDMC+O	Software Defined Mobile Network Control and Orchestration
SDM-O	SDM Orchestrator
SDM-X	SDM Coordinator
SDN	Software Defined Networking
SFC	Service-Function Chain / Service Function Chaining
S-GW	Serving Gateway
SLA	Service-Level Agreement
SoC	System on Chip
SON	Self-Organising Networks
SPTP	Small Packet Transmit Procedure
SW	Software
TeC	Technology Components
TP	Transmission Point
TZ	Trust Zone
UCA	User-centric Connection Area
UDN	Ultra-Dense Network
UE	User Equipment
UEID	User Equipment Identifier
UICC	Universal Integrated Circuit Card
uMTC	Ultra-reliable Machine-Type Communication
URLLC	Ultra-Reliable, Low-Latency Communication
V2X	Vehicle-to-Anything
V-AAA	Virtualised-Authentication Authorization Accounting
VDSL2	VDSL – Very High Speed Digital Subscriber Line 2
veNB	Virtual eNB
VIM	Virtualized Infrastructure Manager
VM	Virtual Machine
VMNO	Virtual Mobile Network Operator
VN	Virtual Network
VNF	Virtual Network Function
VNF-FG	VNF Forwarding Graph

VNFM	Virtual Network Function Manager
VNPaaS	Virtual Network Platform as a Service
WAN	Wide Area Network
ZM	Zone Management

1 Introduction

The 5G NORMA project is committed to design a network architecture that enables new business opportunities. Network services in the future will be highly heterogeneous and differ substantially in their requirements on the underlying communications network. There will be a need for fast and reliable connectivity with virtually zero latency for use cases such as remote control of robots, and support for billions of sensors and things. 5G will also need to provide consistent and high quality connectivity for people and things.

A possible solution to meet the requirements of such heterogeneous services would be to build a dedicated network (including the necessary infrastructure resources) for each service. These networks could differ (among others) with respect to their functional building blocks, the location of these blocks (i.e. their topology) and their dimensioning. E.g. to achieve low latency for automotive applications, network functions should be placed at the network edge, while a central placement of network functions would allow to exploit pooling gains with mobile broadband services.

However, building separate networks per service is highly inefficient from an economic viewpoint. To overcome this problem, 5G NORMA aims to design a novel network architecture that builds on three major innovations:

i) Network customization by adaptive (de-)composition and allocation of network functions;
ii) Network slicing, i.e. sharing a common infrastructure between multiple logical network instances;
iii) Network programmability.

Concepts from Network Function Virtualisation (NFV) and Software-Defined Networks (SDN) are key enablers for the implementation of 5G NORMA's architecture. However, some challenges remain:

- Network slicing shall be realized for the whole end-to-end telecommunication service and not be limited to cloud data centres. In particular, also RAN slices shall be included.
- Network functionality and topology of a network function chain shall be adaptable to the requirements of services and tenants.
- Network management, orchestration and control mechanisms have to support the adaptive allocation of network functions and the automated lifecycle management of network slices.
- SDN principles shall be extended from routers to any network function in mobile networks.
- A northbound interface from the 5G NORMA Management & Orchestration layer, allowing the tenant to interact with the mobile service provider for the deployment and operation of network slices.
- Extensive automation in management, orchestration, and control is a pre-requisite for effective multi-service / multi-tenant operation.

The 5G NORMA project has started with an initial analysis of requirements, documented in Deliverable D2.1 [5GN-D2.1], and a high-level architecture, documented in Deliverable D3.1 [5GN-D3.1]. After this initialisation phase, three design iterations have been foreseen. This deliverable describes the architecture after the first two iterations have been completed.

While the architecture concept described in this deliverable is not yet finalized, some major characteristics of the 5G NORMA architecture have already been defined:

- Multiplexing, multitasking and virtualisation are complementary technologies for sharing resources. In combination, they allow networks to share common resources from end to end.
- Supporting separate logical networks on a common infrastructure requires both dedicated network functions, assigned to a single network slice, as well as common network functions that are shared between multiple network slices.
- Principles for network management and orchestration and software-defined mobile network control (SDMC), e.g., mobility management and QoS/QoE control, have been defined.
- Together with 5G NORMA work packages 4 and 5, an intermediate control and data layer architecture has been designed and integrated into the overall architecture.

- Important points to be investigated in the third design iteration are the interfaces. Since the 5G NORMA project will not be able to define all interfaces in the 5G NORMA architecture, the first step is to identify the most relevant interfaces. This report describes some considerations on this:
- The stakeholder analysis in Section 3 shows how major architecture building blocks can be assigned to different stakeholders. Interfaces between building blocks belonging to different stakeholders are of particular importance, as these may require standardization as well as enhanced security mechanisms.
- According to the SDN principle, SDMC separates decision logic and executing agents in the design of network functions. Since the network programmability paradigm is transferred from routers to any kind of network function in a mobile network, the capabilities of the interfaces between logic and agents have to be extended significantly.
- The northbound interface from the service management entity to the tenant has to become multi-service / multi-tenant capable and highly automated.
- Finally, interfaces between SDM controllers for dedicated and for common network functions need to be investigated.

In these areas, several questions have been raised that have to be analysed and answered in the final architecture design iteration.

Structure of the Document:

The document is structured as follows:

Section 2 reviews the motivation and design principles for 5G NORMA's architecture, looks at the architectural enablers and key concepts of 5G NORMA and identifies some challenges in system design.

Section 3 looks at the infrastructure assets needed for a 5G NORMA network and how different stakeholders can interact in setting up and using a 5G NORMA network. The main stakeholders in this context are tenant, mobile service provider, and infrastructure provider. The identification of interfaces between different stakeholders is very important, as these interfaces may require standardization and enhanced security mechanisms.

Section 4 aims to describe the 5G NORMA architecture in detail. After a review of the high-level architecture defined already in D3.1 [5GN-D3.1], it looks at the design of the control and data layer and at the operational principles behind network management, orchestration, and control.

Section 5 reports on security aspects, covering security threats to a 5G NORMA network, as well as novel countermeasures to ensure the security of logically separated networks on a common infrastructure.

Section 6 describes the methodology for verifying that the 5G NORMA architecture is compliant to the requirements of tenants, mobile service providers, and infrastructure providers, as well as all generic 5G service requirements. It compiles intermediate verification results obtained after the second architecture design iteration.

Section 7 provides a summary and outlook on the work to be conducted in the subsequent final iteration.

Annexes A-C provide further details on the ETSI-NFV Management and Network Orchestration concepts, 5G NORMA Security investigations and architecture verification.

2 Design Principles for 5G NORMA Architecture

As stated in [5GN-D31] the flexible architecture designed by 5G NORMA should be built keeping in mind three main objectives: **Flexibility**, **Scalability,** and **Context-Awareness**. Legacy mobile network architectures were initially designed to provide mostly one type of service and employed a single specific kind of deployment. For example, in 4G, distributed integrated base stations provide users with a single broadband packet data connection. Subsequent amendments to that network architecture still had to maintain ties with those concepts. Now, the introduction of new enabling technologies such as SDN and NFV opens the way to new concepts and possibilities. In the framework of 5G NORMA, we leverage on the concepts of Network Slicing and Network Programmability to achieve the Flexibility, Scalability and Context-Awareness objectives. Throughout this section, we explain the motivation and the rationale behind the choice of employing these methodologies to achieve the envisioned goals.

2.1 The drivers

The 5G network will enable a fully mobile and connected society. The proliferation of wearable devices and wireless connected objects will pave the way to a wide range of new usages and new business models. In addition to more pervasive human centric applications, e.g., virtual reality augmentation, 4k video streaming, etc., 5G networks will support the communication needs of machine-to-machine and machine-to-human type applications for making our life safer and smarter and for automating and mobilizing various sectors of industry and vertical markets (e.g. energy, e-health, smart city, connected cars, industrial manufacturing, etc.). Autonomously communicating devices will create mobile traffic with significantly different characteristics from today's dominant human-to-human traffic. The coexistence of human centric and machine type applications will impose very diverse functional and performance requirements that the 5G network will have to support such as broadband everywhere and enhanced mobility management. The 5G architecture should be future-proof in terms of performance and flexibility to support **multiple network services**, with heterogeneous KPIs, sharing the same infrastructure.

5G will offer operators unique opportunities to address and offer new business models to consumers, enterprises, verticals and third party partners. This large set of stakeholders with current communication solutions is intrinsically difficult to handle due to the large set of varying requirements that need to be addressed at any point in time of deployment. Virtualisation of network functions paves the way for deploying customised network services with different virtualised NFs (VNF) on a common infrastructure, thus realizing network sharing among multiple stakeholders or tenants. **Multi-tenancy** will allow economies of scale expected when hosting multiple logical mobile networks on a single infrastructure. Tenants can range from mobile networks operators to companies from vertical industries. The 5G network architecture should be future-proof for supporting various business models, including vertical markets and multi-tenancy. It will enable an open service ecosystem where different providers can cooperate to satisfy the ever-changing needs of their users.

2.2 The enablers

Given the drivers described above, 5G NORMA designed the proposed network architecture by pivoting around two main concepts: network slicing and network programmability. These two fundamental concepts guide the design for all the elements and enablers of the 5G NORMA architecture. They are thoroughly designed to support the different stakeholder roles, as described in Section 3.

2.2.1 Network customization by adaptive allocation of network functions

Nowadays, very different applications share the same communication infrastructure, but communication networks were not designed with this in mind. With the trend of increased heterogeneity, 5G networks must be designed embracing this from the very beginning. Moreover, the final goal of 5G is not only to support heterogeneous services, but also reduce the costs (OPEX and CAPEX).

Theoretically, this customization goal can be achieved by having several physical networks deployed, one for each service (or even one for each business entity). Isolated services can hence use their resources in an optimal way, avoiding difficult re-configuration of hardware and network entities as well as the need to accommodate possibly conflicting performance objectives. Clearly, this approach cannot be applied to real networks in a cost-efficient manner, therefore, calls for a solution that allows for both an efficient resource sharing and *multi-tenancy* infrastructure utilisation.

An intermediate approach to multi-tenancy, mostly passive, is already standardized [Fri] and applied by many operators that currently share physical cell sites. However, the equipment still belongs to each operator or a joint venture of involved operators, limiting hence the cost reduction. 5G networks will go one step further, pushing for the active sharing of resources within the same infrastructure among different tenants, allowing for the introduction of so-called *vertical sectors*. In such scenarios, also non-operators may utilise network resources and functions to compose their own (virtualised) network instance.

The 5G NORMA innovation of adaptive (de)composition and allocation of mobile network functions that is enabled by the Network Programmability and Network Slicing concepts described in the next two subsections creates completely novel deployment choices. More specifically, 5G NORMA's architecture not only allows for *(re)programming the behavior of individual network functions*, but also for a *adapting the network topology*, i.e., the geographical distribution of network functions. In other words, while the option of re-configuring the behavior of network elements was already available in legacy systems (even though in a limited fashion only), the topology of such networks was static and bound to the geographical distribution of network nodes. With 5G NORMA innovations, the deployment location of a network function (1) can be modified more easily, and (2) can be different for two instances of the very same function, e.g., when instantiated for different network slices.

2.2.2 Service-aware resource sharing with network slicing

To create tenant- or service-specific networks, NGMN has proposed the concept of *network slicing* in [NGMN]. While legacy systems (e.g., 4G mobile networks) hosted multiple telco services (such as, MBB, voice, SMS) on the same mobile network architecture (e.g., LTE/EPC), 5G NORMA builds dedicated networks that exhibit functional architectures customized to the respective telco services, e.g., eMBB, V2X, URLLC, mMTC, cf. Figure-2-1. Moreover, legacy systems are characterized by monolithic network elements that have tightly coupled hardware, software, and functionality. In contrast, 5G NORMA decouples software-based network functions from the underlying infrastructure resources by means of utilizing different resource abstraction technologies. For instance, well-known resource-sharing technologies such as multiplexing and multitasking, e.g., WDM or radio scheduling, can be advantageously complemented by softwarisation techniques such as Network Function Virtualisation (NFV) and Software Defined Networking (SDN). Multitasking and multiplexing (cf. Section 4.2) allow sharing physical infrastructure that is not virtualised. NFV and SDN allow different tenants to share the same general purpose hardware, e.g., Commercial Off-The-Shelf (COTS) servers. In combination, these technologies allow to build fully decoupled end-to-end networks on top of a common, shared infrastructure. Consequently, as depicted in Figure-2-1, multiplexing will not happen on the network level anymore, but on the infrastructure level, yielding better QoE (Quality of Experience) for the subscriber as well as improved levels of network operability for the mobile service provider or mobile network operator.

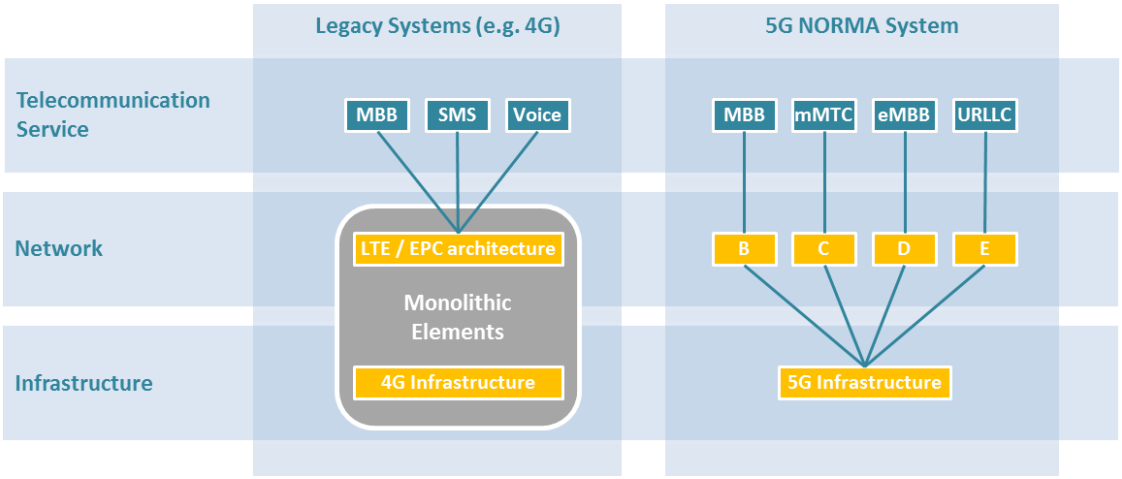


Figure-2-1: Multi-tenancy in legacy networks and slicing-enabled 5G NORMA networks

Generally, a network slice hence comprises a subset of virtual network infrastructure resources and the logical mobile network instance with the associated functions using these resources. It is dedicated to a specific tenant that, in turn, uses it to provide a specific telecommunication service (e.g. eMBB). The decoupling between the virtualised and the physical infrastructure allows for the efficient scaling-in/out/up/down of the slices, suggesting hence the economic viability of this approach that can adapt the used resources *on demand*.

Network slices are created mostly with a business purpose: following the 5G verticals paradigm [5G-PPP], an infrastructure provider will assign the required resources for a network slice, that in turn realizes each service of a service provider portfolio (e.g., the vehicular *URLLC* network slice, the factory of the future *URLLC* network slice, the health network *mMTC* network slice, see

Figure-2-1). The required resources are provided according to different resource commitment models, ranging from rather static reservations to on-demand provisioning (cf. Section 4.5.2.2).

Network slicing calls for a novel architecture capable of flexibly orchestrating and configuring all the resources, functions, and entities used by a network slice. This role is played by the Network Programmability concept described next.

2.2.3 Network programmability for flexible network control

With the introduction of Software-defined Mobile Network Control (SDMC), future 5G networks will bring the concept of *network programmability* beyond SDN. While SDN splits routing and forwarding capabilities of a switch and reassigns the former to an SDN controller, the SDMC performs such split between *logic* and *agent* for any network function in the network. That is, the SDN principles are extended to all control and data layer as well as management functions usually deployed in mobile networks. The following three categories can be identified: (i) networking control functions (in particular mobility management and session management, but also QoS/QoE control); (ii) connectivity control functions (mainly packet forwarding/SDN-based transport); and (iii) wireless control functions (e.g., radio link adaptation and scheduling).

The former two categories are a rather natural extension of the application of SDN principles, while the latter captures the key aspect of 5G NORMA's SDMC concept: selected wireless control functions will not be implemented any more bounded to specialised hardware (e. g., LTE eNB), but rather become independent pieces of software using a software-defined approach. These functions are performed by a (virtualised) programmable and logically centralised controller that abstracts and thus homogenises different network technologies. Such a controller will **make network slices programmable by controlling the topology and functionality of the service chains as well as resource control** inside the network slices.

The SDMC approach implies to have a unique control point for the network: a logically centralised controller that abstracts and thus homogenises different network technologies. By operating

a small number of SDM controllers, network operators reduce the complexity of the network management and control.

Dense wireless networks, as envisioned in 5G, especially benefit from the SDMC approach: the control of mobility support schemes and dynamic radio characteristics is performed by the SDMC that can employ especially tailored algorithms per network slice they are deployed in. Moreover, if needed, VNFs can be deployed closely to the users (e.g., in an automotive network slice) reducing their experienced latency. This feature is particularly desirable for the verticals market, as several network operators can provide their services to verticals by using the SDMC approach.

New services can, hence, be enabled by just modifying the controller functions: services that were not initially included by an operator in its architecture design, can now be introduced by implementing service-specific enhancements. The SDMC behaviour can also be modified to meet specific needs of the application or to better adapt to a specific scenario. A good example is the operation of base station schedulers: as the SDMC has a global view of the network slice, it can optimise (through an SDM-C application) the mid- to long-term behaviour of scheduling algorithms and the resource allocation across them. This concept can be extended to the resource control across network slices. SDMC allows the optimization of network utilisation: a network infrastructure provider may allocate unused resources to demanding network slices, provided that the SLA is satisfied for all the hosted network slices. In this way, more verticals may share the same infrastructure, reducing operational costs as well as avoiding time consuming deployment of dedicated infrastructure.

Another possible usage of the SDMC is mobility management. As stated above, the SDMC is an extension of the SDN concept to any kind of network function in the mobile network. So, a straightforward amendment of the SDN dialect, capable of directly handling user data tunnels, may be used to directly control the gateway entities of the network. However, the same idea can be used to directly control other low-level user flows, steering traffic through network functions implementing Cloud RAN. That is, one centralised, flexible application can control heterogeneous network functions through specialised interfaces. Mobility management is just an example, but verticals may directly provide their customised SDMC application if they desire to do so, focusing on a customised function behaviour. As depicted in Figure 2-2, the SDMC offers a unique interface to several and heterogeneous network functions ranging from wireless control to traffic steering.

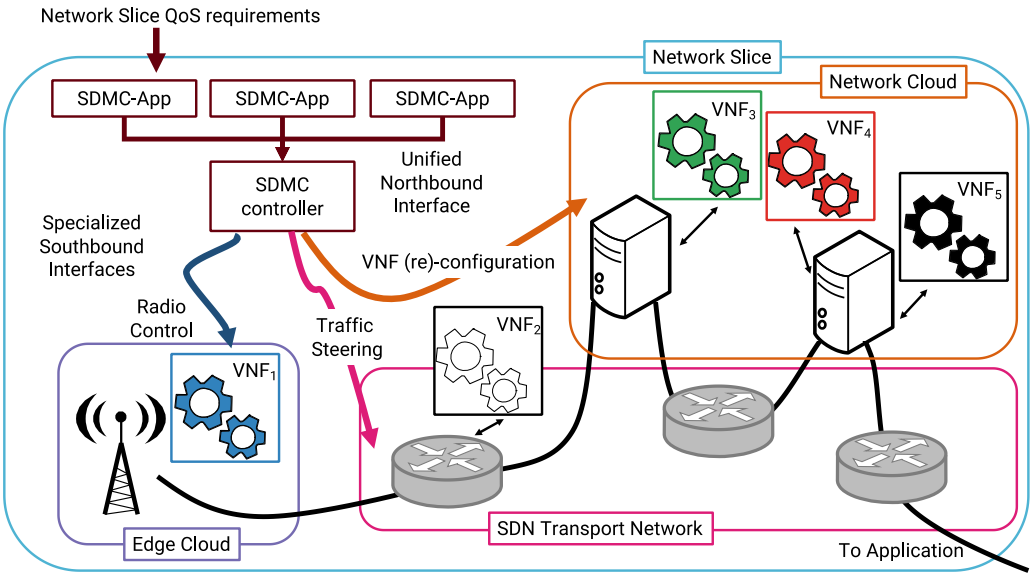


Figure 2-2: An example of the network programmability concept

Therefore, an SDM controller (SDM-C), following the SDN principles, has a northbound and a southbound reference point. The northbound reference point is used by SDM-C applications to exchange high-level messages with the controller. The SDM-C applies these high-level commands to the underlying SDN/NFV-based networks through southbound interfaces to configure

specific resources and functions. With SDMC, service providers will be able to fit the equipment to their needs by simply re-programming the controller using well-defined APIs, and thus enabling new services within a very reduced implementation, test, and deployment footprint. This will lead to a reduced service creation time, as the definition of a standardised and unique northbound reference point will simplify the creation of new network functions, as the low level and vendor-specific characteristics are managed by the SDM controller southbound interfaces.

The advantages of the SDMC concept are manifold. The first one concerns the increased flexibility of the network. By leveraging the programmability of the SDMC approach, operators will be able to match their needs in many cases by simply re-programming the controller and the underlying functions, thus reducing costs. This approach also allows to scale-up and down virtualised functions, enhancing reliability as well. The flexibility is not just exposed to network operators, but also to verticals, that can acquire network resources fulfilling a pre-defined SLA. Programmability also allows to customise the network, enhancing the QoE perceived by users.

2.3 The challenges

Obviously, network slicing and network programmability create several research challenges that are being tackled by the project as a whole (see [5GN-D41] and [5GN-D51]). All the technical innovation developed in Work Packages 4 and 5 are fully compliant with the architecture defined by WP3. Also, some of them are going to be showcased in fully fledged demonstrators, as detailed in [5GN-D61].

Service Aware Orchestration: The work towards a full cloudification of mobile networks is not limited to the architectural issues, but it also involves new research challenges related to the decision about how to allocate functions to nodes. This calls for both an analytical approach to address the corresponding optimization problem and an orchestration architecture capable of supporting this functionality. To this aim, an exhaustive model of both the computational behaviour of nodes and functions as well as the capabilities of transport networks is required. Also, the precise monitoring of the infrastructure utilisation is needed for the maximization of the system performance through the optimal location of VNFs.

QoS/QoE Network Control: The application of the network programmability concept allows the flexible QoS/QoE network control. This entails the definition of a consistent QoS/QoE framework that is finally used to check the status of the virtualised infrastructure, be it computing, networking or storage resources. Conversely to the current NFV MANO architectures, an SDM controller is aware of the domain-specific semantics of each VNF (e.g., 3GPP) running in the network. Therefore, the controller can react to QoS/QoE parameters change, eventually triggering a re-orchestration if more or different resources are needed.

Efficient resource sharing: the solutions needed for efficiently sharing resources among different network slices belonging to multiple tenants are multifold and heterogeneous. How to define the interfaces among different entities in the network, how to decide which resources (or network functions) must be shared among which network slices and how to decide on resource quotas assigned to each slice are open problems to be addressed by according research activities. While many of the technical solutions for these problems are being investigated in WP4 and WP5, the architecture proposed in WP3 allows for their integration and flexible implementation.

Management and control of multi-technology, heterogeneous radio access networks (RAN): 5G RANs will be characterised by a mixture of various system generations and (radio access) technologies. This includes, but is not limited to, multiple connectivity layers, i.e. multiple network layers such as macro and small cells, and multiple RAT layers such as below 6 GHz and mm-wave. The challenge of designing an overall architecture and individual functions to globally and jointly manage these resources is complemented by the need to include mechanisms for flexibly placing and combining network functions. This follows the according overall concepts of 5G NORMA in order to allow for, e.g., different deployments of centralised RAN or to enable multi-connectivity.

Realisation of slicing in radio access networks: The RAN is a typical example of a shared network function controlled by a single authority, where spectrum is shared amongst mobile virtual network and service operators. Consequently, the possibility of adding new instances of virtualised network functions for a newly deployed slice is only available to a limited extent. Rather, RAN slicing requires further mechanisms for multiplexing multiple slices into a shared resource. Depending on the service requirements and infrastructure capabilities, this multiplexing can be implemented at different layers, e.g., multiplexed access to either MAC or lower PHY layer, resulting in (depending on the chosen approach) varying degrees of design complexity, efficiency in radio resource usage and slice-individual adaptability and programmability.

These items are just a sample of the research challenges created by the introduction of these novel concepts into the architecture. We will discuss how the 5G NORMA architecture can simplify the algorithmic design as tackled by WP4 and WP5, while also allowing for a flexible configuration of network elements belonging to different stakeholders. We will broadly discuss this topic in Section 4.

3 The 5G NORMA Ecosystem

3.1 5G NORMA stakeholders and infrastructure

3.1.1 5G NORMA stakeholders

Stakeholders are entities, individuals or organizations, supervising or making decisions that affect how the 5G NORMA ecosystem operates and evolves. Various stakeholders are involved in 5G NORMA, each with their characteristics and interests.

The **5G NORMA mobile service provider (MSP)** is the entity/company that provides Internet connectivity and telecommunication services to subscribers. Furthermore, the MSP offers dedicated mobile network instances (i.e., network slices) that realize a specified telecommunication service (e.g., mMTC) to 5G NORMA tenants (see next stakeholder).

The **5G NORMA tenant,** usually a business entity, buys and leverages on 5G NORMA network slice services provided by the MSP. It can be a Mobile Virtual Network Operator (MVNO), an enterprise (e.g. from a vertical industry) or other organisations that require a telecommunications service for their business operations.

The **5G NORMA mobile subscriber** (hereafter referred to as subscriber) is an individual who consumes services from the MSP or from the tenant. The MSP's subscriber purchases 5G NORMA Internet connectivity and telecommunication services for the purpose of entertainment, communication, etc. The tenant's subscriber consumes business/vertical services. Such business/vertical services might encompass network connectivity and telecommunication services e.g. when the tenant is a MVNO.

The **5G NORMA infrastructure provider (InP)** is the entity/company that owns and manages parts of or all infrastructure of the network. It can be further distinguished between **RAN infrastructure provider** and **datacentre or cloud infrastructure provider.** The former owns the physical infrastructure such as the antenna sites, the HW equipment for the antenna and RRHs (Remote Radio Heads), monolithic base stations, etc. The latter is represented by the collapsed roles of entity/company that owns and manages local and central datacentres. Within 5G NORMA, there are two types of datacentre operators, infrastructure providers acting on small/medium size datacentres (in terms of resources to be deployed and geographical presence) and big players (like Amazon) having big datacentres deployed world-wide.

The **mobile network operator (MNO)** is an entity that operates and owns the mobile network, i.e. it merges the roles of MSP and InP into a single business stakeholder.

The **software vendors** are companies that develop and distribute VNF, management and orchestration, or SDM controller software products (e.g. VNF source and information associated for VNF image creation). They sell their software products to MSPs, MNOs, tenants, and datacentre or cloud InPs.

3.1.2 5G NORMA infrastructure assets

Infrastructure forms the foundation of any network. The following two subsections describe the main properties of mobile and fixed networks: Mobile network infrastructure is indispensable to connect the mobile terminals, and fixed networks accomplish the data transport from the mobile network to the Internet.

3.1.2.1 Mobile networks

The radio access network (RAN) of representative mobile network operators nowadays consists of a mixture of 2G (GSM), 3G (UMTS) and 4G (LTE) platforms. It is to be anticipated that within the timeframe considered by 5G NORMA only the 4G platform is relevant for consideration as a starting point for infrastructure inventory. Figure 3-1 shows the current architecture of 4G networks. The 4G mobile network is optimised for mobile broadband services (MBB) and consists

of a flat architecture including one or more 4G evolved packet core (EPC) and distributed radio nodes (eNodeB / eNB). The eNBs are connected to the ePC via a standardised S1 interface. In practical networks eNBs are backhauled by Ethernet links where transport technology is mostly optical fibre or, more rarely, microwave links. So-called X2 interfaces between eNB are realised on demand by an Automatic Neighbour Relation Function (ANRF) initiated by eNBs. As the S1 backhaul is normally encrypted by IPSec at the eNB (s. Annex B.2), the logical links between eNB typically are routed via the security gateway or an even more central node near the EPC. In rare cases LTE-A (LTE-Advanced) interference coordination schemes may require low latency and hence the installation of direct physical links between the base station sites is needed. In many cases, the backhaul including the aggregation network is leased from fixed network operators. The 4G core (ePC) is supporting this by the bearer concept.

Currently an increasing number of small cells (SC) are deployed leading to so called heterogeneous networks (HetNet). Logically small cell eNB are at the same level as macro eNB but due to lower transmit power and deployment mostly at street furniture (lamp posts, advertising pillars, etc.) their coverage area is much smaller than those of macro cells. The typical cell size of a small cell is below of 20% of the macro cell area. Small cells are backhauled similar to macro nodes. With respect to macro cells at low frequency bands, SC are deployed at co-channel or at dedicated frequencies. Typically, a meaningful number of small cells per macro cell currently can be assumed to be below 5.

For quantitative descriptions of the network topology usually it has to be distinguished between urban, suburban and rural environments, where the average inter-site distance (ISD) or alternatively the cell area of macro sites is the most characterising parameter. Getting access to suitable sites in dense urban areas is also challenging. Within the London sample area used for the evaluation in Section 6 for example, the average cell area of a macro sector antenna accounts for approximately 400000m^2 where we assume that most of the macro sites host 3 sectors (cells).

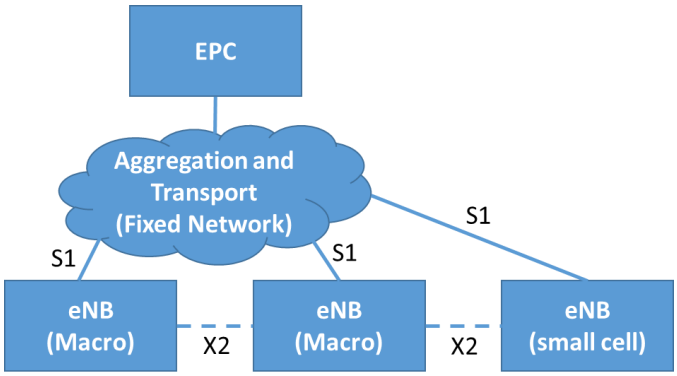


Figure 3-1: Current RAN architecture status

The core network, EPC, is a group of network elements that are usually deployed centrally. For reason of resilience but also because core network functions scale with the number of active PDCP contexts as well as with the U-Plane traffic, it can be assumed that there exist at least two EPC per mobile operator within a nation-wide network.

3.1.2.2 Fixed-line networks

The assets of a fixed-line network operator are organised hierarchically in different network segments, which aggregate traffic from the lower segments to gain advantage of statistical multiplexing.

Customer traffic reaches the network through the access segment, which provides capillarity and a first layer of aggregation to the network. Distance from the access network equipment to the customer premises is variable and depends on the reach of the access technology used, and on the user density in that area. It ranges from several km for PON deployments, to a few hundreds of meters for VDSL2 (Very-High-Bit-Rate Digital Subscriber Line 2, [G. 993.2]) or G.FAST (Fast access to subscriber terminals; the letter G stands for the ITU-T related recommendations, G.9700

[G.9700] and G.9701 [G.9701]) technologies. The mix of deployed access technologies shapes network topology, and also the quantity, density and characteristics of operator Points of Presence where the access network termination points are deployed.

Points of Presence are interconnected by layered technologies to provide bulk capacity and granularity to switch traffic between them.

On one hand, fixed-line network operators deploy an optical layer that provides pure capacity transport across the operator Points of Presence. Parts of this optical layer are the fibre rings and the ROADMs (Reconfigurable Optical Add-Drop Multiplexer) used to inject traffic into these rings. In order to prepare the signals to be injected into the fibre rings and links, C/DWDM transponders and multiplexers are used to interface the customer feeds into the optical layer. The transport capacity obtained with this optical layer is offered to end-customers, such as business customers or vertical industries, while at the same time it is used by the fixed-line operator itself to build on top of that an IP services layer.

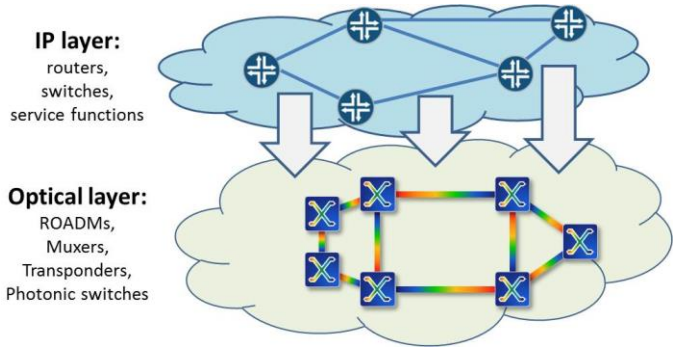


Figure 3-2: Optical and IP layers of a fixed line network operator

The IP layer provides Layer 2 (Ethernet) and Layer 3 (IP) connectivity services. Making use of these connectivity services, the fixed operator builds its offering of residential services (video, voice, Internet) and business services, and these connectivity services can also be used as transport by other operators, like mobile operators. Based on the evolution of the underlying technologies of the IP layer, it has been typically the case that the fixed operators have deployed a metro network and an IP core network to build this IP layer. Metro networks were originally deployed with native Ethernet technologies that could only provide Layer 2 Ethernet services, but evolved to MPLS networks providing Layer 2 VPN services on top. At the same time, IP core networks were already using MPLS technology to provide public IP connectivity and layer 3 and layer 2 VPNs on a regional/national/international scale.

It has been typical also to have different cores dedicated to different services, for instance having a separate IP core for NGN voice networks or enterprise customers. This segmentation of the IP layer in metro and core networks has led over the years to inefficiencies because of the segmented provisioning required and the increased number of required equipment to provide its services.

To increase efficiency, there is a trend in fixed-mobile convergent operators to evolve their networks to a single multiservice IP/MPLS network extending from the edge, where the access nodes and service terminals are directly connected, to the interconnection to external networks. These IP/MPLS networks are composed by the PE nodes (Provider Edge routers) facing the access interfaces to the MPLS backbone, where external systems connect, and the P routers (Provider routers) that provide transit connectivity across the MPLS network.

Figure 3-3 and the accompanying Table 3-1 show a fixed access network, providing transport to various services, and the different kinds of nodes involved from the access to the core segments.

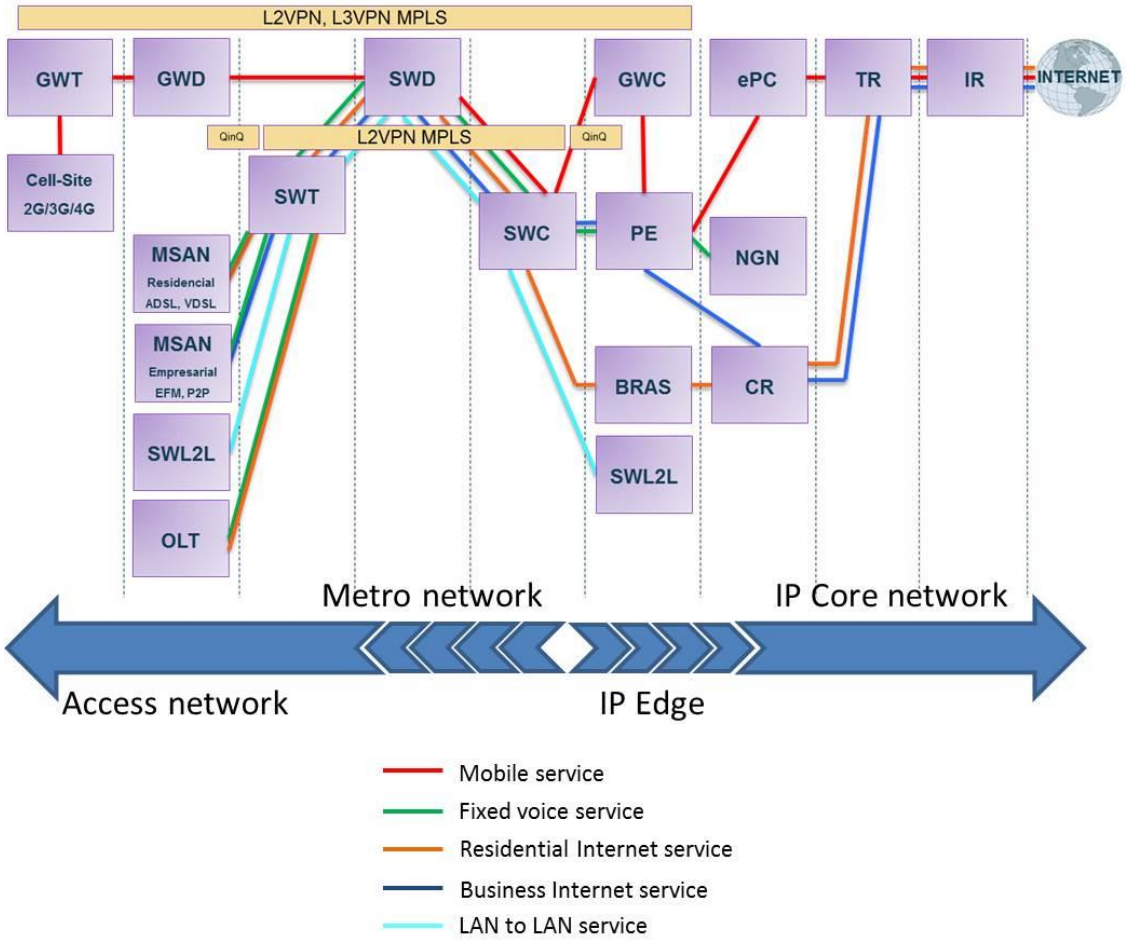


Figure 3-3: Typical Fixed Network Operator node connectivity model at the IP layer

Table 3-1: Legend for Figure 3-3

Acronym	Meaning
GWT, GWD, GWC	Mobile backhaul Terminal, Distribution and Concentration Gateway nodes
SWT	Terminal Switch. Edge of Metro Network, where all current fixed-business nodes are connected
SWD, SWC	Distribution and Concentration Switches. Elements of Metro Network plane
Cell Site	Base Station. It can host any of the following mobile technologies: 2G, 3G and LTE
MSAN	Multi-service Access Node. Access node for all residential and enterprise services
SWL2L	Lan-to-Lan Switch. Specific node for this type of service (where applicable)
OLT	Optical Line Termination. Specific node for this type of service (where applicable)
BRAS	Broadband Remote Access Server
PE	Provider Edge Router
NGN	Next-Generation Network. VoIP Platform
EPC	Evolved Packet Core, including all Mobile Core elements, including RNC and BSC, where co-located
CR	Concentration Router
TR	Transit Router (P Routers)
IR	Interconnection Router (Toll-Gateway)

Transport is typically provided for a combination of different services:

- Mobile services: 2G, 3G, 4G, 5G, WiFi (links marked in red in the diagram)
- Residential Business: Broadband, CMTS (HFC) service (links marked in orange in the diagram), and voice access (marked green),
- Enterprise Business: MPLS-VPN access, VPLS/L2L access, Managed WiFi, Internet access. (links marked navy blue)
- Wholesale Business: IP transit / IPVPN services, Broadband connectivity, Interconnection for LEC (Local Exchange Carriers) / CUG (Close User Group), TV Links, L2L (marked light blue)

With the advent of NFV technologies, fixed line operators are also evolving their networks to include VNFs based on x86 hardware deployed in their Central Offices or Points of Presence.

Centralised control layer nodes such as AAA or IMS are already being virtualised, but the trend is to also virtualise some distributed data plane functions of the IP layer once the performance required is achieved.

Specially tailored for being virtualised are those functionalities that today run on services cards on hardware equipment, like CGNAT (Carrier Grade NAT technology used to apply in network-based equipment NAT policies to subscribers) and firewall functionalities, or functionalities which are customer-aware like BNG functionalities on a PE router.

There are initiatives like the Cloud Central Office [CLOUDCO] of the Broadband Forum to standardise a reference model of how to support network functions of a fixed line operator on a cloud-like infrastructure deployed at its Central Offices.

Therefore, the trend is an evolution towards a fixed network built on 3 layers: an optical layer providing capacity, an IP layer providing basic Layer 2 and Layer 3 VPN services, and an NFV layer deployed on Cloud COs providing feature richness.

On top of these layers, another trend of evolution in the fixed line network operators is the use of the SDN paradigm to achieve the dynamic provisioning of services across all the different layers involved so that service orders/requests can be fulfilled instantaneously by the network.

3.1.3 Relationships between stakeholders

The multi-tenancy requirements of 5G NORMA are the cornerstones of the relationships between stakeholders.

The so-called tenants are MVNOs and vertical market or Over-The-Top (OTT) players as presented in Figure 3-4. The goal of the tenants is to provide services/applications to their end-user subscribers with good quality of experience. For doing this, they have a commercial agreement with the MSP to use an end-to end virtual network, i.e., a network slice. A MSP offers a portfolio of network slice types. The tenant selects the slice type that is most suitable for his purpose. For example, the portfolio of network slice types may cover Network-as-a-Service (NaaS) or Platform-as-a Service (PaaS). It may also be communication services (e.g. voice, messaging, broadband internet or machine type communication) provided by the mobile service provider to end users (subscribers) on behalf of the tenant (in the latter case, the tenant's main role might be restricted to the commercial relationship with the customer).

In 5G NORMA, the MSP's role is central. The MSP is intersecting between tenant and InP. There is no direct relationship between InPs and tenants and the MSP is actually brokering the resources from possibly multiple InPs. The MSP's role is to acquire the necessary resources from one or more InPs to build an end-to-end virtual network (slice) instance according to the needs of the tenant, i.e. a collection of (mobile) network function instances including their required resources necessary to operate an end-to-end (self-contained) logical mobile network. The MSP has to ensure that the SLAs he has with the tenants are satisfied, while being constrained by the availability

of resources rented (bought) from possibly multiple InPs as presented in Figure 3-4. In addition, when also owning the required resources, i.e., (parts of) the infrastructure (e.g. RAN), the MSP acts as an MNO.

In the 5G NORMA relationship model, it is worthwhile to mention that multiple tenants can share both physical and virtualised network functions and their underlying infrastructure resources. A given network slice running for a tenant is composed of network function instances dedicated to the sole tenant's usage and of network function instances shared among multiple tenants (and therefore among multiple slices).

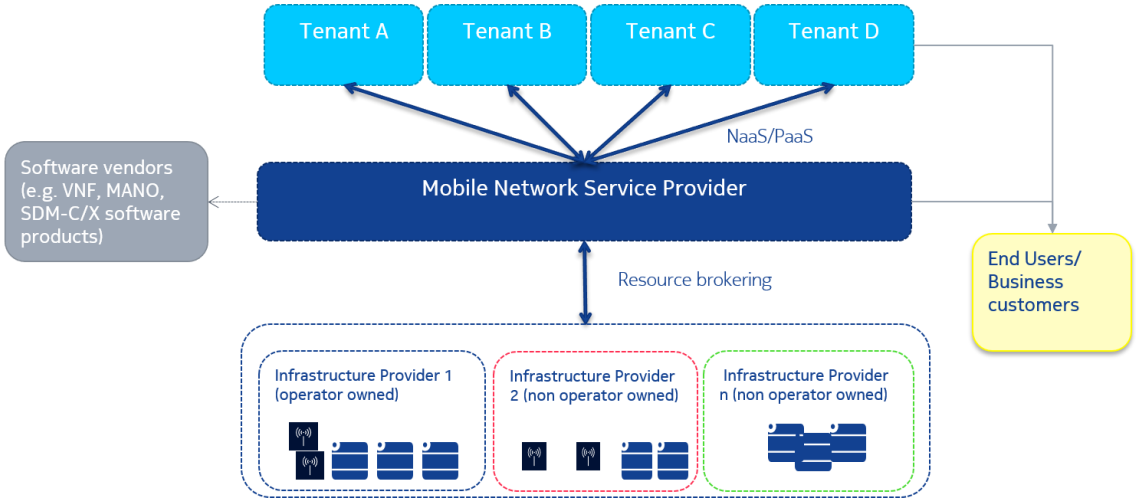


Figure 3-4: Relationship between stakeholders in 5G NORMA with Mobile Service Provider in the core place

The business relationships between MSP, InPs, and tenant have impact on the architecture and the mechanisms investigated in 5G NORMA. Depending on the situation, the entities of the 5G NORMA architecture belong to distinct administrative and technical domains as depicted in Figure 3-5. Thus it is important that the cross domain functional interfaces are carefully designed for possible standardization. Moreover, security aspects at these interfaces must be covered.

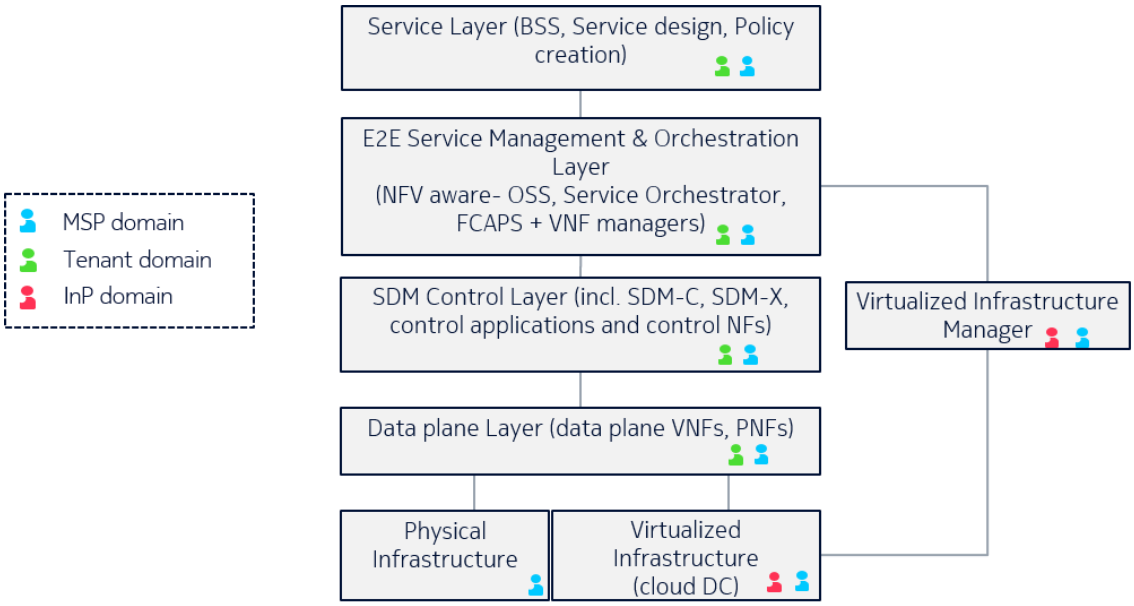


Figure 3-5: Possible domain ownerships of 5G NORMA main functional blocks

The resulting technical dependencies between stakeholders are of diverse nature. For example, the MSP will have to rely on the level of resource availability and utilisation information provided by the InP(s) exposed through APIs from the VIM to control and manage the network function and services in higher or lower granularity. This means, the extent of choice for, e.g., selecting a

data path or the location of any given network function will depend on the granularity of topology information exposed by the VIM of the InP.

Further, the tenant will rely on the SLA it has with the MSP. E.g., the MSP can grant the tenant partial access to 5G NORMA NFV Management and Orchestration (MANO) layer functionality through adequately designed APIs of according entities (e.g. SDM-O, SDM-C, Service Management) or through dedicated instances of these entities controlled/operated by the tenant. Besides, the tenant itself can provide some network function software, possibly accompanied by the respective VNF Manager and network management entities. The 5G NORMA MANO and control framework should handle the case where the SDM-O /SDM-C should coordinate with VNFs and VNF managers belonging to another entity/organization.

Based upon the above considerations, we identified five most relevant reference points for cross domain interfaces in the 5G NORMA architecture main blocks.

- Reference point between tenant-operated service layer functions and MSP-operated service management entity.
- Reference point between InP (or MSP)-operated VIMs and MSP (or tenant) orchestration Layer entities (e.g. SDM-O, VNF managers).
- Reference point between Tenant-operated SDM-O and MSP-operated SDM-O.
- Reference point between tenant-operated SDM-C and SDM-X operated by MSP/MNO.

In the following, we define some scenarios (so-called "Offer Types") in order to analyse in detail the identified stakeholders' relationships to gain insight on the technical and business dependencies among them, and identify potential impact on the components and interfaces in the 5G NORMA architecture.

3.2 5G NORMA offer types and security considerations

Based on the identified stakeholder roles for 5G NORMA, this section analyses three typical so called "offer types" of a mobile service provider for tenants. It shall be explicitly mentioned that this set of offer types, while trying to represent certain categories of similar offer types, does not represent the full bandwidth of scenarios enabled by the 5G NORMA architecture. Rather, they shall depict probable (as conceivable by today) constellations in the stakeholder interaction. Each offer type description covers in detail the identified stakeholders' relationships to gain insight into the technical and business dependencies among them, and identify potential impact on the components and interfaces in the 5G NORMA architecture. Security considerations are taken into account, in particular w.r.t. the inter-stakeholder interfaces. The three analysed offer types include:

- **Offer Type 1: MSP operates slice on behalf of the tenant**
  Here, a tenant requests the commissioning of a network slice by providing the high-level requirements of the telecommunication service to be provided. Operation of the network slice is completely handled by the MSP (or MNO), the tenant only receives coarse-grained performance reports.
- **Offer Type 2: Limited slice configuration and control options for tenant**
  Here, in addition to offer type 1, a tenant can specify more fine-grained configuration options for the requested network slice. Moreover, selected network operations (e.g., subscriber data management, QoS control) are performed by the tenant. Still, the major part of network operation is handled by the MSP or MNO.
- **Offer Type 3: Extended slice configuration and control options for tenant**
  In addition to offer type 2, the tenant has a rather wide control over deployed network functions. This can go as far as the tenant onboarding own network functions for selected

areas, e.g., mobility or session management, contributing own infrastructure, and operating a part of the network slice independent of the MSP (or MNO). Nevertheless, these functions onboarded to the MSP's or MNO's systems have to be certified.

As an additional dimension of the analysis, multiple ownership constellations can be considered in terms of infrastructure and software assets of the various stakeholders. Again, while the 5G NORMA architecture is rather independent of ownership constellations, i.e., it supports a wide variety of such hardware and software ownership scenarios, this deliverable focuses on selected constellations that the consortium deems likely and relevant. The following assumptions on ownership are used in the subsequent analysis:

Software assets:

- The MSP owns all software for running VNFs and telecommunication services. It includes software for element managers, service management, OSS, and other MANO functions except some VIMs depending on the infrastructure assets. Furthermore, the MSP makes available catalogues for system-certified VNF software that are used for operating own network slices or that can be used by tenants.
- The tenant may have its own sets of VNFs software customised/ certified for his business needs. It will have service-level agreements with the mobile service provider for completing the set of needed functions and for provisioning of infrastructure resources.

Infrastructure assets:

- The MSP owns at least the dedicated application-specific HW (PNF) nodes and the edge cloud nodes that comprise the RAN infrastructure[2], possibly also the infrastructure for hosting the central cloud nodes. Hence, it has the role of an MNO according to the definition in Section 3.1.1. Alternatively he can rely on the services of a cloud infrastructure provider for hosting the central cloud nodes. He controls the VIMs for his edge cloud and has an API to the VIM of the central cloud provider.
- It is important to note that the business models to be supported by 5G NORMA and discussed in this chapter at some points require trust between different stakeholders, in the sense that there are no suitable technical means for a stakeholder to secure all of its assets against malicious behaviour of other stakeholders. In particular, trust in the provider of the software assets and the provider and operator of the infrastructure assets is required, as discussed in the following.

Trust in the software vendor:

- The software used by the MSP is assumed to be provided by a software vendor, a party different from the MSP. The MSP must somehow ensure that this code behaves correctly when deployed on cloud infrastructure owned or rented by the MSP. The MSP may carefully inspect and test the code, but realistically, the MSP would not be able to find e.g. carefully embedded backdoors. So trust between the MSP and the software vendor is required.

- Considering long lasting relationships between software vendors and MSPs, it is reasonable for MSPs to assume that the software vendor is generally benign. Still, software errors may lead to misbehaviour of the software. Implantation of malicious code by malicious insiders at the software vendor is also a threat that may not be easy to mitigate.

- These considerations also hold for the NFV software bought and operated by the InP, i.e. also the InP needs to trust its respective software vendors.

[2] The MNO can share the RAN infrastructure with other stakeholders. Any RAN infrastructure resource assigned to a tenant network slice is controlled by the MNO.

Although not in the focus of this discussion, note that also the hardware vendors must be trusted. Malicious hardware vendors could embed hardware backdoors, interception facilities, kill switches etc.

Trust in the infrastructure provider:

- A provider of cloud infrastructure sees all data processed on the infrastructure, those of the MSP as well as those of the tenant - there are no technical means to conceal information that needs to be processed (in clear text) on the infrastructure. Moreover, it is hard for an MSP or tenant to monitor the correct resource assignment.
- While it is reasonable to assume a benign InP (even if it is a third party different from the MSP), still there is the risk of erroneous or ill-configured NFV software or hardware that may allow other customers of the InP to attack the MSP and/or the MSP's tenants.
- It may make a difference for a tenant whether the MSP owns the infrastructure or whether the MSP relies on a 3rd party InP; a tenant may not be willing to accept each possible 3rd party InP.

Next, the three offer types as depicted above are analysed in more detail.

3.2.1 Offer type 1: MSP operates slice on behalf of the tenant

The conventional or legacy scenario covers the case where a MSP creates and operates the network slice to offer full-fledged telecommunications services to its tenant customers.The network slice is provided according to a catalogue without any tenant specific modification.

3.2.1.1 Detailed description of the offer type

In this scenario, the tenant's role is rather limited to be a vertical market customer of the mobile service provider, using telecommunication services for its business service/applications towards its customers. The tenant is only responsible for managing its customer's subscription and billing. The tenant has his own customers to whom it delivers its own applications/services. It manages its own subscribers via its own system for BSS functions. The tenant will provide customer information to the mobile service provider's BSS to register customers. This is for authenticating and authorizing the customer device to access to the network slice and to the related service(s).

The tenant and the mobile service provider have to agree on an SLA contract defining the targeted services and related KPIs targets - a tenant may request several services (e.g. voice call and video streaming services with various service coverage area, user density, etc.) running in a single network slice - The tenant fills the service description with required KPI from a template provided by the MSP. It can select value ranges for service parameters (among possible values included in the SLA). The filled service template is defining (part of) the SLA. The MSP defines the service description templates with possible values range for KPIs. It defines and creates the network slice templates with service & network KPIs based on the service needs expressed in the service description template.

The MSP deploys and operates the customised service on the slice on behalf of the tenant (the tenant has simply an access to monitored KPIs). It acquires needed resource from infrastructure providers to host the logical network slice. It activates/orchestrates the network slice then monitors and optimises it. It procures the tenant with an interface for monitoring service KPIs defined in the SLA.

The tenant monitors the SLA through the interface provided by the mobile service provider, by accessing to monitor service KPIs.

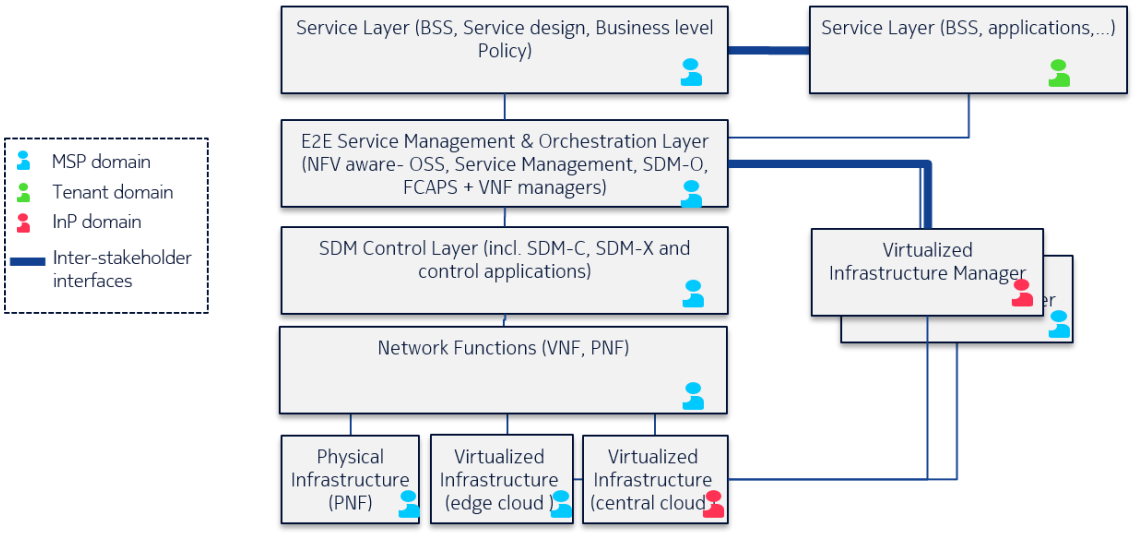


Figure 3-6: Domain ownership of high level functional blocks and relevant cross-domain interfaces in offer type 1

Most relevant cross domain interfaces are depicted in Figure 3-6 and detailed hereafter.

- Interfaces between the service layer functions from the MSP and service layer functions from the tenant:

  These interfaces are produced by the MSP service layer functions and consumed by the tenant's service layer functions. Authorization policies determines the access rights for the tenant. The interfaces are mainly used for defining the service level agreement and the required service KPIs. The interfaces cover the following requirements:

  - Exposure of service description to the tenant with possible KPIs value ranges
  - Management of tenant's request on KPIs targets (tenant's demand and MSP approval)
  - Exposure of monitored SLA KPIs to the tenant (for information)
  - Charging/Billing reports

- Interfaces between cloud InP-VIMs and MSP-Management&Orchestration Layer functions:

  These interfaces are exposed by the VIMs and consumed by the SDM-O (or VNF managers) functions operated by the MSP. Authorization policies determine the access rights for the MSP. The interfaces cover the following requirements:

  - Exposure of available virtualised resources
  - Resource reservation/allocation
  - Exposure of allocated resource monitoring
  - Software images management (adding/updating/deleting)

3.2.1.2 Security considerations

Tenant – MSP relationship:

- The tenant needs to trust the MSP with respect to correct resource assignment, keeping data secret, and maintaining integrity of data. The tenant has no technical means to enforce correct behaviour of the MSP in this respect. The monitoring of the SLA KPIs relies on information provided by the MSP itself and therefore would not help against a fraudulent MSP.

- The MSP can monitor, control and possibly limit everything the tenant's subscribers do in the network.
- The MSP must secure the interfaces provided to the tenant and must do authentication and authorization in order to prevent attackers from gaining information on the tenant's network slice and from making fake requests claiming to be the tenant. Based on the authentication, a security association providing integrity protection and encryption (such as a TLS connection) must be used for the communication at the MSP – tenant interfaces.
- Tenant requests may even be digitally signed – to protect tenants against requests faked by the MSP itself and to give MSPs a proof that a request has been made by a tenant (non-repudiation).
- With these precautions in place, the MSP need not trust the tenant but can maintain the security of the MSP's own services even if the tenant behaves maliciously.

MSP – InP relationship:

- As discussed above, the MSP needs to trust the InP.
- The InP needs to secure the interfaces provided to the MSP, in order to prevent abuse by attackers. This includes authentication, authorization and securing the communication at the interfaces via security associations (such as TLS connections), and possibly even with digital signatures for requests, if non-repudiation is desired.
- The InP must maintain isolation between the resources assigned to the MSP and resources assigned to other parties (other users of the infrastructure).
- With these precautions in place, the InP need not trust the MSP but can maintain the security of the infrastructure even if the MSP behaves maliciously.

3.2.2 Offer type 2: Limited slice configuration and control options for tenant

In this scenario, the tenant is a vertical market customer or a MVNO of the mobile service provider, using telecommunication services for its business service/applications towards its subscribers. The tenant gets extended capabilities for designing a customised slice and a limited option to configure and control a network slice operated by the mobile service provider.

3.2.2.1 Detailed description of the offer type

The tenant has his own subscribers and subscriber database for delivering its own application/service. The tenant manages his own subscribers via his own BSS. The tenant will have access to exposure to the mobile service provider's OSS to trigger that e.g. a new subscriber is configured at the HSS. This is for authorizing the subscriber device to access to the network slice and to the related service(s). The tenant monitors the SLA through the interface provided by the mobile service provider, by accessing to monitored service KPIs.

The tenant can contribute to the definition of the slice with the mobile service provider for customization purposes. The tenant can design/customise the network slice template by composing the VNF graphs from a catalogue of available VNFs or sub-graphs of VNFs. It can also select or design some QoS/QoE policy rules for operating the services in the slice. The network slice can integrate a set of tenant owned functions customised/certified for its needs. The mobile service provider procures the tenant with an access to a network slice design space. The mobile service provider validates the design of the network slice composed by the tenant and its associated rules before integrating it in its catalogue of network slices. It also procures an interface to upload its function software and useful information for mobile service provider to create function images and store in the VNF repository. The mobile service provider is able to check the integrity of such packages. The mobile service provider manages, orchestrates and controls the network slice.

The tenant can control the end-to-end service performance by selecting/changing QoS/QoE policies rules among a set made available/authorised by the mobile service provider. Primarily those customised QoE/QoS policies are set in the design phase and used for the entire life cycle of the network slice. The mobile service provider enforces/applies the policy rules selected by the tenant when orchestrating/controlling the network slice (enforcement is done by SDM-O and SDM-C/X). Optionally, the tenant might be able to modify online the policies within a limited scope and authorised by the mobile service provider. The tenant is provided with an interface to update the policies during the operation time of the network slice. The MSP has means to notify online change of policies to the SDM-O and/or SDM-C/X.

For example, the tenant can customise a slice by adding its certified video optimization function and by defining a policy rule to trigger when the video optimization is used. The mobile service provider validates and authorises the new slice and associated rule. It enforces the modification of the data forwarding path to go through the video optimiser upon a triggering event.

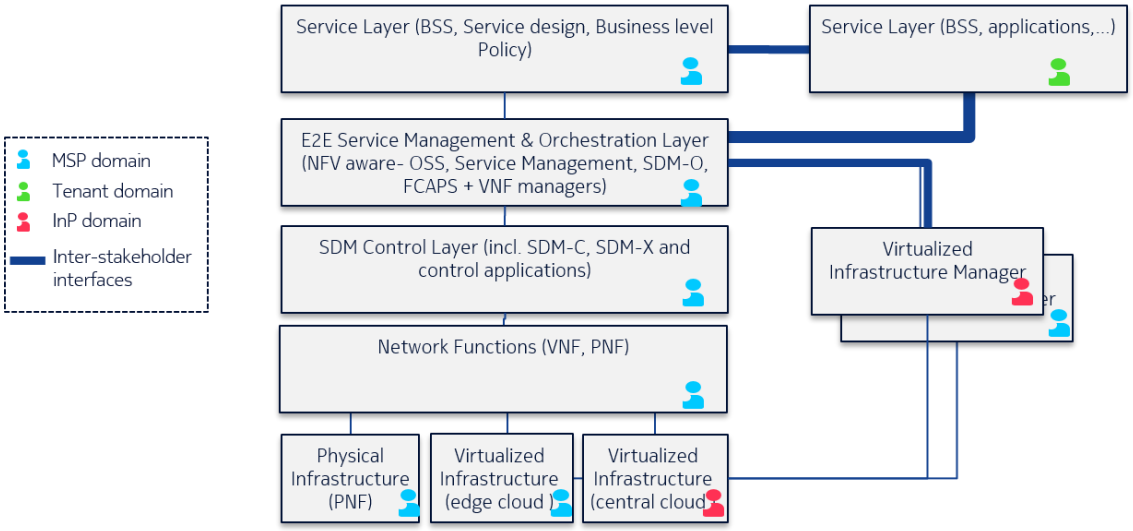


Figure 3-7: Domain ownership of high level functional blocks and relevant cross-domain interfaces in offer type 2

Most relevant cross domain interfaces are depicted in Figure 3-7.

- Interfaces between the service layer functions from the MSP and service layer functions from the tenant:

  These interfaces are produced by the MSP service layer functions and consumed by the tenant's service layer functions. Authorization policies determine the access rights for the tenant. The interfaces are used for defining the service level agreement and the required service KPIs. The interfaces cover the same requirements as for offer type 1. In addition it requires provision of credentials to access to the design space of the MSP platform (e.g. an URL to the MSP platform and a tenant account/password). In contrast to offer type 1, charging/billing reports are not used in offer type 2.

- Interfaces between Service Management from the MSP and Service Layer functions from the tenant:

  These interfaces are exposed by Service Management operated by the MSP and consumed by Service Layer functions from the tenant. Authorization policies determine the access rights for the tenant. The interfaces cover the following requirements:

  - Allowing the tenant to provide subscribers information to be registered in the MSP HSS
  - Reporting Accounting/Charging data
  - Allowing tenant to design/compose a network slice either from a list of available VNF/VNF sub-graphs or by adding its own VNFs (certified for tenant's use)

- Allowing tenant to define set of service policy rules for the network slice operation
- Validating/authorizing the network slice design as well as service policies customised by the tenant
- Providing means to check and incorporate tenant's certified functions in the catalogue of VNF
- Allowing the tenant to order a limited set of network slice operation such as network slice/service activation and stop (this does not cover any network slice management operations such as NS scaling nor FCAPS)
- Exposure of monitored KPIs to the tenant (for changing policies settings)
- Allowing the tenant to change of service policies under network slice operation

- Interfaces between cloud InP-VIMs and MSP-MANO-layer functions:

  These interfaces are exposed by VIM(s) and consumed by SDM-O (or VNF managers) functions operated by the MSP. Authorization policies determine the access rights for the MSP. The interfaces cover the same requirements as for offer type 1.

3.2.2.2 Security considerations

All the security considerations from offer type 1 hold. As the tenant can perform much more activities via the interfaces provided by the MSP, the authorization policies may be more complex than in offer type 1. In addition, the attack surface of the MSP is larger due to the more complex interfaces. For example, software bugs or configuration mistakes may unintentionally open up MSP-internal management functions to tenants.

A tenant has not much better means to enforce any behaviour of the MSP than in offer type 1. A tenant could provide own code for a monitoring component to get a more independent monitoring, but as this component still runs under the control of the MSP and gets its inputs from functions under the control of the MSP, this would not help against fraudulent MSP behavior.

One notable additional consideration holds concerning the usage of code provided by the tenant: The MSP must ensure that such code does not cause harm, similar to the way the MSP checks "its own code" (i.e. the code the MSP has procured from a software vendor, compare discussion above). But in addition, the MSP can apply protection measures such as restricting the interactions of the tenant-provided code with the rest of the system to carefully monitored interfaces.

3.2.3 Offer Type 3: Extended slice configuration and control options for tenant

The scenario is representative of a tenant that operates the slice except for the common functions, operated by the mobile service provider). It reveals the full flexibility afforded by the 5G NORMA architecture to sustain a wide range of business options

3.2.3.1 Detailed description of the offer type

The tenant has his own subscribers and subscriber database for delivering its own application/service. The tenant manages his own subscribers via its own BSS/OSS. The tenant monitors the SLA through the interface provided by the mobile service provider, by accessing to monitored service KPIs. In this scenario, the tenant is offering an end-to-end service to its own subscribers (end users) by using a network slice instance composed of VNF instances provided by the mobile service provider in addition to its own VNF instances. The tenant designs the end-to-end network slice template by integrating sub-graph of VNF template exposed/provided by MSP. The exposition of sub network slice template reflects the agreement between the tenant and the MSP. The tenant has its own MANO/EM/SDM entities and fully manages/orchestrates its slice instance which has been setup and is still controlled by the MSP. The tenant sends a request to the MSP to orchestrate/control the slice according to his needs. The scope of orchestration requests from the

tenant to the MSP is defined/controlled by the MSP. Typically, it covers the request to start and stop the slice instance, and possibly to ask for an authorised change in QoS/QoE or in allocated radio resource. The MSP fully activates and operates the network slice instance by taking into account the orchestration/control request from the tenant. A direct interface between MANO/SDM entities might not be authorised. In such a situation, any control/orchestration request from tenant's entities on slice of the MSP will go through a business-to-business gateway (B2B-GW).

It is assumed that the MNO owns the full infrastructure, including the datacentres of the central cloud. The MNO is responsible for infrastructure resources and owns the VIMs. It provides an interface to the VIM(s) for the tenant to access information on the available provisioned resources and to execute the orchestration of the slice instance according to the decisions of the tenant's NFV orchestrator.

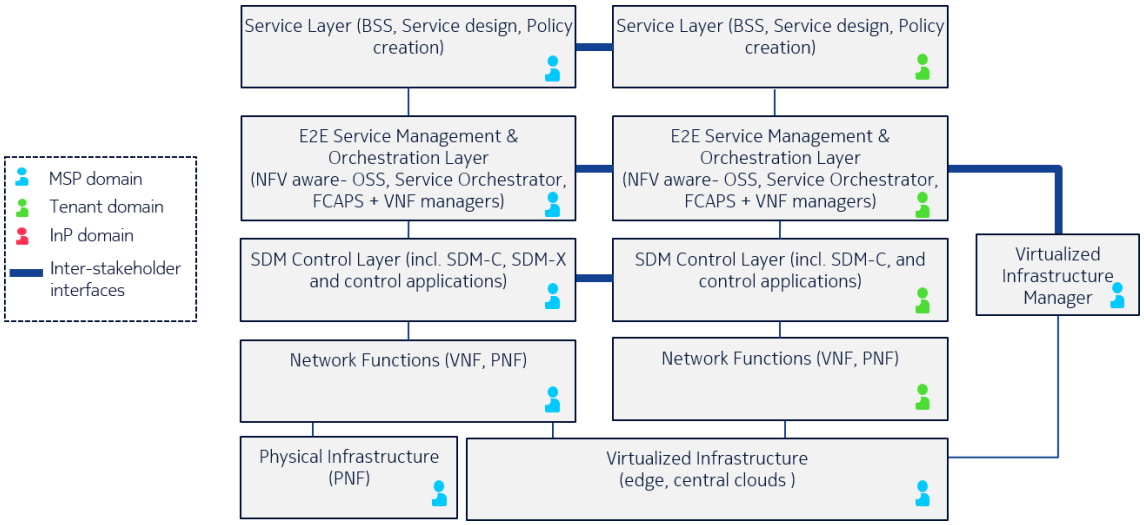


Figure 3-8: Domain ownership of high level functional blocks and relevant cross-domain interfaces in offer type 3

Most relevant cross domain interfaces are depicted in Figure 3-8.

- Interface between the service layer functions from the MSP and service layer functions from the tenant

  These interfaces are produced by the MSP service layer functions and consumed by the tenant's service layer functions. Authorization policies determine the access rights for the tenant. The interface is used for defining the service level agreement and the required service KPIs. The interfaces cover the same requirements as for offer type 1. In addition, it requires provision of credentials to access to the MSP platforms for sub-network slice template sharing and for orchestration/control operation (e.g. providing a URL to the MSP platform and a tenant account/password). In contrast to offer type 1, charging/billing report are not used.

- Interfaces between Service Management & Orchestration layer from the MSP and Service Management & Orchestration layer from the tenant.

  These interfaces are exposed by the Service Management & Orchestration layer from MSP and consumed by Service Management & Orchestration layer functions from the tenant. Authorization policies determine the access rights for the tenant. The interfaces are used for:

  - Exposure and sharing of slice templates
  - Allowing the tenant to order a limited set of network slice operation such as network slice activation/stop/scaling (granted by the MSP Service Management)

- Interfaces between the SDM control layers from MSP and tenant.

These interfaces are exposed by the SDM control layer from MSP and consumed by SDM control layer functions from the tenant. Authorization policies determines the access rights for the tenant. The interfaces cover the following requirements:

- Allowing to interconnect VNFs across domains
- Allowing the tenant to enforce QoS/QoE change under the network slice operation (granted by the MSP Service Management)

- Interfaces between MSP-VIMs and Management & Orchestration Layer functions from the tenant:

  These interfaces are exposed by the VIMs from MSP and consumed by the SDM-O (or VNF managers) functions operated by the tenant. Authorization policies determine the access rights for the tenant. The interfaces cover the same requirements as those described in offer types 1 and 2 for interfaces between InP-VIMs and MSP-SDM-O.

3.2.3.2 Security considerations

Compared to offer types 1 and 2, the tenant has much more control of its network slice. For example, the tenant may operate its own subscriber database, so the MSP will not see the data of the tenant's subscribers. (Still, the infrastructure provider, which could be the MSP, will see the data.) The tenant can to some extent monitor the service provided by the common functions of the MSP. The tenant has a better view on what cloud resources he gets. Depending on the split between MSP functions and tenant functions, most likely the tenant still needs to trust the MSP and cannot enforce every aspect of the SLA.

Like in offer types 1 and 2, any interfaces provided by the MSP to the tenant on the business service layer and on the network management layer need to be secured. Also, interfaces to the VIMs must be secured as described above.

The MSP needs to restrict the interactions of the tenant's network functions with the rest of the system to carefully secured interfaces, e.g. on the SDM-C layer. The MSP can monitor, control and limit every request from the tenant's functions to the MSP's functions and to the MSP-provided VIM interface. The MSP needs to secure these interfaces properly, including authentication and authorization of tenants.

To summarise, the MSP has in this scenario an increased attack surface because of the tighter interaction of MSP network functions and tenant network functions, but the MSP can take suitable measures to secure its network and services against malicious behaviour of tenants. Reversely, tenants have a better view of the MSP provided resources and functions, but still need to trust the MSP (in its role as InP) with respect to correct resource assignment, keeping tenant data secret, and maintaining integrity of tenant data.

4 5G NORMA Architecture

This chapter presents the design of the 5G NORMA mobile network architecture after the second design iteration. The three introductory sections depict a high-level description of the different architecture perspectives, introduce how the combination of different technologies enable end-to-end network slicing, and elaborate on the impact of PNFs and accelerators on the 5G NOMRA architecture. Subsequently, the integrated control and data layer architecture as well as the management & orchestration functionality is described. The final section in detail depicts how the software-defined mobile network control (SDMC) concept is realized, utilizing the example of Mobility management and QoS/QoE control.

4.1 High-level architecture

The 5G NORMA high-level architecture has been defined already in [5GN-D31] and is briefly summarised below. Four perspectives have proven useful to describe the 5G NORMA architecture:

- The functional perspective, showing the functional blocks needed to provide the user data service as such, as well as the related control and management & orchestration functionality.
- Three non-functional perspectives showing different aspects of the underlying infrastructure.

4.1.1 Functional perspective

The functional perspective of the 5G NORMA architecture, depicted in Figure 4-1, defines the architectural elements that deliver the system's functionality. It includes the key functional elements, their responsibilities, the interfaces exposed, and the interactions between them.

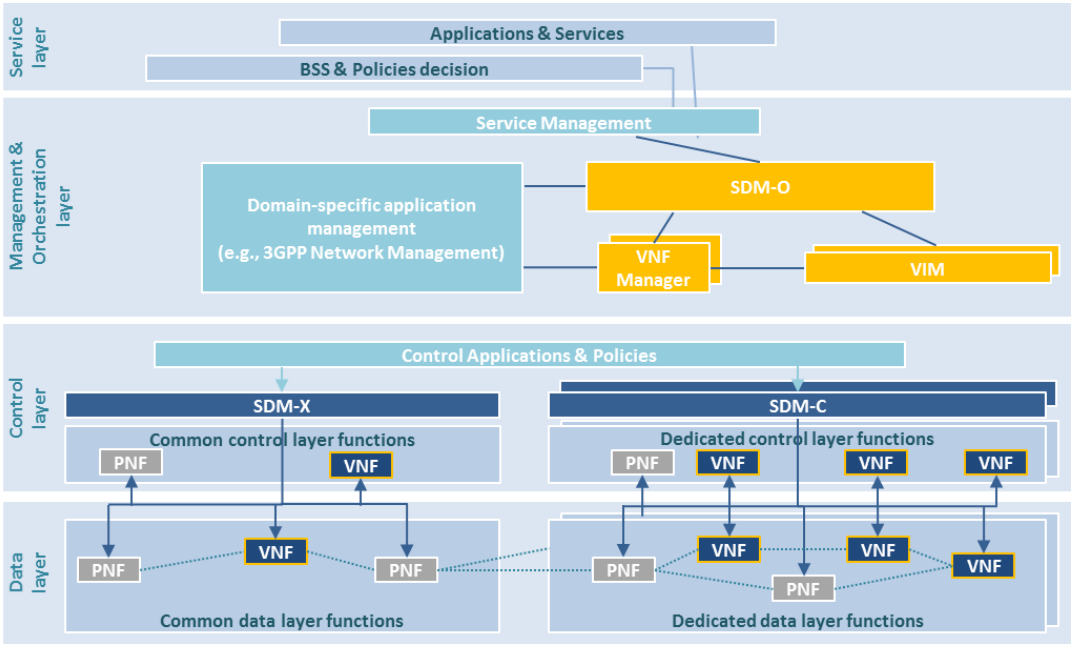


Figure 4-1: Functional perspective of the 5G NORMA architecture

As can be seen, the figure shows four layers of interaction:

- The **service layer** comprises Business Support Systems (BSSs) and business-level Policy and Decision functions as well as applications and services operated by the tenant.
- The **management and orchestration layer** includes 5G NORMA's MANO functions, the VIM, the VNF Manager and the SDM-O, which is further split as follows: an Inter-

slice Resource Broker for cross-slice resource allocation and slice-specific NFV Orchestrator(s). Further, it accommodates domain-specific application management functions. E.g., in the case of 3GPP, this comprises Element Managers (EM) and Network Management (NM) functions which would also implement ETSI NFV MANO interfaces to the VNF Manager and the NFVO. The Service Management is an intermediary function between the service layer and the SDM-O. It transforms consumer-facing service descriptions into resource-facing service descriptions and vice versa. This layer is further explained in Section 4.5.

- The **control layer** accommodates the two main controllers, SDM-X and SDM-C, as well as other control applications. Following the SDN principles, SDM-X and SDM-C translate decisions of the control applications into commands to VNFs and PNFs. SDM-X and SDM-C as well as other control applications can be executed as VNFs or PNFs themselves. Details are provided in Section 4.4
- Finally, the **data layer** comprises the VNFs and PNFs needed to carry and process the user data traffic.

4.1.2 Non-functional perspectives

Function blocks are deployed to different locations in the network, taking into account the different types of required resources as well as well as their geographical availability, i.e. the topology of the network infrastructure.

In [5GN-D31], deployment possibilities, available resources and network topology have been captured by three different non-functional perspectives. Figure 4-2 shows these three non-functional perspectives and how they are related to each other.

4.1.2.1 Deployment perspective

The deployment perspective illustrates one of 5G NORMA's key innovations: The adaptive allocation of network functions depending on the service requirements and deployment needs. It depicts the different possible locations of functional blocks and the mapping of functional blocks to the location in which a block is intended to run. This also includes the possibility that a functional block may be deployed in different locations. In addition, network slices may be represented in this perspective showing different deployment locations of functional blocks based on the service provided in each slice.

The deployment perspective is an abstract representation that shows how different functional blocks with complex runtime dependencies or complex runtime environments have to be placed in one location instead of another (e.g., particular functional blocks need to be placed in the edge cloud or be distributed over a specific number of virtual machines). The deployment view shows neither interfaces nor mapping of functions to layers.

In Figure 4-2, four different deployment types are distinguished: i) (non-virtualised) PNFs characterised by a tight coupling of software and hardware, ii) an edge cloud co-located with the antenna site, iii) an edge cloud within the (access) network, e.g., at an aggregation site, and iv) a central cloud.

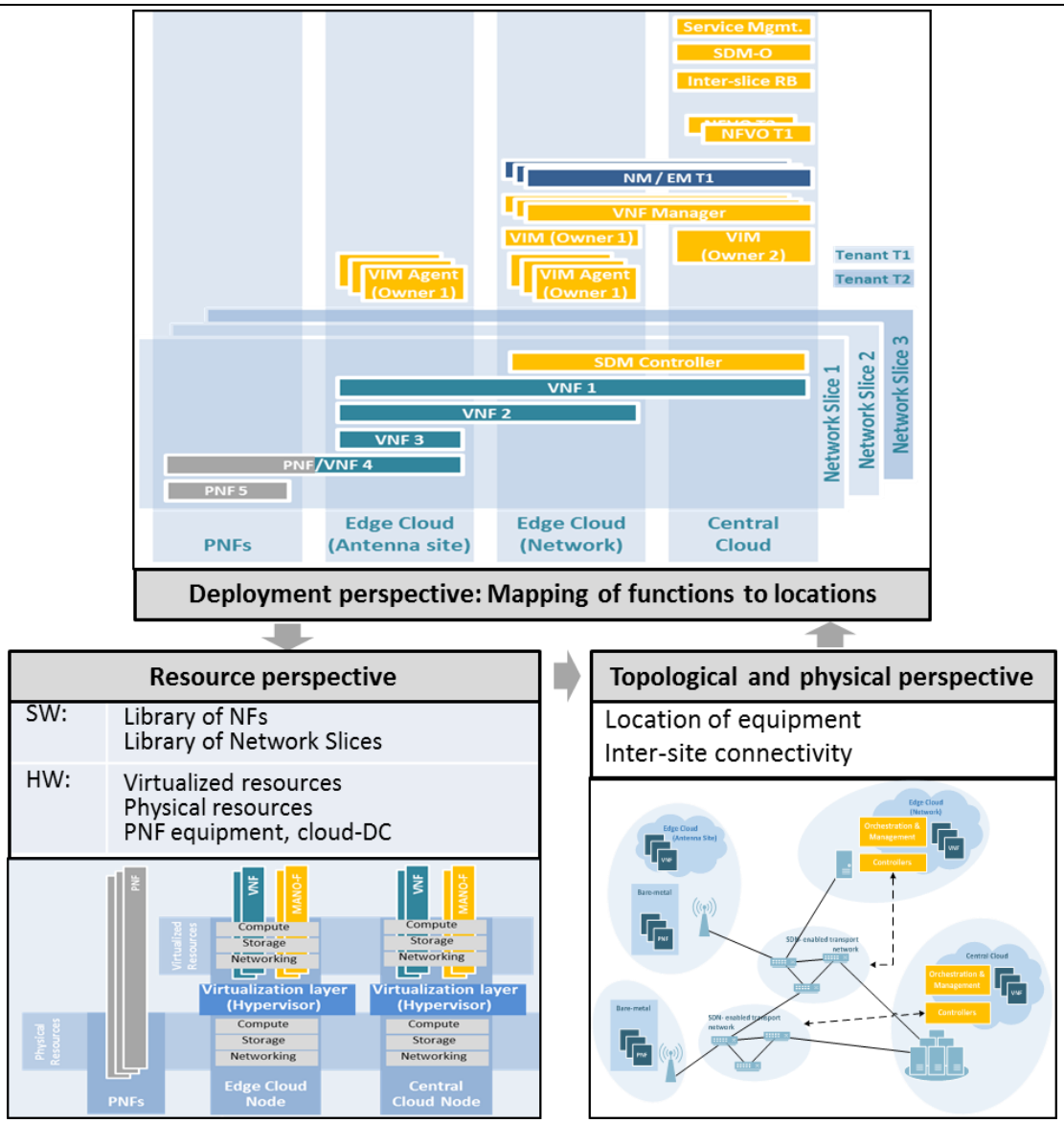


Figure 4-2: Non-functional perspectives of the 5G NORMA architecture

The example shown in Figure 4-2 depicts a single service provider utilizing resources from two infrastructure providers (Owner 1 and Owner 2) and providing services for two different tenants (T1 and T2). Owner 1 provides antenna sites and network with PNF-type nodes and two classes of edge clouds, one co-located with the antenna sites providing minimal latency towards the user terminals and edge clouds within the (access) network. It uses VIM Agents to scale its logical VIM entity to manage its large distributed infrastructure. On the other hand, Infrastructure Owner 2 operates the central cloud only and does not use VIM Agents. Tenant T1 uses two slices to implement its services while tenant T2 only uses one slice for all its services. As a rule, there is one NFVO (as part of the SDM-O) per tenant (operated by the tenant itself)) and a single Inter-slice Resource Broker operated by the mobile service provider.

4.1.2.2 Resource perspective

The resource perspective describes the different categories of infrastructure resources that network management and orchestration entities make use of in order to compose mobile network instances (network slices) for different use cases and tenants. Furthermore, the perspective shows the locations where these resources are provided.

HW and physical resources considered in 5G NORMA include both general purpose and specialised hardware that comprise memory, compute, storage, networking, and other fundamental

capabilities as well as radio spectrum. These HW resources can be made available for either virtualised functions (VNF) or non-virtualised functions (i.e., PNF).

Network Functions: The library of NFs comprises all executable VNF packages including the necessary template and metadata (e.g., resource requirements, supported interfaces, reference points, orchestration, and configuration parameters). It thus supports the creation and management of a VNF via interfaces exposed to other management and orchestration entities. Additionally, the library includes a repository of PNFs that can be orchestrated to be incorporated into a network slice. Similar to VNFs, this includes PNF metadata, such as, PNF location, connectivity to other NFs, performance limits (e.g., capacity), configuration parameters, or sharing and prioritization rules.

Network Slices: The library of network slices comprises descriptions of all executable network slice templates including the necessary metadata such as QoS parameters. A network slice template refers to the set of VNFs and PNFs that should be chained to implement the network service, as well as the NF-FG (network functions forwarding graph) that specifies how these NFs have to be interconnected in order to provide the service properly. In general, a network slice blueprint could contain network service descriptors, link descriptors, and connectivity descriptors.

4.1.2.3 Topological perspective

The topological perspective depicts how the network infrastructure is distributed geographically and thereby includes the notion of distance and associated latency characteristics, which in 5G NORMA determines the main difference between edge and central cloud. The central cloud typically comprises multiple data centres, which may be several hundred kilometres apart and connected through a wide area network (WAN). The WAN also connects the data centres of the central cloud to the data centres of the edge cloud. Considering the RAN requirements and due to a possibly strict delay budget, the maximum distance between edge and central cloud is limited. Furthermore, the topological perspective may also depict bandwidth and latency of transport media between distinct instances of resources in contrast to the deployment view that shows only the generalised class of resources such as edge and central cloud.

4.2 Resource abstraction for E2E network slicing

As explained in Section 2, 5G NORMA uses end-to-end network slicing to provide for multi-tenancy and multi-service capability, i.e., tenant- or service-specific function-chains and slice topology on a common infrastructure resource pool.

The typical enabling technology to create network slices in the core network (CN) domain is virtualisation. In our opinion, network slicing is not limited to those network elements that can be implemented on virtualisable x86 server HW. Instead, network slicing is applicable also to NFs that are executed on DSPs, HW accelerators, and other special purpose HW. In particular, [5GN-D41] explains how RAN functions that greatly benefit from such kinds of HW, can be sliced as well. The key enabler is the understanding that more mechanisms are available for resource sharing than merely the virtualisation of x86 servers.

In the following, we will look at multitasking, virtualisation, and multiplexing in more detail and show that by combining them in an end-to-end abstraction layer, slicing of the whole e2e network is feasible.

4.2.1 Methods for resource sharing

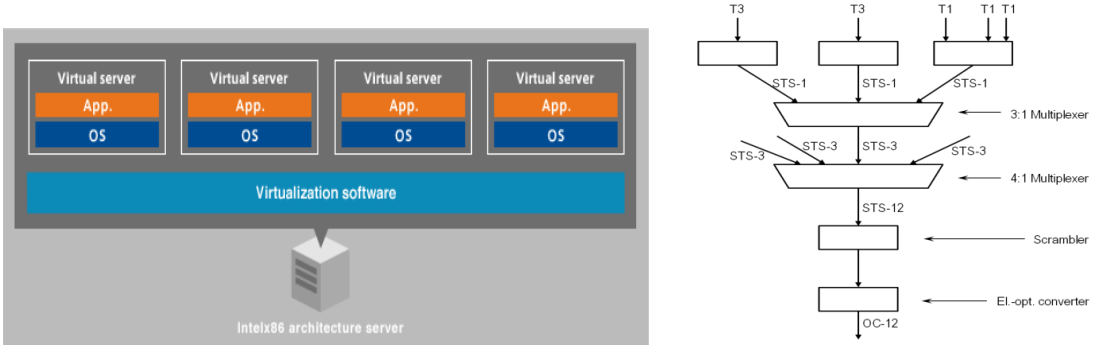
4.2.1.1 Multitasking

In computing, multitasking is a concept of performing multiple tasks over a certain period of time by executing them concurrently. This does not necessarily mean that multiple tasks are executed simultaneously. E.g., in case of a computer with a single CPU, the CPU is actively executing instructions only for one task at any point in time, while other tasks have to wait. Multitasking

solves the problem by scheduling which task may be the one running at any given time, and when another waiting task gets a turn [WP-CMT]. Splitting a resource and assigning it to multiple users is a common principle in virtualisation as well as multiplexing.

4.2.1.2 Virtualisation

Many definitions for virtualisation can be found in the literature, see e.g. [WP-Virt], [NEC], [IBM], [KT]. Basically, *virtualisation* refers to the creation of a virtual machine that acts like a real computer [WP-Virt], or in other words, a virtual computer is a logical representation of a computer in SW [IBM]. The software or firmware that maps virtual machines on the host hardware is called a *hypervisor* [WP-Virt].



(a) (b)

Figure 4-3: a) Virtualisation [NEC] and b) Multiplexing [Tan]

With respect to the 5G NORMA architecture, virtualisation has two important properties:

1. Abstraction of functionality from the underlying execution environment:
   As shown in Figure 4-3a), the virtualisation SW and the Operating System (OS) running in the virtual machine (VM) decouples the Application from the physical server HW. In this way, application functionality becomes independent from the underlying HW. (The possibility that each VM runs its own OS simplifies the adaptation of applications to the underlying HW. While this is highly beneficial in practice, it is not mandatory for creating end-to-end network slices.)
2. Resource partitioning:
   By virtualisation, the resources of a computer are divided into multiple, isolated execution environments that can be assigned to different tenants or services. This aspect is indicated in Figure 4-3a) by the four virtual servers running in parallel on top of the same HW server.

4.2.1.3 Multiplexing

Multiplexing emerged in communications already around 1870. Originally, it means the combining of multiple signals into one combined signal that can be carried over a shared medium [WP-Mux]. Sometimes the term multiplexing is used in a more general sense, e.g. when a display with many pixels is controlled through a much smaller number of lines. Like virtualisation, multiplexing aims for partitioning and sharing an expensive resource among multiple signals or services. Since the resource utilisation depends on multiple signals or services, it is not possible to conclude from a single signal or service onto the state of the resource in total. This can be interpreted in the sense that the functionality of the signals, namely to transport information, is separated from the resource needed to execute this transportation.

4.2.2 E2E abstraction by integration of virtualisation and multiplexing

In the previous section, it has been shown that virtualisation, multitasking and multiplexing enable the sharing of resources between multiple users by i) decoupling the functionality from the resources needed to execute this functionality, and ii) partitioning of resources into isolated execution environments.

This joint property suggests combining hypervisors, multiplexers and multitasking mechanisms in a common abstraction layer. While hypervisors manage the resources of x86-based servers in the central cloud and the network edge cloud, multiplexers and multitasking mechanism perform the same task for other components, like DSPs and accelerators in the base stations and PNF nodes. In this way, partitioning can be applied to all components in the network, resulting in a network slicing from end-to-end.

As pointed out in [5GN-D41], multiplexing and multitasking can be applied on different levels of the ISO-OSI protocol stack, see Figure 4-4.

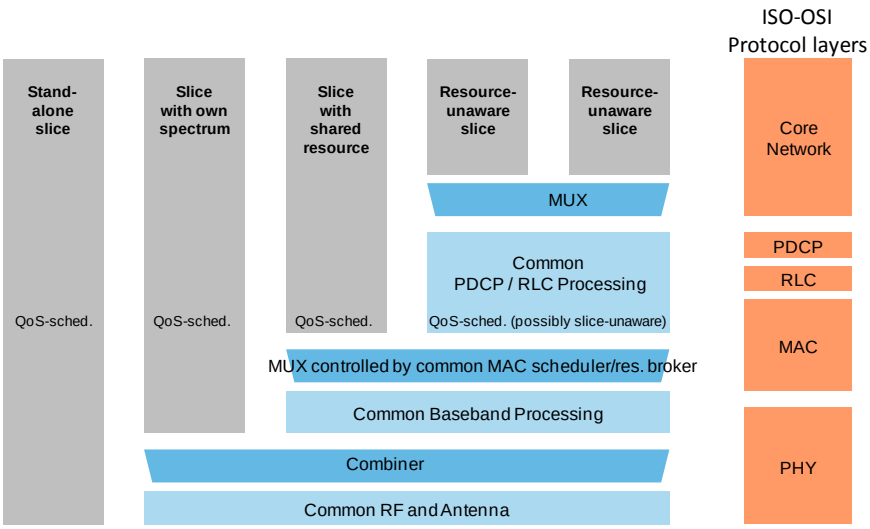


Figure 4-4: Options for slice multiplexing and their relation to the OSI protocol stack

This yields several options for the design of network slices, ranging from standalone slices with own HW and spectrum, to slices that are completely unaware of the resources they are using and hence have no (direct) control on the resource scheduling. The differences between these slice variants will be reflected by the templates of the respective network slices. Using these templates, the SDM-O can set up slices and merge them properly at the described multiplexing point.

Figure 4-5 (user traffic flows are shown from right to left) exemplarily depicts how virtualisation and multiplexing techniques can be integrated in an end-to-end network to implement the principles of resource abstraction and resource partitioning.

In the part above the abstraction layer, the figure shows user traffic flows (solid lines) from PGW (on the right side) through the eNB to the air interface (on the left side) for three different slices. Please note that other flows, e.g., signalling flows originating from the MME or management flows from SDM-O/X/C have been omitted for the sake of readability of the figure. Slices A and B are multiplexed on the MAC layer, thus considering logical channels as shared resources. The data flows of slices A and B are then merged with Slice C by combining signals on the physical layer.

At the bottom, the figure shows the execution environment, consisting of classical IT resources as well as transport network components like optical fibres and RAN components like DSPs, RF HW, and radio spectrum as the most precious resource.

The abstraction of functionality from the underlying resources is achieved by mechanisms that fit to the resources: E.g. hypervisors or simply the OS for CPUs and servers, schedulers for the transmission resources of logical channels, etc. The dotted lines show how incoming user traffic at a functional block triggers the corresponding processing on resource level.

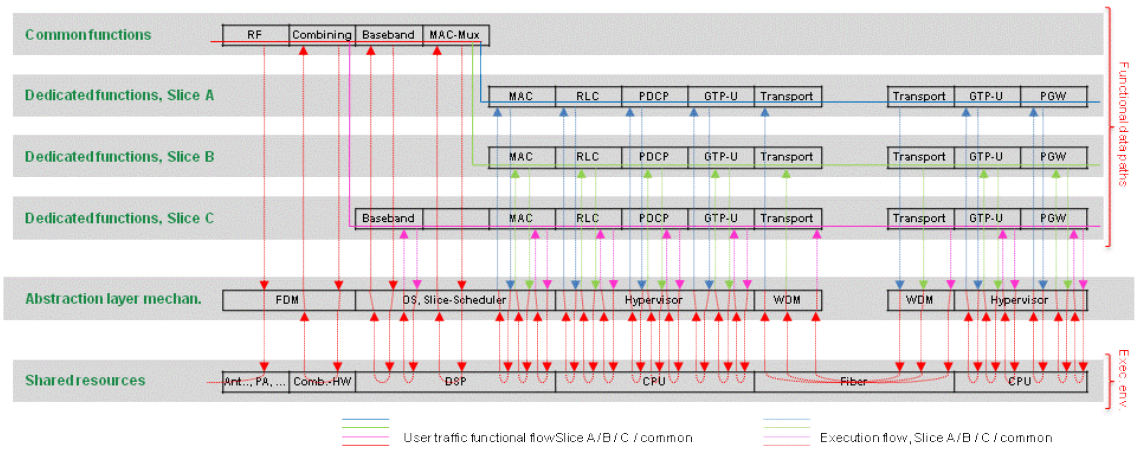


Figure 4-5: E2E network slicing by combining virtualisation and multiplexing

Summarizing Section 4.2, it can be concluded that

- Network slicing is technically feasible from end to end and not limited to certain parts of a network, e.g. the CN.
- The key enabler is the understanding that resource-splitting and multi-tasking are provided not only by virtualisation, but also multiplexing techniques.
- All necessary tools for end-to-end network slicing, namely multiplexing and multitasking, are well known.

Complementing virtualisation by comparable multi-tasking techniques allows for applying network slicing in all network domains, in particular in the RAN. It makes network slicing more applicable in practice and opens migration paths from existing networks to sliced 5G networks.

However the integration of virtualised VNFs and multiplexed PNFs in a single architecture requires that also the corresponding control and management functions are integrated. Furthermore it has turned out that even in an E2E sliced network some functions must be common to multiple slices and cannot be dedicated to a single slice. These challenges will be addressed in Sections 4.4, 4.5 and 4.6 below.

4.3 PNFs in the 5G NORMA architecture

In this section, the role of PNFs and accelerators beside the Virtual Network Functions (VNFs) in 5G NORMA architecture is analysed. As defined in [5GN-D31], PNFs are NFs that exhibit a tight coupling between hardware (HW) and software (SW) systems, which are not flexible. Since PNFs are very much tailored for executing a specific function, they can achieve very high performance, both in terms of latency and computational efficiency, while at the same time being extremely energy-efficient. Examples for dedicated HW include DSPs, FPGAs, ASICs, etc. in which only a very specific part of the NF can be executed. On the other hand, VNFs decouple software from general purpose processing (GPP) hardware and enable a very wide range of possibilities. Furthermore, GPP platform software development is usually faster and portable. Hence, VNFs are building blocks that can be flexibly combined [5GN-D31]. Finally, GPP devices are cheaper than dedicated, function-specific hardware. Accelerators can be combined with VNF/GPP HW, enabling the reduction of the computational load of the GPP itself. Virtualisation comes at a cost, since it frequently introduces some issues, e.g., increased latency, rendering in inadequate for many real-time processing functions, or in terms of power consumption, see [EURECOM]. Latency also is a critical topic considering the emergence of URLLC services, and

consequently 5G NORMA aims to efficiently integrate accelerator technologies into the overall architecture.

In [EURECOM], the cost (in terms of power and latency of the LTE physical layer processing) of a virtualized environment running on GPP HW is compared with dedicated PNF HW. As reported in the paper, real-time functions have strict latency requirements that impose an upper bound for the total processing time of the entire processing chain. Typically, having GPP HW that is not optimised for physical layer signal processing means increasing the number of CPUs to achieve the real-time requirements. To avoid this problem, the processing of some critical functions, such as FFT/IFFT, can be offloaded to a dedicated hardware, i.e., the above-mentioned accelerator. This solution reduces the number of CPUs required to perform a certain task (when compared to GPP HW), thus also reducing the power consumption. In a PNF, every single function is run on the same chip, which also avoids additional latency introduced by hypervisors or for accessing other remote devices. In [EURECOM], some tests are done to compare the power consumption of a PNF, a full GPP system, and a GPP plus accelerator system. This study shows the power consumption per carrier considering different scenarios in an LTE system with 20 MHz channel bandwidth, e.g., varying number of allocated physical resource blocks (PRBs) and different modulation and coding scheme (MCS). The measurements yield a power consumption between 7 and 8 W per carrier for the PNF, between 13 and 18 W per carrier using a GPP plus accelerator system, and about 70 W per carrier for the full GPP system, see [EURECOM].

A similar study has been done on the Azcom hardware eNodeB, [AzENB], based on a DSP platform. According to the scenarios defined in [EURECOM], several tests have been done varying the following parameters in an LTE system with 20 MHz channel bandwidth: number of served users, number of allocated physical resource blocks (PRBs), and utilized modulation and coding scheme. The average measured power consumption of this PNF, for LTE physical layer processing, is between 9 and 11 W per carrier, in line with the study reported in [EURECOM].

On the one hand, the use of accelerators (DSP, FPGA, ASIC) in combination with VNF/GPP HW systems can improve the latency and power consumption of mere GPP-based systems, while not adversely affecting the flexibility introduced by virtualisation. On the other hand, purely PNF-based setups can further increase the efficiency by reducing power consumption and keeping latency within strict boundaries.

At this stage, PNFs continue to be a valid solution for services that have strict requirements in terms of power consumption and latency. In the 5G NORMA architecture, PNFs impact on several architecture design aspects beyond energy efficiency. At the highest level, the orchestrator (SDM-O), during the slice creation, shall select the best matching solution among VNFs, accelerator-supported VNFs, or pure PNFs, depending on the specific telecommunication services provided by a network slice. As described in Section 4.4, the 5G NORMA architecture puts these three variants under the unified control of the Software-Defined Mobile Network Controller (SDM-C) or the Software Defined Mobile Network Coordinator (SDM-X), using a dedicated interface, cf. Figure 4-6. Finally, like VNFs, also PNFs need a lifecycle management entity, that is the Element Manager (EM) which performs FCAPS (fault, configuration, accounting, performance, security) operations. EM is a component of today's 4G network architecture but in 5G NORMA it becomes a VNF-aware entity interconnected also to VNF-M. Moreover, the 5G NORMA architecture allows to merge EM and VNF-M into a single 5G NORMA MANO layer entity, which is responsible for setting up both PNFs and VNFs, reducing the overhead in terms of number of interfaces.

4.4 Integrated control and data layer architecture

5G NORMA introduces a novel concept of network control by extending the software-defined routing (switching) approach to all kinds of mobile NFs from both data and control layer, with a focus on wireless control functions, such as, scheduling or interference control.

For this purpose, 5G NORMA defined controllers that apply the split between the *logic* of the network function and the part that can be controlled (*agent*), implemented by a network function.

Software-Defined Mobile Network Controller (SDM-C) and Software Defined Mobile Network Coordinator (SDM-X) take care of dedicated and shared NFs, respectively.

The SDM-C is designed to abstract technology-specific or implementation-specific aspects of the network ecosystem, with interfaces towards the MANO stack and to different Control Applications implementing, e.g., QoE/QoS control or mobility management [5GN-D51]. The overall extent of the controller is depicted in Figure 4-6.

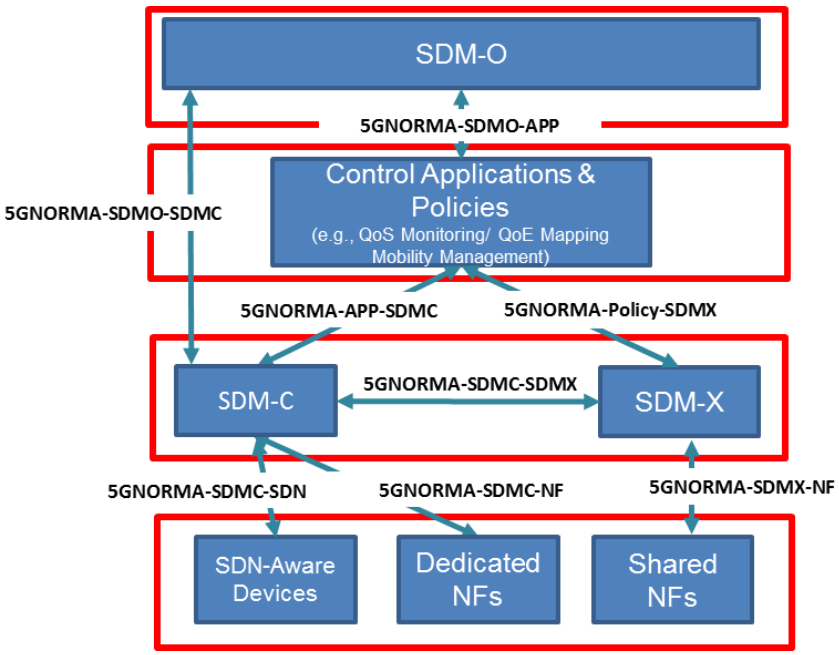


Figure 4-6: SDM-C northbound and southbound interfaces

Interface 5GNORMA-SDMO-APP. This interface is used to convey the Control Application-specific information derived during the translation from high-level tenant requests and established SLAs into the network slice resource provisioning, NFs logic, and lifecycle parameters. E.g., with respect to Mobility Management (MM) Application, this interface can convey the information about the most suitable MM scheme and corresponding network slice template with respect to the agreed SLA and service policies. Depending on the QoS service requirements attached to the network slice, a corresponding mapping onto latency, bandwidth, computing and storage requirements, QoS thresholds to monitor, etc. can be conveyed to the QoS/QoE Application via this interface.

Interface 5GNORMA-APP-SDMC. This interface is used to enforce the conditions defined by the Control Applications that have to be realized for a given traffic identifier on **dedicated** functions and resources in order to fulfil the targeted SLA. E.g. via this interface the MM Application can convey to SDM-C the exact network slice configuration (with right VNFs type selection and right composition and configuration of different VNFs) based on selected network slice blueprint. On the other hand, this interface can convey the information regarding the current slice performance which is reported back to the corresponding Control Application.

Interface 5GNORMA-Policy-SDMX. This interface is used to enforce the conditions defined by the Policies that have to be realized for a given traffic type on **shared** resources and functions in order to fulfil the targeted SLA with respect to the relevant service policy.

Interface 5G NORMA-SDMO-SDMC. In the case that the SLA targets cannot be met, the SDM-C can directly trigger the re-orchestration request on the SDM-O. For that purpose, the SDM-C uses the 5G NORMA-SDMO-SDMC interface.

While the VIM monitors the NFVI resource consumption of VNFs (i.e., CPU load, memory utilisation, networking utilisation), the SDM-C and the related SDM-C applications have knowledge about the VNF semantics and associated current target QoE and QoS constraints. The usual SDM-C application behaviour is related to the control of such VNFs (i.e., changing the VNF parameters)

to keep meeting these constraints. However, when this is not possible by changing VNF parameters anymore, a re-orchestration is needed. Therefore, the SDM-C may raise this request to the SDM-O to obtain a different and better orchestration of the network slice according to the changing conditions.

Interface 5GNORMA-SDMC-SDMX is used for interaction between the SDM-C and SDM-X controllers in a peer communication mode. Based on the resource management policies provided by the SDM-O, the negotiation between SDM-X and SDM-C is established to decide how to fulfil the demands of several partially competing network slices simultaneously. E.g. the SDM-X decides based on the SDM-O policies whether it is necessary or not to modify a network slice's shared resources upon a request coming from SDM-C.

Interface 5GNORMA-SDMC-NF. It controls and configures parameters of the dedicated P/VNFs which implement the NFs on the data path. The 5GNORMA-SDMC-NF interface is hence used to configure and control these (Physical or Virtual) Network Functions. A list of the available NF that can be controlled is listed in Section 4.4.1.1.

Interface 5GNORMA-SDMX-NF. This interface is used to configure and control specific, shared NFs, e.g. shared VNF instances, non-NFVI network resources, accelerators (as FPGA, ASIC, etc.), radio and transmission resources which are shared among different network slices. Some examples are provided in Section 4.4.1.2.

Interface 5GNORMA-SDMC-SDN. It controls SDN-compatible routers (forwarders), building the path(s) that connect the VNFs of a service chain. The interface is equivalent to the southbound interface of an ONF-type SDN controller.

Further considerations on the SDMC-related interfaces can be found in Section 4.4.4.

The c/d-layer architecture is depicted in Figure 4-7, here for the case of RAN slicing Option 2, i.e., with a common MAC layer (cf. Section 3.3 of [5GN-D41] for further details). Network functions are classified whether they belong to the control or data layer. The control layer functions are further classified into i) distributed, ii) common, and iii) dedicated control. Distributed control functions are implemented as VNFs throughout the network, while common and dedicated control functions employ the SDMC concept and run as applications on top of SDM-X and SDM-C, respectively. The depicted function blocks are detailed in the following subsections.

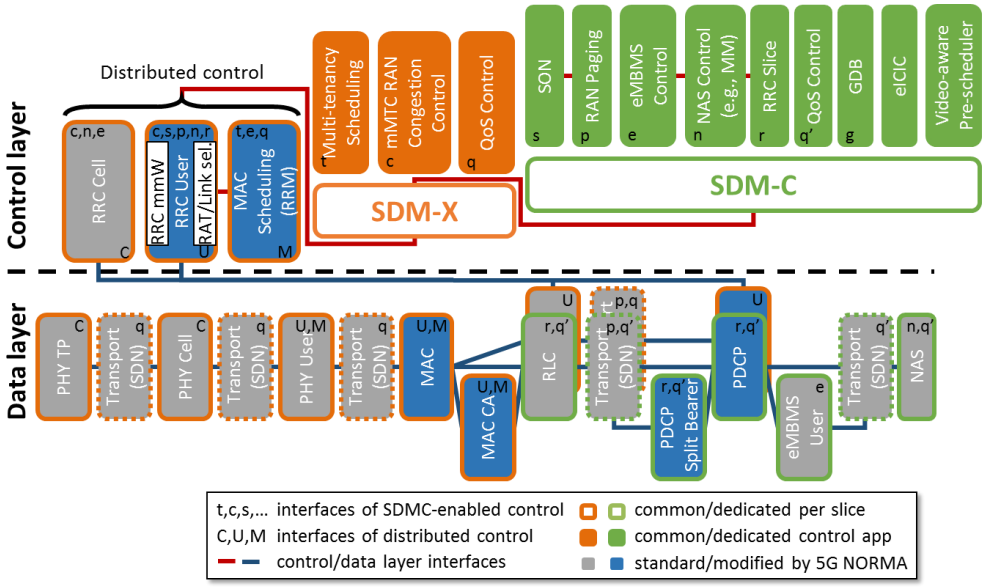


Figure 4-7: 5G NORMA control and data layer functional architecture

4.4.1 Centralised 5G NORMA control layer

4.4.1.1 Software-defined mobile network controller (SDM-C): Control of dedicated functions and resources

Network functions dedicated to a specific network slice are under exclusive control of the slice's own SDM-C. The SDMC concept foresees that the control logic is implemented as part of the an SDMC application, while the controlled functions is part of the so-called agent. This logic has been applied to the following NFs (cf. Figure 4-7):

- **SON.** It covers *i*) Self-configuration, *ii*) Self-Optimisation, *iii*) Self-Healing, and *iv)* User Centric Connection Area (UCA) and mm-wave cluster configuration.
- **RAN Paging.** To reduce signalling messages on the air interface and towards the CN, 5G NORMA employs a RAN paging approach, i.e. inside a User Centric Connection Area (UCA), in addition to paging in a larger tracking area. Further, it sets up and updates connectionless transmission inside a UCA.
- **eMBMS Control.** Performs the admission control and allocation of the radio resources, UE counting procedure, MBMS session management (initiating the MBMS session start and stop procedures), allocation of an identity and the specification of QoS parameters associated with each MBMS session.
- **NAS Control.** It includes the sub blocks
    - **NAS UE specific.** Refers to the user specific NFs and procedures related to the data layer, which are triggered by the NAS UE-specific control layer functions. For example, a mobility management (MM) application that controls the behaviour of gateways following a software defined principle
    - **NAS UE Specific and Data Layer.** Refers to the user specific functions and procedures related to the d-layer, which are triggered by the NAS UE-specific c-layer functions.
    - **NAS UE Specific and Control Layer.** Refers to the user specific functions and procedures related to the non-radio signalling between the UE and MME.
    - **NAS Event-Control Layer.** Refers to the network-side c-layer functions and procedures including those of the interface between RAN and CN (S1-C in 3GPP LTE [36.300]), which are provided to facilitate mobility management.
- **RRC Slice.** Handles the UE management and control related to the slice specific part of the RAN protocol stack.
- **GDB.** Geolocation Database (GDB) stores information linked to geolocation, and makes decisions based on that geolocation information.
- **eICIC.** In 3GPP LTE, eICIC specifically means inter-cell- interference coordination in time domain through almost blank subframes. In 5G NORMA, the function performs interference control by favouring selected flows (e.g., video flows, cf. [sGN-D51]), exploiting their traffic characteristics.
- **Video Aware Pre-Scheduler.** Steers the scheduling pattern to optimise delivery of video flows. According to the considered slicing option there are implications of SDM-X.

The SDMC applications *MM*, *eICIC* and *Video Aware Pre-Scheduler* are described in more detail in [5GN-D51]. Further details on all other SDM-C applications can be found in Part I, Annex A, Sections 6.18 to 6.23 of [5GN-D41].

4.4.1.2 Software-defined mobile network coordinator (SDM-X): Control of shared functions and resources

Network functions that are shared by multiple network slices are put under control of a single SDM-X. It coordinates the control information coming from the different slices' SDM-Cs via the 5GNORMA-SDMC-SDMX interface. Through this interface, each SDM-C controls those network functions that is are shared with other slices up to the extent exposed by the SDM-X through this interface. For this purpose, the SDM-X needs to execute specific rules to be able to make meaningful coordination decisions. For example, SDM-X must authorise the requests of SDM-Cs and reject requests that are not in line with the SLA between MSP and tenant. SDM-X Apps include according policies and can provide such rules via the 5GNORMA-APP-SDMX interface. to generate a single meaningful outcome from the various SDMCs' requests coming in through 5GN-SDMC-SDMX interface before being forwarded towards the agents in the data layer or distributed control layer.

Hence, the SDM-X runs application that exclusively control shared network functions and resources. These applications are run by the stakeholder that determines the set of common NFs, i.e., usually the mobile service provider. These SDM-X applications are (cf. Figure 4-7):

- **Multi-tenancy Scheduling.** Coordinating resource sharing among multiple tenants. It includes functions such as scheduling and ICIC schemes.
- **mMTC RAN Congestion Control.** Grouping mMTC devices into context-based clusters and schedule their RAN procedures in sub-frames, to reduce the RAN congestion rate.
- **QoS Control.** Network monitoring and configuration in real time through open interfaces that interact with the SDM-X and the control radio stack.

A more detailed description of these SDM-X applications can be found in [5GN-D41], Part I, Annex A, Sections 6.15 to 6.17.

4.4.2 Distributed 5G NORMA control layer

Distributed control NFs implement control logic that is too time critical and/or is deemed to be not efficiently implementable in a more centralised way as SDMC application, for example because it would incur too much signalling over the SBI between antenna location and location of that runs the SDM-X. As of now, the radio scheduler *MAC Scheduling (RRM)* is carried out as distributed control NF, as well as RRC NFs *RRC Cell* and *RRC User* to enable fast reconfigurations triggered by the radio scheduler to adapt to the time variant radio channel and interference, without the need for a possibly time consuming detour via SDM-X:

- **RRC Cell.** Handles control plane signalling protocols associated with broadcasting system information, including NAS common information and information relevant to UEs in RRC_IDLE, e.g., cell (re-)selection parameters, neighbouring cell information, and information (also) applicable for UEs in RRC_CONNECTED, e.g., common channel configuration information.
- **RRC User.** Handles the UE management and control including radio bearer setup. It includes the subblocks
    - **RRC mmW.** It adds functionality related to mm-wave transmission points controlled by 5G coverage cell and UCA for short data packet transmission.
    - **RAT/Link Selection.** Enables link selection and packet scheduling if the UE is simultaneously connected to two or more RATs.
- **MAC Scheduling (RRM).** Scheduling the transfer of user data and control signalling in DL and UL subframes over the air interface.

The distributed c-layer NFs are further detailed in [5GN-D41] Part I Annex A Sections 6.12 to 6.14.

4.4.3 5G NORMA data layer

Data layer network functions are inherently distributed due to their purpose of providing data forwarding end-to-end. D-layer NFs are controlled either by distributed c-layer NFs or by SDMC applications running on SDM-X or SDM-C. Whether a data layer NF's control logic runs on SDM-X or SDM-C depends on the deployed RAN slicing option. For example, for RAN slicing Option 2 [5GN-D41] depicted in Figure 4-7, all data layer NFs up to the common *MAC* NF are SDM-X-controlled (however, via distributed "legacy" control functions), while *RLC* and above are dedicated per slice and accordingly SDM-C-controlled.

The only exception is *Transport (SDN)*, which extends the classical SDN forwarding plane and accordingly is controlled through interface 5GNORMA-SDMC-SDN. The 5G NORMA functionally decomposed data layer consists of the following NFs:

- **PHY Transmission Point.** The analogue and mixed signal processing for all signals transmitted (received) via one transmission (reception) point.
- **PHY Cell.** (De-)multiplexing of *PHY User* and baseband signal generation including common PHY signals for one RAT or slice.
- **PHY User.** The generation of the baseband signal (in frequency domain for OFDM-based systems) from user data (DL) and decoding of baseband signals into user data (UL), respectively.
- **MAC.** Provides functionalities such as HARQ, adaptive modulation and coding (AMC) (allow to adapt the modulation and coding to the channel quality), and discontinuous reception (DRX) (allow to improve UE battery life and energy saving).
- **MAC Carrier Aggregation.** Coordinates the exchange of scheduling information as well as feedback information corresponding to the aggregated legs.
- **RLC.** Transparent mode (TM) is dedicated to forwarding RRC information to/from lower layers (RRC messages are passed unmodified to the MAC layer). Acknowledged mode (AM) is dedicated to providing an additional re-transmission process.
- **PDCP Split Bearer.** Executes the functionalities of routing, reordering, and reordering timer
- **PDCP.** Carries out data transfer functions including functions sequence number maintenance, RoHC, (de-)ciphering, integrity protection, and verification.
- **eMBMS.** Performs the transmission of MBMS application data using the IP multicast address with the addition of SYNC protocol to guarantee that radio interface transmissions stay synchronised. Multimedia Broadcast Multicast Services (MBMS) offer support for broadcast and multicast services enabling the transmission of multimedia content (text, pictures, audio and video) and utilizing the available bandwidth intelligently [23.246].
- **NAS.** Performs routing functionality (S-GW, P-GW) and service delivery.
- **Transport (SDN).** Connectionless routing within RAN based on SDN configuration. Within an UCA, a connectionless routing of DL small data packets from the anchor node to the best serving node (and vice versa in UL) has to be established for each UE.

All d-layer NFs are described in more detail in [5GN-D41], Part I Annex A, Sec. 6.1 to 6.11.

4.4.4 Considerations on interfaces between control and data layer

As pointed out above, the separation of control and execution of network functions and the centralisation of the control parts in SDM-C and SDM-X is a major achievement in 5G NORMA. As stated in Section 4.4.1, this separation can be practiced for many different network functions.

The separation of control and execution parts of a network function implies that both parts are connected through an appropriate interface that is able to carry.

- commands from the control part to the execution part,
- acknowledgements to these commands back from the execution part to the control part,
- indications, measurements and status reports from the execution part to the control part.

This kind of interface is shown as Southbound Interfaces (SBIs) in Figure 4-6 with the names 5GNORMA-SDMC-NF and 5GNORMA-SDMC-SDN.

The network functions mentioned in Section 4.4.1.1 differ significantly from each other. Therefore, it is near at hand that their SBIs will require substantially different capabilities. Alternatively, all these interfaces could be bundled in a single southbound interface for the SDM-C. However, this SBI might become very feature-rich and complex.

The properties of these network functions and their requirements on the SBI could not yet be investigated in detail in 5G NORMA. However, some network functions exist for which possibilities to split control and execution parts have been discussed already in literature and for which suitable interfaces have been described:

- **Separation between RRC from lower layer RAN protocols:** A similar split has been practiced already in UMTS, where the RRC control decisions were made in the Radio Network Controller (RNC), while lower layer radio functions like MAC and PHY were executed in the Node B. Communication between RNC and Node B used the NBAP protocol.
- **Mobility Management:** Aside the mobility management schemes standardised for LTE, other schemes like PMIP, VertFor, OFNC and LIME have been discussed in the literature. For some of them e.g. the OpenFlow protocol would be suitable to separate the controller from the router that redirects data packets when mobile terminals move.
- **Routers in the transport network:** Routers in the transport network are the "classical SDN device". Hence, for those the OpenFlow protocol would be suitable as well.

The above examples show that the properties of the southbound interfaces can vary for different network functions. It remains to be seen to which extent the requirements of different network functions on these interfaces can be homogenized. Ideally, all network functions could be served through a single common interface. If such an interface will become too feature-rich and complex, an alternative might be to define several interfaces for groups of network functions with similar requirements. The design of the 5GNORMA-SDMC-NF and 5GNORMA-SDMC-SDN interfaces is thus an open research topic in 5G NORMA.

4.5 SW-defined mobile network management and orchestration

4.5.1 5G NORMA management and orchestration layer

4.5.1.1 Overview of objectives and concepts

The 5GN Management and Orchestration Layer is based on the ETSI NFV MANO framework as described in Annex A.1. One of the two key 5G NORMA innovative functionalities already described in the initial project proposal (Section 1.3.1.2 of the 5G NORMA Proposal Document) [5GN-DoW] is to allow the creation and the dynamic life-cycle management of different network slices. This is very closely related to the management of network services for diverse customer sectors and with different requirements (i.e. diverse vertical industries, mobile network operators, etc.).

The 5G NORMA resource management and orchestration solution shall provide the possibility to independently perform management and orchestration tasks to the following entities of the 5G NORMA ecosystem (cf. Section):

- the infrastructure provider,
- the mobile service provider, and
- the tenant.

The management and orchestration of different network slices with potentially very different requirements represents a big challenge for the 5G mobile service providers, since they will need to deal with a high level of complexity to meet diverse requirements from various customer segments (i.e. MBB, M2M, IoT applications, etc.). In order to address this challenge, 5G NORMA designs a network with a high degree of automation. This approach will allow managing the available resources on-demand with enhanced flexibility and with scarce human intervention. That way we provide a flexible and scalable means of selecting, controlling and deploying the necessary virtualised network functions from different mobile service providers in different network slices.

This is equivalent to use case #3 "Virtualised Network Platform as a Service (VNPaaS)" described in [NFV-UC] and depicted in Figure 4-8.

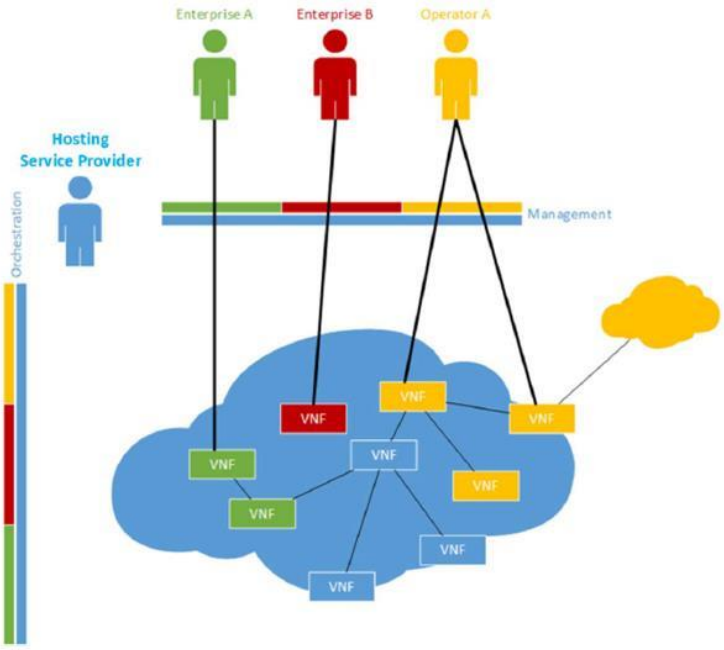


Figure 4-8: Multi-tenant management and orchestration

There are four entities: the so-called Hosting Service Provider (depicted in blue) that owns the infrastructure and offers slices to three third party tenants (depicted in green, red and yellow). 5G NORMA should provide the Hosting Service Provider the possibility to make available a suite of infrastructure and software as a platform on which different tenants could deploy and manage their own network function applications. With this platform, the different tenants could operate their own network slices customised to their business requirements. Specific orchestration and management interfaces exist for each tenant (depicted as horizontal and vertical bars) having the Hosting Service Provider sitting in between to apply specific policies for each entity. Both tenants and the Hosting Service Provider have the possibility to do management and orchestration (MANO) tasks as defined in ETSI NFV MANO (management operations implemented by the management blocks VNF-Manager and VIM and orchestration operations implemented by the NFVO) as well as in 3GPP network management tasks, i.e., fault, configuration, accounting, performance, and security (FCAPS) management for network functions, virtualised and non-virtualised. Replacing the Hosting Service Provider with the MNO (or infrastructure provider and mobile service provider), this VNPaaS use case is very similar to Offer Type 3 as described in Section

3.3.3. With this approach, 5G NORMA is fully aligned and compatible with the ETSI NFV MANO framework, i.e., all the functionality as well as the reference points as defined in ETSI NFV MANO are supported. Extensions include the addition of an Inter-slice Resource Broker, additional interfaces within the specified reference points, e.g., to inject domain-specific data into the SDM-O, as well as the implementation details of MANO entities.

Consequently, in the 5G NORMA context, the definition of a network slice goes significantly beyond the ETSI definition of a Network Service. A slice may contain, in addition to control and data layer VNFs of a Network Service, the management and orchestration system that is local to a specific slice, particularly when Offer Type 3 is realized (cf. Section 3.2.3). Conceptually, this allows for a dedicated NFV MANO system per network slice, thus meeting the demand of certain tenants requiring their own management and orchestration capabilities. However, the same NFV MANO instance can also be assigned to multiple slice instances (e.g., for Offer Types 1 and 2).

4.5.1.2 MANO functional components

While the ETSI NFV MANO framework can be used to meet the objectives outlined in the previous section, it just provides a high level approach to the functional blocks, without providing detailed information about its internal architecture or implementation details.

In this section, we will go more in detail about the implementation of the MANO layer for 5G NORMA, focusing on how it could be implemented to meet our specific requirements for network slicing and how it needs to be extended to allow multi-tenant services and network management & orchestration.

Figure 4-9 shows the different elements in the MANO layer of the 5G NORMA architecture. The depicted scenario consists of a mobile service provider (depicted in green) that, at the same time, also has the role of the infrastructure provider and therefore acts as a mobile network operator (MNO) as defined in Section 3. Further, it shows *n* tenants (yellow, red) that operate their dedicated NFV MANO stacks. In other words, the scenario depicts a single infrastructure domain environment that can consist of one or multiple points-of-presence (PoP).

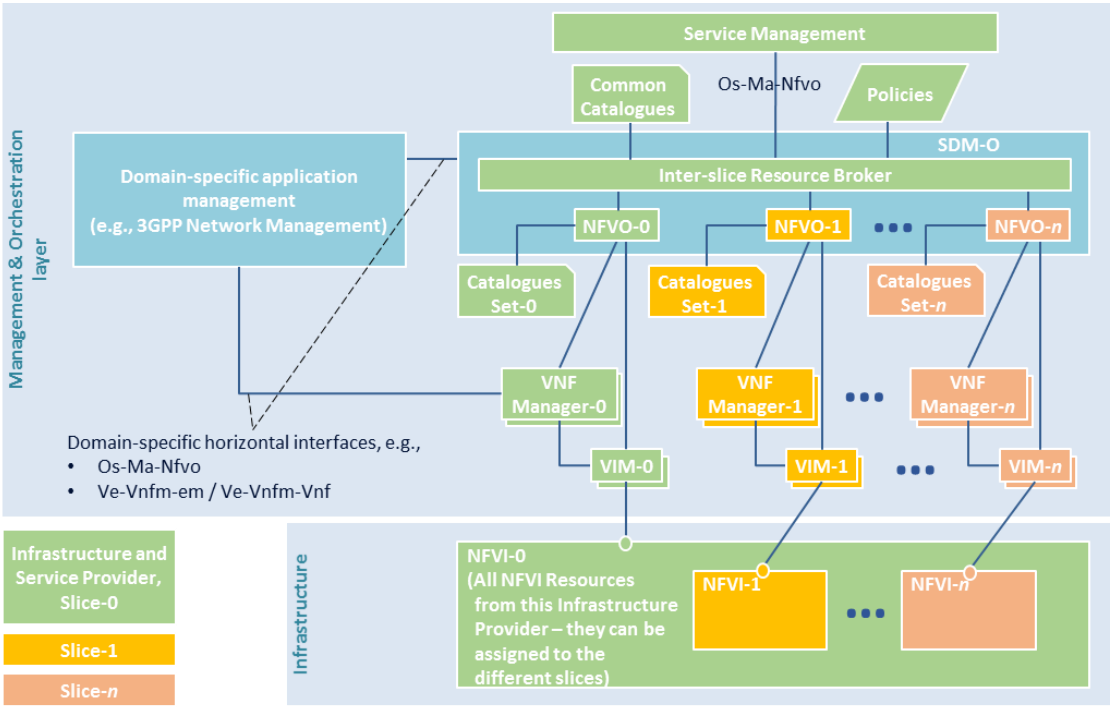


Figure 4-9. 5G NORMA Main Management and Orchestration Blocks

The SDM-O (Software-Defined Mobile network Orchestrator) integrates an Inter-Slice Resource Broker block which can manage different sets of complete ETSI NFV MANO stacks like the one in depicted Figure A-1, i.e., with their corresponding NFVO, VNFM, VIM and Catalogues set.

The core idea is that each network slice (depicted in green, orange, and red) is associated to one of these stacks would be associated to a specific network slice (depicted in green, orange and red in the previous figure). This does not exclude the possibility to associate multiple slices (from the same tenant) to a single NFV MANO stack instance. The associated VIM(s) manage(s) the set of resources assigned to the respective slice, where these resources can come from multiple points of presence (PoP).

However, as depicted in Figure 4-9, the MANO stack-0 (shown in green) has a special role, since it is specifically associated MNO. The stack provides orchestration and management capabilities on the whole NFVI resources. The MNO (or, in the general case, the infrastructure provider) can assign those resources to other slices which belong to different tenants (e.g., to slices 1 to *n* in Figure 4-9). Such assignments are decided by the Inter-slice Resource Broker and enforced by the according VIM(s). Further details on the role of the Inter-slice Resource Broker are given in Section 4.5.1.3. Moreover, the MNO can act in the role of a tenant in case it deploys and orchestrates an own network slice.

So, according to the ETSI MANO framework [NFV-MAN] and the principles described in Deliverable D3.1 [5GN-D31], three main modules are in charge of the orchestration and management for each slice: NFVO (integrated into the SDM-O block), VNFM and VIM. Together with Service Management, these three modules are the core part of the 5G NORMA Management and Orchestration layer. The role of these blocks is to manage the NFVI (including the network control components) and orchestrate the allocation of resources needed by network services and VNFs.

NFVO, VIM and VNF-Manager modules are the same blocks as those defined by the ETSI NFVI MANO specification, including functionality and reference points. In contrast, the SDM-O has been specifically defined for 5G NORMA. As illustrated in the figure, this SDM-O block is internally composed of two different functional elements:

- The slice-specific NFV Orchestrators (NFVO blocks), providing the same functionality as the ETSI MANO NFVO block.
- The Inter-slice Resource Broker, which is the block specifically designed to manage and orchestrate resources allocation for network services and functions across different slices and multiple tenants. It therefore is a key component to realise the 5G NORMA multi-tenant multi-service paradigm. It computes if and to what extent slices (and the associated MANO stacks) are assigned either reserved resources or shared/on demand resources. Such allocation decisions are communicated to the according MANO stacks (in particular the VIMs) for further execution.

Figure 4-9 shows specifically a scenario comprised of three MANO stacks: one of them is operated by the MNO which is at the same time the infrastructure provider (depicted in green). The remaining two stacks (orange & red) that could be assigned to different tenants (each tenant could manage one or more slices). Any slice is associated to exactly one ETSI NFV MANO stack instance, comprised of an NFVO, one or multiple VNF Manager(s) and one or multiple VIM(s). More specifically,

- The MNO has its own slice(s) with a full ETSI NFV MANO stack (green) with a privileged access to the whole infrastructure (all the resources in the NFVI). Further slices (orange & red) would have associated a subset of the infrastructure, which could be managed by their specific ETSI NFV MANO stack instance;
- Each tenant could request slices to be managed by their dedicated NFV MANO stacks. In that case, the VIM(s) are operated by the tenant and manage their own, i.e. allocated, subset of NFVI resources that would be already deployed. The MNO would have the 'complete' view with all the deployed VNFs (via its specific VIM-0). For the 'regular' NFV MANO instances (orange & red), the view would be limited to its own VNFs and the according NFVI resources (e.g., NFVI-1 & *n*). For this purpose, 5G NORMA assumes sufficient isolation between the resources partitions. However, NFVI-1 and NFVI-*n* resources can be extended with resources not belonging to NFVI-0, e.g., 3rd party resources

or resources directly owned by tenant 1 or tenant *n*. The slice-specific VIMs (i.e., VIM-1 & *n*) to manage all this together (note: this is not depicted in Figure 4-9);

- Furthermore, each tenant could operate dedicated VNF Managers (e.g., VNFM-1 & *n*) to apply specific life-cycle management policies on its slice interacting with their corresponding VIM(s);
- Similarly, each tenant could operate a dedicated NFVO (e.g., NFVO-1 & *n*) to orchestrate different network services and associated NF forwarding graphs on its slice(s). The VNFs in these slices could be managed by single or multiple VNF Managers (therefore, an NFVO could orchestrates network services composed of VNFs managed by multiple VNF Managers). In this way, it would be possible to chain VNFs deployed within the same slice to create an end-to-end network slice. Generally, a network slice is comprised of one or multiple network services;
- Each NFV MANO stack instance can either work on a common set of catalogues set for network services and NFs as provided and operated by the Infrastructure Provider/Mobile Service Provider (depicted in green) or on a dedicated catalogues set as on-boarded by the tenant and certified by the provider.
- Both Service Management and Inter-slice Resource Broker are operated by the mobile network operator (depicted in green). The Inter-slice Resource Broker handles the allocation of resources of the different slices, their dynamic provisioning and the management of the shared resources among them within the administrative domain it controls.

Regarding the utilisation of the infrastructure, NFVI-0 represents the resources belonging to the Infrastructure and Mobile Service Provider (green, also MANO stack-0). NFVI-0 contains subsets NFVI-1 to NFVI-*n*. Accordingly, VIM-0 can manage all the infrastructure, including the subsets of Tenant-1 to Tenant-*n*. However, Tenant 1 to Tenant *n* can also use dedicated VIMs (VIM-1… VIM-*n*) to manage the according NFVI subsets. While NFVI-1 to NFVI-*n* depict resources explicitly dedicated to the respective slice, they are not static in nature, i.e., their respective share quota of the overall NFVI (NFVI-0) could be reshaped at runtime. Furthermore, any non-associated NFVI resources that are allocated on demand based on priorities are managed by VIM-0 as well (refer to Section 4.5.1.3).

All the reference points for each MANO stack are the same as those defined in the ETSI NFV MANO specification [NFV-MAN]. However, the *Os-Ma-Nfvo* reference point needs to be extended in order to communicate specific tenant and slice information elements. In 5G NORMA, depending on the scope of the interaction, the '*Os-Ma*' end of this reference point is comprised of the domain-specific application management module (e.g., 3GPP network management systems or enterprise network management systems) or the Inter-slice Resource Broker, cf. Figure 4-9. Further, the ETSI NFV MANO *Ve-Vnfm-em* and *Ve-Vnfm-Vnf* reference points are implemented towards the domain-specific application management module. *SDM-O – Service Management* reference point and the common catalogues set still need to be specified.

4.5.1.3 The role of the Inter-slice Resource Broker

Sharing the same infrastructure across tenants and network slices entails the following trade-off: resource reservation versus flexible resource sharing. A purely static resource allocation approach can provide a certain level of performance guarantees and resource isolation, but it fails to exploit the multiplexing gains envisioned by network slicing.

The 5G NORMA architecture described in the previous section allows for a customizable trade-off between resource reservation and on-demand allocation, particularly for NFVI resources (for other resources such as radio spectrum, such flexible policies are subject to major technical and regulatory constraints). Specific service-level agreements (SLAs) define the concrete embodiment of reservation and on-demand allocation rules. For example, a tenant may request a fixed amount of NFVI resources (in such a case, the NFV MANO stack assigned to a tenant could exclusively manage the allocated quota of resources). In another case, a tenant may agree that a percentage of the associated, but unused resources, may be dynamically allocated to other tenants or slices, thus realizing cost savings.

The role of the Inter-slice Resource Broker is to have a general view of the whole infrastructure that can be offered within a single administrative domain, as well as monitoring the usage of the resource subsets allocated to the tenants. It controls the dimensioning of resources to be assigned to each tenant and their status, including those resources not yet assigned. For instance, when the slice is commissioned, the Inter-slice Resource Broker has to explicitly inform each tenant about the allocated resources.

In this scenario, the assignment of quotas for each slice is performed based on the SLA with the different tenants, i.e., each tenant has an SLA specifying the amount of resources they can use. The general idea is that fixed quotas are initially assigned to the different slices when they are commissioned. While quotas are assigned in a fixed way to each slice, they can be reshaped at runtime if tenants request for that. This also implies that if a slice does not utilise all allocated resources, the idle resources will not automatically be re-allocated to the other slices. Except for the Infrastructure and Mobile Service Provider (depicted in green), tenants are neither aware of the existence nor the resource utilisation level of further tenants. They only have an SLA specifying their right to use certain resources in a certain manner, e.g., special terms in SLAs to allowing a tenant to exceed its assigned quota for certain time and at certain cost. In case a slice is permanently decommissioned, NFVI-1 again contains resources that the Inter-slice Resource Broker could assign to other tenants or to keep them for future use.

The rules for resource utilisation by multiple stakeholders are kept in the policy catalogue shown in Figure 4 10. It is a special catalogue directly connected to the Inter-slice Resource Broker that contains a map of the currently available infrastructure assigned to each slice and their sharing policies (e.g., if they want a guaranteed and scalable resource assignment and to what extent). The idea is that each tenant continuously monitors and reports the status of the resources allocated to a network slice via the ETSI NFV MANO *Or-Vi* reference point (see Figure 7-1). The NFVO then reports to the Inter-slice Resource Broker that may use this information to reshape the current association patterns according to new external triggers, such as a new slice creation request or a re-orchestration request from an already hosted slice. Moreover, the Inter-slice Resource Broker updates and maintains the policies catalogue.

4.5.2 Multi-tenancy- and multi-service-aware 5G NORMA MANO interfaces

Multi-tenancy- and multi-service-aware reference point in the 5G NORMA Management & Orchestration (MANO) layer are comprised of those interfaces that either carry data from multiple tenants or network slices or that convey information from (or to) MANO functions that operate on multi-tenant models, i.e., functions that have an awareness of multiple tenants (slices) sharing the mobile system and infrastructure. The interfaces between

(1) Service Management and Inter-slice Resource Broker (*Os-Ma-Nfvo*) and
(2) Service Management and entities from the 5G NORMA Service Layer (*Sl-Sm*)

belong to this category, see Figure 4-10.

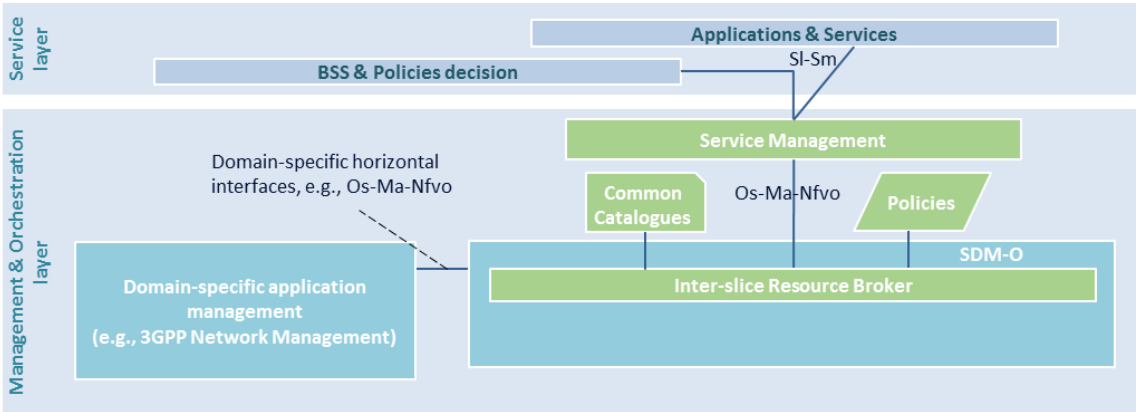


Figure 4-10: Selected functions of the 5G NORMA MANO layer

4.5.2.1 Reference point between Service Management and Inter-slice Resource Broker

One of the central tasks of the Service Management function is to map the service requirements as provided by the tenant via the Sl-Sm reference point to the appropriate network slice template. As a result of this mapping process, Service Management provides a network slice descriptor to the Inter-slice Resource Broker via the *Os-Ma-Nfvo* reference point. As shown in Figure 4-10, the 5G NORMA architecture provides the possibility to commission multiple NFV MANO stack instances, e.g., dedicated to a tenant or a network slice. For this reason, a 5G NORMA network slice descriptor does not only contain information on control and data layer functions, but also on MANO layer functions. Hence, the network slice descriptor is comprised of two major parts that specify the functions, resources, and policies that are required, respectively,

(1) to perform lifecycle management for a network slice and
(2) to realise the network service requested by the tenant.

While (1) comprises a specification of the NFV MANO stack instance (NFVO, VNFM, VIM, NFVI instances, catalogues for network services and functions, etc.) that is dedicated to the lifecycle management of the network slice, (2) includes the network service descriptor(s), i.e., the collection of VNFs and PNFs that, as a whole, form the control and data layer architecture of the particular network slice instance.

According to [28.801] and depicted in Figure 4-11: Lifecycle phases of a network slice instance, lifecycle management is composed for four distinct phases: (i) preparation phase, (ii) instantiation, configuration and activation phase, (iii) run-time phase, and (iv) decommissioning phase. The network slice descriptor as generated by the Service Management therefore contains the necessary information to carry out phases (ii) – (iv) appropriately.

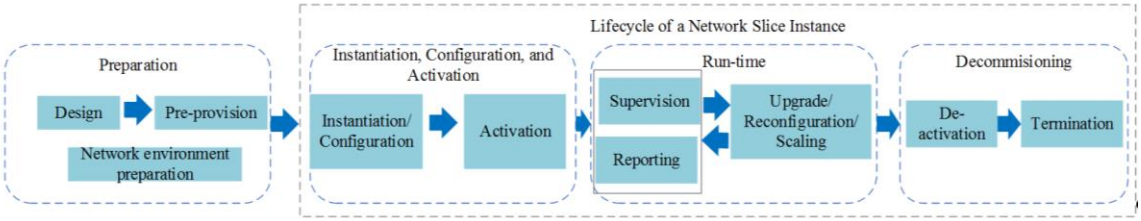


Figure 4-11: Lifecycle phases of a network slice instance

In a first step, the Inter-slice Resource Broker, as part of the SDM-O, uses part (1) of the network slice descriptor, i.e., the NFV MANO descriptor, to commission a new NFV MANO stack. In second step, part (2) of the network slice descriptor is utilised to generate the necessary objects and models that the NFV MANO instance operates on, i.e., NFV service catalogue, VNF/PNF catalogues, NFV instances, and NFVI resources. For the allocation of the NFVI resources that are under control of this MANO stack instance, the Inter-slice Resource Broker uses a combination of the resource commitment models as outlined in Section 4.5.2.2. Commissioning of the network slice control and data layer functions is triggered by the Inter-slice Resource Broker via the Os-Nfvo reference point of the NFVO by providing or referring to the set of network service descriptors to be instantiated. The network slice lifecycle management is now delegated to the NFV MANO instance and the according domain-specific application management functions, cf. Figure 4-10. This includes

- instantiation and configuration of the network services and associated network functions,
- activation of the network slice,
- during runtime: supervision and reporting as well as
- upgrading, reconfiguration, and scaling,
- deactivation and termination of the network slice.

After the NFV MANO stack has taken over network slice lifecycle management, operations are equivalent to a single-tenant environment.

In the northbound direction (i.e., Inter-slice Resource Broker to Service Management), the Inter-slice Resource Broker provides performance, fault, and configuration data about commissioned network slices according to the monitoring rules provided by the Service Management function. These data are used for performance reporting as well as accounting and charging towards the tenant. The monitoring and/or computation of key quality indicators (KQI) to be provisioned is customised according to the SLA specifications of requesting entities from the Service Layer. KQIs cover both high-level objectives (coverage, network sharing, customer satisfaction, interoperability in multi-vendor environments) and technical objectives (general key performance indicators, such as handover failures, and QoE/QoS parameters);

4.5.2.2 Resource commitment models

For resource management procedures, [NFV-IFA010] defines three so-called "resource commitment models"

- reservation model,
- quota model, and
- on-demand model.

While the quota model limits the NFVI resources that a slice can obtain from a particular NFVI-PoP (Point of Presence), the reservation model statically allocates the specified amount of resources to a particular tenant or slice, even if the resources remain idle. Regarding the co-existence of the quota model and the reservation model, a VIM will, as the default behaviour, also apply the slice quota to the slice reservation being made. However, further rules will determine the behaviour of the VIM if a reservation exceeds the specified slice quota [NFV-IFA010]. In 5G NORMA, these rules are determined from the policies as maintained by the Inter-slice Resource Broker. The on-demand resource commitment model does not make any reservation or pre-emptive allocation of resources. Rather, NFVI resources are assigned once they are requested.

The further definition of such policies is part of future work in both work package 3 and work package 5 of 5G NORMA.

4.6 SW-defined mobile network control

4.6.1 Mobility management principles

Envisaged 5G services and slices exhibit different demands for mobility support in terms of e.g., terminal speed, session continuation requirements, and stability of the endpoint address. Mobility management schemes can differ in many ways, e.g., requiring special handover policies and settings in the RAN, flexible mobility anchoring, adaptive gateway relocation rules, or customised network elements (e.g., local gateways or gateways with specific mobility support functionality). The mobility management scheme needs to be selected flexibly according to the context of the service or a network slice.

This section describes principles for design and selection of service-specific mobility management (MM) functionality as well as the implementation within 5G NORMA architecture and interworking with functional elements such as SDM-C, -O, -X.

Criteria for designing a flexible MM function per slice to allow for service tailored MM features is described in Section 4.6.2.1 while binding of MM schemes to a single slice (controlled via SDM-C) or functioning as a common cross-slice NFV (under SDM-X control) are dealt within Section4.6.2.2. Section 4.6.2.3 details implementation of specific MM schemes, both in terms of interfaces towards SDM-C (SDN-App and SBI) as well as application to edge and multipath mobility.

4.6.1.1 Design parameter criteria

Design considerations on the amount and type of parameters to be configured within a service-aware MM function (i.e. describing a VNF template for a specific type of MM scheme which is

chosen during definition of a network slice) depends also on the way of instantiation and the corresponding slice type. Whether a MM scheme is realised as a dedicated or a shared VNF as shown in Figure 4-1 in Section 4.1.1 is mainly governed by exclusive or joint usage of resources. Similarly, the fact whether MM is invoked as a dedicated per-slice function or across multiple slices (see Section 4.6.1.2) depends on the service features to be supported. E.g., a simple on-demand MM would be associated only to a slice requiring low or even no-mobility support (e.g. for MTC, Fixed/Home-Net slice etc.). On the other hand an MM-App supporting a range of terminal speeds and seamless session continuity might apply for multiple sessions/UEs within same regional context (within train/bus, on highway, etc.) but belonging to different (e.g. verticals automotive slices and an eMBB slice). Such a MM App would then be re-used across different slices.

A further design criterion could be the availability and usage of layer sensitive information across layers (e.g. to proactively invoke handover on MAC or IP layer based on signal strength (PHY layer). Such a feature as well as the capability to adjust to variable service demands (e.g. in terms of QoS/QoE) of course would depend on the specifics of the involved access technologies. Another parameter guiding MM-NF design is the potential differentiation for a hierarchical mobility treatment (e.g. local and global mobility). Finally the degree of flexibility which a chosen MM approach supports (e.g. whether a feature may be changed on the move according to changing environments) is restricted to certain variables (e.g. no multi-link or access heterogeneity is possible in case of single-interface devices).

Based on such parameters considered in a mobility support design the correspondingly required effort in terms of process complexity as well as amount and frequency of necessary signalling messages can be estimated. Also the number of network entities included in the message sequences (e.g. only UE and RAN nodes or also core network entities for location, anchoring, or storage of subscription policy information) may determine the required effort of a specific MM application

4.6.1.2 Binding mobility management to network slices

In order to support a service tailored Mobility Management (MM) we aim at designing a network slice which includes specific network functions enabling a specific MM scheme. A way to realise this is to maintain specific, mobility related flavours of network functions and/or specific configurations of network functions and instantiate them according to the mobility related context of the network slice.

The selection of an appropriate mobility management scheme needs to be provided through a binding functionality which is a part of the 5G NORMA Management and Orchestration Layer. More specifically the binding functionality is a part of SDM-O and Service Management. The binding functionality provides the mapping between the mobility related context of the slice, i.e. mobility management requirements and the mobility management scheme that supports the mobility requirements in the most suitable way. Furthermore, it translates this mapping into a concrete configuration of the network slice.

The binding functionality takes into account not only the network slice context but also the predetermined policies in order to select a suitable mobility management scheme.

The binding functionality includes three blocks:

- Binding Policy Management – translates service requirements, operator targets, and KPIs to policies that have to be enforced on the network slice
- Binding Function – selects the mobility management scheme according to the predetermined rules
- Network Slice Selection and Configuration – selection of the right configuration for slice instantiation which will realise the chosen MM scheme

In a nutshell, for given service requirements and network context the binding functionality gives two important outputs: suitable mobility management mechanism/scheme and the fitting template

for slice instantiation/configuration. The resulting slice template is used to instantiate and configure the network slice which implements the selected mobility management scheme.

Based on the slice context and the slice implementation we can envision different levels of sharing MM among slices. E.g. the slices can either have dedicated core network (CN) instances or have different level of CN sharing i.e. either only parts of the CN or the entire CN can be shared among slices. This impacts the exact implementation of MM especially in terms of designing it as dedicated (per slice) MM or common MM across multiple slices. Hereby we can identify three options for mobility management implementation based on level of CN sharing among network slices:

1) Dedicated MM in a dedicated CN for each network slice.
2) "Mixed" MM in CN which comprises both components shared among networks slices and components assigned to dedicated slices. E.g., MM and identity/subscription management are shared between the network slices, while other CN functions, such as session management, are implemented in separate CN instances of network slices.
3) Common mobility manager completely in shared control plane of the CN.

The dedicated MM approach in a dedicated CN although enabling the clean separation of network slices might come with the potential drawback of adding signalling in the network and over the air. On the other hand, designing the MM as a common entity adds implementation complexity and lowers the level of isolation between network slices.

4.6.1.3 SDM-C-based mobility management

Following the design and architectural principles defined in 5G NORMA architecture, the mobility management as a whole can be realised as an SDM-C application. Taking advantage of a unified QoS/QoE control framework like the one described in Section 4.6.2, the management of user mobility is a thorough process that involves network function control and orchestration to achieve an optimised functionality on a per slice basis.

By exploiting these characteristics, the network flexibility is increased: the adaptation of the network slice capacity according to the instantaneous traffic demands and required KPIs entails the re-configuration and re-orchestration of the network at many levels. Therefore, besides the selection of the most appropriate MM algorithm or the parameters that may influence the MM algorithm behaviour, the MM shall be able to control different network configurations seamlessly.

Specific mobility requirements in case of vehicular and low-latency communication can be solved with the edge mobility approach while a highly reliable connectivity even in a mobile environment is provided by multi-path mobility. Both approaches are detailed in [5GN-D51].

One of the key technologies for the enhancement of the flexibility is RAN as a Service (RANaaS) [iJOIN-D53]. This capability, envisioned as one of the future pillars of 5G networks, allows to split the currently monolithic RAN stack into atomic functions that may be orchestrated in different ways, exploiting either the multiplexing gain of baseband processing centralisation or the decentralisation of edge computing. In this very heterogeneous context, an enhanced MM shall i) jointly optimise RAN and Core network functions by leveraging on the centralised network control capabilities of SDM-C and, ii) steer user flows across different network functions according to the RANaaS functional split implemented in the network. The former functionality is implemented within a SDM-C application, while the latter is provided by a set of *plugins* installed on the Southbound Interface of the controller.

We next sketch the overall ideas of a software defined MM algorithm that can cope with the changing environment of a RANaaS-enabled network. According to the selected functional split, different optimization options and network control challenges arise.

PDCP-RRC: this functional split is a pure c-/d-plane split, as PDCP is the highest layer that deals with user data, handling GTP (GPRS Tunnelling Protocol) traffic towards the gateways (and the Internet). Using the SDM-C approach, the functionality currently carried out by NAS, MME and RRC can be centralised and implemented as a SDM-C application. That is, a pool of virtualised

Radio Access Points (RAP), implementing the RAN stack up to the PDCP, may be controlled by a centralised MM application that can take optimal handover decisions according to the load of RAPs. On the other hand, the SDM-C southbound interface needs to interact with both the RAPs and the gateways (that may be joined in a single entity) by managing directly NAS, RRC and GTP session requests from the d-layer network function. This split is depicted in Figure 4-12.

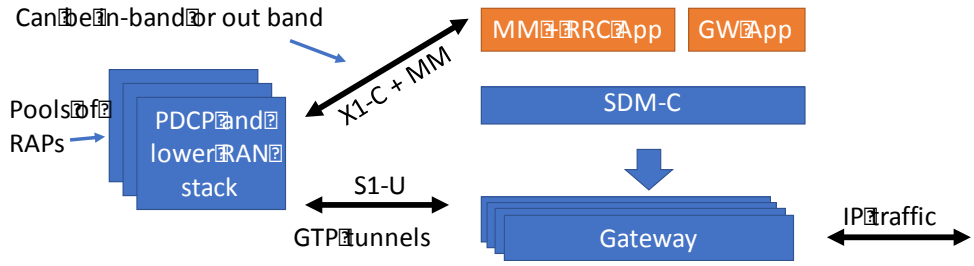


Figure 4-12: PDCP – RRC functional split

RLC-PDCP: this split involves managing directly data radio bearers between a pool of RAPs that implement the RAN stack up to the RLC and a centralised entity that need to perform several functions in addition to the PDCP-RRC including e.g., (de)ciphering. The centralisation of these previously distributed network functions allows for enhanced routing optimization, multipath or radio bearer based mobility. On the other hand, the southbound interface shall be able to manage, among other information, data radio bearers and their mapping to the RLC channels. Figure 4-13 describes this case.

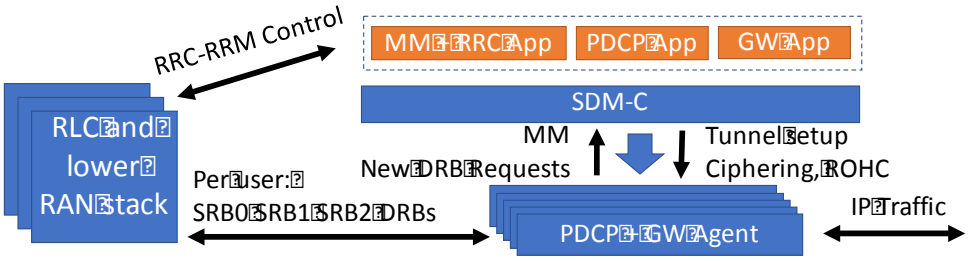


Figure 4-13: RLC-PDCP functional split

These two examples are just an overview of how enhanced MM algorithms can take optimal decision depending on the *cloudification* of the network and the requirement that have to be fulfilled.

4.6.1.4 Function mobility

The 5G NORMA architecture allows for dynamic allocation/instantiation of VNFs, placing them either in the edge cloud or in the central cloud. Users and terminals will be served by different edge clouds as they move around the world, so VNFs might have to migrate across the network as users change their covering edge cloud. Migration entails transmission costs and potentially service outages, while serving a user located in distant edge cloud will increase latency and potentially degrade QoE. The challenge then is to decide if/when to make this migration. A placement decision method should exist that would use different criteria and provide the optimal position for a VNF to be at any given time. One such method is detailed in [5GN-D51].

The SDM-O will have a module for determining the placement (serving edge cloud, central cloud, hosting edge cloud) of a network function or service. As output, the module will provide a placement decision. The SDM-O will then organise the migration using the VNFMs and VIMs.

Migration could mean live migration or duplication and reconfiguration. Live migration consists of the instantiation of the same VNF, followed by the live transfer of memory, storage and network connectivity from the original VNF. It is suitable for a function like Content Caching. Duplication and reconfiguration represents a new instance of the VNF in another location, followed by redirecting all users to using the new instance, and it is best suitable for stateless functions.

4.6.2 QoS/QoE control

QoS is used by network operators to optimize the network in order to provide services with an acceptable level of quality. To compute QoS objective metrics like packet loss, jitter, throughput or delay are commonly used, but experience shows that the whole perception of users is not just based on only certain metrics over individual network elements, but on the overall E2E performance of the system. So, to keep users satisfaction and avoid churn, operators are trying to improve QoS using a more modern approach based on the newest QoE concept.

The ITU-T P.10/G.100 recommendation defines QoE as "the overall acceptability of an application or service, as perceived *subjectively* by the end-user" [ITU P.10]; also, the European Network on QoE in Multimedia Systems and Services defines QoE saying that it can be "influenced by content, network, device, application, user expectations and goals, and context of use" [Qual]. Other less formal definitions are also in the same way, but always, beyond the formal definition, it is assumed that customers experience will not only rely on the network infrastructure, but also on other influence factors which are entirely beyond the control of network operators (UE type, users mood, environment, user profile etc.).

This so general definition of QoE makes things difficult when trying defining a specific implementation for computing QoE: we should be able not only to measure specific objective parameters and assign additional resources when necessary, but also, to guess what could be into the users mind while accessing the services considering subjective influence factors. Also, if we consider the expected 5G capabilities, identifying those influence factors is not a trivial task. Contrary to what happened in the old voice networks on which quality were evaluated mainly from the audio signal received on a specific type of device, the new 5G network is assumed to support a multiplicity of services (voice, MBB, IoT, V2X, gaming, M2M…), devices (smart phones, tablets, remote sensors, vehicles, automation devices…) and network elements (e.g., macro and micro cells). Also, the network will be designed to support new services and devices that could be devised in the future.

Besides, we also need to provide these services in a multi-tenant environment, and each tenant could require a very different approach to the QoE/QoS assessment; i.e., on the same network, we could have certain tenants requiring the deployment of complex high demanding QoE services, while others could only request the traditional QoS approach based on monitoring a small set of certain objective parameters. For instance, while a tenant could request to compute an objective global QoE measurement based on just a couple of physical parameters (e.g., CPU and RAM usage) inside a specific VNF, other tenant could request to deploy a service requiring real-time QoS/QoE measurements from a big set of individual end-users inside certain geofences; this second approach would be probably based on a complex infrastructure with multiple nodes executing real-time DPI and a Big Data Analytics.

Also, if we consider the state of the art regarding QoE/QoS assessment and control we see it is still a vibrant area of research [Alr]. Because we are trying to guess the users subjective experience QoE estimation is not an exact science, but a best engineering effort; this has led to a number different methods considering different media type (video, voice, image…), and for each media, different measurement methods and different computational resources. Different models are under consideration: objective/subjective, intrusive/non-intrusive, different approaches to get the mapping functions…; the main issue is that, while quality can be correlated with different measurable signal features, it is typically not possible to establish a simple relationship between these measurable magnitudes and the E2E quality perceived by a user. For example: services like VoIP, video streaming, on-line gaming or internet browsing has unique performance indicators to measure; but other services, such traditional voice or messaging have other specific KPIs. So, different types of quality estimation and prediction models have been developed for specific application domains and service conditions (e.g., models based on Artificial Neural Networks [Rub], Genetic Algorithms [Gha], Decision Trees [Zha] or Support Vector Machines [Men] among others are some peculiar examples). Consequently, there is no a universal quality model that can be applied for all cases, being the evaluation of QoE heavily depending on the context [ITU G.1011]. This has led to the emergence of different recommendations as shown in the following figure:

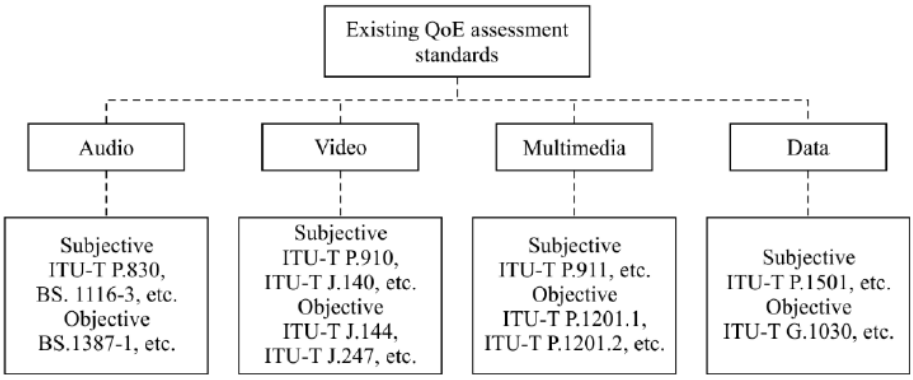


Figure 4-14: ITU Recommendations for the QoE Assessment

In summary, to compute a QoS/QoE metric on the new 5G network is not a trivial task:

- The new 5G network is expected to manage a diversity of services with different features and challenging requirements (e.g. very low latency, high data rates, etc.).
- Influence factors can be objective and subjective, and they could be not clearly defined in advance (they will probably different for each service, UE and boundary conditions).
- The 5G network should be able to allocate different services from multiple tenants and provide them the infrastructure to compute and manage QoS and QoE for their services.
- State of the art regarding QoE control is still an active research area; i.e., there is no a wide accepted general purpose solution. Also, new approaches may appear in the future.

We consider all this makes necessary a completely different approach from the one used to compute just QoS on legacy networks. Although challenging, we think this can be implemented in a flexible and efficient way. In the following sections we describe our approach for that.

4.6.2.1 Overview on QoS/QoE assessment in 5G NORMA

We are talking here about QoS, QoE and related terms, but these concepts are becoming quite common, so they may be defined differently in other documents or context. In our case, for the specific 5GN context, we need to clarify some concepts from the beginning:

- We understand QoS as the degree of adequacy of the service to a number of specified *physical and measurable objective parameters*. We think this is aligned with the common understanding about this (and the common practice in communication networks), but we think it is necessary to set this clear to differentiate it from QoE.
- Regarding QoE, we align with the provided definitions in [ITU P.10] and [LeC], i.e.: "the overall acceptability of an application or service, as perceived *subjectively* by the end-user"; the key here is the inclusion of certain QoE influence factors that are beyond the operator influence (user profile, terminal type, environmental noise…); i.e., the inclusion of a least one influence factor is what produces a QoE metric.
- So, the "QoS/QoE Monitoring and Control System" refers the whole system in charge of monitoring QoS/QoE parameters and controlling QoS and/or QoE. If the monitoring process is just on certain objective measurements the system will control just QoS, but if certain subjective influence factors are monitored the system will control QoE also.
- This "QoS/QoE Assessment and Control System" consists of the following components:
    - The QoS/QoE Management Subsystem, used to define the relevant configuration parameters for the system such as the mapping functions, monitoring methods, possible output values, and so on.
    - The QoS/QoE Monitoring Subsystem, to process the relevant input parameters.
    - The QoS/QoE Mapping Subsystem, to process the input signals provided by the monitoring system in order to generate relevant events (e.g. when an input parameter reaches certain threshold).

- The QoS/QoE Control Subsystem, processing the relevant events generated from the mapping system and executing the proper actions to keep QoS/QoE under required limits.

This last control subsystem will be integrated in the core controller block in 5G-NORMA: the SDM-C block. So, the QoS/QoE Assessment and Control system is split in two: the control system itself (into the SDM-C) and the "QoS/QoE Assessment System", integrating the mentioned management, monitoring and mapping blocks. The following figure shows this idea:

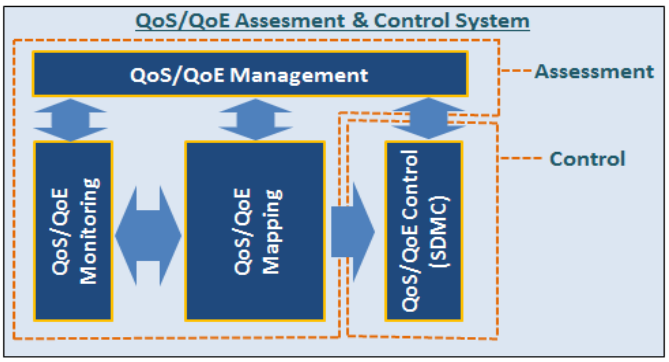


Figure 4-15: QoS/QoE Assessment and Control systems

4.6.2.2 Approach and requirements

From general point of view the QoS/QoE metric computation can be seen just as the application of certain mapping function on certain input parameters; however, this approach is quite simplistic if we have to consider how to integrate it in practice in our 5GN architecture. When considering practical aspects this issue may require very different resources and capacities depending on the specific way the problem arises.

We consider that just pre-selecting a set of relevant KPI's and specific mapping functions to compute QoS/QoE is not the most appropriate approach if we want a general solution; we would run the risk of delivering a too rigid architecture unable to manage certain tenant's requirements or to integrate relevant technical advances.

So, instead of defining a too specific solution, we consider a better approach to define an open framework which can be adopted by the different tenants in order to provide enough flexibility to obtain the desired behaviour. Therefore, instead of a strictly defined set of functional blocks, parameters and mapping functions, our objective here will be to provide 5GN with a flexible way to implement different QoS/QoE control mechanisms. This will enable also a path towards future approaches and strategies that certain tenants could adopt.

From this, we propose the following general requirements for the 5GN QoS/QoE control block:

- It must provide an open framework, allowing the integration of different solutions in function of the specific requirements of each tenant.
- It should be featured to use KPI measurements from different network elements (DPI nodes, mobile terminals, eNB or other network elements). The set of KPIs to use is not restricted; it should be possible to define them in an open way in the scope of each tenant SLA.
- It must be possible to integrate different mapping functions and execute them in parallel (i.e., each tenant may require a different function with different features and parameters).
- It should be based on open interfaces through which it should be possible to flexibly specify the parameters to monitor, mapping functions and the way the output (QoE) is delivered.
- In order to optimise the network resources, it should be possible (if required by a tenant) the continuous monitoring of the QoS and QoE levels for each service.

4.6.2.3 High-level design

To meet the previous requirements, we should start from the beginning; said in a very general way, our initial problem is how to integrate basic QoS/QoE mapping blocks like the one in the following figure in our specific 5GN architecture:

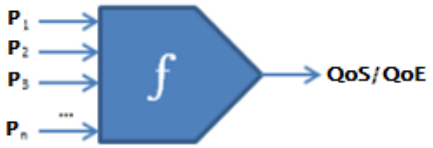


Figure 4-16: Basic QoS/QoE Mapping Functional Block

This figure represents a block that could be used to produce a QoS/QoE output metric by applying a specific mapping function ***f*** to a set of measurable parameters P_1, P_2,… P_n. The output would be interpreted as QoE or QoS depending on the input parameters; QoS could be computed selecting a set of objective network parameters (RTT, jitter, delay…) and applying the corresponding function, while QoE can be obtained using subjective influence factors also.

These mapping functions could be defined in very different ways: one way could be to delegate on an experts team to select the proper set of input parameters and to design the function itself (i.e., the typical top-down computing approach); other possibility is to use the so-called subjective approach, which in some way integrates also the final user's opinion about the service. An example of this subjective approach is the Mean Opinion Score (MOS) method which has been used for decades in phone networks to get direct feedback from users [ITU P.800].

But regardless of how we get the mapping function a first high-level consideration on this is that, on a real deployment, we are not going to have just a single function; on the contrary, we should be able to support the execution of different mapping functions in parallel. So, a first evolution of the simplistic approach in Figure 4-16 should be something like in Figure 4-18.

That is, we should be able to process different mapping functions with different input parameters set in parallel. Each function would provide a different QoS/QoE metric.

Another high level consideration is regarding the input parameters. We could receive them from different sources and expressed in a very different ways. For example, we could require for the same mapping function to process parameters as the jitter expressed as a rational number, or the user mobile terminal type defined as a text string. Information source can be also diverse: RAN parameters, users profile databases, billing information, values from different network interfaces and the parameters format could be different depending on each encoding schema. This makes necessary a normalisation layer prior to the mapping function itself. So, a second step in this high-level consideration could be to add this normalisation layer for each mapping function.

Also, in a similar way, we should have the possibility to normalise the QoS/QoE output values using specific coding schemas (e.g., the resulting QoE could be needed according the legacy MOS quantization scale, or using a rational number according certain sigmoid function).

Figure 4-19 shows the addition of these elements for each mapping function; $N_1 \ldots N_k$ represent the input normalisation functions, while $C_1 \ldots C_k$ are the output codification functions.

Another important point regarding the integration of the QoS/QoE mapping functions into the 5G NORMA architecture is the relationship with other functional blocks. Two of the blocks where a close relationship is required are obviously the above mentioned QoS/QoE Management and the QoS/QoE Monitoring blocks. The first one will provide the definition of the QoS/QoE input parameters agreed in the SLA for each tenant, allowing also the configuration of the different parameters in the QoS/QoE Mapping module; for instance:

1. Number and type of the different mapping functions. Depending on the specific function to be used (objective/subjective approach, etc.) the configuration will be different.
2. The corresponding input parameters normalisation functions.
3. The output encoding functions
4. The set of input parameters to which each mapping function should be bounded

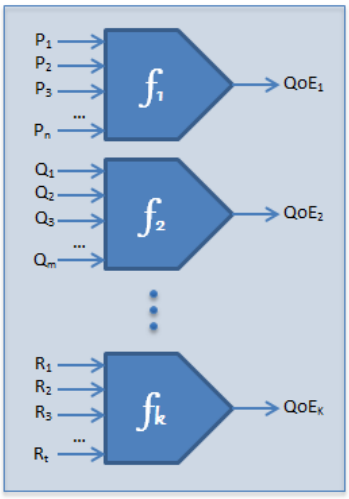


Figure 4-17: Different QoS/QoE Mapping functions in parallel

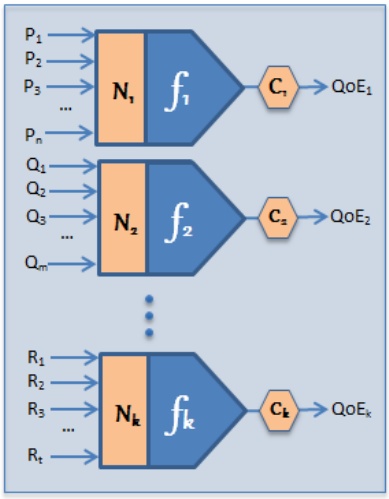


Figure 4-18: Normalisation and Output Codecs

On the other hand, the QoS/QoE Monitoring function will be in charge of monitoring the selected input parameters. Specific configuration parameters should be provided in the SLA for each tenant; for example: the required sampling frequency, the selected monitoring strategy (i.e., reactive, proactive, hybrid…), etc.

This block should be designed to receive messages from the mapping module and to trigger the proper actions for each case (the SDM-C will receive these trigger messages on its North-Bound Interface). We understand that these actions are basically of two types: a) to interact with the virtual infrastructure to request scaling operations for each single VNF; b) to interact with the 5GN management layer to request scaling actions requiring the VNFs forwarding graph update.

Figure 4-20 shows the evolution of the high-level design considering named interfaces. The QoS/QoE Mapping block has three main interfaces:

- A North-Bound Interface (NBI) communicating with the QoS/QoE Management Module. It will be used to configure the mapping module parameters as we previously said (specific functions to use, input parameters, etc.). The QoS/QoE Management module keeps also a close relationship with the QoS/QoE Monitoring module to configure the parameters to monitor and the way each parameter should be monitored (i.e., the monitoring module output and the mapping module input should be perfectly aligned). Also, it is assumed that the QoS/QoE Management module will be part of the Management & Orchestration Layer in the 5G Norma functional architecture, so it will be in close relation with the SDM-O (the entity that interfaces the business domain and handles slices creation requests).

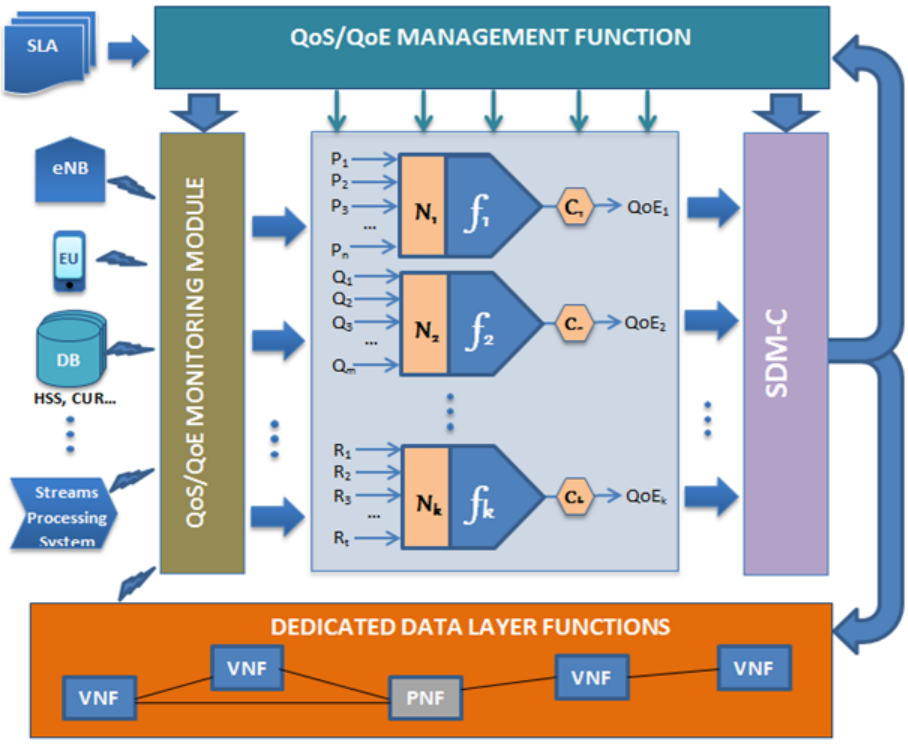


Figure 4-19: QoS/QoE Mapping, Monitoring and Management modules

- A West-Bound Interface (WBI) receiving data from the QoS/QoE Monitoring module. The monitoring module gathers information from different sources that can be relevant for the behaviour of the network itself or the service (e.g., the user's terminal, the Radio Access nodes, different databases containing user's data or the virtual infrastructure itself among others). Continuous monitoring of different parameters of interest would be required.
- An East-Bound Interface (EBI) to send relevant QoS/QoE Mapping events towards the SDM-C; so, the QoS/QoE mapping module will continuously analyse the status according the SLA constraints and will raise QoS/QoE relevant events towards the SDM-C. Based on those events the SDM-C may adapt to the new situation in different ways:
    - By reconfiguring some of the VNFs it manages (e.g., changing the pre-scheduler or asking for a less aggressive MCS).
    - Notifying the QoS/QoE Management block when a management operation is necessary.
    - Reconfiguring some paths using a SDN-alike technique.
    - Asking for more resources to the SDM-O. In this case, network slice reshaping (i.e., scale in/out) or VNF relocation policies could be managed by the orchestrator.

Note that no direct connection between the QoS/QoE Mapping module and the underlying virtual or physical infrastructure is necessary. It is assumed that the SDM-C will work as proxy for this. As we know (Section 4.5) we have an SDM-C instance per network slice. The SDM-C will interact via dedicated plug-ins with the VNFs to control their configuration and resources.

4.6.2.4 Implementation proposal

Until now we have provided a high level description of the QoS/QoE Assessment module and the other blocks to which it has to communicate with. The question now is: How could we design the interfaces to those other systems? The main problem regarding this is that, as we know, the QoS/QoE Assessment module is intentionally not fully defined. As commented, we think that it is better to provide an open framework able to integrate different ways of implementing the required QoS/QoE functions. So, if the QoS/QoE Assessment function is something not clearly defined: How could we define a specific set of interfaces to it?

The different functions we want to integrate in the QoS/QoE Assessment module will be, after all, certain pieces of software code freely defined by the user. They could be something like simple threshold functions, or perhaps something more complex like an ANN already trained with the corresponding weights matrix… but anyway, they will be certain algorithms devised to compute the desired QoS or QoE from certain input parameters.

Hence, our problem here is a problem about *software deployment*. We need to deploy certain software functions (our QoS/QoE mapping, monitoring and management functions) which can be freely defined by the user into our QoS/QoE Assessment system. As a whole, software deployment is regarding all the activities that make a software system available for use; in our case, that "software system" are the QoS/QoE functions and their associated resources (normalisation functions, input parameters, output encoders, etc.). The software industry already provides a well-known solution to deploy and integrate indeterminate (in a certain degree) pieces of code using well defined interfaces. The key ideas are:

- To provide a common execution environment (or software container).
- Considering this execution environment, the user provides the piece of code to be executed together with a set of descriptor files with additional information for the execution.

An example of this approach is the EJB (Enterprise JavaBeans) specification which includes a container for web-related software components [JSR345]. The software components (EJBs) are deployed on a runtime environment using standardised interfaces. Another example specially designed for telco applications is the JAIN SLEE specification (JSR 240) where certain pieces of code (i.e., Service Building Blocks and Resource Adaptors) are deployed also on a software components container. Other examples are the SAP Business Objects deployable units, the IBM Rational ClearCase Deployment Unit Files, the Java Servlets Technology (JSR 315) or even the JAR files used in the Java Language (besides the Java code, the JAR files should contain certain standardised deployment descriptors to properly run in the runtime environment).

Of course, at this point we are not proposing to use EJB's, Servlets or whatever other specific technology for our QoS/QoE Assessment module. These are just examples. What we propose is to use the same conceptual approach. In our case, the pieces of code will be mainly the specific QoS/QoE related functions that will be deployed on a common execution environment according certain fixed rules.

To be more specific, and to be aligned with the description in the previous Figure 4-19, we propose to have the following different building blocks:

- QoS/QoE Mapping Function Blocks (MFB). These are the mapping functions code.
- QoS/QoE Input Adapters (IA). They will gather the relevant parameters from different sources in the architecture, and, after normalizing, they will feed the corresponding MFBs (i.e., they can also include the normalisation stage previous to the mapping function).
- QoS/QoE Output Adapters (OA). This is the mapping function output interface. It will encode the QoE/QoS metric according the required protocol.

These three types of building blocks should be deployed on a common execution environment using the corresponding deployment interface in order to fulfil the specific QoS/QoE approach for each particular case. They would be developed using a specific programming technology (e.g., a general purpose programming language) to meet the specific user's requirements. Each building block would be assigned to a specific slice, tenant or the infrastructure provider. The following Figure 4-20 illustrates this general idea, where different Mapping Function Blocks and Input/Output adapters are deployed on the common execution environment.

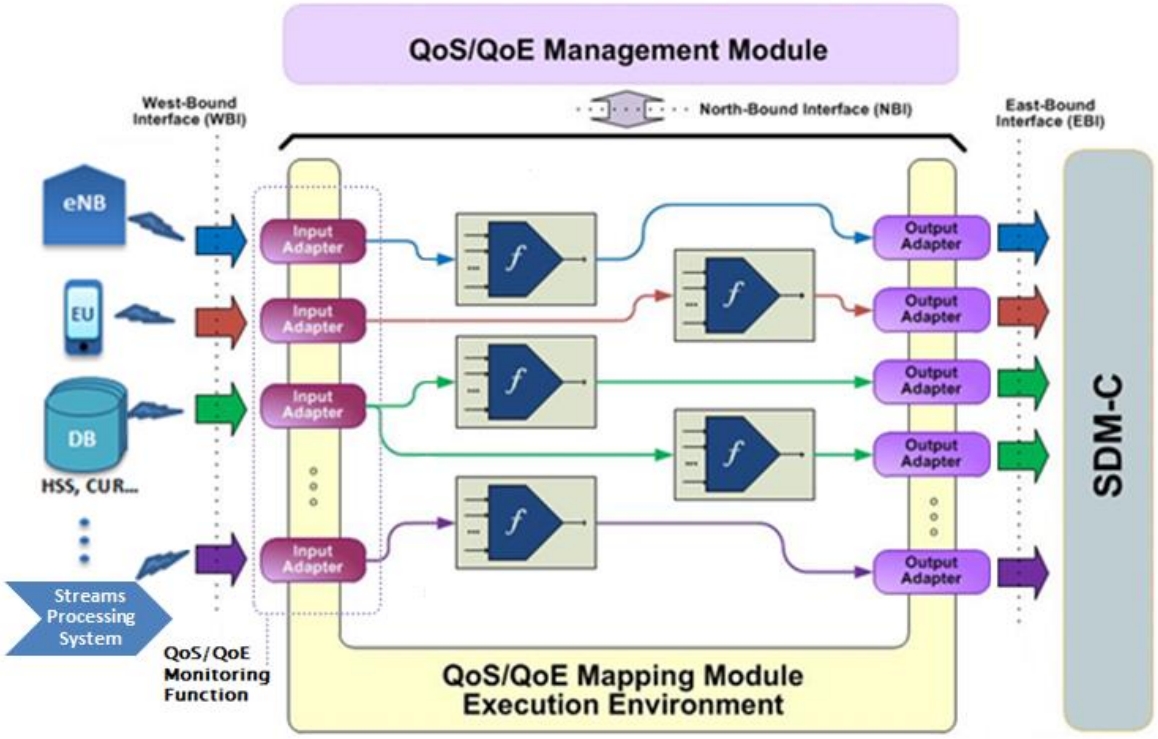


Figure 4-20: QoS/QoE Assessment Execution Environment & Building Blocks

The yellow U-shaped block represents the "Software Container" that works as execution environment for the three possible deployable components: Input Adapters, Mapping Functions and Output Adapters. This execution environment should be understood in a general functional way: it could be implemented as a single node in the network or as a multi-site distributed system. As we can see, the input adapters set (left) are implementing the QoS/QoE Monitoring function. The QoS/QoE Management function is integrated in the QoS/QoE Management module (top). All these building blocks (Input/Output adapters, mapping functions) should have their own life-cycle management primitives and defined deployment descriptors.

In Figure 4-20, we can see also the three main interfaces previously refereed:

- NBI (North Bound Interface or Management Interface), to communicate with the QoS/QoE Management Module in the 5GN Management and Orchestration Layer (Section 4.1).
- WBI (West Bound Interface or Monitoring Interface), to receive incoming input parameters.
- EBI (Est Bound Interface or Output Interface), used to send QoS/QoE Mapping relevant events towards the SDM-C (this interface will connect with the SDM-C's NBI).

In the following subsections we provide a high-level approach to these interfaces. We are not going to enter here in a fine-grain implementation details (programming technology, exact number of parameters, parameters type/range or other similar details), but as initial approach, we'll try to provide the basic functionality that should be supported for each case.

4.6.2.5 Management interface

As mentioned, this interface is used for exchanges between the QoS/QoE Management block and our QoS/QoE Mapping block. The operations supported by this interfaces can be split in two groups: the specific operations for the Execution Environment and the operations for the Deployable Units. Tables 4-1 & 4-2 show a list of possible operations for both:

Table 4-1: Operations for the Execution Environment

Operation Type	Description
`Configuration`	Get/Define the Execution Environment configuration. Explicit configuration parameters will depend on the underlying technology; they could be: – General purpose software configuration parameters (i.e. files location, hostnames, TCP/IP ports…) – Logging facilities configuration – Physical parameters (memory, CPU…) – Security parameters (i.e., access policy)
`Management`	Activate/Deactivate the execution environment itself.
`Licensing`	Install, remove and view the Execution Environment license(s).
`Query`	Query about status of the execution environment. This could include: – Current status (running, stopped, error) – List of already deployed units and their status – Usage statistics

Table 4-2: Operations for the Deployable Units

Operation Type	Description
`Deployment`	Deploy/un-deploy supported deployable units (i.e., Mapping Functions & adapters).
`Management`	Activate/Deactivate already deployed building blocks.
`Configuration`	Operations to configure the deployable units. This could be: – Number and type of inputs for the Input Adapters – Encoding scheme for the Output Adapters – Logging
`Bounding`	Bound/Unbound the different building blocks among them. For example, to connect a specific Input Adapter to a Mapping Function, and that Mapping Function to the desired Output Adapter.
`Licensing`	Install, remove and view the components license (if any).
`Query`	Query the status of deployed units (some status could be: running, stopped, error…). Information about the configuration status can be also provided.

The sequence diagram in Figure 4-21 shows what could be a typical operation to activate the platform and deploy a Mapping Function and the corresponding I/O adapters.

Of course, for something like this to work it would be necessary for each deployable unit to be generated containing the corresponding descriptor files. Those descriptor files should describe the peculiarities of each component; for example, for the Input Adapters, the descriptor files will probably enumerate the different input they will receive from the monitoring system, the sample period for each parameter, ports and host to connect, encoding protocols, etc. Descriptor files could be encoded using broadly accepted languages such XML or JSON.

Figure 4-22 shows an example of what a descriptor file could look like using XML. This is just an example for a hypothetical mapping function (*VideoStreamingQoE_MappingFunction*) which is assumed to compute QoE from a set of relevant video streaming parameters. As shown, the mapping function is bound to an input adapter (*VideoStreamingInputAdaptor*) and an output adapter (*MosOutputAdaptor*) which encodes the output according the MOS scale.

Of course, we are not proposing here to use XML or this so specific format; this is just a conceptual example, but that could be close to a possible real implementation.

4.6.2.6 Output interface

As mentioned in 4.6.2.1.3 the mapping function output encoding could be done in very different ways. This is why we have defined a general-purpose building block (the Output Adapter) to encode the output according the user requirements in a very flexible way. Anyway, this output

has to be sent towards one of the main building blocks in the 5GN architecture: the SDM-C, so there must be a common agreement about how the Output Adapters can generate the output according the SDM-C interface.

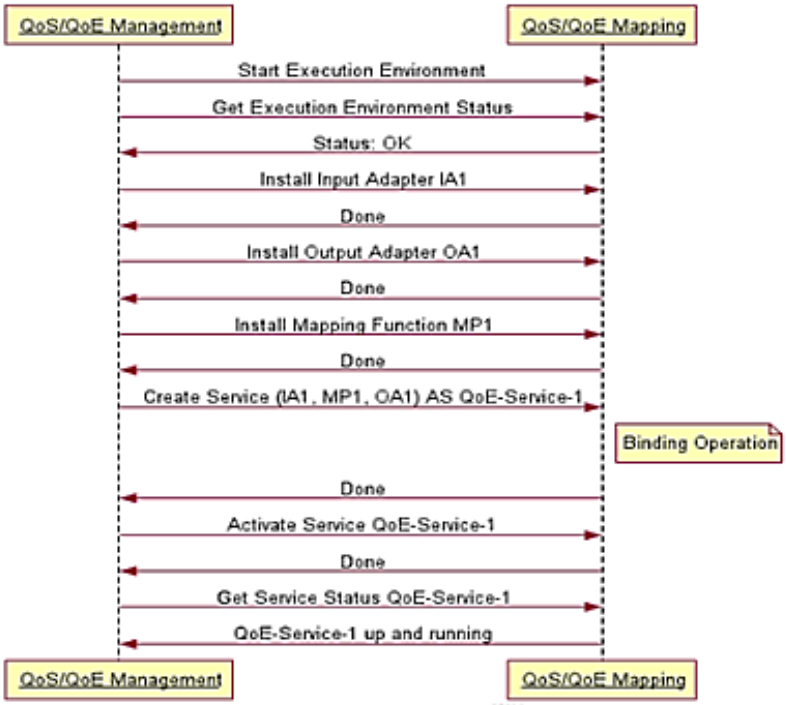


Figure 4-21: QoS/QoE Mapping/Management Modules Interface Example

```
<?xml version="1.0"?>
<mapping-function-jar>
  <mapping-function id="VideoStreamingQoE_mapping-function">
       <mapping-function-name>VideoStreamingQoE_MappingFunction</mapping-function-name>
       <mapping-function-vendor>5GNorma</mapping-function-vendor>
       <mapping-function-version>1.0</mapping-function-version>

       <input-params>
          <input-param type="float" min="0.0" max="1.0">Jitter</input-param>
          <input-param type="int" min="0.0" max="1e6">BitRate</input-param>
          <input-param type="String">ScreenSize</input-param>
          <input-param type="String">ScreenResolution</input-param>
          <input-param type="String" values="MPEG2, MPEG4">Encoding</input-param>
          <input-param type="float">Delay</input-param>
          <input-param type="int">PacketLoss</input-param>
       </input-params>

       <input-adaptor-binding>
        <input-adaptor-ref>
          <input-adaptor-name>VideoStreamingInputAdaptor</input-adaptor-name>
          <input-adaptor-version>2.1.1</input-adaptor-version>
        </input-adaptor-ref>
       </input-adaptor-binding>

       <output-adaptor-binding>
        <output-adaptor-ref>
          <output-adaptor-name>MosOutputAdaptor</output-adaptor-name>
          <output-adaptor-version>1.1</output-adaptor-version>
        </output-adaptor-ref>
       </output-adaptor-binding>
  </mapping-function>
</mapping-function-jar>
```

Figure 4-22: QoS/QoE Mapping/Management Descriptor Example

To avoid losing generality our proposal for this is to use the well know event-driven software architecture (EDA); i.e., using a communication pattern based on events [Cha]. In this architecture an event is simply defined as "a significant change in state". In our case, this can be a change in

the QoE Mapping Function which is relevant somehow (for example, the computed QoE could reach a certain threshold). From a practical point of view, this "change of state" is communicated by means of a (typically asynchronous) message: the "event notification".

The event emitter will be the QoS/QoE Mapping System (using the Output Adaptors), while the events consumer will be the SDM-C. The mapping system will have the responsibility to detect and transfer events, while the SDM-C will have the responsibility of applying a reaction as soon as an event is received. For the communication to be possible, we also need *events channels*, which are the conduits in which events are transmitted from the Output Adaptors towards the SDM-C. The practical implementation of event channels could be based on traditional components such as message-oriented middleware or point-to-point communication.

To ensure communication, the semantic of event notifications will be decided by the Output Adapter designer, but following certain common rules to ensure that the SDM-C can understand and process the notification once received. Event messages could trigger the usual scaling operations in virtual environments (e.g., scale in/out) or the VM power up/down operations. This would allow implementing the typical elasticity operations from QoS/QoE measurements.

Event notifications will have the common structure normally used in Event-driven architectures; they are usually with two main parts: an "event-notification-header" and an "event-notification-body". The header should include the most relevant information to process the event (e.g., event name and type, timestamp). The event body can be used to provide more detailed information about the event. Table 4-3 could be an example of an event notification definition:

Table 4-3: Example of an Event Notification Definition

Header:	EventName	This is a free text agreed between the OutputAdapter and the SDM-C. A set of pre-defined event names could be defined, so the SDM-C could know what to do for each case according to that. Examples: – Change_in_MOS_Scale – 80_percent_threshold_reached
	EventType	Different types can be considered (threshold, MOS…)
	Priority	Different priority levels could be defined (e.g. CRITICAL, MAJOR, LOW…).
	TimeStamp	It will be probably necessary to prioritise or track events.
Body:	Description	This can be an optional informational element (free text)
	Subscriptor	To link the event to a specific subscriptor (MSISDN, IMSI, IP addess…)
	Tenant	Tenant identifier
	Slice	Slice Identifier
	Source	The Output Adapter raising this event.
	QoE Metric	Distortion, blurring, freezing, noiseless, echo…
	ContentType	Voice, speech, music, video, 3D movie.
	AppType	Multimedia, gaming, augmented/virtual reality.
	UserType	VIP User / Regular User

4.6.2.7 Monitoring interface

As mentioned, the monitoring interface specific implementation will rely on the QoS/QoE Input Adapters. As presented in Figure 4-21 the QoS/QoE Monitoring function can be understood as the set of the specific Input Adapters deployed on the network.

Input Adapters are the components that constitute abstract interfaces with external resources. For example, a simple Input Adapter might bind to a network socket and propagate incoming network messages into to the QoS/QoE execution environment. As a whole, they represent and interact with other systems outside the QoS/QoE execution environment, such as network devices, protocol stacks, directories and databases.

Since QoS/QoE relevant information could come from heterogeneous sources (RAN, UE, billing systems, databases…) each IA should be specifically designed for each case; anyway, to provide modularity and generality the relevant information towards the internal Mapping Function Blocks should be provided in a common well defined way (e.g., using also an events based interface IAs could accept arriving protocol messages and fire specific events towards the internal MFBs in a common and well defined way).

For instance, in a regular deployment we could have an 'LDAP-IA' to interface with external databases using the LDAP protocol, a 'Diameter-IA' to interface with a billing system using the Diameter protocol, and so on; anyway, all IAs should generate signals towards the internal MFBs in the common well defined interface. This way different IAs could be easily connected to different mapping functions, and also, different IAs and MFBs coming from different vendors could work together into this common framework.

The Input Adapters layer decouples the network integration from the mapping functions deployed within the execution environment. Furthermore, different monitoring strategies (reactive, proactive or hybrid…) could be implemented for each specific IA to cover different QoS/QoE approaches.

4.6.2.8 QoS/QoE assessment function placement

Once we have a general view about the QoS/QoE Assessment block and its main interface systems another important question arises: Where, in the 5G NORMA architecture, this QoS/QoE Assessment module should be located?

It is clear this function belongs to the 5GN control layer, but this is true only from the functional point of view. If we consider how the QoS/QoE Assessment function could be physically deployed things are not so straightforward. Depending on the requirements the QoS/QoE Assessment module could have quite different aspects. Let us recall our two extreme examples previously mentioned:

- To compute an objective global QoE measurement from just a couple of physical parameters monitored inside a specific VNF (e.g., CPU and RAM usage).
- To compute real-time QoE measurements for a big set of individual end-users with a big number of input parameters and using DPI and a Big Data Analytics infrastructure.

For the first case the requirements are not too demanding. Probably the assessment function could be executed locally on each deployed NF. A simple threshold function executed on the NF could be used to raise an alarm towards the SDM-C module (or even some actions could be executed locally on the NF itself without involving other systems).

On the other hand, for the second case, it would be necessary to deploy a dedicated node (or even a set of distributed nodes) to perform DPI, BigData Analytics and real-time stream processing. Probably a complete CEP (Complex Events Processing) architecture would be necessary in order to trigger the events towards the SDM-C. Furthermore, the deployment of such complex set of nodes should be performed according the provider specifications; e.g., some nodes should be required to work in the central network, while others should be placed in the edge network to get the best performance.

Our view is that 5GN should provide a general solution for this, and not only a specific approach. Ideally, the three mentioned interfaces (north, east and west) should be implemented. However, to give more generality, its implementation should not be always mandatory; this will depend on the implementation needs. The most evident case is the first case above, where the parameters monitoring is so simple that it can be performed into the NF itself, so no WBI is really necessary. Also, even the QoS/QoE Management function could be probably omitted, since the QoS/QoE Mapping & Monitoring blocks can probably work as a stand-alone process with no special management functions required.

So, for the placement of the QoE/QoS functions we could consider the following options:

1. Deploy the QoS/QoE Assessment function as a stand-alone process into the individual NFs (physical or virtual). The full QoS/QoE Assessment function could be deployed on a single

NF (even if the NF-FG comprises more than one), or redundantly, on each NF composing the NFG. In practice this QoS/QoE Assessment function would be probably devised as a specific algorithm able to generate the QoS/QoE mapping events towards the SDM-C. WBI should be implemented locally in the NF to receive the monitored parameters from the QoS/QoE Monitoring block (the monitoring function could be executed internally also; in that case the WBI is not used). Also, the NF could optionally implement the NBI to communicate to the QoS/QoE Management module (if not implemented, no management functions will be provided).

2. Deploy the QoS/QoE Assessment module as a dedicated service in a dedicated node. The complexity of that service is variable, based on the tenant requirements (it can be a single node or a more complex service with a distributed set of nodes). Anyway, the basic idea is the same: the service should implement at least the WBI to communicate towards the SDM-C (est/nord interfaces are optional depending on the specific implementation).
3. Implement the QoS/QoE Assesment functionality into the SDM-C. In some cases, it could be simpler to have a kind of "monolithic" SDM-C including the QoS/QoE Assessment capabilities. This way the QoS/QoE Assessment Module work as an internal SDM-C module, but preserving its functional independence. Generation of QoS/QoE Mapping events "towards" the SDM-C is internal, so even this interface is not mandatory here.

Figure 4-23 represents these options in the 5GN functional architecture:

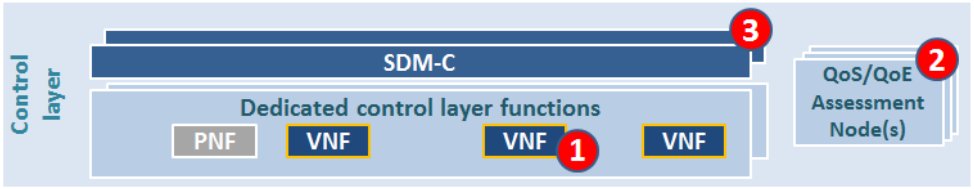


Figure 4-23: QoS/QoE Assessment Function Possible Placements

5 Security

Security is of paramount importance for future 5G networks. 5G NORMA has substantiated the need for security by analysing important 5G use cases and setting up (black-box) security requirements in [5GN-D21]. In a next step, taking into account the envisaged architectural principles of 5G NORMA, dedicated security requirements have been specified in [5GN-D31].

The present document now describes in detail the results of the 5G NORMA security work so far. While we discuss certain security aspects also in other parts of the present document, this chapter bundles the results and gives a comprehensive description of 5G NORMA security.

Section 5.1 provides additional motivation for the security work by showing the high socio-economic importance of providing a supreme level of security in 5G communication networks. Section 5.2 gives an analysis of potential security risks associated to new concepts and procedures defined by 5G NORMA, and gives guidelines in order to make sure that these risks are suitably mitigated. In Section 5.3 we investigate the applicability of the security mechanisms used in current LTE networks to the 5G NORMA architecture.

Most importantly, Section 5.4 describes new or enhanced security concepts tailored to the 5G NORMA architecture that have been investigated and specified in the framework of this project. Finally, Section 5.5 concludes the chapter.

Note that more detailed information on various security aspects are provided in Annex B, to which this chapter refers at various places.

5.1 Study on impact of security breaches

High profile announcements from European Governments and Industry in the context of 5G and IoT demonstrate that perhaps more than ever before the opportunity for 5G is significant to secure substantial levels of investment due to renewal of national critical infrastructure to underpin the Digital Strategy of Europe. The imperative to assure and provide evidence that technologies are being created that can be trusted is emphasized by Governments. In this section we present some examples from the communications and IT industry where security has been compromised in order to ascertain the impact of such security breaches, and to examine if there are particular elements of the 5G NORMA architecture that would make this more, or less, susceptible to such threats.

Since 2013 security company, Gemalto, have published a bi-annual report on publicly disclosed security breaches. Their latest Breach Level Index (BLI) report [BLI2016] indicates that 974 publicly disclosed security breaches occurred, involving over 500 million data records in the first half of 2016. As shown in Figure 5-1, there is a tendency for more breaches across all sectors over the past 3 years.

The 2015 security breach survey commissioned by the UK Government [UKG_SBS], broadly echoes the findings of the above report and further notes:

- For companies employing over 500 people, the ‘starting point’ for breach costs – which includes elements such as business disruption, lost sales, recovery of assets, and fines & compensation now commences at £1.46 million, up from £600,000 the previous year
- The telecoms sector had a sharp increase in security spending in 2015 – more than doubling the percentage of their IT budget spent on security from 13% in 2014 to 28% in 2015 (whereas financial services and were in-line with 2014)
- 50% of the worst breaches were caused by “inadvertent” human error
- 30% of large organisations were hit by DoS attacks in the last year
- Nearly 9 out of 10 large organisations surveyed now suffer some form of security breach – suggesting that these incidents are now a near certainty.

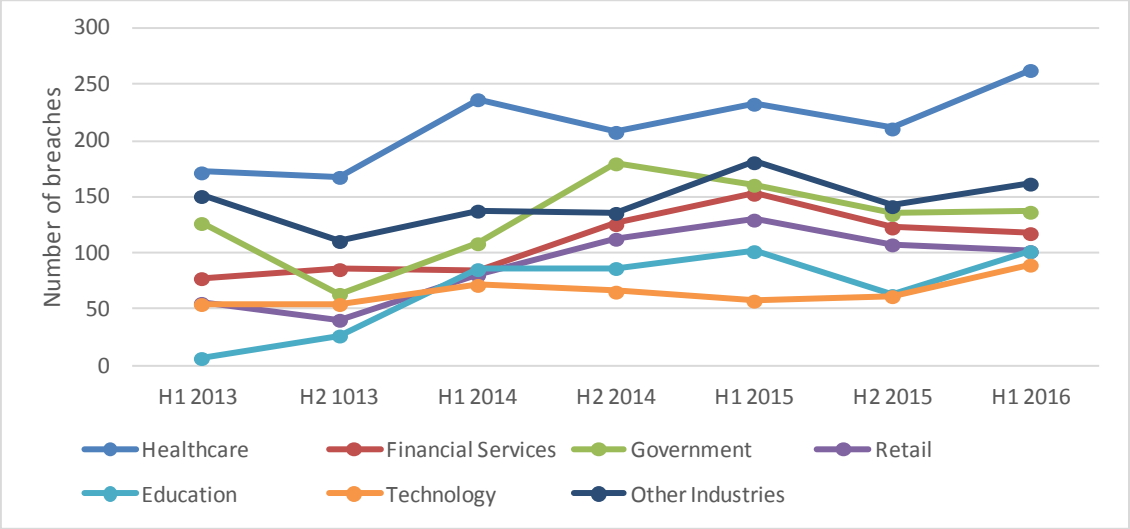


Figure 5-1: Number of worldwide publicly disclosed data breaches by sector (data from BLI2016).

The key lessons reported by Gemalto, and supported by the UK Government/PWC report, is not to rely on stopping threats at a network perimeter with boundary security, but to accept that breaches will occur and to work at securing data and minimising the impact of breaches that will occur.

The sharp increase in security spending reported above may be a result of the wide reported breach at UK-based MVNO TalkTalk which demonstrates the economic impact on a telecom operator of a simple security breach, see text box below.

> In 2015, the personal details of approximately 4% (157,000) TalkTalk's customers were breached – including the bank account details of approximately 15,000 customers [GUARD]. TalkTalk was fined £400,000 by the ICO (Information Commissioner's Office) for poor web-site security [BBC]. In addition to the significant negative publicity the attack results in the loss of approximately 2.5% of their customer base, and cost £60M to address the breach. Whilst revenues increased during the quarter in which the attack occurred the performance of the business was below analysts' expectations, and the share price dropped by 20%. Additional costs will include reputational damage and customer acquisition to regain lost customers – significant costs in a relatively mature market.

In moving towards a network to support multiple services and tenants, using virtualisation and cloud computing, the 5G architecture should consider the generic threats posed to mobile, tele-communication and ICT networks, together with non-generic features, such as shared infrastruc-ture network orchestration.

Kaspersky Lab report [CSO] a misperception that exists in virtualisation and cloud architectures. Whilst people believe that cloud environments can be more secure, the recovery costs of breaches are double that in a traditional IT environment, since malware is may be able to hop from one virtual machine to another companies (for example by exploiting hypervisor flaws). Companies are less prepared for failure disaster when it comes to virtualizsation compared to traditional in-frastructure. Virtualisation introduces new components (hypervisors, management servers and VM guests) that need to be patched, monitored and managed, on top of the host physical infra-structure. Nokia [NOK-NFV] recognise a potential threat of virtualisation, compared to tradi-tional networks, this is mainly due to reliance on additional software which creates a longer chain of trust, reduced isolation of network functions and fate-sharing due to multi-tenancy.

A further risk with network sharing with the flexible network architecture envisioned by 5G NORMA is the flexible service orchestration by different tenants, on a shared infrastructure. This requires 3rd parties to provision services at short notice – but necessarily allows a level of control

of a (different) network slice to a 3rd party, potentially on shared infrastructure, with the risks of malware transferring between virtual machines and functions.

So in the following, we discuss mitigation of generic and specific threats (Section 5.2), and propose new, enhanced security approaches (Section 5.4) in order to ensure that networks according to the 5G NORMA architecture will have the level of security that is required by the society and the economy in the future.

5.2 Mitigating security threats to the 5G NORMA architecture

5.2.1 Multi-tenancy

5G NORMA builds on the use of central and edge clouds, and supports multi-tenancy in the sense that two or more different networks may share a common cloud infrastructure. The basic and obvious threat in multi-tenant cloud (or NFV) environments is the failure to maintain strict tenant isolation.

For example, a tenant may manage to grab an arbitrary amount of resources (like computing time, virtual RAM allocation or disk memory) with the consequence that other tenants do not get the resources that are legally assigned to them, resulting in a DoS condition for other tenants. As another example, by exploiting flaws in a hypervisor, a VNF belonging to tenant A may be able to read or even modify memory allocated to a VNF belonging to tenant B.

The capability to provide isolation of tenants is a fundamental feature in cloud computing. Assuming that the relevant telco cloud software is designed, implemented, configured and operated with highest care in order to minimise the number of errors and thus the vulnerability, it can be concluded that tenant isolation works in telco clouds. An extra level of security may be achieved if the cloud offers the option for physical VNF separation (as required according to [5GN-D41], Section 3.3): VNFs of different tenants can be physically separated, so attacks between them via the local hypervisor are excluded. Note however that the need for physical separation reduces the flexibility in placing VNFs, so the hardware may be used less efficiently. An edge cloud site with a low overall number of separated physical units, e.g. processing boards, physical separation may not always be feasible

Multi-tenancy in the RAN is not restricted to edge clouds, but also affects "bare metal" RAN equipment. Equipment specific mechanisms need to facilitate multi-tenancy and provide proper isolation. For example, a radio scheduler running on a non-virtualised base station may not be aware of the different tenants but may be configurable to distinguish between groups of UE sessions and ensure a certain amount of radio resources per group. So if the groups are defined in a way that each tenant's sessions form one group, the radio scheduler will provide the proper resource isolation between tenants.

A more comprehensive discussion of security issues of multi-tenancy is given in Annex B.1.

5.2.2 Network slicing

The previous discussion on isolation-related threats deliberately referred to the tenant as the entity that requires isolation. However, there is a second notion where an isolation requirement applies: The network slice, as discussed in [5GN-D41], Section 3.3. In many cases, there may be a one-to-one mapping between tenants and slices. But even if several slices are operated by a single tenant, isolation between them may be required. For example, in one slice the tenant may deploy experimental or otherwise less trusted software than in other slices, so it is important to prevent that possible misbehaviour of the less trusted software in one network slice affects other network slices of the same tenant.

Another scenario is that a single tenant operates different slices for different end user services that provide different security levels and may be subject to different risks. If for example one slice is

more likely to become the target of a resource exhaustion attack, it is important to prevent such an attack from affecting other slices.

As these examples suggest, the threats and their respective mitigation discussed for the co-existence of multiple tenants on a common infrastructure also apply to the co-existence of multiple slices – whether they are operated by different tenants or not. The actual likelihood and impact of each specific threat may depend on whether the affected slices belong to different tenants or to the same tenant. An assessment of threats on this level of detail at this point of time (without sufficient experience with actual deployments) may not lead to tangible results and is out of the scope of the present document.

5.2.3 Multi-connectivity

Multi-Connectivity obviously increases the number of links that may be attacked. If multi-connectivity is used for increasing the reliability, and thus information is redundantly transmitted via more than one link, it is therefore somewhat more endangered. In case of different RATs or different fixed network transport techniques, security features may not be equally strong for all links. This can lead to attackers trying to trick mobiles into using mostly the RAT with the weakest security, or causing a DoS on a secure link in order to divert communication to a weaker link. An example for such an attack in existing networks is when an attacker makes UEs connect to GSM for calls rather than to UMTS and subsequently exploits the well-known security weaknesses of GSM to attack the user traffic. In mobile networks, such "bidding down" attacks can typically be prevented by providing cryptographic integrity protection for messages used to exchange network and UE capabilities. In case of a RAT not supporting such integrity protection (e.g. GSM in the example above), an effective policy for a UE would be to refuse to attach to networks via this RAT. But note that we do not expect that 5G networks use RATs with weak security. Moreover, when 5G networks use access networks with doubtful security, we expect that possible threats are mitigated by extra protection measures, similar what is done in LTE for non-trusted non-3GPP access networks.

A possible threat could be that an attacker who succeeds in breaking the cipher on one link may be able to use the knowledge gained by this to successfully attack also the other links. This threat can clearly be mitigated by using independent keys (or other cryptographic information) on different links, if the threat is considered relevant for two different RATs.

A difference in the security level may not be caused by different technical properties of different links, but rather by the fact that different organizations with different security postures may operate different network parts. For example, a network operator may make use of heterogeneous access networks, provided by more or less trusted third party organizations, and thus may require means to bring the security up to a satisfying level for all these different access networks. However, this issue is not specific to the novel 5G NORMA architecture. For example, already in LTE there is support for distinguishing between "trusted" and "untrusted" access networks, and specific security measures for "untrusted" access networks have been specified.

It can be concluded that for the time being it seems feasible to mitigate the specific threats of multi-connectivity in heterogeneous environments by means similar to those used in LTE.

5.2.4 Network virtualisation

The fundamental problem of virtualisation technology is the lack of visibility from the host operating system (OS) to guest virtual machine (VM), low adoptability of multi-purpose guest virtual machines (VM), and insufficiency in maintaining the consistency of security attributes to guest machines. In fact, it also introduces a number of potential vulnerabilities when MNOs virtualise or cloudify their infrastructure. For example, a malicious software could be planted into one of the guest VMs in the MNO's virtual network infrastructure. The state-of-the-art implementation of the host OS's virtual machine monitor (VMM) would not have enough information to differentiate a malicious software or genuine software, and whether affected processes or unaffected

processes are running in the guest VM. Even if the VMM implemented a functionality to differentiate the affected and unaffected processes it would still be unable to identify to which Tenant they belong. The malicious software infections in NFV can be more devastating than in the physical world. Those infections can even be propagated faster in a virtualised network infrastructure. Not only can they induce a number of unknown damages, chain reactions and havoc, but also bring upon more side effects than in the physical network infrastructure.

Even though, the VMM / hypervisor has a large amount of low-level information about the internal processes of the guest VM and the VM context switching, e.g., all memory page table mapping requests, the OS states and the privilege of the hardware states, however, due to the nature of micro-kernels, which have only the basic processes information embedded inside the memory page table, it is very difficult to check, extract or reconstruct all the guest VM actions from the basic processes information in the memory page table. Furthermore, the VMM does not have a capability to understand the meaning behind those actions and lacks the knowledge to diagnose those actions. The VMM should have an ability to understand those actions and to identify the responsible owner of those actions. Therefore, the VMM should increase the semantic awareness in the virtual machine introspection (VMI) and find a method to bridge this semantic gap [CHE01]. There is a suggestion of embedding the process owner information into process that proposed in Section 5.4.3.

5.2.5 Software-defined mobile network control

The landscape of SDN is increasingly exposed to a number of legacy and new network infrastructures. However, on the other hand, it also introduces a single point of failure at the heart of the network switching and routing. For instance, if virtual switches can request flows from the SDN controller, in order to switch or manipulate a flow, this actually opens the doors for vulnerabilities towards the SDN-based network infrastructure. In fact, when 5G NORMA architecture and design are applied, threats like hijacking of network resources or modification of routing policies might also be introduced to MNO infrastructure.

Particularly, 5G NORMA architecture relies on the SDM-C as the main architecture components to construct the flexible RANs and isolate the resources between Tenant's network slices. SDM-C requires to interact and exchange messages with the other network entities, i.e. SDM-X, SDM-O, PIM, VIM, PNF and VNF etc., to obtain the accuracy of allocation resources to each Tenant and maintain the optimum network performance. However, the current implementation of OpenFlow does not have sufficient security mechanisms to protect the interactions and communications between SDM-C and the other physical or virtual network entities, and to minimise the DoS attack on the communication channels. A security association must be established to protect the interactions and communications between SDM-C and the other physical or virtual network entities on southbound as well as on northbound interfaces, and to minimise the risk of being compromised by the DoS attacks.

5G NORMA proposes that the implementation of SDN protocol should be based on stream control transmission protocol (SCTP) [STE]. Integrated with datagram/transport layer security (DTLS/TLS) [DIE][JUN][TUX], they can provide secure interactions and communication channels between network entities, and have a better defence mechanism against DoS attacks than the Transmission Control Protocol (TCP) when the session is being established. Furthermore, SCTP has multi-homing capabilities and allows the southbound network entities to have redundancy links to another inactive SDN controller. Once the active SDN controller has been compromised, the southbound network can silently switch to the inactive SDN controller without losing the control plane communications and affecting the data plane.

5.2.6 Resource abuse

Typically, cyber-attacks are hidden in the bit streams that propagate through the network devices using communication pipelines. Most of the fundamental protocol design requirements are aimed at resolving security issues (e.g. denial-of-service attack). However, protocol-level defence is not

always sufficient to protect the network and to understand the hidden attacks. Furthermore, multistep cyber-attacks [CHE03] are not just hidden from the bit streams; they can also be hidden from the sequences of network functions. For example, an attack could start with gathering information from the first few stages of the sub-network functions. Once it gathers enough information and also reaches the core of the network function or the core of virtual network service chain, it could trigger the attack. Especially, when 5G NORMA architecture and design are applied, a MNO provisions a flexible RAN with multi-tenancy and multi-network slice services that requires defence mechanisms to detect the attacks and protect the tenants and as well as the MNO infrastructures. Even with a well-known proven reactive method, deep packet inspection (DPI) [LIN][SMA][DHA] deployed as part of network defence mechanism to scrutinise the bit stream pattern and grammar of an application, that is still insufficient to avoid the insider attacks and resources being abused by Tenants and their end users. Although DPI can be used to detect the flow pattern passing through a network and to look for applicative signatures, and a MNO can use service level agreements to set boundaries on the Tenant's resources, there is an insufficiency to protect the resources against the Tenants who initiate insider multistep and co-resident attacks. For instance, the current commercial cloud policy of treating the oversubscription of network resources is to allocate new VMs to the server with the most VMs. The Tenant could place co-resident attacks to oversubscribe the network resources. Consequently, it forces the MNO to process, manipulate and reallocate VMs and network resources that might affect the network performance, if the Tenant constantly sends requests to trigger such processes. Therefore, when the Tenant initiates a co-resident attack [HAN] by oversubscribing virtual resources, e.g. memories, CPU and VM etc, the infrastructure network performance will be affected even if the attacks fail. DPI and SLAs would be ineffective to deal with such attacks. Therefore, 5G NORMA requires a method to recognise co-resident attacks and service-chaining oversubscription attacks. The 5G NORMA architecture should provide functions to increase the visibility of the underlying virtual infrastructure and the awareness of the virtual network behaviour and potential anomalies. A possible solution is proposed in Section 5.4.2.

5.3 Applicability of LTE security concepts

The security measures for an LTE network are described in some detail in Annex B.2. They can be classified as follows:

- 3GPP specified security measures relevant for 3GPP specified reference points
- 3GPP platform security requirements and security assurance methods
- Non-standardised network and network element security measures

We shortly discuss the applicability of these mechanisms for 5G NORMA in the following.

5.3.1 3GPP-specified security for reference points

The reference point between the UE and network and the need to secure it clearly remains. From the viewpoint of the UE, for the traditional mobile Internet access use case, the LTE security mechanisms seem reusable. However, changes in the network architecture have impact on this reference point. In particular, the 5G NORMA architecture requires a more flexible access stratum key hierarchy, because the RAN functions will no longer necessarily be placed into a single entity. Moreover, while the specified LTE crypto-algorithms are still considered fully secure, the multi-service capability calls for a broader set of algorithms, in particular new, highly energy-efficient algorithms for massive IoT use cases.

In LTE, mutual authentication is performed between UE and serving MME. While the basic principle is applicable, 5G NORMA needs a more flexible authentication architecture and for this introduces the V-AAA approach.

New use cases may require additional changes of the security concepts for the UE-network reference point, which may include new authentication algorithms, new ways to identify subscribers and new ways to store credentials on mobiles. Also, to achieve higher robustness against attacks requires additional features, such as jamming protection. While such necessary changes are not

in the main focus of 5G NORMA, the flexible architecture is supposed to facilitate the integration of new procedures and algorithms implementing the required new security features.

For backhaul and core interfaces, 3GPP specifies the use of IPsec. But as the network functions become VNFs in core and edge clouds, secure communication between them may simply be achieved by regular cloud mechanisms, that may use encryption on any suitable protocol layer, not necessarily on the IP layer. Still, for inter-operator interfaces, security on a specific protocol layer may be specified, e.g. the IPsec approach of LTE may be re-used.

5.3.2 3GPP platform security and security assurance methods

Platform security requirements as for the eNB may still be applicable for all-in-one 5G base station "boxes", but not so much for virtualised base stations, where edge cloud deployments may have a lower risk of illegal physical access. Also assurance methods need to be adapted to the virtualisation scenario, with increased complexity, as boxes are replaced by VNFs that may be executed on various different NFV SW platforms, running on various different hardware platforms. However, these aspects are not in the focus of 5G NORMA, as the project aims at a new architecture rather than on a concrete implementation and its security hardening aspects.

5.3.3 Non-standardised security measures

Such methods are harder to capture, as it is not exactly known what kind of measures are typically implemented. Still, some of them will be heavily influenced by virtualisation, e.g. physical separation and zoning concepts may need to be replaced by purely logical separation. Also network element security measures will have to take into account the major change brought up by virtualisation. Others, like perimeter security, network internal traffic inspection, secure network element management protocols and secure operation of network services may be applied as before.

5.4 5G NORMA enhanced security concepts

5.4.1 Virtualised-Authentication Authorization Accounting (V-AAA)

The mobile telecommunication security architecture has been developed in 3GPP authentication and key agreement (AKA) and non-3GPP Extensible Authentication Protocol (EAP) for protecting the UEs and network elements [33.401, 33.402]. On the other hand, based on OpenStack Keystone [NFV-Sec], ETSI NFV uses security management catalogue to support Tenant authentication. However, both 3GPP AKA and non-3GPP EAP, and ETSI NFV have not got a complete security approach solution for supporting nor tracking mobile subscribers and NFV-based Tenants. 5G NORMA would suggest these two standards are independent system but should provide a single platform to oversee the following objectives

i) to keep the centralised governance of Tenants and mobile subscribers in the core network (central cloud),
ii) to allow a degree of freedom of tenant governance their subscribers at the access network (edge cloud),
iii) to use the AKA generated data to track and authenticate the mobile subscribers from the access network (edge cloud) and core network (central cloud),
iv) to maintain the mobile subscribers and Tenants with the point of attachment in the core network (central cloud) as well as in the access network (edge cloud),
v) to remain subscriber services even when the access network (edge cloud) is disconnected from the core network (central cloud), and
vi) to design a trust platform collecting the trust value from the mobile subscribers to the NFV-based Tenants, and from the physical network entities to the virtual network entities.

The design of Virtualised-Authentication Authorization Accounting (V-AAA) is based on the 3GPP AKA and non-3GPP EAP-AKA, and ETSI NFV, reuse of the existing cryptographic functions in 3GPP and tokenisation technique in OpenStack KeyStone, and provision of the remote method invocation to other network entities (see Trust Zone) and billing platform for subscribers, tenants and tenant's subscribers. Therefore, apart from the above 5G NORMA suggested objectives, there are additional objectives of the V-AAA. Firstly, to assist Tenant isolation and Tenant data isolation. Secondly, to support Tenant's data replication in many-to-one manner from the access network (edge cloud) to core network (central cloud) and local bi-directional replication approaches within tenant's network slice. Furthermore, another objective is to maintain the central governance in the MNO's core network as well as the decision when a Tenant is authorised to have a full control of their network slice which allowing theTenant to manage his subscribers in the access network (edge cloud). In this situation, the V-AAA may or may not need to share and move some of the core network functionalities (e.g., mobility management entity (MME) or home subscriber server (HSS)) to the access network (edge cloud) which depends on the SLA. It may also provide a better accuracy of UE point of attachment information under the complex 5G NORMA multi-tenancy, multi-network slicing, multi-level service and multi-connectivity environment. Last but not least, the additional objective is to increase the mobility efficiency of subscribers at the access network (edge cloud) and to reduce the traffic between access network (edge cloud) and core network (central cloud).

Subsequently, the V-AAA approach converts the traditional macro management to a per-region based, per-network slice based or even per-Tenant based micro management. For example traditionally, the MNO applies and uses the 3GPP AKA as an enforcement of the overall security management that remains at the core network. In contrast, the V-AAA takes a hierarchical, distributed and dedicated security management approach that can be located within the current LTE e Node B and only responsible for security management within the e Node B region. It can also be located within Tenant's network slice; then the security management responsibility scope is the entire Tenant's network slice. Furthermore, it depends on the MNO and Tenant to configure the scope of security management and to locate the V-AAA entity. This V-AAA approach also enhances the flexibility in security management, the accuracy of tracking information (i.e. mobility and billing information etc.) and the isolation of a Tenant's end-user based on its own geolocation database, which is equivalent to the current LTE eNodeB locations. For example, the first security goal in 3GPP (authentication) is to verify the UE's identity. While legacy networks perform this verification in the core network, the functionality could be shifted to the edge cloud. A comprehensive explanation of the hierarchical and distributed databases approach for subscribers, tenant and tenant's subscribers, and under different network integrations of V-AAA into 5G NORMA architecture framework are given in Annex B.4.4.

5.4.2 Shielded network behaviour

Section 5.2.6 described threats and vulnerabilities caused by the lack of visibility into the underlying network infrastructure [KRE][YAN], thus limiting the MNO's awareness of the overall network behaviour, and potential threats, risks and vulnerabilities. Even though ETSI NFV management and orchestration (MANO) [NFV-MAN1] has the knowledge of the allocated and unallocated resources to each isolated tenant to prevent such co-resident attacks, those attacks can adversely reduce the overall network performance. In these cases, DPI and SLA would become ineffective to deal with such attacks. Therefore, a proactive supervised learning for fault pattern detection [HOO1] can be applied to NFV to detect the network anomalies. This proactive detection can be integrated with VNF changing requests or SDN controller and MANO interaction patterns as information input to the inference engine for proactive defence means [HOO2][COL]. Additionally, a supervised learning of pattern detection of network anomalies to prevent service disruptions is required. This pattern detection method is an enhancement of those reactive methods to proactive method. This proactive defence method uses pattern recognition techniques to detect network anomalies and to evaluate the suitability of these methods within the given service provider business model. Tenant traffic patterns can also be diagnosed proactively. Hence, improving the sensitivity to threat detection to prevent intrusions or malicious attacks and increasing

the transparency of network behaviour. could potentially harden the security of 5G NORMA architecture.

5.4.3 Role-based VM introspection

Threats and vulnerabilities by lack of visibility from the host OS to guest VM, low adoptability of multi-purpose guest VMs, and the insufficiency in maintaining the consistency of security attributes to guest machines are elaborated in Section 5.2.4. In this section, an innovative solution of bridging the semantic gap is presented. 5G NORMA suggests a method to bridge the semantic gap by using role-based access control (RBAC) to verify the access right on the specific VM, and embedding the Tenant identity information and access token into the micro-kernel process data type declaration. Therefore, when Tenant requests to create a VNF, those Tenant identity information and access token would be inserted into the micro-kernel process and would give extra information when digital forensics diagnosis takes place. This technique allows the host OS to increase the level of visibility to the VM memory page tables. On the other hand, this technique can be integrated into the VMI to strengthen the protection of Tenants. Figure 5-2 illustrates how the Tenant's information embedded into the virtual memory page tables and the memory mapping of virtual to physical memories. Those memory page tables could belong to one of the Tenant's VNF. In fact, this is one of the objectives of 5G NORMA security approach by increasing the level of visibility from the roots of virtualisation technology to protect Tenants and reinforcing the consistency of the security status in the network.

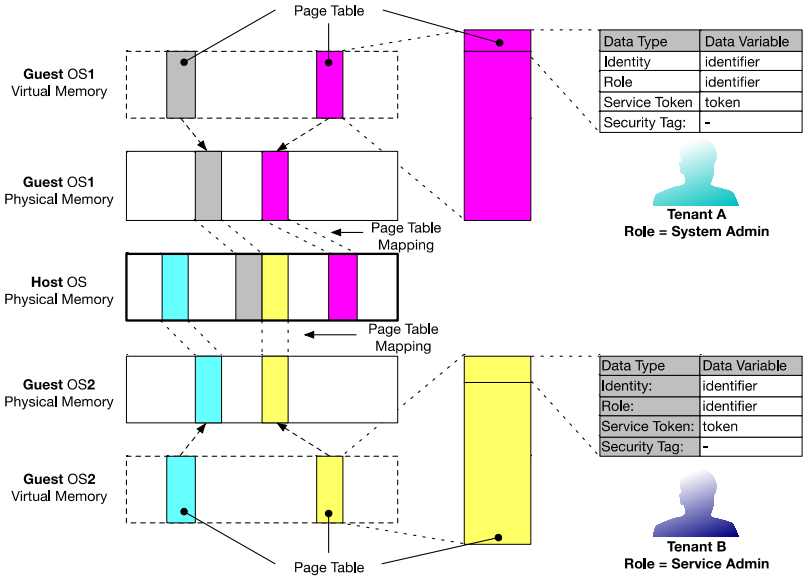


Figure 5-2: Physical and virtual memory mapping of virtual machine introspection (VMI) with role-based access control (RBAC)

5.4.4 5G NORMA RAN security concepts

5.4.4.1 A new Access Stratum security concept supporting multiple dynamically allocated radio interface security termination points

As described in Annex B.2, in LTE the radio interface is terminated at a single eNB (or at most at two, in LTE Dual Connectivity). In contrast, 5G NORMA features a much more flexible RAN comprising edge clouds as well as bare metal equipment, and RAN functions can flexibly and dynamically be allocated. Also, multi-connectivity is a native 5G NORMA feature. This requires a far more flexible Access Stratum (AS) security concept, which is described in detail in Annex B.3.1 and summarised in the following.

The Figure below visualises the proposed new AS security concept.

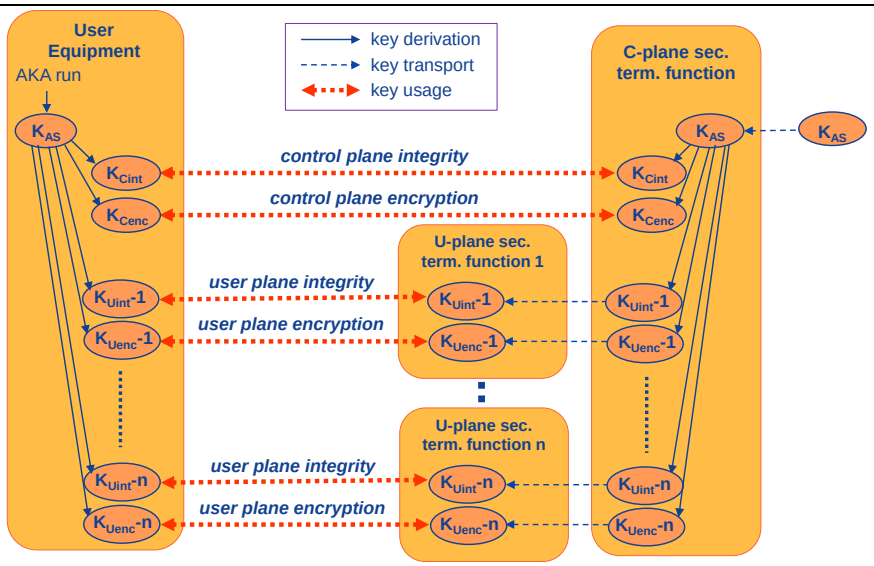


Figure 5-3: Flexible 5G NORMA Access Stratum Security Approach

Here, based on a key K_{AS} derived in the core, first encryption and integrity keys for the control plane are derived. Moreover, multiple key pairs for multiple possible user plane termination functions (that may be allocated at different physical entities) are derived to support encryption and integrity protection for user plane radio legs (where both encryption and integrity may be optional and for example be replaced by application layer security, if network policies allow this).

It should be noted that these procedures ensure that a compromised user plane function has no means to decipher or fake control plane messages. The user plane function may thus be allocated also on physically exposed entities, very close to the antenna, without endangering the security of the control plane.

The proposed setup allows to refresh a key pair for one user plane radio leg while keeping all other keys unchanged. A user plane security termination function that needs to refresh a key pair (e.g. to prevent a repetition of the key stream) can trigger the control plane security termination function to perform the refreshing.

A new key may also be needed when a user plane security function is relocated, in order to prevent that consequences of a possible security breach at one entity are propagated to other entities. Relocation may be necessary due to mobility of the user, but also due to network-side reconfiguration. Relocation may also require a change of the security algorithms, if the platform to which the termination function is relocated does not support the current algorithms or has other preferences concerning algorithms. All this can be executed individually per user plane radio leg.

A relocation may also be required for the entity terminating control plane security. In this case, like in an LTE handover, it may be reasonable to refresh the complete AS key hierarchy. A new K_{AS} may either be received from the core, to avoid any dependency on the previous one. However, to optimise speed, a new K_{AS} may also be derived from the old K_{AS}. Clearly, care must be taken that sufficiently often an independent new K_{AS} is generated.

5.4.4.2 Support of multiple network slices in the RAN

From the UE point of view, the presence of several slices does not affect the AS security procedures, as long as the UE connects to one slice only at a time. Otherwise, it depends on what slicing approach is taken. We discuss three possible approaches below:

- Approach 1: A single slice with respect to the AS security functions, but several slices otherwise: This may influence the way how K_{AS} is established. However, independently of how it works, the procedure must result in one K_{AS} to be used in the single set of AS security functions, so the key hierarchy and the procedures described above are fully applicable.

- Approach 2: Several RAN slices with independent AS security functions: The key hierarchy and procedures described above are applicable per slice. This means, that for each slice, an own K_{AS} needs to be established, and all key management is done independently per slice, based on the slice-specific K_{AS}. In this scenario, the number and location of the user plane security termination functions may vary between slices. A UE, which is connected to several such independent slices simultaneously, must obviously distinguish between these slices, e.g. maintain separated key hierarchies for the different slices.
- Approach 3: One may also imagine a slicing approach where different RAN slices are present that share AS security related functions. Most significantly, there may be a common control plane in the AS, but user plane radio connections may belong to different slices. This setup is also very well supported by the key hierarchy and the procedures proposed above: Security termination points belonging to different slices would be considered different (even if they were located at the same entity, e.g. at a single edge cloud deployment), so different, independent key pairs would be derived for them.

On the network side, the presence of several network slices will affect a range of procedures, whether UEs can connect simultaneously to multiple slices or not. This includes security procedures such as AKA. However, this does not affect the applicability of the proposed new AS security approach: As pointed out above, it is capable to support the various different multi-slice approaches quite smoothly.

5.4.4.3 Multi-service support by tailored radio interface security algorithms

As described in Annex B.2, LTE specifies three different pairs of crypto algorithms, where each pair comprises an encryption and an integrity protection algorithm. The intention of providing more than one algorithm pair is to have some diversity and fallback options in case one of the algorithms should be broken during the expected lifetime of LTE systems. So these multiple algorithms are all targeting the same, single LTE use case, i.e. mobile Internet access.

In contrast, 5G NORMA aims at supporting a very broad range of services. For example, this comprises massive IoT services, where mobile equipment may have a very low energy budget, as a single battery is required to last many years. To provide security for data transmissions by such devices, new, highly energy-efficient crypto algorithms are required, so called lightweight crypto algorithms. Another example is low latency cryptography, supporting services with ultra low latency requirements.

Annex B.3.2 discusses possible options for a choice of crypto algorithms that will allow to tailor radio interface security optimally for multiple different services supported by the 5G NORMA architecture.

5.4.4.4 Securing 5G NORMA RAN entities and interfaces

As described in Annex B.2, an LTE eNB needs to provide a "secure environment" to protect e.g. the keys it holds. In 5G NORMA, we assume that the respective functions are mostly running in cloud environments that obviously must provide a level of security that allows to maintain security relevant data such as keys and execute cryptographic operations. Such cloud security is ensured by the protection measures discussed in the previous sections.

Only in case of "LTE-eNB-like" deployments of bare metal RAN equipment, where radio interface security is terminated at a physically exposed location, there is a need for providing specific protection like the secure environment required in LTE. While solutions to provide such a secure environment exist, they clearly cause extra costs, which may lead to implementations that do not provide a sufficient level of security. Therefore, the recommended option is clearly to deploy the RAN functions critical for security at physically protected locations.

With a RAN implemented mostly by VNFs in secure cloud environments, communication between the RAN function is expected to be mostly secured by cloud security mechanisms. (See Annex B.3.3 for a more detailed discussion.)

The picture below illustrates the two deployment options for the RAN function terminating radio interface security and the necessary security features for each of the options. As the illustration suggests, terminating radio interface security in a secure cloud and relying on cloud security for all network interfaces avoids the need for a secure environment in physically exposed RAN equipment and a specific backhaul link security solution and is therefore the preferred option.

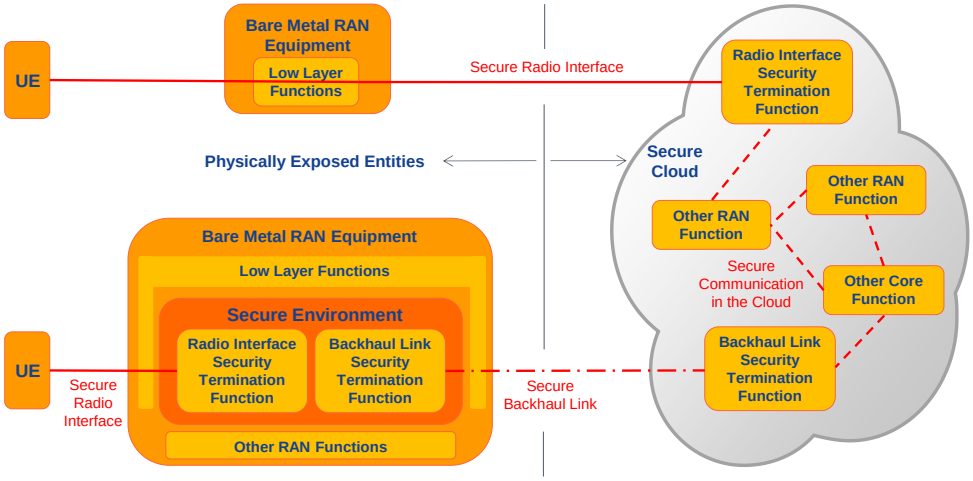


Figure 5-4: Security implications of different RAN function deployment options

5.4.5 Trustzone

5.4.5.1 Background and concept

Edge clouds are seen as one of the most important functions provided by 5G NORMA to verticals. Compared to central clouds, they are generally supposed to have more nodes, providing customer-oriented services with higher heterogeneity. Especially, they are expected to implement various policies, including access control, privacy, data protection, security functions, auditability, etc. Although most of these functions are provided by mobile networks already today, they are typically not specified for autonomous operation, i.e., they depend on the central security functions, which are probably provided by the central clouds in the 5G NORMA architecture. However, a connection between edge cloud and central cloud might - intentionally or unintentionally – become very poor, or in an even worse case, completely lost. In this case, the functional availability of the edge cloud can be significantly damaged.

The concept of the so-called “Trust Zone” is motivated by this risk. A Trust Zone is defined as a geographical area served by a local base station i.e. an edge cloud, where different policies are autonomously implemented to ensure data security, while as many services as possible can be provided, regardless of the connection status between this edge cloud and the central cloud. Due to the concern of tenant dependency requirements in security functions such as authorization, authentication and auditing, a Trust Zone (TZ) is a tenant-dependent function, i.e. each tenant can have its individual TZ in an edge cloud.

As the TZ strongly relies on distributed security functions, it is tightly integrated with the V-AAA framework. Generally, a TZ is an edge cloud V-AAA server extended with network monitoring function and emergency services.

In the framework of Trust Zones, 5G NORMA will define functions that allow dynamic implementations of security policies and procedures, mechanisms to protect these implementations against attacks.

5.4.5.2 Use scenarios of Trust Zones

In 5G NORMA D2.1, twelve different reference use cases are defined, including Industry Control, Enhanced Mobile Broadband, Emergency Communications, V2X Communications, Sensor Networks Monitoring, Traffic Jam, Real-Time Remote Computing, Massive Nomadic/Mobile MTC, Quality-Aware Communications, Fixed-Mobile Convergence, Blind Spots and Open Air Festival. Among them, the following ones are highly related to concerns of data security in normal situations as well as emergency situations, where edge clouds can be disconnected from central clouds:

- Industry Control
- Emergency Communications
- V2X Communications
- Sensor Networks Monitoring
- Massive Nomadic/Mobile MTC

5.4.5.3 State model of Trust Zone

As stated above, the activity of a TZ depends on the status of connection between the edge cloud and the central cloud. Hence, a state model can be built to describe the behaving pattern of TZs, including five different TZ states: Connected (C), Reconnecting (R), Weak Connection (C), Disconnecting (D) and Connection Lost (L) (see in Figure 5-5). In the "Connected" state, the TZ works in a centralised security mode. In the "Connection Lost" and "Weak Connection" states, the TZ works in a fully/partially decentralised security mode, respectively . The states "reconnecting" and "disconnecting" are transient states. A TZ can suffer from a weak connection due to attacks or disasters (C-W), and eventually totally lose its connection to the central cloud (W-D-L). Or, very rarely, it may also lose the connection completely in a sudden (C-D-L). The connection can be recovered from offline to weak (L-W), or from a weak sate to normal (W-R-C), but never completely rebuilt in a sudden. Hence, the L-R transition is invalid in this model.

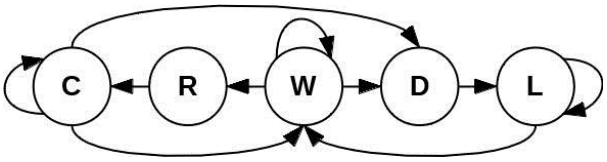


Figure 5-5: State model of Trust Zone. State abbreviations C, R, W, D and L stand for Connected, Reconnecting, Weak Connection, Disconnecting and Connection Lost, resp.

5.4.5.4 Functionalities required for Trust Zones

In this section, we will list and expand on different functionalities necessary for the implementation of local TZs.

5.4.5.4.1 Functionality list

A TZ includes the following functional modules:

- Central Cloud Connection Monitoring (CCCM)
- Local Access Assistant (LAA)
- Zone Management (ZM)
- Security Auditing (SA)
- Emergency Services (ES)

5.4.5.4.2 Central cloud connection monitoring

Depending on the connection status between the edge cloud (EC) and the central cloud (CC), the policies can be implemented in different ways. To automatically select and implement the most appropriate policy implementations, the connection status should be monitored in real time, which is the responsibility of the CCCM module. CCCM periodically visits the OSS and the SDM-O to evaluate the EC-CC connection status. When a change in the connection status is detected, it

informs the ZM to trigger the TZ state transition. Especially, when a disconnection or connection quality decrease occurs, it tries to diagnose the problem. The result can be utilised by other 5G NORMA modules such as the SDM-X to support a connection recovery. During a disconnection, it keeps monitoring the EC-CC connection to detect reconnection on time.

5.4.5.4.3 Local access assistant

Users expect local secure communications to remain if the edge cloud gets disconnected. This translates to allowing the edge cloud to provide functionality similar to LTE's MME. The edge cloud should be able to derive new keys as necessary. Hence, the 5G NORMA architecture is expected as able to partially move the security services, which are executed in core network blocks such as HSS and MME in LTE/LTE-A, to edge clouds. To enable this, a TZ provides the modules LAA, ZM and SA, which together construct an edge-cloud V-AAA server for distributed authorization, authentication and auditing services. Among them, LAA is responsible for supporting a central-cloud-independent local user access procedure in the AS domain. In scenarios where centralised secure AAA is available and preferred, LAA is deactivated. Otherwise, it performs as part of a decentralised security centre, executing the tasks of the MME that can be moved to edge clouds. Its functionality is strictly limited to the NAS domain due to the security consideration that the NAS keys can only be generated in the central cloud.

5.4.5.4.4 Zone management

ZM works as the core module of edge-cloud V-AAA server. Connected with all other TZ entities, it triggers and coordinates the state transition of the entire TZ, when it receives a report of EC-CC connection status changing. Integrated with interfaces to the central cloud and to the UEs, it collaborates the security procedures in the local base station. When the connection to CC is available, ZM cooperates with the central cloud V-AAA server to provide normal security services. When decentralised AAA services are needed, e.g. when the edge cloud is disconnected from the central cloud, ZM collaborates with LAA instead of the central cloud.

5.4.5.4.5 Security auditing

The policy management and their implementation will be subject to attacks. They will be under higher risk when disconnected, because the auditing server in the central cloud is in this case unavailable. An attacker may create a disconnection, compromise the edge cloud and get the access to the central cloud in the reconnecting process. An important aspect of TZ security then must be to implement local auditability capabilities. The local edge cloud will keep a record of all security events, and when the connection to the central cloud is re-established, all centralised security services will be able to access those records. Losing the local audit information could represent a security risk, as the central cloud would not be able to keep track of all the changes in data and communication.

The SA module is designed for this. In the decentralised AAA mode, it is responsible for keeping track of all security-critical events and operations in the local edge cloud, and recording them locally. In the centralised AAA mode, it is deactivated. These local auditing records should be uploaded back to the central cloud. Two different schemes are available for this. Either in an active approach, i.e. SA pushes the records to the central cloud when the connection quality allows; or passively, i.e. the central cloud pulls the records from SA when it needs. Moreover, an auditing centre is suggested to be implemented in the central cloud, so that the records from different edge clouds can be collected and examined. This centre is supposed to be located in BSS.

5.4.5.4.6 Emergency services

In many special emergency scenarios, such like earthquake, fire and explosions, an edge cloud can be suddenly disconnected from the central cloud, or the connection may be overloaded by a traffic burst, before the TZ is able to fully prepare the security data for the decentralised AAA mode, so that most services will become unavailable. On the other hand, some communication

services, e.g. emergency call, SMS, positioning, may greatly help people in these catastrophes. Concerning about this, the ES module provides a set of special services, which are accessible even with unauthenticated devices, in order to help under emergency circumstances.

Some emergency services such as emergency calls should remain available regardless of the central cloud connection status. The others, in contrast, are only activated only with certain diagnose of connection error.

EM provides an interface to IoT networks, through which it can receive reports from "disaster alarm" systems, which are supposed to provide information about public emergency events. This information is forwarded to CCCM, in order to help it diagnose the damage to the connection between edge cloud and central cloud.

In most cases, the public emergency events only reduce the connection quality instead of totally cutting off the connection. In this case, ES is responsible for deploying backup resources or approaches (e.g. rerouting, satellites, network redundancy, etc.), to attempt retaining a minimal connection or an alternative connection to the central cloud. It also manages the connection resources with a higher preference to promise the emergency services than others.

5.5 Summary and conclusion

By building on new networking paradigms such as NFV and SDN, and by introducing multi-tenant, multi-service and multi-connectivity concepts, the 5G NORMA architecture doubtlessly introduces new risks into mobile networks. In this chapter, we have analysed these risks carefully and proposed ways how to mitigate them. We have further investigated and specified a number of innovative security approaches that can be integrated into the new architecture (although they are mostly applicable also in more general contexts): Virtualised AAA, shielded network behaviour, role based VM introspection, a new AS security approach supporting flexible allocation of RAN security functions and the Trust Zone approach. At the time being, we are confident that the proposed measures, applied carefully, together with other relevant security measures that are not in the focus of the project (as a complete coverage would require an effort at much larger scale), can make 5G NORMA networks highly secure.

In the remaining period of the project, along with the finalisation of the overall architecture, we will revisit the threat assessments and, where necessary, adapt and enhance the novel security concepts to ensure that implementations of the finalised 5G NORMA architecture can result in highly secure networks that comply with the challenging security requirements raised by the expected 5G use cases.

6 Verification

6.1 Methodology

6.1.1 Objectives and expected results

Architecture design verification, which has been already introduced in deliverable D3.1 [5GN-D31], is guiding a two-step approach for architecture design iteration. Thereby, it finally contributes to a proof concept (PoC) of the 5G NORMA key innovations multi-tenancy as well as multi-service and context aware adaptation of network functions. As a modification in contrast to D3.1, meanwhile we distinguish between use case validation and architecture design verification. Whereas use case validation will revise the KPI fulfilment from a user centric view directly related to the use cases defined in deliverable D2.1 [5GN-D21], architecture design verification checks fulfilment of different requirements from system point of view based on generic 5G services as defined in METIS [MET14-D66] and evaluation cases that are introduced later on in this section. With this deliverable, we are located at the end of the first design iteration. Based on 6 evaluation criteria depicted in Figure 6-1, we will figure out the fulfilment of requirements and identify open issues that might be addressed during the second design iteration until the end of the project.

Figure 6-1: 5G NORMA evaluation criteria

Performance and functional requirements are defined in D2.1 and can be mapped to the generic 5G services. Whereas eMBB and mMTC are dominantly described by "enhanced Mobile Broadband" and "Massive nomadic / mobile machine type communications", uMTC services reach from "critical industry control" with very low latency to "real time remote computing" including low latency and at the same time minimum throughput requirements. Hence, uMTC spans a wide range of requirements and use cases. 5G NORMA evaluations for this category will focus on V2X services described by the use case "Vehicle communications" which allows for more tangible results.

Operational and security requirements depend heavily on the use cases to be supported. For our evaluation, this includes evaluation case specific requirement sets. The same applies for studies on economic feasibility conducted by WP2 and the check of certain soft KPIs that are detailed Section 6.2.5. As an overview, Table 6-1 outlines where results of the verification in the first design iteration are compiled. A concise conclusion on the different evaluation results is given in Section 6.2.

The check of performance and functional requirements should include all aspects that are essential for operation of the generic 5G services. Unfortunately, not all KPIs (e.g. peak data rates, cell capacities, inclusion of device networks…) are covered by 5G NORMA activities. Hence in order

to provide a comprehensive assessment as far as possible 5G NORMA will integrate results of other R&D projects.

Table 6-1: Results at the present state of architecture design

Result	Evaluation Criterion	Source
Performance requirements eMBB	Performance requirements	Section 6.2.1
Functional requirements eMBB, mMTC, V2X	Functional requirements	Section 6.2.2
Operational requirements	Operational requirements	Section 6.2.3
Multi-tenant isolation, Security issues with different services	Security requirements	Section 6.2.4
Sufficiency of interfaces between Service, Management and Orchestration Layer and within Management and Orchestration	Soft KPI	Section 6.2.5.1
Scalability of centrally arranged management functions	Soft KPI	Section 6.2.5.2
Feasibility of increasing network complexity with growing number of slices	Soft KPI	Section 6.2.5.3

Tools that are available to support the verification activities are described in Annex C.1.2. It is worth mentioning that besides system level simulations, demonstrators and protocol analysis tools a methodology for calculation of area wide network capacity has been developed.

6.1.2 Evaluation cases

For verification purposes, WP2 defined at the beginning of the project scenarios that combine a set of use cases happening at the same time [5GN-D21]. Meanwhile, WP3 has introduced stakeholder roles and respective interfaces. Hence, besides different deployment scenarios and architecture options even feasibility of stakeholder roles are to be included into verification activities.

Evaluation cases defined in the following combine deployment scenarios, architecture options and 5G NORMA ecosystem aspects (cf. Section 3) so that evaluations span over distinctive features describing these three elements. Concrete parameters of the three evaluation cases are detailed in Annex C.1.3. In order to keep evaluation effort and topics clearly arranged, evaluation cases try to address the most important questions and put them into evaluation stories[3] that are intended to replace the scenarios defined at project start. In Table 6-2 to Table 6-4, the differentiators describing deployment scenarios, architecture options, and 5G NORMA ecosystem are compiled.

Table 6-2: Deployment scenario

Differentiator	Dependencies	Distinctive feature
Macro node density	Directly related to traffic demand and/or indoor coverage requirements	Nodes per sqkm Classification into Urban, Suburban and Rural
Transport network options	Depend on operator specific preferences, technical requirements & economic feasibility	Bandwidth, Latency Technologies can be Fibre based, or microwave

[3] Evaluation story in this context means that we are running a map exercise for roll out of the described network looking at those concrete examples observing appearing challenges trying to solve them.

Network Coverage	Depends on used frequency bands, kind of multi-connectivity, air interface properties, has big economic impact	Coverage probability Also use case specific (e.g. MBB =nationwide with normal requirements of 95%)
Use case	End user requirements depends from the specific use case	User TP, coverage, latency, reliability

Table 6-3: Architecture options

Differentiator	Dependencies	Distinctive feature
RAN split alternative	Placement of RAN network functions in in different deployment types depend on traffic demand, transport network features	Central cloud, edge cloud, bare metal node
Core function placement	Placement of core network functions in different deployment types depend on end user requirements (e.g. latency), achievable multiplexing gains	Central cloud, edge cloud, edge cloud @ bare metal node
Technical solutions	Novel 5G enablers like Multi-RAT, mmW/ HetNets, M-MIMO will have specific requirements to transport networks and/or functional split, can sometimes be applied alternatively	Multi-RAT, HetNet, M-MIMO

Table 6-4: 5G NORMA Ecosystem

Differentiator	Dependencies	Distinctive feature
Services	KPI and system requirements may be directly mapped to services	eMBB, mMTC, V2X
Stakeholder roles	Stakeholders are current or future business roles that may implement their own business and interact with each other in order to provide services to customers	Mobile telco service provider, Infrastructure provider, MNO, Software vendors
Customers	There will be a lot of novel customer roles, most important from system point of view are those that directly benefit from network services	Tenants, end user

In the following sub-sections, three evaluation cases are described that will cover comprehensively the PoC of the 5G NORMA key innovations. Evaluation cases in this context are intended to provide a common playground toolset to for check of

- Performance requirements
- Functional requirements
- Operational requirements
- Security requirements
- Economic feasibility and
- Soft KPI.

All evaluation stories play in the common sample areas as proposed in deliverable D2.2 [5GN-D22], which are specifically the areas of Westminster, City of London, Kensington, and Chelsea (Figure 6-2).

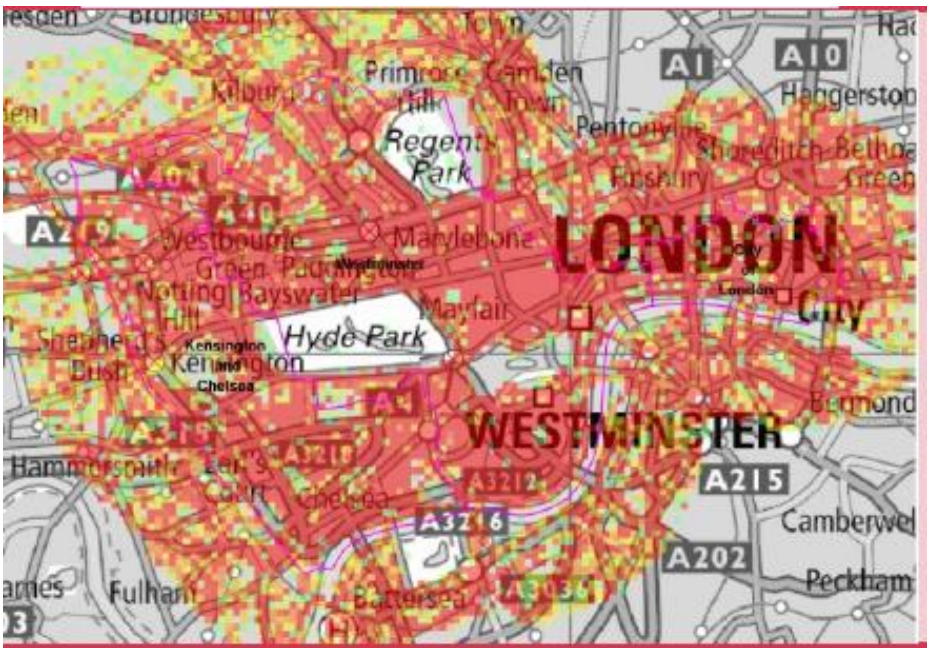


Figure 6-2: Modelled Traffic distribution in the sample area in Central London

The evaluation cases described in the following build upon each other and describe successively for the time span between 2020 and 2030

- how networks for MBB services can be deployed,
- how they can be extended to multi-operator or multi-tenant networks and
- how these networks may be expanded providing novel 5G services like massive machine type communication (mMTC) and vehicular to anything (V2X).

6.1.2.1 Baseline evaluation case

The baseline case is a deployment of an MBB radio access network in the London sample area where we compare the deployment of a LTE-A network with 5G NORMA network deployment. Capacity extension need is determined assuming the rules for network development described in Annex C.1.3.1 and applying the methodology for capacity verification described in Annex C.1.2.4.

Objectives of baseline evaluation are

- to provide a baseline economic cost case evaluation for MBB services deploying legacy LTE-A Pro technologies within the years 2020 to 2030,
- to compare this legacy cost case with the deployment of 5G technologies, identifying most important differences, and
- to check performance, functional and operational conditions in the London sample area against respective requirements for MBB originating in D2.1 and D3.1.

Most significant drivers for the development of (e)MBB are

- Peak bit rates (indoors and outdoors 10s of Gbps)
- Low latency, e.g. for interactive video conferencing (a few 10s ms)
- Traffic density (indoors and outdoors at least Tbps/km^2)
- High mobility with bit rates and latency (500 km/h, 10 Mbps, 10 ms)

Peak data rates indoors as well as capacity provisioning may dominantly be done by WiFi or 5G femto cells. We assume that those indoor networks are out of scope of 5G NORMA. But even at outdoor spots, small cells at higher frequency bands will contribute to higher peak data rates.

Realizing low latency for interactive video services, WP4 enablers like U-plane enhancements but also many novelties provided by other R&D projects [FANT5G] will contribute. Future MBB services will cause increased traffic densities and realizing the capacity demand is one of the main challenges for future MBB services providers. How this impacts the RAN infrastructure is described in more detail in Annex C.1.3.1. In general, it is to be anticipated that with respect to MBB there will be no significant performance differences between LTE-A pro and 5G technologies. Evaluating the 5G NORMA network considering the definitions given in Section 0, we assume that the mobile network operator (MNO) owns all infrastructure including RAN and edge as well as central clouds. Hence, for interfaces between Service, Management and Orchestration and within Management and Orchestration layer there exist baseline requirements in this case as all nodes are owned by the same stakeholder. Details on methodology and the parameters that we assume for this evaluation can be found in the Annex C.1.3.1. Results of performance KPI analysis can be found in Section 6.2.

In this sense, the baseline evaluation case provides the basic assumptions for the evolution of a MBB slice over the time span under consideration. It identifies most significant differences with respect to performance and cost compared to legacy alternatives and allows checking fulfilment of functional and operational requirements directly related to the MBB service.

6.1.2.2 Multi-tenant evaluation case

In the multi-tenant evaluation case the baseline 5G NORMA network for MBB shall be extended to a multi-operator network.

Objectives of multi-tenant evaluations are

- Constructing on 5G NORMA network considered in the baseline evaluation compare cost and deployment of MBB networks for single and multiple operators 5G NORMA networks
- Identify key benefits of multi-operator deployments
- Check for suitability of service and management / orchestration layers for multi-tenancy applying the roles described below.

While service offering and deployment is nearly identical to the baseline evaluation case, the focus of this investigation is to consider multi-operator networks. Moreover, we want to find out what would happen if 2 or 4 operators offering their MBB services in the sample area at the same time sharing parts of the network infrastructure and making use of the novelties provided by 5G NORMA architectures. Compared to single operator 5G NORMA networks, there would be the following differences

- Instead of a mobile network operator owning all physical and cloud infrastructure resources for evaluation, we select a variant of offer type 3 described in Section 3.2.3. The mobile service provider is offering mobile services to tenants (former mobile operators) that own their RAN infrastructures. Hence these former mobile operators change their roles into RAN infrastructure providers. At the same time, they rely on SLAs with mobile service providers for provisioning of their end-to-end MBB slices.
- Mobile service providers may have multiple SLAs with infrastructure providers (cloud and RAN infrastructure).
- Tenants would have to rent mobile services from mobile service providers but own their own software including element managers, service management, OSS, and other MANO functions (except VIMs).
- Tenants orchestrate the telecommunication services (compose, start, stop) for its own business.
- Alternatively, tenants own and operate their own spectrum exclusively or put a part of the spectrum into a spectrum pool sharing. Hence, they would share spectrum resources with other tenants according to [Singh_2014]. Alternatively, they could also use parts of the spectrum in common making use of SDM-X functionalities on joint resource management. Important benefits are also expected if antennas would be used jointly (s. Section 6.2.1).

In this sense a check of feasibility of selected stakeholder roles as well as security requirements are in focus of the technical evaluations. From an economical point of view, potential opportunities for cost savings are of interest.

6.1.2.3 Multi-service evaluation case

Even the multi-service evaluation case is building upon the 5G NORMA network considered in the baseline case. Now we trial run an extension of the network by two additional slices namely a slice for mMTC and another slice for V2X.

Objectives of the multi-service evaluation case are

- Comparison of 5G NORMA multi-service networks with legacy single stovepipe and identification of performance deviations
- Identification of multi-service cost benefits
- Check for suitability of service and management / orchestration layers for multi-service including defined stakeholder relation
- Check for feasibility and scalability of central management functions
- Estimation of cost impact for increased coverage, reliability and latency

Constructing on the baseline deployment of the 5G NORMA network described in Section 6.1.2.1, this evaluation is elaborating on extending the network by slices for massive machine type communication reflecting the use case "Massive nomadic/mobile machine type communications" and ultra-reliable communication as well as the use case "Vehicle communications".

The following performance requirements for massive nomadic/mobile machine type communications have to be fulfilled.

- The network should be able to handle a density of up to 200.000 active sensor connections per /km^2.
- The network should be able to handle small data payloads in the order of 20-125 bytes per message of sensor-type devices with often moderate latency requirements in the range of one second.
- The reception of those messages should be feasible for velocities of up to 500 km/h of objects the sensors are mounted on.
- The network should have increased coverage and availability compared to MBB.

Achieving the needed coverage and availability for mMTC our map exercise network is assumed to be deployed at sub 1 GHz frequency bands. The 5G NORMA radio protocol stack developed by WP4 will allow for optimizations for accommodating a large number of small data packets. In that sense, there should be an extra network slice configured for mMTC services. Realisation of high connection density could be facilitated by the deployment of cluster heads (device networks) collocated or integrated within small cell equipment. The operation of these configurable cluster heads would improve coverage and availability. It is not anticipated that in urban areas velocities of up to 500 km/h may appear. However, sensor applications e.g. at high speed trains will require air interface conditions that enable robust transmissions over the air interface. The new radios (NR) considered in other EU projects (e.g. FANTASTIC-5G [FANT5G]) will fulfil these requirements.

uMTC as originally defined in the METIS project [MET14-D66] imposes a broad spectrum of different requirements. To be comprehensive for our evaluation we just want to introduce low latency and high reliability requirements. V2X services as described in the respective D2.1 use case fit properly for this purpose. The key performance indicators for V2X services are

- Very low latency (1 to 5 milliseconds end-to-end latency) for critical services and applications.
- High reliability (nearly 100%) for critical services and applications.
- Very high mobility (absolute speed more than 200 km/h while relative speed more than 400 km/h).
- High positioning accuracy (0.1 to 1 meters).

- High density of connections for vehicles up to 10000 per square kilometre in scenarios with multiple lanes and multiple levels and types of roads).

Basically, the V2X services splits into 3 sub-groups

- mMTC sensors that generate small data payloads to be transmitted with moderate latency requirements;
- uMTC sensors that generate small data payloads to be transmitted with low latency and high reliability;
- Infotainment services with requirements similar to MBB.

From mobile service provider view, a service provided to a certain tenant should integrate the whole span of sub services into the provided network slice. Hence, mMTC, uMTC and MBB should be provided by a separate tenant specific slice. Hence, cluster heads that are already rolled out for the mMTC slice should be able to connect to more than one slice. The portion of uMTC traffic that is related to V2I communication can be linked via NR air interfaces enabling low latency and high reliability [FANT5G]. Antenna sites should allow for local breakouts for uMTC services. Hence, edge clouds at antenna sites are mandatory. The infotainment services may require bigger portion of the radio spectrum. Hence, to improve spectrum efficiency it would be favourable if the slices are not partitioning the spectrum into sub domains. In order to improve reliability and availability for V2I communication, the service should be covered by multiple RATs. With NR, low latency packet transmission may be realised together with MBB communication; hence, it would be favourable if not only sub 1 GHz but also low and high frequency bands (1800-2600 MHz) would be included into the V2X slice.

With respect to stakeholder roles, it is assumed that the mobile network operator (MNO) operating the MBB slice (as in baseline) owns all infrastructure including RAN and cloud as well as the software. In that sense, he combines the roles of mobile service and infrastructure provider. Vertical industries rent the slices for mmTC and V2X from the mobile network operator with offer type 1.

6.2 Intermediate verification results

6.2.1 Performance requirements

Performance requirements have been defined by WP2 in a way that they can be mapped to the generic 5G services. As already mentioned, relevant use cases from [5GN-D21] are

- enhanced Mobile Broadband (eMBB)
- Massive nomadic / mobile machine type communications (mMTC) and
- Vehicle communications (V2X)

Related KPIs are listed in the description of the evaluation cases in Sections 6.1.2.1 to 6.1.2.3. In a first step of verification, we focus on MBB requirements. Fulfilment of those performance requirements includes a couple of aspects that are not in the scope 5G NORMA. Sources for verification as well as a comparison between defined requirements and current available results are compiled in Table 6-5.

Table 6-5: Input sources for 5G MBB performance assessment

eMBB-performance KPI	D2.1	Fulfilled	Reference
Peak Data Rate	10s of Gbps	12.35 Gbps	[MAG-D41]
Latency	a few 10s of ms	10-20 ms	Operator statements
Capacity Density	at least Tbps/km^2	80 Gbps/km^2	C.2.1
Mobility	10 Mbps@500 km/h	NA	[FANT-D31]

6.2.1.1 Peak data rates

As defined in Annex C.1.1.4, peak data rate is the highest theoretical single user data rate, e.g., including reflecting the MIMO capabilities. In [MAG-D41], a reference for peak data rate calculations has been given (Table 6-6). As this calculation for FBMC uses a numerology valid for frequencies above 40 GHz, which does not apply MIMO schemes, we may multiply the data rates with the number of spatial steams provided by a M-MIMO transmission. Information on device capabilities currently is missing but the given figures per stream suggest to state that peak data rate requirements are fulfilled.

Table 6-6: Example for calculation of peak data rates [MAG-D41]

Occupied bandwidth (GHz)	0.270	0.540	1.08	2.16
$N_{subframes}$	10			
$N_{preamble}$	4			
$N_{payload}$	84			
FFT size	512	1024	2048	4096
Sampling frequency (GHz)	0.4608	0.9126	1.843	3.686
Number of subcarriers used	300	600	1200	2400
T_{symbol}, µs	2.77			
$T_{preamble}$, µs	6.11			
$T_{payload}$, µs	95			
$T_{subframe}$, µs	99.44			
T_{frame}, µs	979.4			
Bit Peak data rate (QPSK), Gbps	0.514	1.03	2.06	4.12
Bit Peak data rate (64-QAM), Gbps	1.54	3.09	6.17	12.35

6.2.1.2 Latency

Some MBB services like video conferencing will need e2e latency in the range of a few 10 ms. The e2e latency measured at the interface between layer 2 and layer 3 as proposed by [PA-5GPPP] is depending on multiple influencing factors so that a generally mature statement is difficult to give. Roundtrip times in current fixed national operator networks are in the range of 10-20 ms and it is quite likely that future 5G networks will provide improved figures. With respect to RAN and CN, there will be the following 5G NORMA enablers improving the e2e latency [5GN-D41]

- MAC level connectivity,
- adaptive use of master-slave RRC connections,
- Geolocation database and geolocation-based management, and
- service-aware allocation of network functions.

6.2.1.3 Capacity density

Capacity density is a performance indicator that is most significantly influencing economic KPIs, e.g., CAPEX and OPEX. Hence, there has been a close collaboration between WP2 and WP3 in order to maturely estimate economic effort achieving the required network capacity in the London sample area. Considering capacity and traffic density (s. C.1.1.1), at the first glance, it seems obvious that the demanded Tbps/km^2 could easily be realised locally using technology providing the peak data rates compiled in Table 6-6. From an operator perspective, however, network capacity is to be interpreted in an area wide manner. This means that a classified service area should be served with a certain data volume during a defined busy hour (s.cf. Annex C.1.1.1). Section C.1.2.4 and C.1.3.1 describe a methodology as well as basic assumptions for RAN development that allow for estimating the radio node density as a function of that kind of traffic demand.

In the model, the RAN is realised by three cellular radio layers

- Macro layer,
- SC layer at low and medium bands, and
- SC layer at high bands,

and in addition, by a WiFi or femto layer providing MBB services indoors in residential areas or at hot spots.

In order to span a spread between minimum and maximum traffic demand in a real network, we have investigated hexagonal site grids with macro site inter-site distances (ISD) of

- 200 m for serving high traffic demand,
- 500 m for serving medium traffic demand, and
- 900 m for serving low traffic demand in urban areas.

For calculation of wide area capacity we consider one single cell area. For extension of the macro layer, we assume that spectrum is used as soon as it is available and rely on the assumptions for spectrum availability in accordance to [5GN-D21]. In addition, improvements of spectrum efficiency by higher order MIMO (4x4) is considered. According to our model described in section C.1.2.4, the macro layer contributes with its average cell throughput directly to network capacity. Unlike to macro-cells, due their spotty coverage small cell layers contribute to network capacity either by their average cell throughput (capacity) or by their capability of off-loading the macro layer; – where the minimum of both is contributing to the area wide network capacity. Off-load of cellular networks by the WiFi has been investigated by several market studies, e.g. [OV15]. Similar references for the capability of small cell layers to off-load cellular macro layers are currently not available. Assumptions for this investigation have been compiled in Annex C.1.3.1 (Figure C-5).

The most important results of our assessment are (figures scf. Annex C.2.1):

- The major part of the MBB traffic (ca. 80%) is carried by the WiFi (or indoor-femto) layer.
- Depending on the load scenario, the portion of the MBB traffic carried by the macro layer is decreasing in the considered time frame to less than 10%. Remaining traffic portion is served by the two small cell layers.
- Due to different size of cell areas the off-load capability of small cell layers at low/medium bands is much higher than those at high bands. Therefore increasing the spectrum efficiency of small cells at low/ medium bands would make sense because off-load capability and cell capacity is in the same order of magnitude.
- Due to much higher spectrum availability the cell capacity of small cells at high frequency bands will be at least an order of magnitude higher than their capability to off-load the macro layer.
- The target architecture for MBB services elaborated in Annex C.1.3.1 reveals that deploying all spectrum available in 2030 will lead to bottlenecks at macro antenna sites. EMF issues will limit antenna deployment and multi-operator site sharing. Concepts for multi-operator RAN sharing that include virtual, bare metal (antennas) and spectrum resources could mitigate these limitations and contribute more efficient RAN deployment including CAPEX and OPEX savings as well as coverage improvements.
- It will be very costly to realise a uniform capacity density of Tbps/km^2 intended by [5GN-D21] in an area wide manner.

6.2.1.4 Mobility

Extensive work on enhanced mobility solutions is done by the H2020 project FANTASTC 5G [FANT-D31]. In high mobility scenarios, pulse-shaped OFDM is outperforming CP-OFDM robustness due to the time and frequency distortions caused by high mobility. Most work on enhanced mobility in terms of the definitions described in Annex C.1.1 is currently ongoing. Hence,

final results on mobility performance of the air interface are still missing. In terms of mobility management, 5G NORMA investigates the following enablers

- VNF Routing & Chaining with Mobility Support [5GN-D51] and
- 5G NORMA mobility management [5GN-D51].

Both will support user mobility by enhanced 5G network functionalities.

6.2.2 Functional requirements

Similar to performance requirements, also functional requirements groups are defined in D2.1 [5GN-D21] in a way that they can be mapped to the generic 5G services. The requirements groups defined for use cases "enhanced Mobile Broadband" (eMBB), "Massive nomadic / mobile machine type communications" (mMTC), and "Vehicle communications" (V2X) are analysed in the following subsections.

6.2.2.1 eMBB

Table 6-7 depicts the functional requirement groups that are relevant in the eMBB use case and shows which 5G NORMA technologies support the fulfilment of these requirements.

Table 6-7: Mapping of eMBB-related requirement groups and associated 5G NORMA technologies

eMBB-related functional requirement groups	Associated 5G NORMA technologies
Flexible capacity allocation on cell (sector) level	Multi-tenant dynamic resource allocation; Centralised RRM with virtual cells
Efficient backhaul utilisation	SDM-C-based routing; Flexible allocation of functions (SDM-O)
Application awareness	QoE-based routing QoE-aware eICIC
Multi-layer and multi-RAT connectivity	Multi-RAT integration (RRC for mm-Wave & UCA); 5G NORMA multi-connectivity architecture

6.2.2.1.1 Flexible, on-demand capacity allocation on cell and sector level

5G NORMA has developed a centralised RRM concept for virtual cells. It enables dynamic frame alteration of each eNodeB in addition to allowing the UEs to use the available sub-frames of multiple eNodeBs. The approach allows for the utilisation of neighbouring cell radio resources in case of congestion of the serving cell. In case of favourable locations of UEs (cell edge), available cell/sector capacity can be allocated in a more flexible manner by a dynamic alteration of TDD patterns. Furthermore, WP4 has developed a multi-tenant dynamic resource allocation scheme that introduces novel criteria and mechanisms for sharing radio resources among operators, tenants, network slices and individual subscribers. Finally, virtualized RAN protocol stack functions requiring high computational effort can be assigned to antenna site or aggregation site edge clouds depending on resource demand and availability, exploiting multiplexing gains between different radio nodes [5GN-D41].

6.2.2.1.2 Efficient backhaul utilisation

Efficient utilisation of transport capacity is mainly achieved by two 5G NORMA functionalities. First, the SDMC concept centralizes the routing decisions, extending ONF SDN concepts and thus exploiting the broad availability of relevant information at a single controller. Second, the decomposition of RAN and CN functions allows for more fine-grained deployment decisions by the SDM-O for such functions. SDM-O does not only take into account the available processing capacity at different infrastructure nodes, but also the utilisation of transport network paths. This

includes, but is not limited to, a varying degree of centralisation of the RAN protocol stack as well as the deployment of CN functions at the network edge [5GN-D51].

6.2.2.1.3 Application awareness

WP5 of 5G NORMA has developed a general framework for QoS/ QoE service control according to the SDMC paradigm, i.e., utilizing an SDMC application and the SDM controller for QoE/QoS mapping as well as QoS/QoE monitoring and enforcement. On the one hand, application awareness is realized by the QoE-aware eICIC scheme that integrates a direct measurement of user satisfaction (namely QoE) in the utility function used within a LTE framework performing a dynamic eICIC optimization in multi-RAT multi-layer networks. On the other hand, the novel QoE-based routing introduces a feedback-based Q-learning scheme to adapt traffic flows depending on the comparison between the link quality required by the application or service and the link quality available on different routes in the network [5GN-D51].

6.2.2.1.4 Multi-layer and multi-RAT connectivity

As part of the 5G NORMA RAN architecture, the *RRC User* is extended by *RRC for mm-Wave*, controlling UE multi-RAT connectivity to a macro cell (e.g., 5G macro cell) and a mm-Wave RAT transmission point. The so-called user-centric connection area (UCA) enables a fast re-selection of mm-Wave transmission points. Furthermore, the proposed RAN multi-connectivity architecture involves the use of an edge cloud where the RRC (control) and the PDCP layer will be located. The remaining protocol stacks will remain at the involved eNodeB sites. The approach offers the advantage of hiding frequent mobility between small cells (i.e., on the small cell layer) from the core network and facilitates either higher throughput or data duplication (and thus increased reliability) by transmitting data via multiple cells [5GN-D41].

6.2.2.2 mMTC

Table 6-8 depicts the functional requirement groups that are relevant in the mMTC use case and shows which 5G NORMA technologies support the fulfilment of these requirements.

Table 6-8: Mapping of mMTC-related requirement groups and associated 5G NORMA technologies

mMTC-related functional requirement groups	Associated 5G NORMA technologies
Management of a massive number of devices	User-centric connection area (UCA)
Device type aware RAT selection	Inter-RAT link selection function; RAN multi-connectivity architecture
Context aware mobility management	Adaptive (programmable) mobility management schemes

Moreover, D2.1 identified additional functional requirements that are relevant for the mMTC use case, namely low signalling overhead, support of sensors-type devices, connectivity support over cluster heads or relays, device-to-device (D2D) connectivity between sensors including network controlled local links, unidirectional as well as bidirectional communication between sensors, and flexible security and authentication procedures. These additional requirements are not explicitly tackled by any of the 5G NORMA innovations and concepts. However, they are in the scope of other 5G-PPP projects. For example, access procedures suitable for a huge number of devices are investigated by, among others, FANTASTIC 5G [FANT5G]. This also applies to D2D communications including connectivity over cluster heads as well as dedicated credential provisioning and authentication procedures. 5G NORMA will provide an overview in D3.3 (due September 2017) of how the respective solutions from other 5G-PPP projects address mMTC requirements.

6.2.2.2.1 The management of a massive number of devices

The expected high numbers of mMTC-type devices require multiple changes to next-generation mobile network architecture and protocols. Here, 5G NORMA contributes with the user centric

connection area (UCA) concept developed by WP4. Devices that are characterised by very sporadic transmissions (e.g., sensors, but also devices with messaging applications) will be transferred to a sub-state with low power consumption avoiding any reporting and measurements. Cell reselection is executed with low signalling overhead and only requires infrequent involvement of CN mobility management. Further, addressing of such UEs is simplified, such that it allows for managing a large number of devices [5GN-D41].

6.2.2.2.2 Device type aware RAT selection

Besides the *Inter-RAT link selection* block, which enables service- and device-aware link selection and packet scheduling if the UE is simultaneously connected to two or more RATs, 5G NORMA has designed a generic multi-connectivity RAN architecture. Most importantly, this architecture allows the UE to set up multiple links using radio legs from the same or multiple RATs. Once these multiple links are established, traffic flows can be flexibly distributed across the links, thus matching device type and traffic flow requirements with the link/RAT characteristics [5GN-D41]. The developed architecture can thus integrate dedicated MTC radio technologies such NB-IoT and select them for specific devices or services.

6.2.2.2.3 Context aware mobility management

While legacy mobile networks implement a rather static mobility management scheme, one of the central objectives 5G NORMA is provisioning of adaptable mobility management (MM) procedures. To this end, WP5 has developed a novel MM concept that not only allows to select the MM scheme to be selected and parameterized per network slice, but also to change the MM configuration during run time of a network slice, depending on the monitored UE movement pattern as well as the observed service requests and data traffic patterns. For this purpose, WP5 realizes MM as an application on top of SDM-C (or SDM-X), following the programmable network approach. More specifically, the developed mobility management scheme design, identification, and selection procedures allow to (i) exchange one MM application by another (this is executed by the SDM-O) or (ii) adapt the behaviour of the operating MM application by re-configuring parameters based on a pre-defined (and extensible) rule set or policy [5GN-D51].

6.2.2.3 V2X

Table 6-9 depicts the functional requirement groups that are relevant in the V2X use case and shows which 5G NORMA technologies support the fulfilment of these requirements.

Table 6-9: Mapping of V2X-related requirement groups and associated 5G NORMA technologies

V2X-related functional requirement groups	Associated 5G NORMA technologies
Coexistence of mMTC and URLLC vehicular applications	Management and orchestration of e2e network slicing; RAN and CN slicing
Fast and targeted dissemination of safety messages	Multi-connectivity RAN architecture; Multi-path routing
Optimizations for control plane and data plane functions	Decomposed and adaptive control and data layer architecture RAN slicing
Provisioning of intrinsic security mechanisms	RAN security architecture, NFV-native security functions, Trust Zone

Beyond Table 6-9, D2.1 has identified further functional requirements that are relevant in the context of V2X communication services. These include the creation of ad-hoc subnetworks, linking specific nodes and allowing for local access; retrieval of network information to be processed by external applications; ability to keep track of devices; ability to discover the topology of V2V networks (even if links have been established using non 5G links); support of content discovery for certain types of information (e.g., traffic conditions in the roads); prediction of robustness

against changing traffic conditions or other factors; and functions assuring confidentiality, integrity, and availability. Clearly, 5G NORMA cannot address all of these requirements. However, since many of them are in the scope of other 5G-PPP projects, the project will provide an overview in D3.3 (due September 2017) of how the respective solutions from these other projects complement 5G NORMA concepts in addressing V2X requirements in the context of an end-to-end mobile network system.

6.2.2.3.1 Coexistence of mMTC and URLLC vehicular applications

Vehicular communication applications exhibit different traffic characteristics and therefore require both mMTC services and URLLC services. On the one hand, 5G NORMA differentiates between the control of dedicated and shared functions by designing different controllers for the two categories. Accordingly, network slice templates include both SDMC applications for the control of dedicated functions and SDM-X policies that coordinate the usage of shared resources and functions. On the other hand, WPs 3 and 5 have developed Management and Orchestration (MANO) layer functions supervising the concurrent operation of multiple network slices on a shared infrastructure. These MANO functions enforce different resource commitment models (from fixed reservations to on-demand allocations) and prioritize resource allocation in cases of bottlenecks based on policies and SLAs. Furthermore, the 5G NORMA MANO design realizes a dedicated slice lifecycle management under control of the tenant as well as operator-controlled inter-slice resource brokering and optimization.

6.2.2.3.2 Fast and targeted dissemination of safety messages

Reliability allowing for fast targeted dissemination of safety messages will particularly be enabled by multi-connectivity concepts investigated by in WP4 and by SDM-C-controlled multipath routing concepts in WP5. While RAN multi-connectivity increases the reliability of message delivery over the air by means of packet duplication across multiple radio legs, the multi-path routing approach connects VNF appliances using multiple transport network paths. Furthermore, 5G NORMA has defined a closed loop control layer (involving SDM-C and SDM-O) for QoS/QoE-based re-orchestration. Several triggers for re-orchestration have been defined, ranging from shortage of IT resources (compute or storage capabilities) to more complex reliability and latency triggers in which assistance of the QoS monitoring and QoE mapping modules are needed [5GN-D41] [5GN-D51].

6.2.2.3.3 Optimizations for control plane and data plane functions

5G NORMA has developed a novel control and data layer architecture that does not only allow for varying degrees of centralisation versus distribution of selected NFs from RAN and CN, but also for a fast and programmable optimization of NF behavior. More specifically, for a wide range of control layer functions, the 5G NORMA controllers SDM-C and SDM-X can be combined with control applications that implement the desired network behavior. For example, EPC mobility management can be replaced with an application implementing a distributed MM approach. Data layer functions in both RAN and CN are modified by the controllers according to the logic realized by the respective application. Furthermore, WP4 has defined multiple network slicing options for the RAN, each of them exhibiting specific benefits that match selected use cases, such as, V2X [5GN-D41] [5GN-D51].

6.2.2.3.4 Provisioning of intrinsic security mechanisms

T3.3 of WP3 has continuously evaluated security threads of the current architecture design and checked where extensions of 4G security mechanisms is necessary. To this end, cryptography algorithms for efficient integrity protection and encryption have been investigated (see Annex B.3.2), addressing the need for adaptive, use case-tailored cryptography solutions. Moreover, NFV-native approach to implement security as VNFs increases robustness (e.g., against faulty behaviour of specific hardware platforms) as it allows, if required, to instantiate the functions flexibly on multiple hardware platforms or NFVI PoPs. The developed Trust Zone Concept (see

Section 5.4.5) is a specific security concept for isolated (disconnected) RAN parts and therefore applicable to V2I networks.

6.2.3 Operational requirements

Operational requirements as far as considered in this project phase have been defined in deliverable D3.1 [5GN-D31]. They are mainly related to deployment of multi-tenant and multi-service networks. Further requirements originating from [NGMN] will be considered in deliverable D3.3. Table 6-10 associates operational requirements to 5G NORMA enablers supporting them.

Table 6-10: Mapping of operational requirements and associated 5G NORMA technologies

Operational requirements	Associated 5G NORMA technologies
Multi-tenant dynamic resource allocation	Intra-slice resource broker, Section 4.5.1, Annex C.1.2.1
Saving of operational and capital expenditures	Multiplexing gains by joint usage of network functions and virtual resources as well as bare-metal and spectrum resources, sensitivity analysis to be reported in D2.3
Service specific and context-aware derivation of service requirements, adaptation and placement of VNF	Proper design of slice templates, Section 4.1.2
Flexible vertical-specific and service-specific detection of traffic and dynamic network monitoring	QoS/QoE Assessment in 5G Norma, Section 4.6.2

6.2.3.1 Multi-tenant dynamic resource allocation

Network resources such as communication, storage, processing, and function resources are provided in a sliced manner based on different service requirements. This will be performed using a pool of resources, which are reserved for a given network slice to achieve specific performance goals. Thus, a resource optimization mechanism is required to optimally allocate resources optimizing metrics such as spectral efficiency or network energy consumption. Distinct tenant requests can result in different profits. The prioritization of a multi-tenancy system is very important to enable new business models where the mobile service provider has the role of a mediator. In such case, based on the resource availability, the main objective of the mobile service provider, i.e., the admission controller is to admit multi-tenant requests in order to optimise the global revenue. This will result in potential unfairness, which must be properly handled.

The basis for flexible allocation of resources is described in Section 4.5.1 where it was assumed that in every multi-tenant network, there is one privileged slice belonging to the mobile service provider that includes a pool of resources and flexibly can be flexibly allocated to the other slices on demand. In Section C.1.2.1, a simulation methodology is sketched that describes WP5 activities investigating multi-tenant dynamic resource allocation. Hence, the topic will be addressed in the remaining project time.

6.2.3.2 Saving of operational and capital expenditures

An important requirement is represented by OPEX and CAPEX reduction. Shared utilisation of resources and network equipment, e.g., to accommodate and balance tenants' capacity requests, helps to realise multiplexing gains and reducing costs significantly.

OPEX and CAPEX saving potential introduced by multi-tenant and / or multi-service networks will be investigated by WP2 (cf. upcoming deliverable D2.3). Main sources for cost savings are expected by multiplexing gains due to slice specific different traffic behaviour or by utilisation of slice- overarching utilisation of radio resources. The saving potential will heavily depend on assumptions regarding multi-slice traffic concurrency. Furthermore, there might be different saving

potentials dependent from the location within the network topology, i.e., centrally, at edge clouds, or at antenna sites. As spectrum is a limited resource that cannot be extended arbitrarily or migrated to other locations, investigations on the benefit of shared spectrum usage are of interest. Antenna installation and EMF limitations will significantly limit the expansion potential of future antenna sites. Hence, joint usage of bare metal resources by multiple network slices may be highly beneficial. As these topics have to be investigated from technical as well as economic view, there will be joint efforts of WP2 and WP3 in the remaining project time.

6.2.3.3 Service specific and context-aware derivation of service requirements, adaptation and placement of VNF

Requirements on QoE and QoS, mobility, security, and other requirements can be derived dynamically based on detected services as well as network context. For instance, mobility management requirements may be identified according to UE type and class, UE mobility pattern, the detected service characteristics in terms of reliability and continuity, and RAT capabilities. QoE and QoS requirements may be derived based on dynamic policies as well as real-time user plane traffic monitoring. The VNF selection, placement, and configuration can be adapted based on the derived requirements of the detected service to enable service and in-service differentiation. VNFs may be located at the central cloud for the detected services having relaxed latency requirements, while a VNF may be placed at the edge of the network for detected services requiring low latencies. For instance, multi-connectivity functionality may be configured to increase coverage and throughput for MBB services, and alternatively it could also be configured to provide redundant connectivity, e.g. for uMTC services.

All these static service specific requirements are to be addressed by proper design of the slice templates during the slice preparation phase.

6.2.3.4 Flexible vertical-specific and service-specific detection of traffic and dynamic network monitoring

Different services may require different service detection methods, e.g. IP-based Multimedia Services (IMS) and other operator-provided services may be identified via control plane signalling, while Over-the-Top (OTT) or Internet services may require user plane traffic monitoring to allow for detecting specific application flows. A service detection function should be designed sufficiently flexible and extensible to enable different service detection methods for any foreseeable service. Continuous monitoring will enable service- and context-aware NF adaptation. For instance, they might provide monitoring of available network resources for more efficient mobility and multi-path control, and real-time monitoring of user traffic flows to enable application-specific and context-aware QoE/QoS management and dynamic routing control. Furthermore, it might provide monitoring of processing and radio resource allocation schemes, e.g., for better slice management and verification/ or planning of the performance of slices. Vitally, they might also assist the detection of security-related vulnerabilities or violations. Network signalling due to dynamic monitoring and reporting should be minimised in order to increase resource usage efficiency.

These requirements add a dynamic component to the static requirements treated in the last sub section and have three tangential points with the 5G NORMA architecture

- The orchestration of "probe" VNFs that perform DPI in the Gateway elements
- The processing of this information at the QoE/QoS monitoring level or, in case of IMS, by specific SDM-C applications
- Transformation of service, slice and context awareness in the control plane

The QoS/QoE Assessment and Control systems described in Section 4.6.2.2 may build a basis for implementation of service specific and context-aware QoE and QoS monitoring and management. Currently, there are no specific activities on service specific traffic detection but 5G NORMA systems basically will establish a framework for fulfilment of these requirements.

6.2.4 Security requirements

Assessing the threats and risks present in a complex system and the compliance with security requirements of the system is a difficult task, even if it relates to one concrete implementation only. Mostly, this is done by expert assessments, in a kind of unscientific way, as there is no exact reproducibility of such assessments – different experts may come to different results. Even less tangible is the theoretic evaluation of compliance to security requirements for a general architecture (as opposed to one concrete implementation). So there is no way of doing a verification in the strongest sense of this word.

Still, in the following we give an assessment of the compliance of the architecture to security requirements stated in deliverable [5GN-D31].

Tenant isolation: As discussed in section 5.2.1, tenant isolation can be assured by NFV specific mechanisms, assuming careful implementation of the NFV platform. Tenant isolation on bare metal equipment must be ensured by equipment specific mechanisms.

Secure Software Defined Mobile Network Control: As presented in WP6 demo #3 [5GN-D61] a secured authorization and delegation access control can be assured in the 5G NORMA SDM-C/X/O deployment. A proof of secured authorization and delegation concept demonstration uses OpenID Connect and open authentication protocol version 2 to access a simple network resource that can be interpreted as SDM-C/X/O.

Physical VNF separation: The project did not focus on the physical VNF separation in terms of practical and theoretical measurements. In general, the flexible allocation of functions in the 5G NORMA architecture leverages also physical separation of functions while it is at the same type able to cope with the possible restrictions of the freedom in function placement imposed by such physical separation requirements, given suitable pools of physical resources.

Flexible security: The dynamic allocation of security functions allows flexibly adapting and optimizing the setup for the various use cases. A choice of crypto algorithms has been investigated that can be flexibly applied depending on the needs of the applications to be supported. The access stratum security concepts support flexible allocation of functions terminating the radio interface communication.

Support of reactive security controls: The project did not focus on this kind of security measures. Still, the flexible 5G NORMA architecture is an excellent basis for the deployment of such innovative security controls.

Security orchestration: The project did not focus on security management and orchestration, which is clearly a challenge in complex architectures making use of NFV and SDN. However, work on this is carried out throughout the respective research community, and the results are expected to be applicable to a 5G NORMA network.

The project did not focus on reliable fallback alarm system design - work on this is carried out in the product security research community. Still, networks implementing the 5G NORMA architecture are expected to accommodate such solutions.

6.2.5 Soft KPI

Deliverable D3.1 has defined “soft-KPIs” that measure sufficient flexibility, manageable network complexity, sustainable standardisation effort, sufficiency of interface definitions, and scalability, in particular for centrally arranged management functions, in a qualitative way. More general, by check of this soft-KPI, it shall be ensured that architecture design provides mature results that can be handed over to next step realisation activities. References to respective 5G NORMA activities are given in Table 6-11.

Table 6-11: References to Soft KPI related 5G NORMA activities

Soft KPI	5G NORMA activities
Interfaces between Service Management and Management and Orchestration	D3.2, Sections 3.1.1 to 3.2.3, 4.5.2
Scalability of centrally arranged management and control functions	Adopt results of scalability studies available in context with SDN controllers
Feasibility of growing number of slices	Bottleneck considerations in D3.2 Section 6.2.5.3

6.2.5.1 Sufficiency of interfaces between service layer, management & orchestration layer and within management & orchestration

The 5G NORMA ecosystem is characterised by a multitude of stakeholders and their heterogeneous technical requirements when interacting with each other via dedicated system interfaces. Section 3 has selected and analysed the most relevant offer types of mobile service providers/ (MSPs (or MNOs, respectively) to external tenants that want to deploy their own telecommunication services. The system design (as described in Section 4) reflects the identified requirements in the 5G NORMA functional architecture. Generally, the design foresees that both multi-tenancy and multi-services are realised by different network slices that are customised according to the service characteristics provided by multiple tenants.

In particular, the design of the 5G NORMA Management & Orchestration layer (cf. Sec. 4.5) explicitly analyses where interfaces need to support multi-tenancy and where a single-tenant environment can be assumed. For example (cf. Sec. 4.4.1), the interface between Service Management and the Inter-slice Resource Broker carries information on network slices that has to be extended with tenant-specific meta-information. This covers tenant-specific commissioning information, management and control options, as well as SLA-related information. This enables the Inter-slice Resource Broker to define and maintain sharing policy catalogues, which are used as a decision base for allocating resources to different tenants (multi-tenancy) and different service or slices (multi-service aspects) during operation. The SLAs between a tenant and MSP/MNO define the amount of resources allocated to tenants as well as the level of control executed by the tenant, ranging from a "monitoring only" option (cf. Sec. 3.3.1) to far-reaching slice configuration and control options delegated to the tenant (cf. Sec. 3.3.3). The same logic applies to domain-specific application management functions. For example, in case of 3GPP, mobile network management functions (such as the Element Manager) decide how to split available resources among multiple tenants and how to prioritise traffic from different slices, e.g., when sharing an eNB. While Service Management and Inter-slice Resource Broker are multi-tenant-capable entities, the individual instances of the NFV MANO stacks (NFVO, VNFM, VIM) operate as if they were placed in a single-tenant environment. Hence, they implement the reference points as given by ETSI NFV MANO and do not require any extensions on their models and objects to support multi-tenancy. With respect to security, MSPs/MNOs expose a certain part of their overall system, in particular, the Service Management and Orchestration functions, to external parties and therefore need to establish appropriate access control (authentication and authorization) mechanisms (cf. Sec. 5.4.1 and Annex C.3.4.). On the other hand, tenants have to trust the MSP/MNO that network slices are operated according to the SLAs and data integrity is maintained.

In summary, the 5G NORMA interfaces between Service-, Management and Orchestration Layer and within Management and Orchestration utilise existing concepts from single-service and single-tenant environments and extend them with specific functionality that have awareness of multiple tenants/services. On a conceptual level, the current design therefore provides the basic means to sufficiently support multi-service/multi-tenant environments and to fulfil the related operational and security requirements. The in-depth verification of the sufficiency of the relevant interfaces remains to be done as part of the final design iteration of 5G NORMA.

6.2.5.2 Scalability of centrally arranged management and control functions

Moving functionality from the edge to the center of the network certainly entails scalability issues. Operations that are currently supported by distributed elements in the current architecture such as handover decisions performed by eNBs, will be centralized and performed by a SDM-C application running on top of a centralized SDM-C controller.

On the one hand, this certainly relieves the network from performing distributed coordination actions, (e.g., the actions currently performed supported by the X2 interface), but, on the other hand, centralizing functions may harm the scalability of the system, especially on a multi-site virtualized infrastructure.

However, scalability in the context of SDN controllers is already a widely studied topic, so similar solutions can be applied to SDM-C and SDM-X. Moreover, since SDM controllers are VNFs themselves, the hosting virtual machine(s) can, if necessary, be distributed over multiple nodes in both the central and the edge cloud in a relatively easy manner, thus adapting to the signalling load offered to the controller.

6.2.5.3 Feasibility of increasing network complexity with growing number of slices

The increased complexity posed by the management of several network slices operated by different tenants over a multi-layered cloud infrastructure has been taken into account by 5G NORMA while defining its architecture. Stating how the impact of this increased complexity is difficult at this stage, especially because we lack a clear view on how many network slices can be hosted in the same physical infrastructure managed by an InP.

Managing multiple network slices poses significant pressure on the inter-slice resource broker, which has to accommodate all the network slice requests into a set of federated infrastructure elements. Identifying the best configuration when dealing with several, heterogeneous KPIs that have to be fulfilled is a problem whose difficulty increases with the number of slices.

However, even without providing a defined number, we claim that the number of network slices is going to be limited by the scarce bottleneck resources and elements of the infrastructure, (i.e., spectrum, backhaul capacity for rural areas). Hence, even if the complexity for the correct management of a high number of slices is not negligible, we expect to have a low number of slices to be managed. Furthermore, orchestration and management algorithms may be configured to perform low complexity fall back configuration if, in some exceptional cases, the number of slices outgrows. Finally, the growing level of automation for many low- and medium- level network operations procedures is expected to somewhat counterbalance the increasing network complexity from network slicing.

6.3 Summary and next steps

At end of the second architecture design iteration, in this document we reported about the fulfilment of

- performance requirements for MBB,
- functional requirements for all generic 5G services,
- operational requirements,
- security requirements, and
- soft-KPIs.

Performance verification

Fulfilment of MBB KPIs listed in Table 6-5 is mainly driven by other 5GPPP projects. Maturity of results will improve until the end of the project hence an update is expected in the next report.

Fulfilment of traffic requirements should be investigated in more detail. In the following, we conclude the most significant results.

An area wide capacity density in the range of Tbps/km^2 would lead to huge economic effort extending the small cell layer at higher frequency bands. On the other hand, according to current traffic forecast [Cisco_2016], it probably will not be needed within the considered time frame. Local hot spots of course may be served by mmW nodes that will provide the intended capacity.

Due to the expected cell sizes, capacity contributions of small cells at low and medium frequencies will probably not be limited by their capability to offload macro layers, hence it would make sense to improve their spectrum efficiency beyond the assumed values (Figure C-4).

Usually, this will not apply for small cells at high frequency bands. Due to their high cell throughput (system bandwidth) and small cell areas, the capability to offload the macro layer is the limiting factor.

Concepts for multi-operator RAN sharing that include virtualised, PNF (e.g., antennas), and spectrum resources could mitigate limitations due to EMF issues and contribute to more efficient RAN deployments, including CAPEX and OPEX savings as well as coverage improvements. With respect to performance verification the following topics should be covered in the remaining project time

- Update of performance verification for mMTC and V2X services
- Investigation of multi-connectivity
- Investigation of reliability concepts.

Verification of functional requirements for all generic 5G services

The set of functional requirements for all generic 5G services is of considerable magnitude so that their fulfilment can only be realized by means of a joint effort across multiple R&D activities in the industrial and academic area. Nevertheless, 5G NORMA has begun to contribute to selected and important subsets of these requirements.

For the generic 5G service “eMBB”, all relevant functional requirements as identified in D2.1 are fulfilled by current design status of the 5G NORMA architecture. In particular, the design has contributed to flexible capacity allocation on cell and sector level, efficient backhaul utilisation, application awareness, and multi-layer / multi-RAT connectivity.

For the generic 5G services “mMTC” and “URLLC” (represented by V2X in 5G NORMA), the set of functional requirements is even more heterogeneous. Hence, the 5G NORMA concepts and solutions only address a subset of these and have to be complemented by results from other efforts, e.g., other 5G-PPP projects.

More specifically, for the mMTC use case, the project has developed functions and algorithms that contribute to the management of a massive number of devices, device-type aware RAT selection, and context-aware mobility management. For the V2X use case, 5G NORMA solutions tackle the following requirement groups: coexistence of mMTC and URLLC vehicular applications, fast and targeted dissemination of safety messages, and optimizations for control plane and data plane functions.

Moreover, analysis of the functional behaviour of 5G networks will be supported by the following simulation and protocol analysis activities (cf. Annex C.1.2)

- Investigation of network programmability,
- Investigation of QoE based routing and network agility,
- Investigation of Edge function mobility, and
- Assessment of mobility concepts.

The results of these activities as well as a consolidated functional requirement fulfilment analysis, taking into account results from other 5G-PPP projects, will be provided in D3.3.

Verification of operational requirements

Operational requirements with respect to multi-tenant and multi-service networks are addressed to a big extent by current project activities. Multi-tenant dynamic resource allocation will be investigated by simulations in WP5 (s.cf. Section C.1.2.1).

Investigation of cost savings in the context with of multi-tenant and multi-service networks will require close cooperation between WP2 economic and WP3 technical considerations. Aspects such as traffic diversity and resource sharing including physical (bare metal) and spectrum should be included.

Service specific and context-aware derivation of service requirements including rules for adaptation and placement of VNFs as well as dynamic service specific monitoring and adjustment based on QoS/QoE assessment and control framework are important topics that should be elaborated in more detail in the remaining project time.

In the next design iteration phase we will consider operational requirements published in [NGMN]. In addition, results of other 5GPPP projects on backhaul properties will be summarised.

The following topics should be considered in the remaining project time

- Opportunities for RAN sharing (virtual, bare metal & spectrum resources) including economic evaluations
- Backhaul aspects including results of other 5GPPP projects
- Rules for adaptation and placement of VNFs as well as dynamic service specific monitoring and adjustment based on QoS/QoE assessment and control framework.

Verification of security requirements

As far as security verification is possible on the pure architectural level (as opposed to an assessment of a concrete implementation of the architecture), it can be concluded that implementations of the 5G NORMA architecture can comply with the security requirements. As described in Section 5, possible security threats imposed by the architecture can be mitigated. Moreover, innovative security solutions inside the framework of the 5G NORMA architecture have been proposed. Indeed, the flexible architecture based on NFV and SDN is an excellent basis for integrating also future security solutions addressing those issues that are not in the immediate focus of the 5G NORMA security work.

In the remaining period of the project, along with the finalization of the overall architecture, we will continuously monitor this work, and, where necessary, adapt and enhance the security concepts to ensure that implementations of the finalized 5G NORMA architecture can comply with the security requirements.

Verification of soft-KPIs

The check of Soft-KPIs shall make sure that architecture design at the end of the project hands over mature results to next step realisation activities. In that sense, it was to be expected that at this stage of the project no final results would be available. Important to mention with respect to architecture design so far is

- The current architecture design has been done being aware of different stakeholder relations that need for automatic communication among each other.
- Scalability of SDN controllers is already widely studied. Results can be applied to SDM-C and SDM-X where also these network elements can be considered as logical centralized but virtually distributed.

Final statements on the feasibility of multiple-slice networks cannot be made up to now. For sure, with a growing number of slices the complexity and also inter-slice resource management will become more and more challenging. But in general, it is to be anticipated that spectrum and virtual resources will limit the number of feasible slices more than growing complexity. The following topics should be addressed in the remaining project time

- Internal and external interfaces, comparison 4G/5G interfaces
- Learnings from demonstrator implementations

- Trail runs that emulate implementation of the multi-tenant and multi-service networks including the defined stakeholder roles.

To give an overview all topics are compiled in Table 6-12.

Table 6-12: Topics identified for next architecture design iteration phase.

Topic	Check of fulfilment of
Update of performance requirements (mMTC, V2X, e2e latency)	Performance requirements
Multi-connectivity	Performance requirements
Opportunities for RAN sharing (virtual, bare metal & spectrum resources)	Operational requirements
Backhaul aspects	Operational requirements
Adaptation and placement of VNFs	Operational requirements
Investigation of network programmability	Functional requirements
Investigation of QoE based routing and network agility	Functional requirements
Investigation of Edge function mobility	Functional requirements
Assessment of mobility concepts	Functional requirements
Protocol overhead analysis	Functional requirements
Reliability concepts, reliability prediction	Functional requirements
Update of security requirements	Security requirements
Internal and external interfaces, comparison 4G/5G interfaces	Soft KPI
Demonstrator learnings	Soft KPI
Trial runs implementing multi-tenant and multi-service networks	Soft KPI
Economic evaluations (WP2 part of verification)	Economic feasibility

7 Summary and conclusions

This deliverable "D3.2 5G NORMA network architecture – intermediate report" has covered the status and intermediate results of Work Package 3 "Multi-service Network Architecture". The work in WP3 is organized by three design iterations. This report has described the architecture after completion of the first two iteration phases.

5G mobile networks are faced with unprecedented challenges. Besides increasing performance requirements, e.g., in terms of higher throughput and lower latency, these networks are expected to exhibit great flexibility and versatility to accommodate highly varying requirements from different services at the same time. For example, while applications from the V2X (vehicular-to-anything) communication domain have a need for very low end-to-end latency, virtual reality applications additionally require high throughput capacity on the air interface. For obvious reasons, building dedicated network infrastructures for each of these service classes is not a viable option. Therefore, 5G NORMA has the goal to design a mobile network architecture that supports the multiplexing of multiple services and multiple tenants to a single, shared mobile network infrastructure.

For this purpose, 5G NORMA has performed a thorough analysis of the specific 5G requirements and the according technological and economic environment [5GN-D21] and outlined an initial architecture [5GN-D31]. This work has been continued and extended by a stakeholder analysis, indicating which interfaces need to be open and standardised in order to allow multiple stakeholders in a 5G ecosystem to cooperate effectively.

Building on this work, concepts and methodologies for the 5G NORMA architecture have been defined in the following three main areas:

- Network Slicing:

  5G NORMA has understood that virtualisation, multitasking and multiplexing are comparable techniques in the sense that they all allow sharing of resources, while being complementary in that they can be applied to different kinds of infrastructure equipment. The 5G NORMA architecture integrates these techniques, allowing to setup slices on all network components. In this way, network slices can be created in a true E2E manner, i.e. stretching from one end of the service function chain to the other.

- Network Programmability:

  In 5G NORMA, the Network Programmability concept, known e.g. from OpenFlow and IEEE ForCES, is transferred from routers to any kind of network element, in particular to network elements in the RAN. For this purpose, the RAN data layer has been decomposed into functional blocks, and these blocks have been analysed with respect to their interactions to the control layer. In the control layer, functional blocks have been categorized into three groups of blocks that can be executed i) as common applications (policies) of multiple slices running on the SDM-X, ii) as dedicated applications of a single slice running on the SDM-C and iii) independent from SDM-X and SDM-C. Furthermore, principles for mobility management and QoS management have been defined.

- Network Management & Orchestration:

  The 5G NORMA network management & orchestration builds on the well-known concepts from ETSI-NFV MANO. These concepts have been extended for multi-service and multi-tenant operation by the SDM-O entity. The SDM-O combines the slice-specific NFVO orchestrators with overarching orchestration functionality and an inter-slice resource brokering mechanism.

The work on 5G NORMA's functional architecture has been complemented by investigations on security aspects. Starting with a study of security breaches in the recent past, security-critical points to the 5G NORMA architecture have been identified, namely the i) isolation between network slices, ii) multi-connectivity offering an enlarged target surface and allowing an attacker to

select the weaker connection for his attack, iii) virtualisation technology allowing to attacks to propagate faster within a network, iv) SDM controllers as single points of failure, and v) resource sharing. The applicability of LTE security measures to cope with these threats is investigated, and additional countermeasures as enhancements have been proposed.

Finally, the compliance of the 5G NORMA architecture with requirements has been studied. The verification methodology builds on three evaluation cases to verify multi-service and multi-tenancy capabilities of the 5G NORMA architecture. Intermediate evaluation results have been presented for performance requirements, functional, operational and economic requirements, security requirements and so-called Soft KPIs. The architecture design work is not completed yet, and hence it is not possible to state that all requirements are already fulfilled. Several measures have been defined to close these gaps in the upcoming final design iteration.

This final design iteration phase will continue in WP3 until M27, when D3.3 "5G NORMA network architecture – final report" will be delivered. It will include the final network architecture as well as further network security analyses and solutions. In particular, it will establish and integrate the results from the current developments within WP4 and WP5 in an end-to-end context.

References

[23.246] 3GPP: "TS 23.246 V14.1.0; Technical Specification Group Services and System Aspects; Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Rel. 14)". Dec. 2016; http://www.3gpp.org/ftp//Specs/archive/23_series/23.246/23246-e10.zip

[28.801] 3GPP: "TR 28.801 V0.3.0, Technical Specification Group Services and System Aspects; Telecommunication management; Study on management and orchestration of network slicing for next generation network (Rel. 14)". Nov. 2016; http://www.3gpp.org/ftp//Specs/archive/28_series/28.801/28801-030.zip

[33.401] 3GPP: "TS 33.401 v14.1.0: 3GPP System Architecture Evolution (SAE): Security Architecture (Rel. 14)". Dec. 2016; http://www.3gpp.org/ftp//Specs/archive/33_series/33.401/33401-e10.zip

[33.402] 3GPP: "TS 33.402 v14.0.0: 3GPP System Architecture Evolution (SAE): Security aspects of non-3GPP accesses (Rel. 14)". Dec. 2016; http://www.3gpp.org/ftp//Specs/archive/33_series/33.402/33402-e00.zip

[36.300] 3GPP: "TS 36.300 v14.1.0; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Rel. 14)". Dec. 2016; http://www.3gpp.org/ftp//Specs/archive/36_series/36.300/36300-e10.zip

[5GN-D21] EU H2020 5G NORMA, "D2.1 Use cases, scenarios and requirements", Nov. 2015

[5GN-D22] EU H2020 5G NORMA, "D2.2 Evaluation methodology and KPIs, evaluation architecture design and socio-economic analysis – intermediate report", Oct. 2016

[5GN-D31] EU H2020 5G NORMA, "D 3.1: Functional Network Architecture and Security Requirements", Dec. 2015

[5GN-D41] EU H2020 5G NORMA, "D 4.1 RAN architecture components – intermediate report", Nov. 2016

[5GN-D51] EU H2020 5G NORMA, "D 5.1 Definition of connectivity and QoE/QoS management mechanisms – intermediate report", Nov. 2016

[5GN-D61] EU H2020 5G NORMA, "D6.1 Demonstrator design, implementation and initial set of experiments", Oct. 2016

[5GN-DoW] 5G NORMA Project Proposal

[5GN-IR41] EU H2020 5G NORMA, "RAN architecture components – Preliminary concepts", April 2016

[5G-PPP] 5G-PPP White Paper: "5G empowering verticals". https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

[ABA+16] D. Aziz, H. Bakker, A. Ambrosy, and Q. Liao, "Signalling minimisation framework for short data packet transmission in 5G," accepted for publication at IEEE Vehicular Technology Conference, Montreal, Canada, September 2016.

[Adr] David Adrian, Karthikeyan Bhargavan, Zakir Durumeric, Pierrick Gaudry, Matthew Green, J. Alex Halderman, Nadia Heninger, Drew Springall, Emmanuel Thomé, Luke Valenta, Benjamin VanderSloot, Eric Wustrow, Santiago Zanella-Béguelin, and Paul Zimmermann. "Imperfect forward secrecy: How Diffie-Hellman fails in practice". In 22nd ACM Conference on Computer and Communications Security, October 2015.

[Alr] Alreshoodi, M., & Woods, J. (2013). "Survey on QoE\QoS Correlation Models For Multimedia Services". International Journal of Distributed and Parallel systems, 4(3), 53-72.

[AzENB] Azcom Technology baseband board, http://www.azcom.it/index.php/products/hardware-platforms/small-cell-bbu/

[Bat] Lejla Batina, Nele Mentens, Kazuo Sakiyama, Bart Preneel, Ingrid Verbauwhede. "Low-Cost Elliptic Curve Cryptography for Wireless Sensor Networks" Chapter Security and Privacy in Ad-Hoc and Sensor Networks, Volume 4357 of the series Lecture Notes in Computer Science, pp. 6-17

[BLI2016] "First half findings from the 2016 breach level index", http://breachlevelindex.com/assets/Breach-Level-Index-Report-H12016.pdf

[CHA] Chandy, K. Mani (2006). Event-Driven Applications. Costs, Benefits and Design Approaches. Presented at the Gartner Application Integration and Web Services Summit, San Diego, CA.

[CHE01] Peter M. Chen and Brian D. Noble, “When Virtual Is Better Than Real”, Proceedings of the Eighth Workshop on Hot Topics in Operating System 2001, pages: 133 -138.

[CHE03] Steven Cheung, Ulf Lindqvist, Martin W. Fong: “Modeling Multistep Cyber Attacks for Scenario Recognition”, In Proceedings of the Third DARPA Information Survivability Conference and Exposition (DISCEX III), Washington, D.C., April 22–24, 2003, Volume I, pages 284–292.

[Cid] C. Cid, S. Murphy, and M. J. Robshaw. “Small Scale Variants of the AES”. In Fast Software Encryption, 12th International Workshop | FSE 2005, volume 3557 of Lecture Notes in Computer Science, pages 145{162}. Springer, 2005.

[Cisco_2016] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update. February 2016.

[CLOUDCO] https://www.broadband-forum.org/standards-and-software/major-projects/cloud-central-office. Reviewed Dec’16.

[COL] Richard Colbaugh and Kristin Glass, "Proactive Defense for Evolving Cyber Threats", IEEE International Conference Intelligence and Security Informatics (ISI), Year: 2011, Page(s): 125 - 130.

[CSO] http://www.csoonline.com/article/2974712/disaster-recovery/report-virtualization-doubles-cost-of-security-breach.html

[DHA] Sarang Dharmapurikar, Praveen Krishnamurthy, Todd S. Sproull and John W. Lockwood, “Deep Packet Inspection Using Parallel Bloom Filters”, IEEE Micro, Vol.: 24, Issues: 1,Year: 2004, JANUARY–FEBRUARY, Page(s): 52 - 61.

[DIE] T. Dierks and E. Rescorla, “The Transport Layer Security (TLS) Protocol Version 1.2”, Internet Engineering Task Force (IETF) Request for Comments: 5246, August 2008.

[Eis] T. Eisenbarth and S. Kumar “A survey of lightweight-cryptography implementations”. Design Test of Computers, IEEE, 24(6):522--533, Nov 2007.

[EURECOM] EURECOM: “Processing radio access network functions in the cloud: critical issues and modeling”, MobiCom, 2015

[FANT5G] Scheich F, et.al., FANTASTIC-5G, 5G-PPP Project on 5G Air Interface Below 6 GHz,

[FANT-D31] FANTASTIC-5G: Preliminary results for multi-service support in link solution adaptation, July 2015

[Fri] T. Frisanco, P. Tafertshofer, P. Lurin and R. Ang, "Infrastructure Sharing for Mobile Network Operators; From a Deployment and Operations View," 2008 International Conference on Information Networking, Busan, 2008, pp. 1-5.

[G. 993.2] ITU-T Study Group 15: Very high speed digital subscriber line transceivers 2 (VDSL2); Recommendation ITU-T G.993.2. 2015-01-13; http://handle.itu.int/11.1002/1000/12370

[G.9700] ITU-T Study Group 15: Fast access to subscriber terminals (G.fast) - Power spectral density specification; Recommendation ITU-T G.9700. 2014-04-04; http://handle.itu.int/11.1002/1000/12010

[G.9701] ITU-T Study Group 15: Fast access to subscriber terminals (G.fast) - Physical layer specification; Recommendation ITU-T G.9701. 2014-12-05; http://handle.itu.int/11.1002/1000/12090

[Gau] G. Gaubatz, J. P. Kaps, E. Ozturk and B. Sunar, “State of the art in ultra-low power public key cryptography for wireless sensor networks," Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on, Kauai Island, HI, 2005, pp. 146-150.

[Gha] Ghalut, T., Larijani, H., & Shahrabi, A. (2016). QoE-aware Optimization of Video Stream Downlink Scheduling Over LTE Networks Using RNNs and Genetic Algorithm. Procedia Computer Science, 94, 232-239.

[Gom] G. Gomez et al., “Towards a QoE-Driven Resource Control in LTE and LTE-A Networks,” J. Computer Networks and Commun., 2013.

[HAN] Yi Han, Jeffrey Chan, Tansu Alpcan and Christopher Leckie, "Virtual Machine Allocation Policies against Co-resident Attacks in Cloud Computing", IEEE ICC 2014 - Communication and Information Systems Security Symposium, Year 2014, page(s):786-792.

[HAR] Ed. D. Hardt, “The OAuth 2.0 Authorization Framework”, Internet Engineering Task Force (IETF) Request for Comments: 6749, October 2012

[HOO1] Cynthia S. Hood and Chuanyi Ji, "Intelligent Agents for proactive Fault Detection", IEEE Internet Computing, Vol.: 2, Issue: 2, Year: 1998, Page(s):65 - 72.

[HOO2] Cynthia S. Hood and Chuanyi Ji, "Proactive Network Fault Detection", IEEE INFOCOM, Vol.: 3, Year 1997. Page(s) 1147 - 1155.

[IBM] IBM: “Virtualization in Education”. IBM Global Education White Paper, 2007; http://www-07.ibm.com/solutions/in/education/download/Virtualization%20in%20Education.pdf

[iJOIN-D53] iJOIN, Final definition of iJOIN architecture, 2015, http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D5.3.pdf

[ISO] ISO/IEC 29192-2:2012 Information technology -- Security techniques -- Lightweight cryptography – Parts 2, 3: Block ciphers, Stream ciphers

[ITU G.1011] ITU-T. Recommendation G.1011. Reference Guide to Quality of Experience Assessment Methodologies. 2010. https://www.itu.int/rec/T-REC-G.1011/en.

[ITU P.10] ITU-T. Recommendation P.10/G.100 Amendment 5. Vocabulary for performance and quality of service. 07/2016. https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.10-201607-I!Amd5!PDF-E&type=items.

[ITU P.800] ITU-T Recommendation, (2003) “ITU-T Rec. P.800.1: Mean Opinion Score (MOS) Terminology”.

[JSR345] Java Specification Request 345: Enterprise JavaBeans, https://jcp.org/en/jsr/detail?id=345

[JUN] A. Jungmaier, E. Rescorla and M. Tuexen, “Transport Layer Security over Stream Control Transmission Protocol”, Internet Engineering Task Force (IETF) Request for Comments: 3436, December 2002.

[KRE] Diego Kreutz, Fernando M. V. Ramos, Paulo Verissimo, Christian Esteve Rothenberg, Siamak Azodolmolky and Steve Uhlig, “Software-Defined Networking: A Comprehensive Survey”, Proceedings of the IEEE, Vol.: 103, Issue: 1, Year: 2015, Page(s): 14 – 76.

[KRE] Diego Kreutz, Fernando M. V. Ramos, Paulo Verissimo, Christian Esteve Rothenberg, Siamak Azodolmolky and Steve Uhlig, "Software-Defined Networking: A Comprehensive Survey", Proceedings of the IEEE, Vol.: 103, Issue: 1, Year: 2015, Page(s): 14 – 76.

[KT] http://www.kernelthread.com/publications/virtualization

[LeC] Le Callet P., Möller S. and Perkis A. (03/2013). Qualinet White Paper on Definitions of Quality of Experience. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Eds. Lausanne, Switzerland.

[LIN] Po-Ching Lin, Ying-Dar Lin, and Tsern-Huei Lee, Yuan-Cheng Lai, “Using String Matching for Deep Packet Inspection”, IEEE Computer Vol. 41, Issue: 4, Year: 2008 April, Page(s): 23 - 28.

[LTESecB] Dan Forsberg, Günther Horn, Wolf-Dietrich Moeller, Valtteri Niemi: "LTE Security", 2nd Edition, Wiley 2012

[MAG-D41] mmMAGIC Deliverable 4.1, “Preliminary radio interface concepts for mm-wave mobile communications”, June 2016

[Men] Menkovski, V., Oredope, A., Liotta, A., & Sánchez, A. C. (2009). Predicting quality of experience in multimedia streaming. Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia - MoMM '09.

[MET14-D66] METIS D6.6 “Final report on the METIS 5G system concept and technology roadmap,” ICT-317669 METIS Deliverable 6.6, Version 1, May 2014

[MPTCP] A. Ford, C. Raiciu, M. Handley, O. Bonaventure: "TCP Extensions for Multipath Operation with Multiple Addresses". Internet Engineering Task Force (IETF), Request for Comments: 6824, Jan. 2013.

[NEC] NEC: “What is server virtualization?” http://www.nec.com/en/global/solutions/servervirtualization/merit.html

[NetSecB] Eric Cole: Network Security Bible. 2nd Edition, Wiley 2009

[NFV-IFA010] ETSI NFV IFA, “ETSI GS NFV-IFA 010 V2.1.1, Network Functions Virtualisation (NFV); Management and Orchestration; Functional requirements specification”, April, 2016 (http://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/010/02.01.01_60/gs_NFV-IFA010v020101p.pdf)

[NFV-MAN1] ETSI GS NFV-MAN 001 V1.1.1 (2014-12): Network functions virtualisation (NFV); management and orchestration.

[NFV-Sec] ETSI GS NFV-SEC 002 V1.1.1 (2015-08): Network Functions Virtualisation (NFV); NFV Security; Cataloguing security features in management software

[NFV-UC] ETSI, “Network Functions Virtualisation (NFV); Use Cases”, ETSI GS NFV 001, v1.1.1, Oct. 2013.

[NFV-WP] ETSI, “Network Functions Virtualisation, Introductory White Paper”, ETSI Technical Report, Oct. 2012.

[NGMN] Rachid El Hattachi/ Javan Erfanian: "NGMN 5G Whitepaper". February 2015; https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[NOK-NFV] https://insight.nokia.com/how-manage-security-nfv-environments

[OV15] Whitepaper: Smartphone & tablet usage trends & insights, 4G LTE and Wi-Fi powering data consumption

[PA-5GPPP] M. Maternia et. Al., 5G PPP use cases and performance evaluation models, Version 1.0, April 2016

[Rub] Rubino, G., Tirilly, P., & Varela, M. (2006). Evaluating Users’ Satisfaction in Packet Networks Using Random Neural Networks. Artificial Neural Networks – ICANN 2006 Lecture Notes in Computer Science,303-312..

[SAK] N. Sakimura, J. Bradley, M. Jones, B. de Medeiros, C. Mortimore, “OpenID Connect Core 1.0 incorporating errata set 1”, OpenID Specification: http://openid.net/specs/openid-connect-core-1_0.html

[SCTP] R. Stewart, Ed.: "Stream Control Transmission Protocol". Internet Engineering Task Force (IETF), Request for Comments: 4960, Sept. 2007.

[Singh_2014] Bikramjit Singh, Konstantinos Koufos, Olav Tirkkonen, Coordination protocol for inter-operator spectrum sharing based on spectrum usage favours, https://arxiv.org/pdf/1505.02898

[SMA] Daniel Smallwood and Andrew Vance, “Intrusion Analysis with Deep Packet Inspection Increasing Efficiency of Packet Based Investigations”, International Conference on Cloud and Service Computing, 2011, Page(s): 343 - 347.

[SSP+16] K. Samdanis, R. Shirvastava, A Prasad, D. Grace, and X. Costa-Perez, Samdanis, TD-LTE virtual cells: An SDN architecture for user-centric multi-eNodeB elastic resource management. *Computer Communications*, *83*, 1-15.

[STE] Randall R. Stewart, “Stream Control Transmission Protocol”, Internet Engineering Task Force (IETF) Request for Comments: 4960, September 2007.

[SWS15] S. Saur, A. Weber, and G. Schreiber, "Radio access protocols and preamble design for machine-type communications in 5G," in Signals, Systems and Computers, 2015 49th Asilomar Conference on, Nov 2015, pp. 3-7.

[Tan] A. Tanenbaum: Computer Networks. 3rd ed., Prentice Hall, 1996

[TUX] Michael Tuexen, R. Seggelmann, E. Rescorla, “Datagram Transport Layer Security (DTLS) for Stream Control Transmission Protocol (SCTP)”, Internet Engineering Task Force (IETF) Request for Comments: 6083, January 2011.

[UKG_SBS] “2015 Information security breaches survey”, by PWC on behalf of UK Government. http://www.pwc.co.uk/assets/pdf/2015-isbs-technical-report-blue-03.pdf

[Wan] Arvinderpal S. Wander, Nils Gura, Hans Eberle, Vipul Gupta, Sheueling Chang Shantz. “Energy analysis of Public-Key Cryptography for wireless sensor networks” Proceedings of the 3rd IEEE Int’l Conf. on Pervasive Computing and Communications (PerCom 2005)

[Wen] Erich Wenger, Johann Großschädl. “An 8-bit AVR-Based Elliptic Curve Cryptographic RISC Processor for the Internet of Things”. In 45th Annual IEEE/ACM International Symposium on Microarchitecture Workshops, 2012, pp: 39-46.

[WP-CMT] Wikipedia: “Computer multitasking”. https://en.wikipedia.org/wiki/Computer_multitasking

[WP-Mux] Wikipedia: “Multiplexing”. https://en.wikipedia.org/wiki/Multiplexing

[WP-Virt] Wikipedia: “Virtualization”. https://en.wikipedia.org/wiki/Virtualization

[YAN] Qiao Yan, F. Richard Yu, Qingxiang Gong, and Jianqiang Li, “Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges”, IEEE Communications Surveys & Tutorials, Vol.: 18, Issue:1, Year 2016, Page(s): 602 – 622.

[Zha] Zhang, Y., Yue, T., Wang, H., & Wei, A. (2014). Predicting the Quality of Experience for Internet Video with Fuzzy Decision Tree. 2014 IEEE 17th International Conference on Computational Science and Engineering.

[23.246] 3GPP: "TS 23.246 V14.1.0; Technical Specification Group Services and System Aspects; Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Rel. 14)". Dec. 2016; http://www.3gpp.org/ftp//Specs/archive/23_series/23.246/23246-e10.zip

[28.801] 3GPP: "TR 28.801 V0.3.0, Technical Specification Group Services and System Aspects; Telecommunication management; Study on management and orchestration of network slicing for next generation network (Rel. 14)". Nov. 2016; http://www.3gpp.org/ftp//Specs/archive/28_series/28.801/28801-030.zip

Annex A 5G NORMA Management & Orchestration Layer Fundamentals

A.1 ETSI NFV MANO architecture

Network Functions Virtualisation offers the opportunity to design, deploy and manage networking services by decoupling the network functions (e.g., S-GW, P-GW, MME…) from specific hardware appliances [NFV-WP]. That way, network functions become elementary software building blocks which can be managed and orchestrated in a very agile way as software components.

Once the network functions are available as VNFs, there could be, of course, different ways to manage and operate them. A widely accepted general framework for doing this is the ETSI NFV MANO reference architecture [NFV-MAN]. This is the ETSI-defined framework for the management and orchestration of all resources in a cloud data centre, including computing, networking and storage resources. It addresses various aspects of management and orchestration, which are specific to NFV such as lifecycle management, operations management (e.g., provisioning, scaling…), information elements, architecture and the interaction with legacy operational and management systems (e.g., Operations and Business Support Systems -OSS/BSS-), but above all, it offers the decoupling of network functions software from the hardware.

As it is well known, the ETSI is a key in developing standards for information and communications technologies in Europe; in this case, a specific group in charge of developing the requirements and architecture for network functions virtualisation within telecoms networks has been created, i.e., the ETSI Industry Specification Group for Network Functions Virtualisation (ETSI ISG NFV). The ETSI NFV MANO specification comes from this group.

Figure A-1 represents the ETSI NFV MANO architecture.

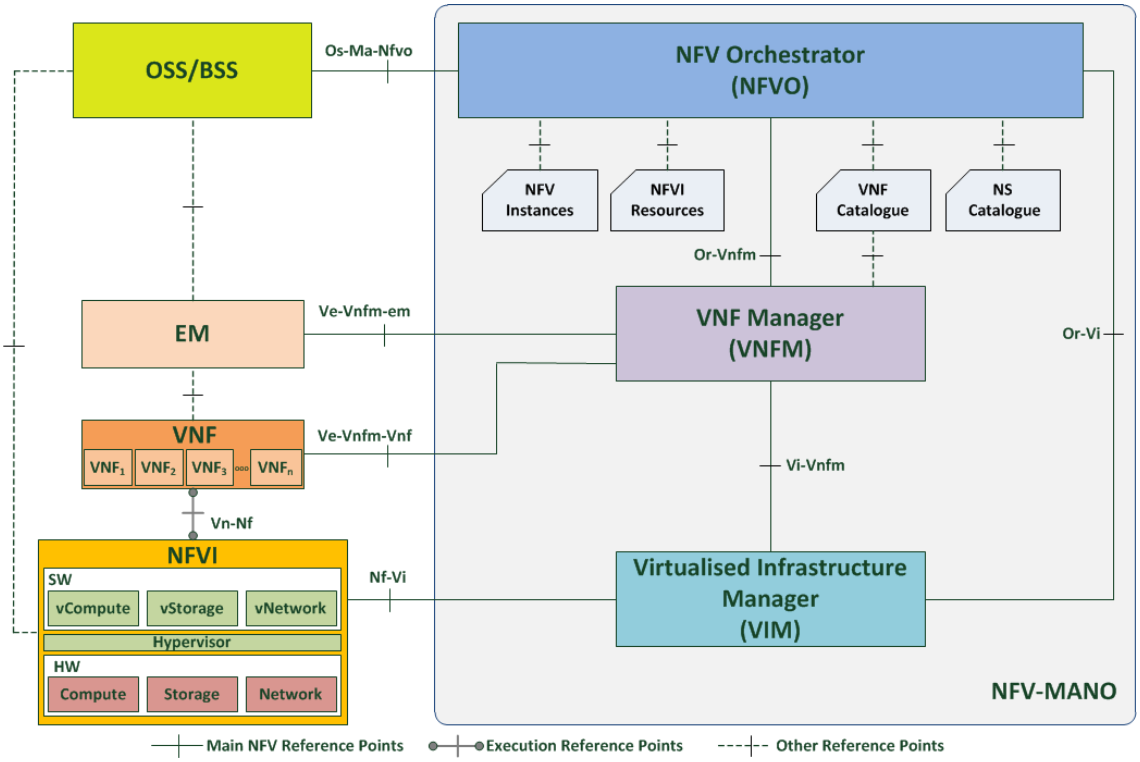


Figure A-1: ETSI NFV MANO Architecture

As we can see the NFV MANO block comprises the blue tone blocks on the right. The other four blocks on the left (OSS/BSS, EM, VNF and NFVI) are outside the scope of the NFV MANO

itself, although high level interfaces (or "Reference Points" as they are defined in the ETSI specification) are defined.

This is a brief description of the functional blocks in the figure (cf. [NFV-MAN] for a more detailed description):

- Network Functions Virtualisation Infrastructure (NFVI). This block represents all the hardware components (e.g. compute, storage, and networking) and software components (e.g. hypervisors and virtualised compute, storage and networking resources) that together provide the infrastructure resources where VNFs are deployed.
- Virtual Network Function (VNF). This block represents the set of Virtualised Network Functions. These functions will be executed on the virtualised NFVI resources.
- Element Management (EM). It is responsible for 3GPP fault, configuration, accounting, performance and security (FCAPS) management for a VNF. It collaborates with the VNF Manager to perform those functions that require exchange of information regarding the NFVI Resources associated with the VNF. Its main functions are:
  - Domain-specific (3GPP) configuration of the function provided by the VNF.
  - Fault management of functions provided by the VNF.
  - Accounting for the usage of VNFs.
  - Collect performance measurement results for the function provided by the VNF.
  - Security management for the VNF.
- Virtual Infrastructure Manager (VIM). This is the module responsible for managing the virtualised infrastructure. Main functions are:
  - Control and manage the NFVI compute, storage and network resources.
  - Collect performance measurements and relevant events from the VI resources.
  - Keep an inventory of the allocation of virtual resources to physical resources.
  - Organise virtual links, networks, subnets, and ports.
  - Manage a repository of NFVI hardware resources (compute, storage and networking) and software resources (hypervisors), and discover the capabilities and features to optimise the use of such resources.
- VNF Manager (VNFM). The VNFM is responsible for the lifecycle management of VNFs under the control of the Network Functions Virtualisation Orchestrator (NFVO), which it performed by instructing the VIM. VNFM operations include:
  - Instantiation & termination of VNFs (lifecycle management).
  - VNFs scaling.
  - Updating or upgrading VNFs.
  - Overall coordination and adaptation role for configuration and event reporting between NFVI and EM.
- Network Functions Virtualisation Orchestrator (NFVO). Performs top level resources orchestration and network service orchestration. It connects different functions to create an end-to-end, resource-coordinated service. The orchestrator provides a high level logical abstraction view of the Network Services (NS) deployed on the infrastructure (e.g., from here, it is possible to define complex services specifying different forwarding graphs among the existing network functions). As you can see the orchestrator can access different catalogues (databases) in the NFV MANO scope. The NFVO main functions are:
  - On-boarding new Network Service, VNF Forwarding Graphs (FG) and VNF Packages.
  - NS lifecycle management, including instantiation, scaling, performance measurements, event correlation and termination.
  - Policy management for NS instances.
  - Global resources management, validation and authorization of NFVI resource requests.
- Operations and Business Support Systems OSS/BSS. These are the combination of the operator's other operations and business support functions, which are not captured in the ETSI MANO framework but are expected to have information exchange with the ETSI

MANO functional blocks. OSS/BSS functions may provide management and orchestration of legacy systems and may have full E2E visibility of services provided by legacy network functions in an operator's network.

Although all the main blocks and their interfaces are well defined, the ETSI NFV MANO specification does not mandate any specific realisation of the NFV MANO architectural framework.

A.2 VNF Life-cycle management

5G NORMA, as a multi-tenancy mobile network, should enable the creation and dynamic life-cycle management of different network slices for the different tenants. This section describes slice management operations considering slice VNF allocation and slice VNF life-cycle management. In a similar way VMs are allocated in a cloud environment, VNFs forming a network slice in a 5G network need to be provisioned, configured, monitored, scaled, updated/upgraded and decommissioned along with the lifecycle of the associated service. Virtualisation capabilities (dynamic addition, removal, or updating of services, and dynamic mapping of different network resources to services) are used to achieve dynamic management of the deployed VNFs.

VNF allocation in the NFV infrastructure is a complex task since many requirements and constraints usually need to be met at the same time. Also, VNF allocation on telco networks adds higher complexity compared to the common IT resource allocation strategies. For instance, some VNFs requiring low latency or high bandwidth could be preferred to be physically allocated on Edge Cloud nodes, while traditional MBB resources should be kept in the Central Cloud. This is not common in conventional IT approaches where it does not matter very much where a specific function is allocated. In addition, allocation and release of resources can be a very dynamic process, so frequent VNF allocations and releases may be needed along the VNF lifetime.

The functions for VNF life-cycle management are described in Annex A.1, i.e., the ETSI NFV MANO functions used to manage and orchestrate NFs and network services. In other words, for each slice, the VNF allocation and life-cycle management and network service orchestration operations are the same as those described for ETSI NFV MANO [NFV-MAN]. The module directly involved in VNF allocation & life-cycle management is the VNF Manager (VNFM). Specifically, a VNFM does the following:

- It manages a VNF's life cycle. That is, it creates, maintains and terminates VNF instances (which are installed on the VMs which the VIM creates and manages).
- It is responsible for the typical FCAPS functions for VNFs (i.e. Fault, Configuration, Accounting, Performance, and Security Management of VNFs).
- It scales the infrastructure (up/down, by adding/removing additional VNF instances, or in/out, by adding/removing intra-VNF resources such as CPU or memory).

As depicted in Figure 4-9, there can be multiple VNFMs in a typical NFV MANO stack deployment (and possibly from different vendors), each one managing its own set of VNFs. This creates an additional challenge: the management and coordination of end-to-end services involving VNFs from different VNFM domains. For each slice, this challenge is handled by the respective NFVO as follows:

- NFVO creates end-to-end network services, which contain different VNFs, by coordinating directly with the respective VNFMs (i.e., it does not need to talk to each VNF directly).
- NFVO instantiates new VNFMs where it is applicable.
- NFVO specifies the topology of the network services instances (i.e., the so-called VNF-FG).

Annex B Security Aspects

B.1 Mitigating multi-tenancy-related security threats to the 5G NORMA Architecture

B.1.1 Multi-tenancy in central and edge clouds

5G NORMA relies on usage of central clouds and edge clouds to enable network multi-tenancy. In traditional mobile networks, core network entities were typically located in a small number of "central offices". When adopting NFV technology, it is a logical step to replace these "central offices" by central data centres providing a telco cloud, or in other words an NFV environment in which the core functions can be implemented as VNFs. Taking this step, obviously, a mobile network needs no longer necessarily its own, dedicated hardware, but central data centre infrastructures can be shared by different networks, leading to multi-tenancy in the core network infrastructure.

A probably even more significant evolution step is the virtualisation of the RAN. While the LTE RAN is basically a set of a (typically very high) number of eNBs, each connected via the S1 interface to the core, with additional X2 interfaces between geographically neighbouring eNBs, 5G NORMA assumes – according to [5GN-D31], Section 5.3 - a RAN consisting of edge clouds and "bare metal" at the antenna sites. Two flavours of edge clouds are considered, either implemented directly at the location of the antenna site, or somewhat less distributed, within the access network. [5GN-IR41] emphasizes the scenario where radio resource control and the packet data convergence protocol are located "in a central location", referring to an edge cloud in the access network. But most importantly, this new RAN infrastructure is expected to host multiple tenants, i.e. different mobile networks.

The figure below visualises a setup where two virtualised networks share core and RAN cloud infrastructure, as well as the RAN bare metal equipment, such as remote radio heads.

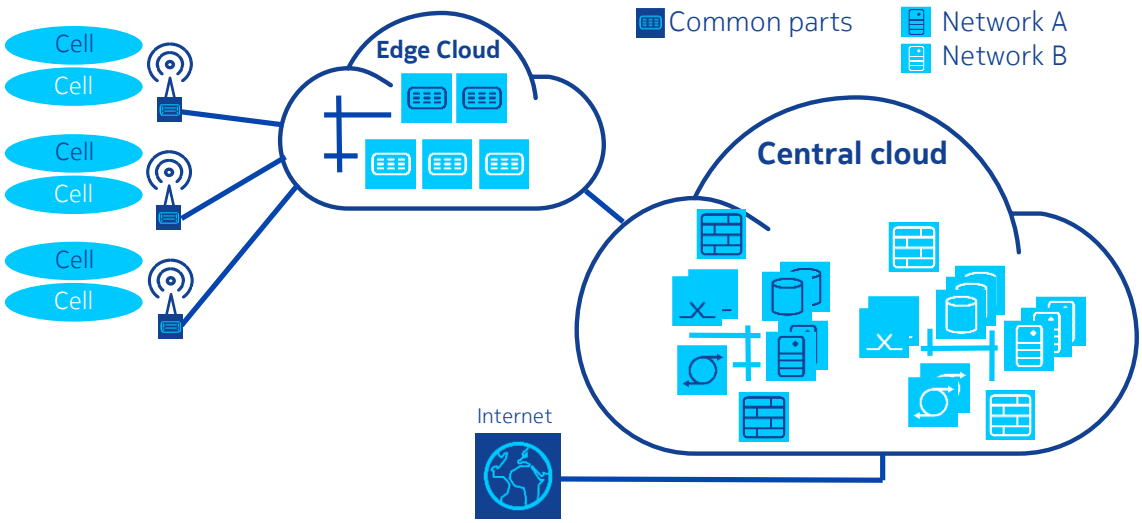


Figure B-1: Multi-tenancy example: Two mobile networks sharing core and RAN infrastructure

Common, shared infrastructure needs to be provided and managed by an infrastructure provider. The separation of the infrastructure provider on the one hand and the mobile service provider as a tenant on the provided infrastructure on the other hand makes the security and availability of the network dependent on the well-behaviour of the infrastructure provider – a strong trust relationship is required. For a more detailed discussion, see Section 0.

The basic and obvious threat in multi-tenant cloud (or NFV) environments is the failure to maintain strict tenant isolation. Isolation has two aspects:

- Resource isolation means that resources dedicated to one tenant cannot be taken away by another tenant.
- Security isolation means that a tenants data or traffic cannot be intercepted or modified by another tenant. Another threat is that one tenant may manage to exceed its agreed resource quota, with the consequence that other tenants don't get the resources that are legally assigned to them, resulting in a DoS condition.

Lack of resource isolation may allow one first tenant to grab an arbitrary amount of resources (like computing time or disk memory) with the consequence that other tenants do not get the resources that are legally assigned to them, resulting in a DoS condition for other tenants. Note that the first tenant in this example may not act maliciously on purpose. Rather, it may itself be subject of a DoS attack, that drives it to scale up its resources, thus propagating the DoS attack to other tenants.

As an example for lack of security isolation, by exploiting flaws in a hypervisor, a VNF belonging to tenant A may be able to read or even modify memory allocated to a VNF belonging to tenant B.

Note that neither of the tenants nor any infrastructure provider may act maliciously on purpose. Still, as a result of errors or misconfigurations such "attack situations" may happen. In this context, the increased complexity caused by the various options how to allocate core and RAN functions must be considered. Theoretically, this complexity increases the surface for purposeful attacks, but more realistically, it only increases the potential of mistakes in setting up the network, which may result in non-optimised or even defective network operation.

The capability to provide isolation of tenants is a fundamental feature in cloud computing. Assuming that the relevant telco cloud software is designed, implemented, configured and operated with highest care in order to minimise the number of errors and thus its vulnerability, it can be concluded that tenant isolation works in telco clouds. An extra level of security may be achieved if the cloud offers the option for physical VNF separation (as required according to [5GN-D31], Section 3.3.): VNFs of different tenants can be physically separated, so attacks between them via the local hypervisor are excluded. Note however that physical separation may not be feasible at an edge cloud site with limited overall resources.

Central or aggregation level data centers are typically located within secure buildings and in addition supervised constantly by human staff. In contrast, edge cloud deployments may suffer from much more physical exposure. This would primarily apply to equipment located "at the lamp post", but also to equipment that may be placed in locked but unsupervised rooms in solid buildings. While this may be mitigated to some extent by suitable remote supervision, including the rising of alerts, e.g., when doors are opened, it still seems to increase the likelihood for certain successful attacks based on physical access to the equipment. It is however out of scope of this project to investigate protection measures against attacks facilitated by physical access to cloud infrastructure.

B.1.2 Multi-tenancy on bare metal equipment

Multi-tenancy in the RAN is not restricted to RAN infrastructure that provides an NFV platform, but also affects "bare metal" RAN equipment. Depending on the nature of the equipment, it may or may not be aware of the different tenants. In the former case, equipment specific mechanisms need to facilitate multi-tenancy and provide proper isolation.

Again, the party providing and managing the bare metal equipment must be trusted by the tenants to assign appropriate resources to each tenant and prevent any inter-tenant attacks via the shared equipment. It can be noted however, that most likely, traffic handled by RAN bare metal equipment is cryptographically protected (by the security layer between user equipment and RAN), thus preventing inter-tenant traffic interception or injection.

B.2 LTE security concepts

The security architecture specified by 3GPP for LTE is considered to be very sound - no major flaws have shown up until now. The figure below visualises the most important elements of this architecture.

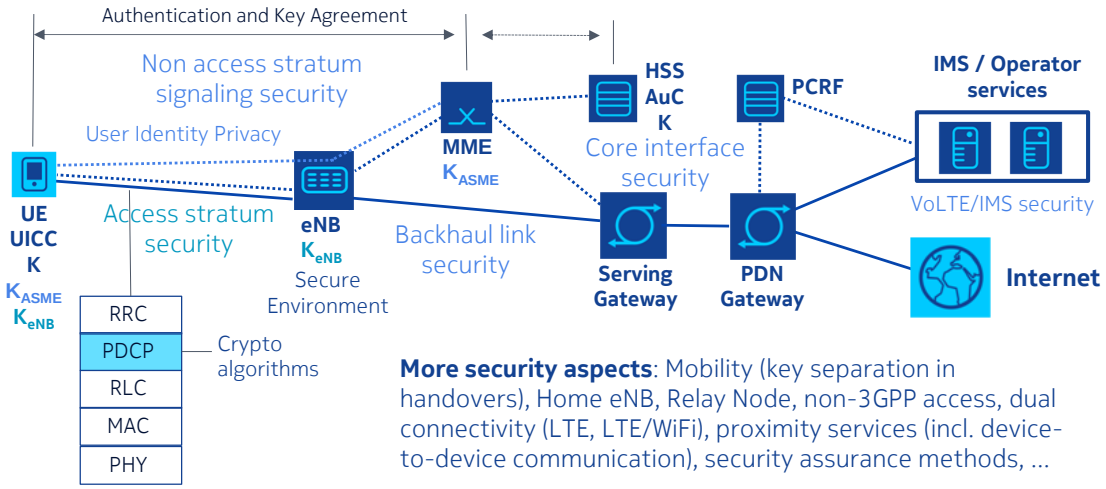


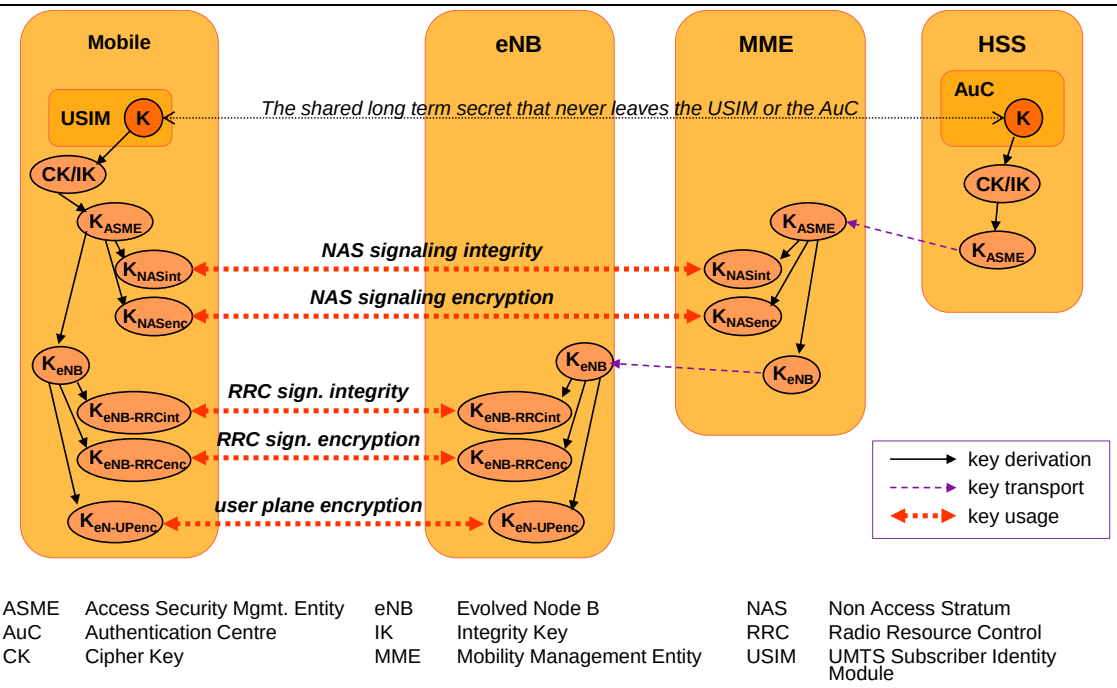
Figure B-2: Elements of the LTE Security Architecture

In the following, we describe some parts of this architecture in more detail.

When a UE attaches to an LTE network, EPS AKA (Authentication and Key Agreement) is executed between UE and the core network, namely the MME. As a result, keys are derived both on the UE and the MME that are used to secure the so called NAS (Non Access Stratum) signalling between these two entities. In addition, the so called K_{eNB} is derived at the UE and the MME and is passed from the MME to the eNB to which the UE is connecting. The K_{eNB} is specific for one session of an UE at one eNB and is used as the basis of AS (Access Stratum) security, i.e. security between the UE and the eNB.

The K_{eNB} is not directly used to protect traffic. Instead, three keys are derived from it, one for control plane integrity protection, one for control plane encryption and one for user plane encryption. There is no general user plane integrity protection in LTE, except for the support of the so called relay nodes (eNBs that have no fixed connection to the network but use the LTE radio interface to connect to another eNB and thus relay traffic of UEs to this eNB), where user plane integrity protection is required and is based on a fourth key derived from K_{eNB}.

The LTE key hierarchy is illustrated in Figure B-3.



ASME	Access Security Mgmt. Entity	eNB	Evolved Node B	NAS	Non Access Stratum
AuC	Authentication Centre	IK	Integrity Key	RRC	Radio Resource Control
CK	Cipher Key	MME	Mobility Management Entity	USIM	UMTS Subscriber Identity Module

Figure B-3: LTE Key Hierarchy

For LTE, three different pairs of crypto algorithms are specified (where each pair comprises an encryption and an integrity protection algorithm). The intention of providing more than one algorithm pair is to have some diversity and fallback options in case one of the algorithms should be broken during the expected lifetime of LTE systems. Both for NAS and AS security, procedures are specified to negotiate which algorithm pair is to be used.

To protect traffic on the backhaul link (S1 interface between eNB and core) and between eNBs (X2 interface), IKE/IPsec is specified to be used. Mostly, on the core side the IPsec tunnels are terminated by a dedicated security gateway (SEG), but they may also be terminated at the MME (for the control plane) or at the serving gateway (for the user plane).

LTE backhaul link security is illustrated by the figure below.

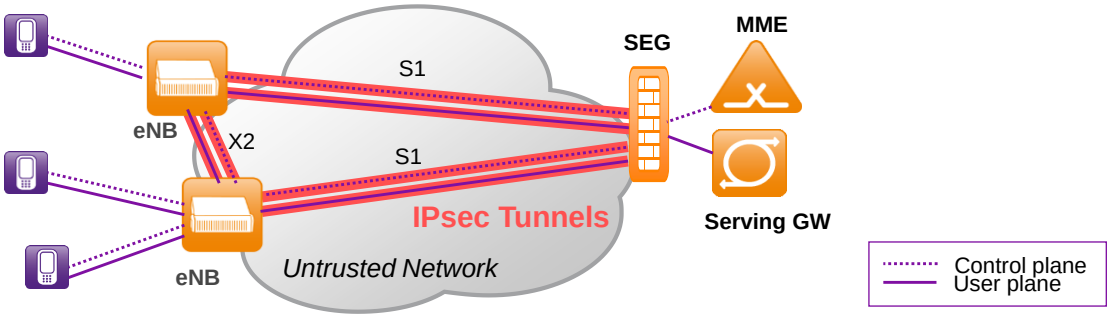


Figure B-4: Backhaul Link Security in LTE

In LTE, the eNB is considered as a physically exposed entity, with a somewhat notable risk of being compromised by an attacker with physical access. An eNB has no access to the NAS signalling, as it is protected end-to-end between UE and MME. However, it has access to the user plane traffic, as it is decrypted and re-encrypted at the eNB between radio interface and backhaul link. To mitigate the threat of eNBs being compromised via physical access, 3GPP requires a "secure environment" inside the eNB, where keys are stored, crypto operations are executed etc. and which does not allow any unauthorised access. Despite of this requirement, security versus cost balancing may lead to implementations where the secure environment is too weak to resist a determined attacker.

As in LTE compromise of an eNB is considered a notable risk, measures are used to prevent that a compromised eNB affects communication running via other eNBs. The preferred approach for a handover to another eNB is to derive a new K_{eNB} that is independent of the old K_{eNB}. This requires interaction with the MME, and is therefore not possible in a so-called X2-handover that is done without involving the MME. In this case, the old eNB will derive a new K_{eNB} and pass it to the new eNB, so it will be able to intercept the traffic between UE and new eNB (but not vice versa). However, measures are specified to ensure that if a subsequent X2-handover takes place, a third K_{eNB} is derived, which can't be derived by the first eNB. Thus, a compromise of an eNB does not affect the security of the communication via the second-next eNB (in a chain of handovers).

In LTE, traffic protection keys cannot be used for an infinite amount of data. Therefore, a method is specified to re-fresh the K_{eNB} "on-the-fly". For this, the eNB triggers an intra-cell handover by which a new K_{eNB} is derived and new traffic protection keys are derived from it.

LTE supports dual connectivity (DC), e.g. parallel connectivity of an UE to a master eNB (MeNB) and a secondary eNB (SeNB). In one DC option, security is terminated at the MeNB only, and no specific security provisions are required for the SeNB. In another DC option, user plane security may be terminated at the SeNB. In this case, a dedicated K_{SeNB} for the SeNB is generated by the MeNB (from its own K_{eNB} and other input) and passed to the SeNB. The MeNB sends the UE information allowing it to also derive the K_{SeNB}.

It is out of the scope of this document to explain more details. The interested reader is referred to TS [33.401], [33.402] and [LTESecB].

As natural for standardization work, this security architecture focuses on what is most relevant for interoperability, i.e. the interfaces, with the interface between UE and network as the most prominent one. In some cases, also platform security is touched. For example, security requirements for the eNB are specified. Moreover, the specification of security assurance methods has been started, with the MME as the first target entity, and others like S-GW or eNB to follow.

For an overall security concept for an LTE network, additional, non-standardised network security measures must be implemented, such as

- Perimeter security, traffic filtering between network internal security zones
- Traffic separation (separate data plane, control plane, management plane, …)
- Secure Operation and Maintenance (O&M)
- Secure operation of services/protocols like DNS, NTP, IP routing etc.

Moreover, each single network element must be secured, first of all by applying a product creation process that includes steps such as a threat analysis, specifying the security requirements, designing and implementing the security features, security testing and hardening (e.g. eliminating any functions that are not required in a specific setup).

For details on this type of security measures, the interested reader is referred to general literature on (network) security, such as [NetSecB].

B.3 Securing the 5G NORMA RAN

B.3.1 Overall radio interface security concept

B.3.1.1 Basic key hierarchy

For the 5G NORMA security architecture, we assume in this section that some AKA mechanisms are in place and that they deliver a key on which AS security can be based, called K_{AS}. On the network side, K_{AS} is passed from the entity that has derived it to the entity that constitutes the AS signalling peer of the UE, which we also call the C-plane security function. Depending on the RAN deployment scenario, this may be a VNF running in an edge cloud, but it could also be a

piece of software running on a bare metal RAN equipment. In the following, a 5G (NORMA) AS key hierarchy is proposed based on K_{AS} as the root key.

Like in LTE, we assume that cryptographic protection (integrity and confidentiality) is required for the control plane over the radio interface. So similar to LTE, a key K_{Cint} for control plane integrity protection and K_{Cenc} for control plane encryption are both derived from K_{AS}, using a key derivation function (KDF) with K_{AS} and specific, different constants as an input. Assuming that there will be a choice of security algorithms, AS signalling procedures must be in place that ensure that the network knows the UE's capabilities and preferences and that allow to agree between network and UE on the security algorithms to be used. After this step, AS security is setup for the control plane and AS signalling messages are protected against attacks.

For the user plane, similarly a key K_{Uint} for user plane integrity protection and K_{Uenc} for user plane encryption are derived from K_{AS}. Note that we assume that in 5G NORMA networks, both user plane encryption and integrity protection may be "mandatory to support" by the network, while it may not be "mandatory to use". Instead, UE and network may negotiate individually whether to use one or both of these security features in a specific session.

Support of multi-connectivity may require the use of different key pairs (K_{Uenc}, K_{Uint}) in parallel. It may also be necessary to distinguish between key pairs that are sequentially used. For these reasons, an "AS user plane key set identifier" is used as an additional input when deriving a specific pair (K_{Uenc}, K_{Uint}). The AS user plane key set identifier must be communicated to the UE to allow the UE to derive the respective key pair. (See Section B.3.1.4 below for a further discussion of multi-connectivity support.)

The security algorithms to be used in the user plane (including "null-algorithms", i.e. not applying security at all) may differ from those used in the control plane, as depending on the intended service or application, different algorithms may be preferable (see Section B.3.2). Based on the UE capabilities and preferences, and taking into account the capabilities of the entity terminating user plane security in the network, the network will select suitable algorithms and inform the UE about this selection, e.g. in the message that also conveys the AS user plane key set identifier.

Flexible allocation of RAN functions means in particular that the entity terminating user plane security needs not be collocated with the respective entity for the control plane. Obviously, some communication protocol between the control plane entity and the user plane entity must be defined for this situation. This protocol can be used to pass K_{Uenc} an K_{Uint} to the user plane entity. (It is assumed that the interface between control plane entity and user plane entity is sufficiently secured, such as any network internal signalling interface.)

Figure 5-3 (in the main text above) illustrates the proposed AS security key hierarchy.

B.3.1.2 Key refreshing

Clearly, procedures to refresh the traffic protection keys will be required. Similar to LTE, the complete AS key hierarchy may be refreshed starting from a new K_{AS}. (Note that a new, independent K_{AS} may be derived without a new AKA-run, assuming that like in LTE there is still a key maintained in the core network that resulted from the previous AKA-run, is unknown to the RAN entities and can therefore be used to derive a new, independent K_{AS}.) Considering the fact that there may be by orders of magnitude more user plane traffic than control plane traffic for certain applications, it is reasonable to specify also a procedure for solely refreshing K_{Uenc} and K_{Uint}. For this, the control plane entity will provide new keys (based on a new AS user plane key set identifier) to the user plane entity and trigger the UE via a control plane message (containing the new AS user plane key set identifier) to switch keys for the user plane.

The figure below illustrates the key refreshing procedure for user plane keys. Note that the figure only shows the principle of the procedure, but does not give a complete description of the message exchange which may for example require additional acknowledge messages.

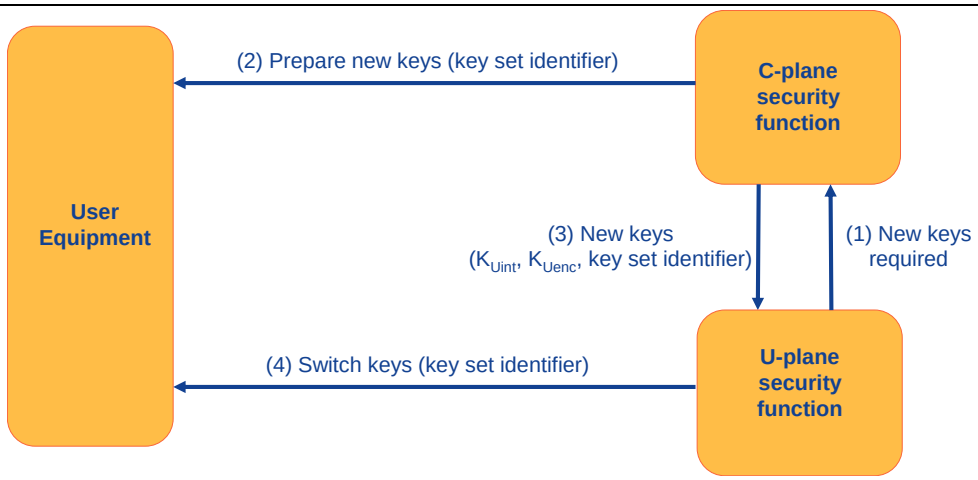


Figure B-5: Refreshing User Plane Keys

B.3.1.3 Relocation of security termination points

The flexible and adaptive 5G NORMA RAN architecture clearly supports the option to relocate the RAN functions terminating user and control plane security. This may be related to the UE moving from one bare metal RAN equipment to another (i.e. a handover is required), but may also happen independently of UE mobility, e.g. in order to re-distribute the network load.

The key hierarchy and the key refreshing procedures described above support the relocation of the entity terminating user plane security in two variants:

- Relocation plus key refresh: This option is reasonable if the function is shifted between two entities where at least one of them is physically exposed and has a notable risk of being compromised. Changing the keys will prevent that a compromised RAN user plane security entity can decrypt or fake communication running via the other entity.
- Relocation without key refresh: If the relocation takes place within one security domain, i.e. an area where it is assumed that a typical security breach in this area affects the whole area, then key refreshing may not increase the security at all. If on the other hand key refreshing has disadvantages that are considered relevant for an application, e.g. introducing additional delay in the user plane, it may be reasonable to avoid the key refreshing.

A relocation of the entity terminating user plane security may also require a change of the security algorithms applied in the user plane, for example if the currently used algorithms are not supported by the platform to which the entity is transferred. It is recommended in this case that the network triggers a relocation including a key refresh and indicates to the UE the new algorithms together with the new AS security key set identifier.

A relocation may also be required for the entity terminating control plane security. The same two options as for the user plane may be supported. If the relocation means just shifting a VNF within a cloud environment constituting a single security domain, the same keys may be maintained. Otherwise, e.g. when the function is shifted from one physically exposed bare metal RAN equipment to another during handover, refreshing the complete AS key hierarchy is required. In this case, generating an independent new K_{AS} would be the preferred approach. As discussed above, this could be done without a new AKA-run, based on a key maintained in the core network that resulted from the previous AKA-run.

Fast handovers without involvement of the core network may be also be supported, where the new K_{AS} is derived from the old one. However, to limit the possible impact of security breaches of individual base stations, care must be taken that sufficiently often an independent new K_{AS} is generated when a chain of handovers is carried out. The policy for this could be similar to LTE that has a mechanism that ensures that an independent key is used at least after each second handover, as described above.

B.3.1.4 Support of multi-connectivity

[5GN-D41] describes different types of multi-connectivity. Not all of these require specific provisions with respect to AS security. In particular, multi-connectivity may be handled in a low radio protocol stack layer, below the protocol layer handling security and possibly transparent to it. Also, a UE may be attached to two access networks at a time, e.g. "3GPP 5G" plus WiFi and use two independent, complete AS security associations for the two access networks. Multi-connectivity may then be handled on the (IETF) transport layer, e.g. via SCTP [SCTP] or Multipath TCP [MPTCP]. From a security point of view, it would also be possible to maintain different AS security associations to (different entities in) one access network in parallel. Whether such a multi-connectivity scenarios makes sense is out of the scope of this section, but handling parallel, independent security associations would anyhow be more an issue for UE implementations, and probably not affect the AS security principles.

Specific security provisions are required for supporting a multi-connectivity approach where within a single overall AS security association (with a single K_{AS}), several different entities terminate security for the user plane for different radio legs. As described above, this can be covered by introducing the AS user plane key set identifier and deriving different user plane key pairs for different terminating entities. Obviously, the UE must be informed via AS signaling, which key pair applies to which radio leg, and what value is to be used as the AS user plane key set identifier for each radio leg. Based on this information, the UE will be able to derive and apply the correct keys for the multiple user plane radio legs.

As discussed above, different keys make in particular sense if the entities holding the keys may be compromised individually. If several different entities within the same security domain terminate the security for different radio legs, also identical key pairs could be used from a security point of view if other differentiating input is used when performing the actual crypto operations, e.g. a bearer or service flow id that is unique for a session. It seems however that there is not much gain in not using different key pairs, as the overall number of different legs is assumed to be very small.

The procedures of refreshing a user plane security key pair and of relocating an entity terminating user plane security for a specific radio leg can be applied individually for each such entity. Note also, that the key pairs for different entities are independent, i.e. a compromised user plane entity has no way to derive the keys used by another user plane entity.

B.3.1.5 Support of multiple RATs

It is assumed that the security mechanisms operate on a protocol layer well above the physical layer and independent of it. Any RAT like mmWave, cmWave or sub-6GHz can thus be secured in the same way. This applies also to "non-3GPP" RATs, in particular WiFi, assuming that such RATs can carry the PDUs (Protocol Data Units) of the protocol providing radio interface security.

As an example, in the LTE-LWA approach (see [33.401], Annex G), PDCP (Packet Data Convergence Protocol) PDUs are transmitted between the eNB and the UE over the WiFi access link, thus allowing to apply LTE radio interface security also to WiFi.

Clearly, when new approaches to use "non-3GPP" RATs are specified, it is not excluded that they may raise additional security issues and may require additional protection mechanisms that will need to be specified together with such new approaches.

To conclude the discussion on multiple RATs, we assume that common radio interface security mechanisms will exist and that the considerations concerning the key handling for these mechanisms as discussed in the preceding sections apply independently of the actually used RAT.

B.3.2 Radio interface security algorithms

We now focus on the available options for crypto-algorithm selection to accommodate the plethora of applications supported by the 5G NORMA architecture. In general, we identify three distinct families of mechanisms, as shown in Figure B-6.

1 The first one refers to traditional cryptographic methods, already deployed in commercial telecommunication systems. In this category belong the well-established symmetric and asymmetric cryptography, already in use in 2G, 3G, and 4G systems. The main concept behind symmetric cryptography is the a-priori establishment of a common key between the service provider and the UE. Building on top of this master key, a series of other keys are derived via the so called "key-derivation functions" (KDF) to mechanise encryption, decryption, authentication, and key-agreement procedures between communicating parties. Asymmetric cryptography on the other hand does not require any previous communication between two parties and provides the means for flexible device enrolment, authentication, non-repudiation and public-key infrastructures, among others.

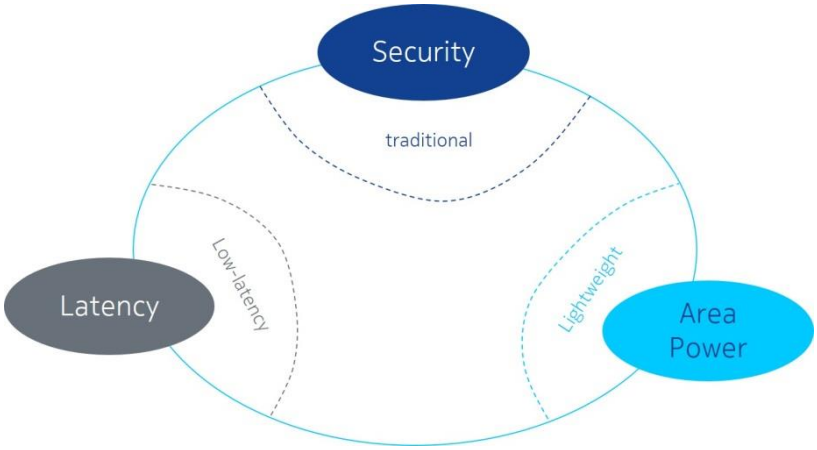


Figure B-6: Trade-offs in crypto-algorithm selection

2 The second one is known by the term Lightweight Cryptography (LWC). LWC was originally devised to provide the means for low-area and low-power cryptographic primitives, targeting specifically the IoT area.

3 The third category refers to low-latency cryptography, a branch of LWC. In low-latency cryptography the cryptographic primitives achieve extremely low number of clock cycles and fast response times to address the stringent delay requirements of critical applications like cMTC (critical Machine Type Communication) or D2D.

Before presenting the crypto-algorithmic options for these cases, it is useful to distinguish software from hardware implementations, as they usually have different and sometimes contradicting properties. Take as an example bit permutations; in hardware this corresponds to simple re-wiring so it comes for free, while in software it's usually a painful operation to perform. Substitution tables on the contrary are easy to implement in software, but plague hardware realisations. In general, it is envisioned that some end devices might be able to accommodate general-purpose micro-processors. In such cases efficient software implementations should be considered. On the other hand, low-cost devices can embed only Application-Specific Integrated Circuits (ASICs), mainly due to small power consumption and reduced fabrication costs. In such cases, efficient hardware implementations are desired.

B.3.2.1 Traditional cryptography

It is envisioned that symmetric cryptography will play an important role in 5G NORMA networks. Its core functions are expected to remain more or less the same, while an increase in the keys' word-length to 256-bits seems a viable scenario. Although current implementations make use of 128-bit keys, the KDFs already employ a 256-bit architecture and subsequently this is a straightforward improvement.

As already mentioned, symmetric ciphers are usually employed in authentication, integrity checks, and payload encryption/decryption. Asymmetric ciphers on the other hand offer better key-management mechanisms and non-repudiation. The computational complexity though of asymmetric algorithms is by far higher (in fact the difference is 2-3 orders of magnitude). This

implies that for fast payload data encryption or authentication procedures that are executed frequently, a symmetric solution would be the option. For more flexible key-establishment scenarios or authentication services conducted in a more sporadic fashion, asymmetric cryptography is probably the way to go forward.

A possible scenario could be to make use of the Diffie-Hellman (DH) scheme (and its Elliptic Curve derivatives) for key-agreement. In this case, a DH key-agreement procedure establishes a common key between two parties without prior knowledge of any other key. This key can be subsequently used in an efficient symmetric algorithm for payload encryption/decryption or in KDFs for other security procedures. DH, although extensively analyzed and attacked, is still considered a safe practice in telecommunications and IT security. Recently, vulnerabilities have been revealed in terms of server misconfigurations, but from cryptanalytic/number-theoretic point of view the scheme is considered safe [Adr].

Note here that, among the main asymmetric schemes employed, namely RSA, Diffie-Hellman, and Elliptic Curve Cryptography (ECC), it is envisioned that ECC will play an increasingly important role, since it offers the smallest key size for equivalent levels of security compared to RSA and DH (see Table B-1 below). The reason is that for RSA a sub-exponential attack is known (meaning an exponential increase in key-lengths over time in order to maintain the security level), while for ECC no such attack is known until now. Interestingly enough, ECC is the only asymmetric algorithm considered for lightweight implementations so far [Wen][Wan][Bat][Gau].

However, in view of the promising era of pervasive computing or IoT, security is facing new challenges. In general, a massive deployment of small wearables, sensors for home use, car sensors for enhanced driving experience and assistance, RFID tags, application-critical devices for industrial use, robotics, etc, to mention only a few, is well anticipated. The main characteristic of these devices is the tight cost constraints - an inherent characteristic in mass deployment cases.

Table B-1: Comparison of symmetric and asymmetric algorithms (algorithms in the same line have similar strength)

Security (bits)	Symmetric encryption algorithm	Public-key	
		RSA	ECC
80	Skipjack	1024	160
112	3DES	2048	224
128	AES-128	3072	256
192	AES-192	7680	384
256	AES-256	15360	512

B.3.2.2 Lightweight cryptography

IoT devices are expected to operate in resource-constrained environments in terms of area, processing power, memory, and of course power consumption. These have been traditionally the main axes of the digital design community to define the efficiency trade-offs of chip manufacturing. But here security is emerging as an extra factor and in fact it is usually contradicting the other. One could achieve high levels of security but would require larger chips with high power dissipation or perhaps adhere to less secure standards in favor of better performance. In this context, the necessity for cryptographic components that could efficiently address the resource and price constraints, but on the same time offer acceptable levels of security is gaining pace. We refer to those implementations specifically tailored for area and power-efficient implementations as Lightweight Cryptography (LWC) [ISO].

LWC mainly provides symmetric algorithms, which usually belong to the Substitution Permutation Network (SPN) algorithm family. In hardware implementations the keys vary from 56-bits for short-term security or very low-cost applications up to 184-bits for higher security levels. Main options in this category are, among others, derivatives of the well-known DES algorithm, namely DESL and DESXL, AES, PRINCE, PRESENT and CLEFIA. The latter two are well-studied about their security and implementation and they are under consideration in ISO/IEC 29192 "Lightweight Cryptography" [ISO]. PRINCE is not yet standardised and NIST has just recently initiated discussions on the standardization of lightweight algorithms.

Considering software implementations, it is useful to take into account the properties of modern programmable micro-controllers, which can enter a wide range of power-down and power-save states. The Tiny Encryption Algorithm (TEA) and the International Data Encryption Algorithm (IDEA), as well as the stream ciphers like Salsa20 or LEX are possible candidates in this category. In terms of trade-offs, the stream ciphers seem to be a good choice. LEX and Salsa20 perform well in terms of code size and throughput, but they are good choices only if the encrypted payload is sufficiently large [Eis]. Otherwise, they produce considerable computational overhead due to their huge block length and setup phase. When code size is extremely critical, TEA, IDEA or even Present seem to be reasonable choices. For most other cases, AES shows again its resilience and strength in software environments [Eis].

B.3.2.3 Low latency cryptography

As already mentioned, the term "lightweight" is used explicitly for area and power constrained environments while recently the term "low-latency" cryptography has been proposed [Kne]. The latter defines cases for which an extremely small number of clock-cycles and fast response-time is required. This feature has been identified as a predicate requirement in critical applications like MTC, cMTC, Car2X, etc. It has been shown that the obtained results for latency, area, power, and energy consumption are strongly influenced by the design properties such as the number of rounds, the round's complexity, and the similarity between encryption and decryption procedures. LWC algorithms are good candidates for low-latency cryptography as well. As an example, PRINCE and PRESENT algorithms as well as a small-scale variant of AES called MINI-AES [Cid] present some interesting low-latency properties, but the field is relatively new and further research is required.

B.3.2.4 Conclusions on crypto algorithm selection in the 5G NORMA architecture

Apparently, selecting the appropriate strategy is a game of trade-offs, which the prospect designer and system architect needs to play in order to meet the application requirements. It has been shown that optimal support of different services requires a flexible security architecture supporting a reasonable set of crypto algorithms with suitable properties. The optimal choice may not only depend on the application and the capability of the mobile devices, but also on the allocation of the security functions within the different network parts, such as bare metal versus cloud environment. Support for dynamic re-allocation of these functions will consequently require procedures that do not only allow the initial security algorithm negotiation but also re-negotiation and seamless change of security algorithms.

B.3.3 Security for other RAN interfaces

This section discusses security for all RAN interfaces except the radio interface. In LTE, these are in particular the S1 and X2 interfaces. In the 5G NORMA architecture, there is a significant difference to LTE, as RAN and core functions will mostly be implemented by VNFs running in distributed or central cloud environments. Moreover, the flexible allocation supported by 5G NORMA means that two communicating functions may in one situation be running in the same edge cloud, maybe even on the same board or processor, and in another situation only one VNF may run in the edge cloud, while the other VNF is allocated in a central data centre.

Tracking the location of the different VNFs and adapting the security features applied to the traffic between them to the actual deployment (e.g. applying a security protocol when two functions are running in different data centres) seems to be an inconvenient approach. On the other hand, always applying security protocols between each pair of VNFs – or even virtual machines (VMs), as a VNF can consist of several distributed VMs – seems a very inefficient approach, that would require a lot of effort for multiple encryption and decryption of data, as well as a for the key management.

The solution for this issue is that the cloud environment provides the required security. Each data centre or edge cloud deployment needs to be secure against attackers. A cloud environment consisting of multiple, distributed sites must take care that traffic between these sites is also secure, and apply suitable, efficient security protocols for the typically large amounts of data exchanged between the sites. Assuming fibre connections, an option could be encryption on the optical layer.

So it is assumed that in a 5G NORMA network, the RAN network interface security is covered mostly by cloud security mechanisms. Only in case of "LTE-eNB-like" deployments of bare metal RAN equipment, where radio interface security is terminated at a physically exposed location, a dedicated backhaul link exists and requires appropriate security. In this case, using backhaul link security measures as in LTE would be a straightforward option.

B.4 5G NORMA V-AAA hierarchical and distributed databases

5G NORMA is expected to develop a flexible RAN and to provide the necessary adaptability for handling the fluctuations in the traffic demands. It is also intended to deliver an independent control of logical network slice and to provide an isolatable network resource for the tenants with their plethoric network services. However, most of the commercial cloud provides only weak performance of isolation between Tenants and applies the multi-tenant resource allocation either on VM allocations or hard rate limits.

In 5G NORMA, the Tenant isolation is not just about isolation of Tenants resource block per-VM nor hard rate limits via SDM-O's orchestration. Resource block of a Tenant can be differentiated from the lowest level as a memory/storage block, or as a single black box with a set of memory/storage blocks, or partitioned into VNF/PNF or VNF service chain, or distinguished between plethoric network services. On the other hand, tenant isolation can also use the weight to assign the resource blocks. The weight can be defined by the type of services, the volume of end-users, or the generated traffics between Tenants. Once the resource block has been assigned to the tenant, the resource block is entered into a lifecycle. It can ensure the resource block to disjoint from other resource blocks and this particular resource block will not be assigned to any other Tenants until it gets released.

Once the differentiations of isolation have been formulated and the resource blocks have been assigned to the tenant, the hierarchical and local distributed replicas can be selected according to the resource blocks location. Therefore, the hierarchical and local distributed replicas can only be operated or exchanged information between the specific allocated resource blocks.

Under this hierarchical and local distributed database architecture approach, we have two main types of data: subscriber data and Tenant VNF logging data. These two datasets require to be protected and replicated across the resource blocks within the Tenant network slice and under such architecture. Both these datasets must retain the MNO central governance while Tenant can manage their own subscriber dataset and VNF logging dataset in the access network (edge cloud). The MNO and Tenant must fulfil the data privacy protecting individual's privacy preferences, data integrity assuring the accuracy and consistency of data when database transitions take place, and data isolation ensuring the data is not visible to other subscribers and Tenants, while processing the data.

While 5G NORMA bringing the flexibility of the network infrastructure to the telecommunication network but also increases the possibility of risk and vulnerability in different aspects, or misallocated the resources that are not belong to a particular Tenant. By applying a complex database approach ensures on tracking the origin of operation at the edge cloud, cooperating the unpredictable flexible network infrastructure at the edge cloud and making information available to the MNO as well as Tenant on the access network (edge cloud). Furthermore, this complex database approach collaborates with V-AAA on delivering the network slice isolation, Tenant isolation and Tenants data isolation when 5G NORMA architecture is in operating.

B.4.1 5G NORMA tenant database

This is a general role based access control (RBAC) database for storing Tenant information, their network slice SLA information and network resources provisioning logs, and interworking with the network slicing and network resources i.e. physical and virtual network functions, provisioning platform. It only locates in the core network (central cloud) as a member of service layer entities within the business support system, which is illustrated in figure 8-11a.

B.4.2 5G NORMA subscriber database

Historically, subscriber data or information has significant value in the telecommunication system. For example, the subscriber data has been used for real-time subscription management or obtaining the UE point of attachment in the network.

5G NORMA proposes a new subscriber hierarchical and distributed database architecture which assists the flexibility of the network infrastructure, provides the real-time subscription management according to the traditional approach in the core network (central cloud) as well as in the access network (edge cloud), analyses the spectrum of allocation and collaborates with the V-AAA in delivering the local AAA functionalities. This subscriber database architecture has two levels. The first level of the database servers sit on the access network (edge cloud), it securely and smartly replicates the necessary subscriber data across the resource block within Tenant's network slice. For instance, the subscriber data replication policies can be set to replicate data in the regional base. The replicas could also exchange with the local subscriber dataset when necessary.

The second level of database server sits on the core network (central cloud), it is managed by MNO as the central subscriber database and stores all the MNO subscriber as well as Tenant's subscriber data and information. This database server can also be seen as HSS or Local Subscriber Server (LSS). It uses many-to-one synchronization to obtain the data from many access network databases. For instance, the hierarchical database architecture can be configured to have one core network database server and many access network database servers, or many core network database and many access network database servers in the operations. In this section, we only consider the replication of the data from many access network database servers to a single core network database server. If one of the distributed database server fails, the other distributed database servers or the adjacent regional database servers can operate normally and pick up the slack. When the database server is back online, it will catch up using replica on and ready to sync. Especially, the database servers in the edge cloud are geographically distributed in several physical sites across the edge cloud. This practice helps the system to alleviate the burden and provides Tenant data isolation. The many-to-one synchronization ensures tenants data only replicate to the core network, but not to the other Tenants. Furthermore, MNO would not push any data back to the Tenant database. Therefore, this many-to-one synchronization is a one-way replication, and the MNO only has the read-only privilege toward the Tenants data. Figure B-7 shows the hierarchical and distributed database architecture and two levels of replications.

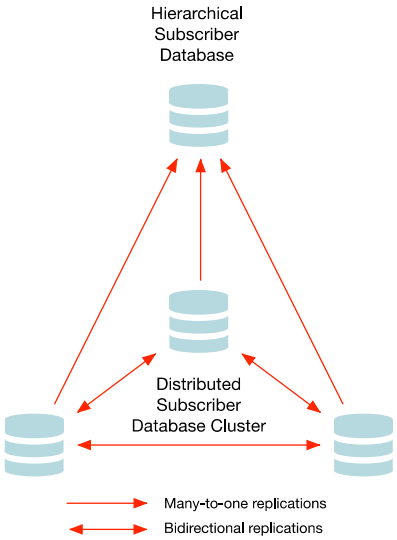


Figure B-7: Hierarchical subscriber's database and distributed subscriber's database cluster network for a tenant

B.4.3 5G NORMA tenant VNF service database

Typically, Tenant data or information has great economic value in the information technology system. For example, the Tenant data has been used for access control, subscription management, services provisioning or to analyze the patterns of fraudulent activities.

5G NORMA proposes a new Tenant hierarchical and distributed database architecture which assists the elasticity of the network slice in the central cloud and edge cloud infrastructures, provides the access control of proofing the identity of Tenant, stores services log for analysing the patterns of the fraudulent activities, and collaborates with the V-AAA in delivering the OpenStack KeyStone functionalities. This Tenant database architecture is the same as the subscriber database which has two levels. The first level of the database servers sit on the access network (edge cloud), it securely and smartly replicates the necessary Tenant's data across the resource block within the Tenant's network slice. In practice, a Tenant could generate many types of dataset, for instance, the identity token from an Identity Server (IS) and the access token from an Authorization Server (AuS), network entities and VNF event logs etc. Additionally, the Tenant data replication helps to increase data availability for digital forensics, and the replication polices could be based on the dataset characteristics. The replicas could also exchange data when necessary.

The second level of the database server can be referred to the second level of subscriber database in the previous section. They have exactly the same characteristics and the same method of replicating the data. However, this database can also be an IS and an AuS. Moreover, it uses many-to-one synchronization to obtain the data from many access network databases. These similarities are illustrated in Figure B-7 and Figure B-8.

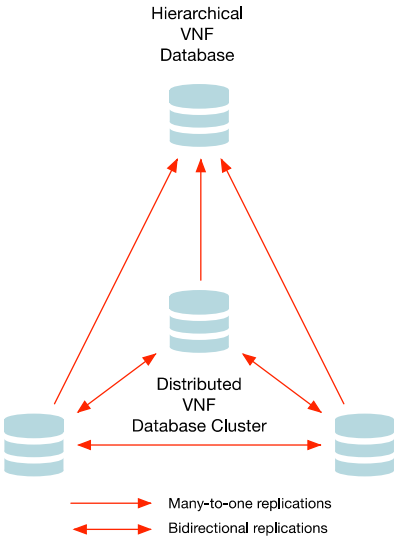


Figure B-8: Hierarchical Tenant VNF database and distributed Tenant VNF database cluster network for a tenant

B.4.4 5G NORMA architecture and V-AAA integration

The V-AAA approach provides a solution to cope with the flexible RAN and network virtualisation elasticity. It can also be applied to the 5G NORMA architecture as a VNF whenever the subscriber densities increase or reduce in a specific region. The Tenant might demand to add or remove the V-AAA service in the edge cloud. In this section, we present two scenarios about the 5G NORMA architecture which provides V-AAA, and hierarchical and distributed databases services. We assume Tenant always has the option to take full control of their network slice or let MNO to have full control of their network slice. The V-AAA capabilities and functionalities fully relies on such decision due to its high flexibility and adaptability of the network infrastructure environment. Furthermore, the interactions between V-AAA and other network entities are also relied on such decision.

In the previous section, the hierarchical and distributed databases and the methodology of replications are presented. It provides a new flexible architecture to handle different datasets under the 5G NORMA architecture whenever Tenant demands V-AAA service.

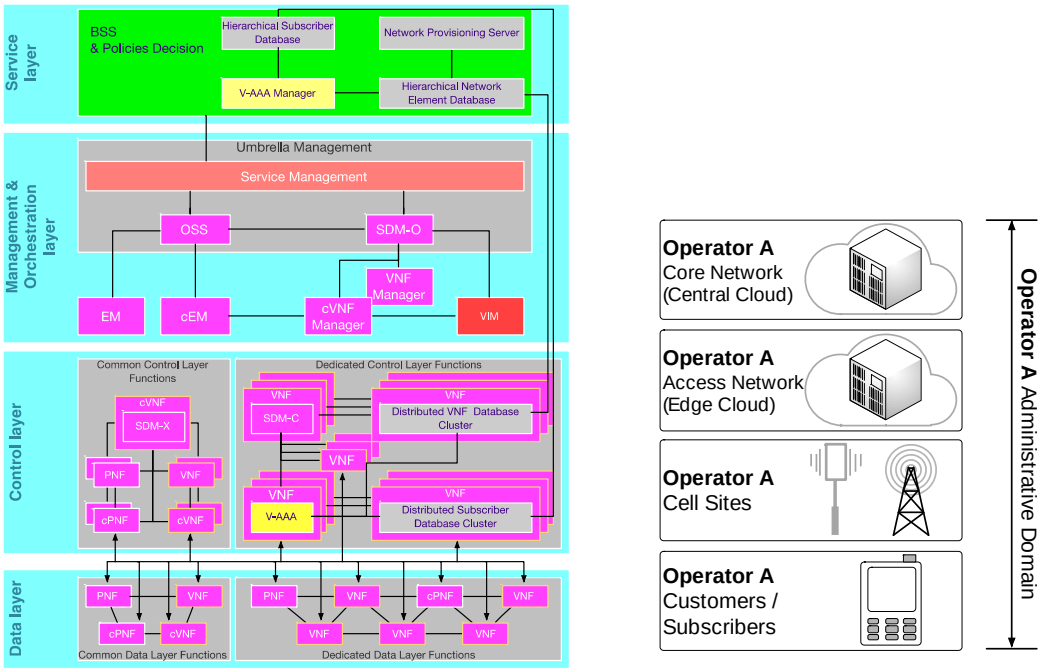
B.4.4.1 5G NORMA architecture and V-AAA integration with monolithic MNO

In this section, a scenario is based on a MNO that has evolved the infrastructure to 5G NORMA and provides flexible RAN. However, MNO does not configure any network slice and there is no Tenant to subscribe network slice. Basically, it cloudifies their core network and access network to apply 5G NORMA like architecture to provide 5G services. It also be the most basic one with no network slice and multi-tenancy. Nevertheless, it is just a cloudification of LTE infrastructure using 5G NORMA architecture. Figure B-9 illustrates the V-AAA approach integrated in 5G NORMA framework for control, management and orchestration of network functions without Tenant and network slice.

In the central cloud, a V-AAA Manager oversees two independent international standards authentication, authorisation and accounting system. On one hand, the V-AAA Manager discreetly oversees the traditional 3GPP AKA functionalities in handling the subscribers. On the other hand, it tactfully oversees the ETSI NFV OpenStack KeyStone to handle the virtual network function services. Due to the independence of development and evolvement of these two international standards, the role of V-AAA Manager is to harmonise these two international standard systems whenever they have new development. They can be independently applied and evolved the overall

services. Furthermore, the V-AAA Manager takes the softwarization advantage of all functionalities which can be plugged-in or unplugged whenever it is necessary. This is the main characteristic of the V-AAA approach. It provides flexibilities and adaptabilities to any new development of those international standards in the core network as well as it oversees each process. In this scenario, the MNO or Tenant opens the blueprint catalogue network slice provisioning platform to choose a blueprint of network slice. The V-AAA oversees the blueprint catalogue network slice provisioning platform that initiates a network resource i.e. network slice. It requests for obtaining the identity token and access token from the IS and AuS for provisioning and deploying the network slice from the core network.

In the edge cloud, an optional V-AAA is presented for emergency purposes which could handle identification, authentication, authorisation or delegation when the edge cloud is disconnected from the central cloud. During the emergency periods, the edge cloud requires some MME and HSS functionalities to maintain the subscriber point of attachment that allows others to connect with the subscribers. Therefore, V-AAA might include the MME and HSS functionalities during the emergency periods.



a) V-AAA approach integrates with 5G NORMA Architecture

b) A logical illustration of the scenario with no tenant and no network slice.

Figure B-9: 5G NORMA cloudification no Tenant and network slice framework for control, management and orchestration of network functions

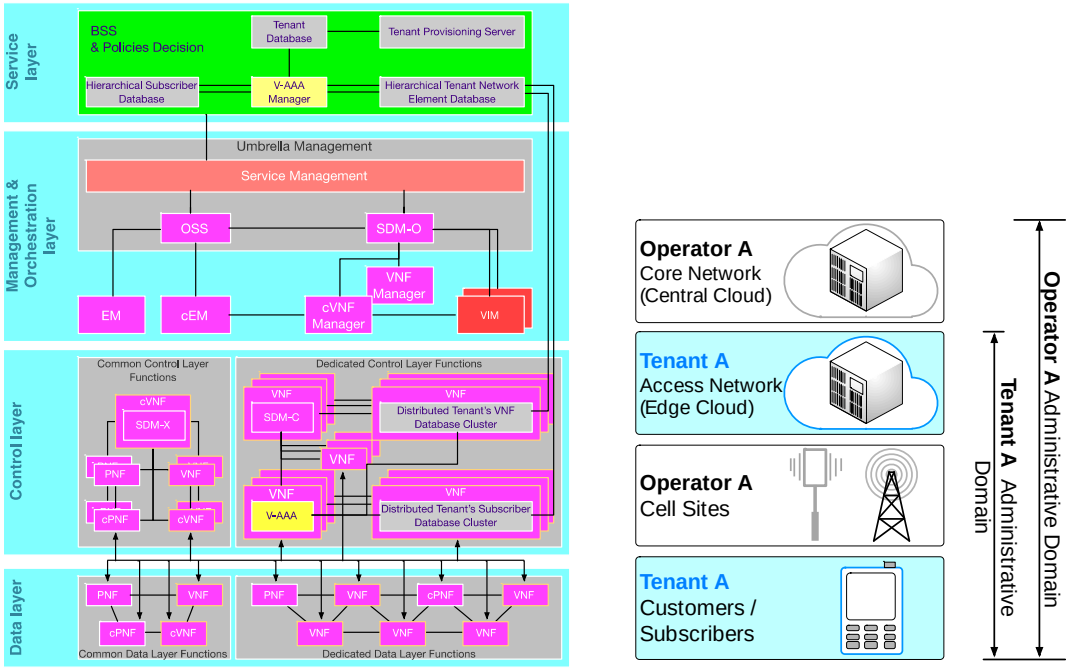
B.4.4.2 5G NORMA architecture and V-AAA integration with monolithic MNO and single tenant

In this section, a scenario is based on a MNO that has evolved from the LTE network infrastructure to 5G NORMA and provides a flexible access network and core network in the cloud environment. Precisely, MNO configures the network infrastructure to support network slices and to provide a single Tenant subscription of the network slice. MNO basically cloudifies their core network and provides a single Tenant edge cloud subscription. Therefore, this scenario has a single tenant and network slice. The level of security awareness is raised into two levels and two administrative domains that is illustrated in figure 8-10b. The tenant might prefer to manage its own subscribers and network slice. On the other hand, MNO still provides the central cloud functionality and maintains the central governance of the entire network including the Tenant's sub-

scribers, therefore, subscriber authentication still retains in the core network. Figure B-10 illustrates the V-AAA approach integrated in 5G NORMA framework for control, management and orchestration of network functions with a single Tenant and network slice.

In the central cloud, a V-AAA Manager oversees two independent international standards authentication, authorization and accounting system which is similar to the previous scenario. However, in this scenario, the ETSI NFV OpenStack KeyStone catalogue could be customised to handle the MNO, Tenant, network slice and VNF services. Since this scenario builds on top of the previous scenario, the V-AAA Manager basically has the same functionalities in the central cloud as the previous scenario. However, there are some add-on functionalities. The complexity of this scenario has been increased from the RBAC that support two different organizations and multiple access policies within this two groups i.e. MNO and Tenant, to manage the two levels of identity proven process when the member of these two organizations are trying to access the network resources. The V-AAA Manager oversees the authorization and delegation processes that initiate a request from MNO and Tenant to IS and AuS, and all the way to access the network resources in edge cloud.

In the edge cloud, the optional V-AAA is still presented to the Tenant's network slice. Additionally, the Tenant might prefer to manage their own network slice and subscribers at the edge cloud. MNO can use the hierarchical database approaches to collect data and the hierarchical V-AAA architecture to oversees the Tenant's network slice as well as to give the Tenant freedom on manipulating the network resources at the edge cloud within the SLA scope. When V-AAA applies in the edge cloud, it gives the edge cloud more flexibility and efficiency in dealing with the real-time network resource register and deregister in demand due to the elasticity of network slice. It also requires to deploy the Tenant IS and AuS as a VNF for proofing the Tenant's administrative group or role's identities, authorizing an access to a specific network resource and delegating a particular service in a specific edge cloud location and time period. On the other hand, during the emergency or disaster event take place, the edge cloud requires some MME and LSS functionalities to maintain the subscriber point of attachment that allows others to connect the subscriber which is illustrated in the Trust Zone, Section 5.4.5.



a) V-AAA approach integrates with 5G NORMA Architecture

b) A logical illustration of the scenario with single tenant and single network slice

Figure B-10: 5G NORMA cloudification a single Tenant and network slice framework for control, management and orchestration of network functions

Annex C Details of Verification Analysis

C.1 Methodology

C.1.1 KPI definitions

C.1.1.1 Capacity and traffic density

Area wide capacity density describes the ability of a network to provide an amount of data volume per service area during the hour with the highest traffic load (busy hour). As the performance of mobile networks depends on many influencing factors for simplicity it is assumed that during busy hour all radio resources are occupied (fully loaded system). The spectral efficiency is defined as the aggregate uplink / downlink cell full buffer cell throughput per spectrum block assignment bandwidth. Hence cell capacity [Mbit] can be calculated from spectral efficiency by multiplication with system bandwidth and time duration (1 hour).

As macro cells provide full coverage in the whole cell range (except for small percentage described by coverage probability) the capacity contribution by macro cells is sufficiently characterised by macro cell capacity as described above.

Heterogeneous networks consist of different network layer (macro, small cell, WiFi) where some of those network layers are not available in the whole cell (e.g. small cell coverage is assumed to be only a small fraction of the macro cell). Hence the amount of data volume carried by those layers in addition of the node capabilities depends on the positioning of nodes and user distribution (a node at a location without traffic demand cannot contribute to network capacity). The capacity provided by layers with spotty coverage (small cells, Wifi) depends on the number of nodes within the service area and may be determined by measurements or by expert estimation.

Traffic density characterises the demand of data volume during busy hour per service area.

C.1.1.2 User plane Latency

UP latency is defined as the one way transmission time of a packet between the transmitter and the availability of this packet in the receiver. The measurement reference is the MAC layer in both transmitter and receiver side. Analysis must distinguish between UP latency in an infrastructure-based communications and in a direct device-to-device (D2D) communication [PA-5GPPP].

C.1.1.3 E2E latency

Different types of latency are relevant for different applications. E2E latency, or one trip time (OTT) latency, refers to the time it takes from when a data packet is sent from the transmitting end to when it is received at the receiving entity, e.g., internet server or other device. Another latency measure is the round trip time (RTT) latency which refers to the time from when a data packet is sent from the transmitting end until acknowledgements are received from the receiving entity. The measurement reference in both cases is the interface between Layer 2 and 3 [PA-5GPPP].

C.1.1.4 Peak data rate

The peak data rate is the highest theoretical single user data rate, i.e., assuming error-free transmission conditions, when all available radio resources for the corresponding link direction are utilised (i.e., excluding radio resources that are used for physical layer synchronization, reference signals or pilots, guard bands and guard times). Peak data rate calculation shall include the details on the assumed MIMO configuration and bandwidth [PA-5GPPP].

C.1.1.5 Mobility

Mobility refers to the system's ability to provide seamless service experience to users that are moving at a certain speed. Mobility requirements may be specified by a maximum percentage decrease of user throughput that is caused by increasing the device velocity [NGMN].

C.1.1.6 Device density

Device density denotes the number of devices per service area that are connected to the network. For connection oriented services devices must be in active or idle mode. For connectionless services devices just have to be within coverage area of the network.

C.1.1.7 Reliability

The reliability of a communication is characterised by its reliability rate, defined as follows: the amount of sent packets successfully delivered to the destination within the time constraint required by the targeted service, divided by the total number of sent packets. Note that the reliability rate is evaluated only when the network is available [NGMN].

C.1.1.8 Availability

The availability in percentage is defined as the number of places (related to a predefined area unit or pixel size) where the QoE level requested by the end-user is achieved divided by the total coverage area of a single radio cell or multi-cell area (equal to the total number of pixels) times 100.

(Note: FANTASTIC-5G defines availability as equal to (1 – service blocking probability), where service blocking probability is due to lack of enough resources to access, grant and provide the service, even in case of adequate coverage) [PA-5GPPP].

C.1.1.9 Coverage

Coverage probability refers to geographical locations and indicates the percentage of locations with respect to the whole service area where a certain service can be provided.

C.1.2 Verification tools

In the following verification tools are described that will support our evaluation activities during the second design iteration loop.

C.1.2.1 System level simulations

Multi-connectivity

Performance requirements like user throughput, user plane latency, frame error rate etc. can be checked by system level simulations. More specifically inter- (LTE + 5G) and intra-RAT (5G cells) multi-connectivity will be investigated. The simulations will involve heterogeneous networks involving both wide-area and ultra-dense small-cell deployments. Results will include user throughput improvements and reliability performance which demonstrates a trade-of between reliability and aggregation mode.

Network programmability

Based on simulation of a congestion issue that shall be resolved by the SDMC controller there will be an assessment of network programmability.

QoE based routing, network agility

Routing will take into account the QoE feedback received from the users, and dynamically determine and enforce the optimal route the flow needs to take to improve that feedback. Reinforcement learning (Q-learning) will be the basis for this routing control. Simulations in 5G NORMA

will be based on a discrete event network simulator (ns-3) to evaluate the contribution in two ways

- Study how reactive our system is (how long does it take for the QoE to improve after changes in feedback and how do the various algorithm parameters affect the speed of convergence, how are changes in delay and packet which affect the QoE handled by the routing control)
- Compare a QoS-based routing scheme with a shortest-path routing scheme.

Edge function mobility

The 5G NORMA architecture will allow for network functions to be allocated dynamically either in the edge cloud or in the centralised network cloud. When a user moves away from the area covered by an edge cloud, three decisions can be made:

- The network function continues to run at the original edge cloud,
- the function is reallocated to a new edge cloud, or
- reallocated to the network cloud.

Factors like delay requirements, communication overhead, reallocation costs and user QoE influence this placement decision. The plan for this simulation is to create a module for determining the placement of a network functions, taking into consideration the user's location, network conditions and QoE demands. A migration and transmission cost estimation will be performed, and the output is a placement sequence that the Service Orchestrator can follow. It is planned to simulate a group of cells and a random walk user mobility model. The goal is to compare the placement decision model to other approaches, for instance never migrate (functions will never move, and user requests are routed to the edge cloud running the functions), always migrate (functions follow the user), always migrate first to the network cloud.

Multi-tenant dynamic resource allocation

Via simulation the amount of requests that can be admitted by the network while satisfying the given QoS constraints will be evaluated. This implies evaluating the maximum number of users of a given class that can be admitted to the network while satisfying the corresponding SLA requirements. Based on the above results, an algorithm is designed that decides the requests can be admitted and the corresponding resource sharing in order to maximise the overall utility. The approach will be extended to Multi-RAT scenarios. Along the above lines, the contribution focuses on the definition of a new sharing criterion that allows allocating the resources among tenants in more flexible way. Given the following input: (i) a set of resources required from the different tenants, (ii) the period during which tenants require these resources, (iii) the QoS constraints, and (iv) the probability of not satisfying the given constraints, the algorithm will allocate the available resources maximizing the resources' utilization as well as the infrastructure provider revenue.

Assessment of mobility concepts

Assessment of mobility concepts will cover an analytical qualitative evaluation of proposed concepts (e.g. protocol efficiency and granularity vs. use cases and signalling overhead estimation/comparison) in terms of breakdown of procedures for service requirements and infrastructure capabilities' mapping for selected use cases to MSCs (Message Sequence Charts) (e.g. in case of mobility management protocol design, Task 5.1). Comparison of different interface definitions between functional building blocks within WP5 and assessment of possible Network Function chaining models (e.g. for coordinating and orchestrating service aware resource allocation between 5G network slices) maybe the approach for , Task 5.2 and Task 5.3, respectively. Depending on the expense in terms of optional and mandatory information exchange and required additional headers and defined options a rough impact on expected performance can be examined – potentially including consideration of scalability and granularity.

C.1.2.2 Demonstrators

In addition to simulations, proof-of-concept prototypes are very useful to assess the feasibility of the proposed architecture and evaluate it under realistic conditions. WP6 partners will provide an

integrated MANO implementation for real-world proof-of-concept that has to be adopted to the different evaluation cases as well as possible. In particular, the Atos demonstration will consist in a MANO implementation that will include the developments carried out in 5G-NORMA. This MANO implementation will include all the modules and interfaces of the 5G-NORMA architecture. In addition to this, Atos will contribute also with the virtualised infrastructure aspects. Therefore, evaluations on the impact of virtualization on the 5G-NORMA architecture will be carried out. Within the architecture design verification WP6 contributes primarily to PoC by reporting on practical experience implementing the demonstrators.

C.1.2.3 Protocol analysis

Protocol overhead analysis

Theoretical / mathematical tools shall be used to validate 5G NORMA concepts. Processes will be investigated where the behaviour of 5G NORMA system to external triggers is monitored. Four different kind of analysis will be performed:

- Based on the evaluation cases, protocol overhead analysis will provide information on minimum achievable latencies and delay that is being introduced by the network. Expected output meaning - data transmission time if no connection establishment is required (e.g. for transmission of sensor data) and minimum roundtrip time with established connection (UE – application server – UE; w/o access to public internet) as well as impact of connection establishment on roundtrip time.
- Based on generic interface definition from WP4 protocol overhead on these interfaces will be evaluated. Expected result will be the relative amount of overhead compared to the transported payload data.
- It is expected that similar to legacy mobility management schemes, some control signalling will be required. The assumption here is that the air interface and the SDMC server will likely be the bottlenecks in managing this control traffic. Expected input from WP5 is placement of SDMC servers, definition of control messages, frequency by which type of message is send and assumption of traffic load, number of terminals per cell etc. Outcome will be the number of control messages at the SDMC server as well as control traffic demand.
- The amount of data to be transferred during the execution of a terminal handover and a function re-allocation shall be determined. In case of terminal handover, this can be compared to a handover in LTE. It is expected that in case of terminal handover, the mechanism for traffic forwarding will have the highest influence. Expected input by WP5 is Flow of signaling messages in case of terminal / function re-allocation, Size of the signaling messages, Amount of traffic to be forwarded and resp. amount of data to be forwarded in case of function re-allocation (e.g. executable SW image or only function state). Results of this investigation will be additional traffic volume created by the handover, in comparison for SDMC and legacy LTE and additional traffic volume created by function re-allocation, in comparison for multiple re-allocation schemes.

Inter-function communication and information availability

In 5G NORMA, interfaces and protocols are to be defined (exemplarily) as part of the overall project goals in WP3.

For user-centric radio access in WP4 will provide input to simulation assumptions in respect to which information is available to processing at the different antenna sites and with what latency. Latency will be derived from Message sequence charts (MSC) in conjunction with assumptions on inter-network function communication latencies (transport) for different deployments scenarios. For 5G NORMA, this will yield the required capacity demands on the access network between the different NFs, to be derived from protocol definitions.

Primary focus is on massive broadband in an ultra-dense urban heterogeneous network with (partly) centralised processing in edge clouds and (near) antenna sites. As second service, in the

same deployment, likely low latency mission critical will be considered. As second deployment, rural areas with decentralised processing may be considered.

Expected output will be interface requirements and respective possible latencies and capacity demand.

C.1.2.4 Capacity verification methodology

The evaluation area (Figure C-) consists of three hexagons containing a macro site in the middle. Cell areas depend on the inter-site-distance (ISD) of an assumed regular hexagonal grid of base station sites and can be calculated by

$$A = \frac{ISD^2}{2\sqrt{3}} \quad (1)$$

For capacity verification three typical site types in the London sample area are considered

- Low traffic sites with typical ISD of 900m.
- Medium traffic sites with typical ISD of 500m
- High traffic sites with typical ISD of 200m

Four different capacity layers are considered

- The macro layer including services at sub 1 GHz, low and medium frequency bands
- The small cell layer at low and medium (unpaired) frequency bands
- The small cell layer at high frequency bands (< 6 GHz)
- MBB traffic carried by fixed network over WiFi

In public or commercial hot spots and in residential areas Wifi is carrying a significant portion of the MBB traffic (s. Table C-). Reliable information on these off-load factors is collected by data meter apps running on smart phones.

Multiplying the total MBB traffic demand by this off-load factor we get the cellular traffic demand. As suggested by WP2 use case definition we measure the traffic demand T related to the cell area A in [Mbps/km^2]

$$T_{Cellular} = T_{total} \cdot \rho_{WiFi} \quad (2)$$

ρ_{WiFi} denotes the off-load factor by the WiFi layer and T the respective traffic densities.

The cellular fraction of the traffic demand has to be carried by the macro and the two small cell layers introduced above. Developing the RAN for MBB over the years we assume that the macro layer is available at any location and hence his capability to carry traffic depends on the spectrum efficiency ε of respective radio technologies and the bandwidth B

$$C_{macro} = \frac{B \cdot \varepsilon_{macro}}{A} \quad (3)$$

C_{macro} denotes the capacity density provided by the macro layer. The evolution of spectrum efficiency ε is depicted in Figure C-4 where a distinction was made between sub 1 GHz / low bands where higher order MIMO was assumed (max. 4 antenna ports) and medium frequency bands where we assume the deployment of more enhanced full dimension and massive MIMO technologies (max. 64 antenna ports) as well as between the different small cell layers below and above 6 GHz.

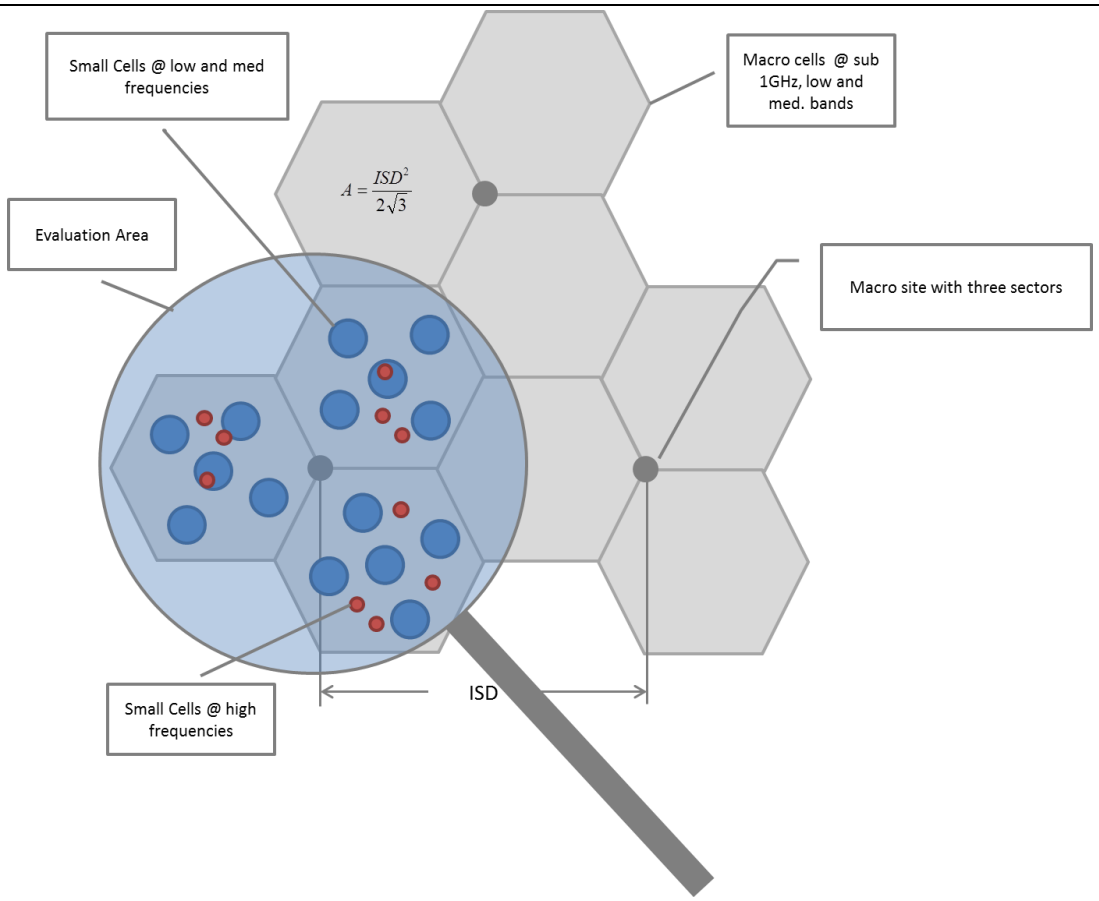


Figure C-1: Methodology for capacity verification

Small cell layers normally provide spotty coverage. Hence the capability of small cells to serve traffic demand depends heavily from the user distribution and their positioning with respect to this timely changing traffic demand. Therefore in the model we assume that small cells off-load[4] the cellular traffic demand by certain fraction that depends on their cell range (coverage area) and the number of small cells per macro sector. Assumptions on this off-load capabilities of small cells at low and medium as well as high frequency bands are depicted in Figure C-5. Furthermore, we assume that the capacity contributed by a small cell layer is as long identical to the traffic demand based on the small cell off-load capability

$$T_{pico} = T_{Cellular} \cdot \rho_{pico} \quad (4)$$

as the aggregated sector capacity density

$$C_{pico} = n_{pico} \cdot \frac{B\varepsilon_{pico}}{A} \quad (5)$$

in the evaluation area with n_{pico} SC not exceeded.

Our strategy for extending the network consist of the following steps: The different layers are extended according to the spectrum availability depicted in Figure C-3.

- Adjusting the number of small cells per sector at low/medium and high bands we try to meet the traffic density defined by start value in 2020 and CAGR. In order to avoid poor performance due to interference the number of SC at med. and low frequencies has to be restricted to ≤5[5]. As the coverage range at high bands is much less, the number of SC may be incremented as long as the capacity target is reached.

[4] Similar to the WiFi layer.

[5] This as well as the restriction to dedicated small cell frequency bands could be refined if assume introduction of enhanced interference management as provided by 5G NORMA

- The spectrum efficiency as depicted in Figure C-4 increases with time by increasing penetration of devices with more antennas and improved releases of radio node technology.
- First the number of small cells at low / medium frequencies can be increased to a limit of 5.
- If cellular traffic demand (2) is not met SC at high frequency band are added as long as needed.

C.1.3 Evaluation cases

C.1.3.1 Baseline evaluation case

This section describes how an LTE or alternatively a 5G NORMA network could evolve in order to serve the requirements for MBB in the time span between 2020 and 2030. In addition most important assumptions for baseline evaluation are compiled.

For evolution to LTE-A Pro with respect to MBB we assume that there is no need to change the architecture depicted in Figure 3-1. Compared to LTE, 5G NORMA however will allow for more flexible site chains as introduced in deliverable D2.2 [5GN-D22]. Besides the classical base station variant 1 where antenna sites are connected directly or via edge cloud to central clouds (D-RAN) there exist options of restricting antenna sites to the function of remote radio heads (RRU) that are connected to edge clouds via CPRI/ORI interfaces (configuration 2 in D2.2).

In case of D-RAN there might be a need of direct physical links between antenna sites to enable base station cooperation. These direct physical links may be realised by breakout from the IPSec tunnels at aggregation switches (s. Figure B-4) or by deployment of links between antenna sites.

Optional antenna sites could integrate edge cloud capabilities enabling support of processing of higher radio protocol stack elements (configuration 3 in D2.2). With respect to MBB the latest option would be favourable enable small cell interference management or functional splits at higher protocol stack layers applicable in case of non-available CPRI or ORI transport technology. For MBB we skip edge clouds at the antenna sites and assume that antenna sites optionally host D-RAN or remote radio heads (RRU) (Figure C-2).

.According to WP2 evaluations the average and peak traffic demand density for MBB services in the sample area in 2020 is anticipated to range between approximately 1800 Mbps/sqkm and 20000 Mbps/sqkm respectively.

It is assumed that capacity demand grows according to prediction of Cisco VNI [Cisco_2016] with a CAGR of 30%.

For the sample area we assume an average macro cell ISD of 570 m as a starting point for the year 2020. In some areas it is possible to have a minimum ISD of 250 m. Within the timeframe up to 2030 it is assumed that due to increased traffic demand it will be possible to densify the macro base stations to the maximum base station density on demand.

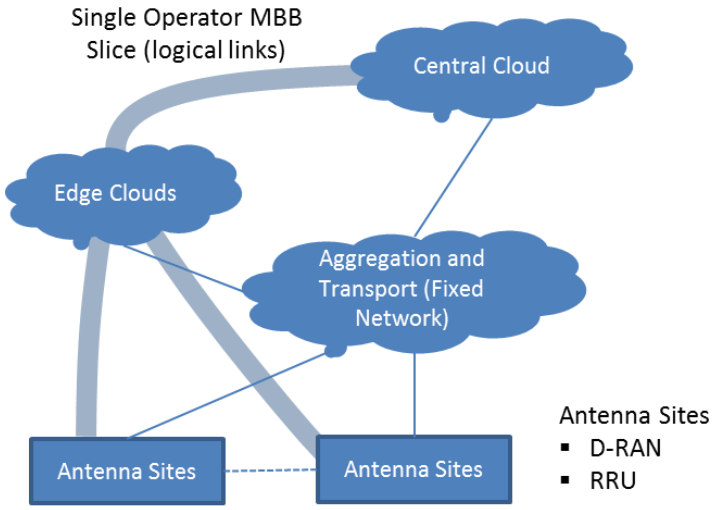


Figure C-2: Target architecture for baseline evaluation

In line with assumptions from deliverable D2.2 (s. Figure C-3) in 2020 spectrum deployment per operator is

- Sub 1 GHz (700-900 MHz) 10 MHz paired[6] (Macro)
- Low frequency band (1800-2600 MHz) 40 MHz paired (Macro)
- Low band (1800-2600 MHz) 20 MHz unpaired (Small Cell)

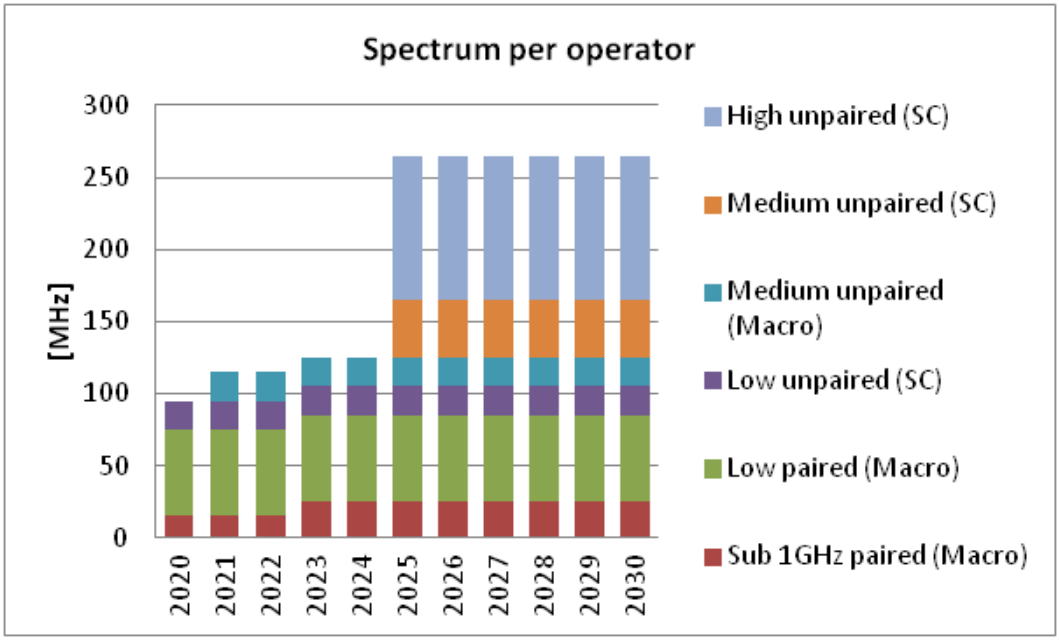


Figure C-3: Spectrum availability per operator (according to D2.2)

We assume that until end of 2030 the following spectrum per operator will be available

- Sub 1 GHz (700-900 MHz) 20 MHz paired (Macro)
- Low frequency band (1800-2600 MHz) 60 MHz paired (Macro)
- Low band (1800-2600 MHz) 20 MHz unpaired (Small Cell)
- Medium bands (2800-3800 MHz) 40 MHz unpaired (Macro)
- Medium bands (2800-3800 MHz) 40 MHz unpaired (Small Cell)
- High band (> 6000 MHz) 200 MHz unpaired (Small Cell)

[6] Paired means this figures refers to downlink spectrum, for uplink the same amount of spectrum is available.

Whereas low band frequencies provide a capacity layer sub 1GHz bands enable deep indoor coverage and serve indoor mobile traffic demand that is not covered by WiFi or indoor femto cells. According to market studies [OV15] end of 2014 around 80% of smart phone traffic is offloaded by WiFi. In our evaluation we assume that this traffic share is in saturation.

The transmitted power density at macro base stations is 4W per MHz at sub 1 GHz bands and 2 W per MHz at low and medium bands[7]. Small cells at low and medium spectrum bands are limited to total EIRP of 40 dBm and are deployed at dedicated (unpaired) spectrum in order to avoid interference of the macro layers. For high band small cells we assume that the maximum transmitted power is limited to 30 dBm EIRP. Hence high band small cells will be limited to lone-of-sight (LOS) coverage.

Macro- site antennas are designed for multiple frequency bands where currently sub 1GHz and low bands need extra sector antenna arrays. The antenna arrays consist of a single column with two cross-polarised antenna ports. Three sector antennas are arranged around an antenna pole and remote radio heads (RRH) are mounted below the antenna panels. These RRH are connected to the base band units via fibre fronthaul links (CPRI/ORI) that carry high bitrate I/Q samples per antenna port[8]. We assume that for deployment reasons antennas at sub 1 GHz and low bands have to be consolidated into a single multi-band antenna panel. Each of this antenna panels comprise 2 columns per band with cross polarised antenna elements providing 4 antenna ports for each band (sub 1 GHz and low band). Antennas at medium unpaired frequency bands may enable full dimension (3D) MIMO including quad column antenna arrays with up to 64 antenna ports at macro base stations. Due to the beamforming features these antennas normally cannot be used for multiple frequency bands. In order to keep EMF safety distances spectrum deployment is to be restricted up to 100 MHz per site per operator. In total, two operators may share those macro sites each deploying one antenna plane for sub 1GHz and low band and one antenna plane for medium bands.

The following options for capacity extension are considered

- More spectrum
- Increase of spectrum efficiency
- Improved MIMO schemes (4x4, 32x2 and 64x4 depending on the frequency band)
- 6 sectors (utilizing horizontal beamforming splitting a conventional sector) per macro site
- New macro sites carrying sub 1GHz and low and medium bands
- Small cell sites at low and medium frequency bands
- Small cells at high bands (at traffic hot spots and points of interest)

The increase in spectrum efficiency for sub 1GHz and low bands will be available independent of 5G networks (s. Figure C-4). It is realised by application of improved MIMO schemes (4x4) or increased sectorisation using up to 2 column antennas where 2 column and UE penetration cause the temporal development with improving spectrum efficiency. The increase in spectrum efficiency for medium frequency bands is estimated assuming deployment of M-MIMO with up to 64 antenna elements at macro sites. Due to power limitations at small cells and corresponding low power density at medium frequency bands it does not make sense deploying the entire spectrum at small cells. Hence in our map exercise medium spectrum bands (2800 -3800 MHz) have to be deployed partly at small cells and partly at macro sites. Due to enhanced MIMO techniques (full dimension / massive MIMO) that can be deployed favourably at unpaired spectrum (channel reciprocity for TDD systems) spectrum efficiency is assumed to be improved compared to macros at low and sub 1GHz bands. It is assumed that with 64 antenna elements applying single-user

[7] These values are needed for evaluation in order to calculate safety distances that limit spectrum deployment at macro and small cell sites.

[8] Antenna ports allow for data transmission to devices over spatially separated radio channels and are not to be confused with antenna elements. Antenna elements are interconnected by a radio distribution networks enabling analogue beamforming (e.g. forming the vertical pattern).

(SU) and multi-user MIMO up to 7 times the spectrum efficiency of low band cells can be achieved.

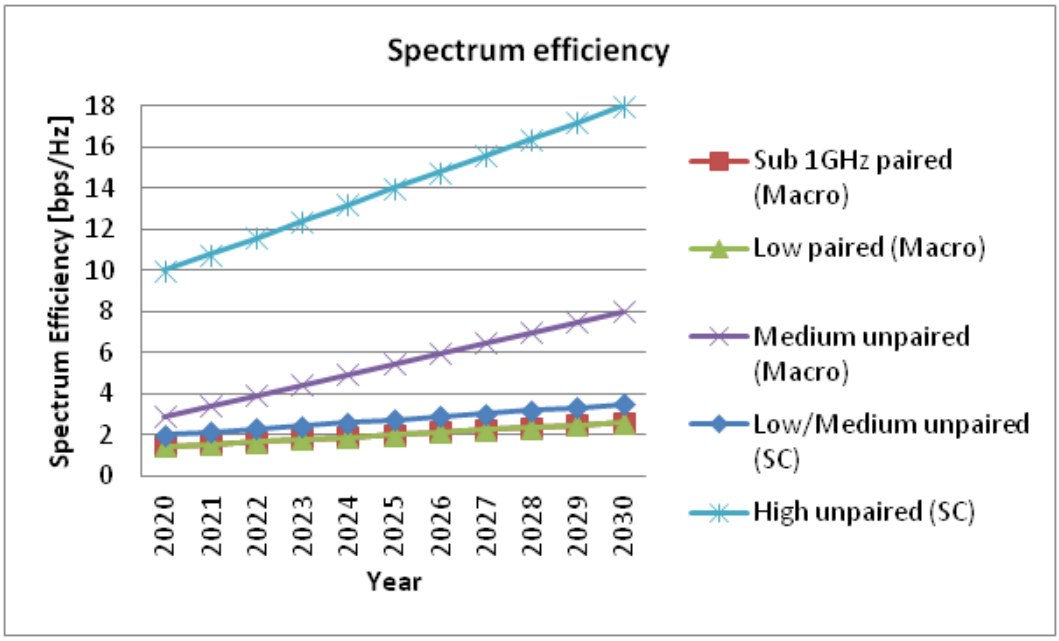


Figure C-4: Average spectrum efficiency for different cellular layers

Small cells in general (at low, medium and high frequency bands) will have spotty coverage. Hence their contribution to capacity extensions cannot be measured in terms of average cell throughput as done for the macro layer (s. Annex C.1.2.4). Instead it is assumed that small cells contribute to capacity extension in the same way than WiFi by their traffic offload. Traffic offload is defined as the portion of the total traffic that is carried by small cells (or WiFi) and may be described as a function of the number of small cells per macro sector rather than the possible user throughput. The assumed different off-load capabilities depicted in Figure C-5 are substantiated by different cell areas of small cells at low/med and high bands.

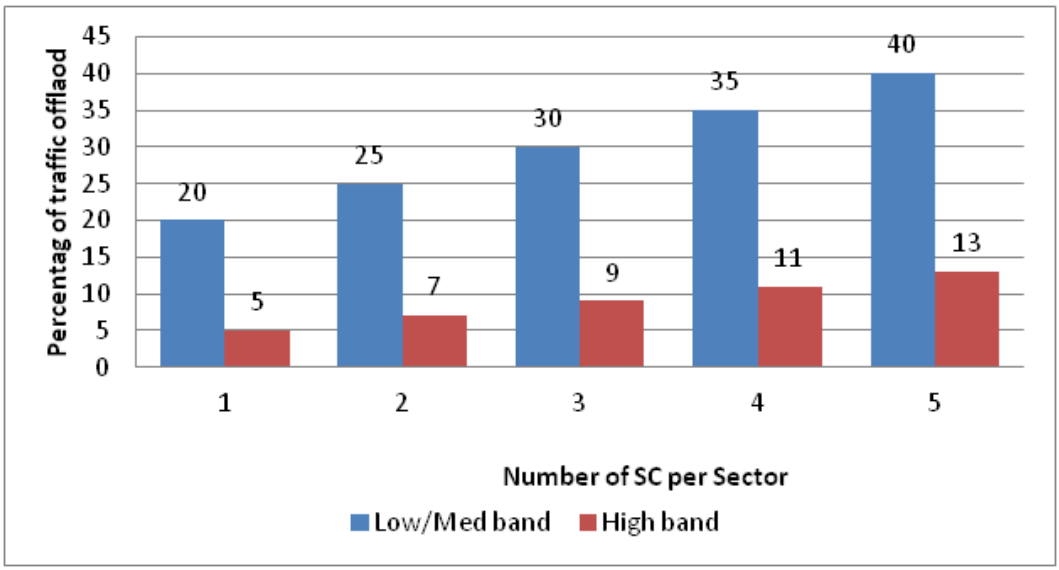


Figure C-5: Traffic offload by small cells

Small cells are deployable on street furniture – such as lampposts and bus stops, available several meters above street level. Per site up to two sectors may be operated and the number of small cells sites per macro sector should be limited to ≤5 for low and medium bands. Small cells are favourable for capacity extension at traffic hot spots.

According to deliverable D2.1 [5GN-D21] coverage probability for MBB has to be at least 95% which fit to current status of operator networks. Hence there will be no need for further base station densification due to increased coverage demand for MBB.

In order to provide higher per-user data rates, serve higher user densities small cells at high frequency bands > 6000 MHz may be operated. SC at mmWave frequencies may be deployed indoors as well outdoors and will benefit from novelties investigated in the H2020 project mmMAGIC [MAG-D41]. Not only wider bandwidth but also the application of M-MIMO leading to increased spectrum efficiency promise user data rates in the range of several Gbps and will contribute to eMBB as envisaged in D2.1.

With respect to MBB there will be no significant performance deviations between LTE-A pro and 5G. Edge cloud have to be sufficiently dense in order to make sure that CPRI links to antenna sites fulfil the latency requirements of < 100µs. Regarding central clouds the same conditions apply as for EPC in case of 4G networks. For reason of resilience there should be at least 2 central clouds per operator. Most important assumptions are compiled in Table C-.

Table C-1: Assumptions for baseline evaluation

Business model parameter	**4G / LTE-A Pro**	**5G NORMA**
Services	MBB	MBB
Offer Type	MNO owns all infrastructure	MNO owns all infrastructure
Stakeholder / Customers	MNO / End users	MNO / End users
Deployment Scenario Parameters	4G / LTE-A Pro	5G NORMA
Radio node density	Sites in existing London area + on demand densification up to current max. density	Sites in existing London area + on demand densification up to current max. density
Spectrum efficiency (only relevant for macro cells)	Macros @sub 1GHz and low bands: transitions from 2x2 to 4x4 MIMO Macros @ med bands: transition from 32x2 to 64x4	Macros @sub 1GHz and low bands: transitions from 2x2 to 4x4 MIMO Macros @ med bands: transition from 32x2 to 64x4
Transport network	Backhaul macro + SC: Fibre Aggregation network: Fibre	WAN[9] Network: Fibre Fronthaul: Fibre Backhaul: Fibre
MBB requirements (according to use case A2, Enhanced mobile broadband)	Coverage probability: 95% Capacity demand: according to Cisco VNI report	Coverage probability: 95% Capacity demand: according to Cisco VNI report Peak data rate: 10 Gbps
Indoor / Outdoor traffic demand	80% / 20%	80% / 20%
Traffic offload by indoor SC+WiFi	79% (1% to be served by outdoor macro + SC)	79% (1% to be served by outdoor macro + SC)
Spectrum deployment	Max. 100 MHz per site But 4 operators share the whole spectrum	Max. 100 MHz per site But 4 operators share the whole spectrum
Architecture Options	4G / LTE-A Pro	5G NORMA
D-RAN radio stack completely BTS	D-RAN (D2,.2 configuration1)	D-RAN (D2.2 configuration 1)

[9] Wide Area Network (WAN) connecting network clouds

C-RAN @ central offices		C-RAN (D2.2 configuration 2) C-RAN density[10]: (tbd) per sqkm,
HetNet	≤5 SCs per sector @ low and med. frequency bands Indoor SCs[11] >6 GHz incl. M-MIMO	≤5 SCs per sector @ low and med. frequency bands Indoor SCs >6 GHz incl. M-MIMO
Operator scenarios	4 single-operator networks	4 single-tenant networks

C.1.3.2 Multi-tenant evaluation case

This section describes the target architecture for the multi-tenant evaluation case and compiles the most important assumptions.

In the multi-tenant evaluation case the envisaged service as well as the architecture is identical to baseline. Hence from architectural point of view there is no big difference between single and multi-tenant networks. According to different priorities of the tenants even in case of identical services there could be different density of edge clouds and antenna sites.

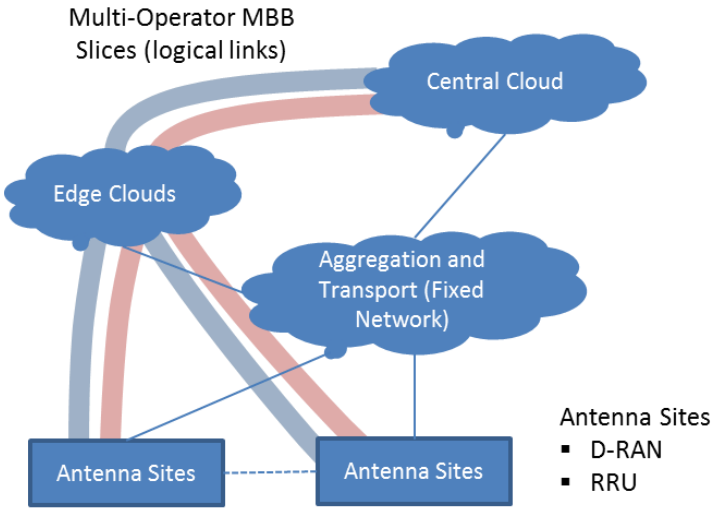


Figure C-6: Target architecture for multi-tenant evaluation

A possible target architecture is depicted in Figure C-6 and most important assumptions are compiled in Table C-2.

Table C-2: Assumptions for multi-tenant evaluation

Business model parameter	**4G / LTE-A Pro**	**5G NORMA**
Services	MBB	MBB
Offer Type	MNO owns all infrastructure	3 with asset assumptions from Section 6.1.2.2
Stakeholder / Customers	2/4 MNOs/ End users	1 mobile service provider, 2/4 RAN infrastructure pro-

[10] Apply C-RANs in dense urban, D-RAN in suburban

[11] Indoor SC in residential buildings assumed to be subscriber owned, in public buildings operator owned

		viders, several cloud infrastructure providers 2/4 tenants / End users
Deployment Scenario Parameters	5G NORMA	5G NORMA
Radio node density	Sites in existing London area + on demand densification up to current max. density	Sites in existing London area + on demand densification up to current max. density
Spectrum efficiency (only relevant for macro cells)	Macros @sub 1GHz and low bands: transitions from 2x2 to 4x4 MIMO Macros @ med bands: transition from 32x2 to 64x4	Macros @sub 1GHz and low bands: transitions from 2x2 to 4x4 MIMO Macros @ med bands: transition from 32x2 to 64x4
Transport network	WAN Network: Fibre Fronthaul: Fibre Backhaul: Fibre	WAN Network: Fibre Fronthaul: Fibre Backhaul: Fibre
MBB requirements (according to use case A2, Enhanced mobile broadband)	Coverage probability: 95% Capacity demand: according to Cisco VNI report Peak data rate: 10 Gbps	Coverage probability: 95% Capacity demand: according to Cisco VNI report Peak data rate: 10 Gbps
Indoor / Outdoor traffic demand	80% / 20%	80% / 20%
Traffic offload by indoor SC	79% (1% to be served by outdoor macro + SC)	79% (1% to be served by outdoor macro + SC)
Spectrum deployment	Max. 100 MHz per site 4 operators share the whole spectrum	Max. 100 MHz per site 4 operators share the whole spectrum and utilise part of the spectrum jointly
Architecture Options	5G NORMA	5G NORMA
D-RAN radio stack completely BTS *C-RAN @ central offices*	D-RAN (D2.2 configuration 1) C-RAN (D2.2 configuration 2) C-RAN density: (tbd) per sqkm,	D-RAN (D2.2 configuration 1) C-RAN (D2.2 configuration 2) C-RAN density: (tbd) per sqkm,
HetNet:	≤5 SCs per sector @ low and med. frequency bands Indoor SCs[12] >6 GHz incl. M-MIMO	≤5 SCs per sector @ low and med. frequency bands Indoor SCs >6 GHz incl. M-MIMO
Operator scenarios	2/4 single-tenant networks	Multi-tenant network

C.1.3.3 Multi-service evaluation case

This section describes the target architecture for the multi-service evaluation case and compiles the most important assumptions.

In the multi-service evaluation case we add services as mMTC and V2X to the baseline MBB service. Most important difference to MBB target architectures is caused by different requirements of the additional services.

mMTC would require improved link budgets especially in uplink in order to achieve similar of better coverage even for low power low cost devices. These link budget improvements can be

[12] Indoor SC in residential buildings assumed to be subscriber owned, in public buildings operator owned

achieved by increasing the system bandwidth and/or densification of the network nodes. In order to make this possible an operator providing combined services like MBB and mMTC would have an increased motivation for small cell deployment as he can combine these small cell equipment with cluster heads operating at low or sub 1 GHz bands.. Alternatively cluster heads could also be installed at street furniture equipped with solar panels and in band backhauled (device networks).

V2I will also raise the requirement for edge clouds directly at the antenna sites (Figure C-7). Edge clouds at antenna sites would promote the introduction of attractive site chain options (e.g. D2.2 configuration 3) including functional splits adopted to available transport technology as well as alleviate deployment of enhanced interference coordination (CoMP, eICIC, SIC,…).

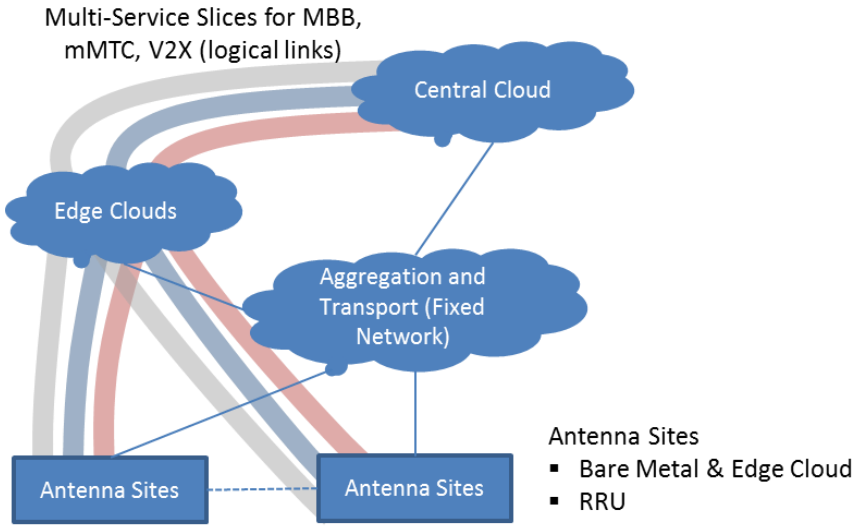


Figure C-7: Target architecture for Multi-service evaluation

In case of D-RAN (bare metal antenna sites) but also in case of operation of synchronous network functions enabling enhanced interference condition like in the baseline case we need direct low latency high capacity links between antenna sites. Most important assumptions are compiled in Table C-3.

Table C-3: Assumptions for multi-service evaluation

Business model parameter	**4G / LTE-A Pro, LTE-V, NB-IoT**	**5G NORMA**
Services	MBB, mMTC, V2X	MBB, mMTC, V2X
Offer Type	MNO owns all infrastructure	1
Stakeholder / Customers	1 MNO / end user	1 MNO, 3 tenants / end users
Deployment Scenario Parameters	4G / LTE-A Pro, LTE-V, NB-IoT	5G NORMA
Radio node density	Sites in existing London area + on demand densification up to current max. density	Sites in existing London area + on demand densification up to current max. density
Spectrum efficiency (only relevant for macro cells)	Macros @sub 1GHz and low bands: transitions from 2x2 to 4x4 MIMO Macros @ med bands: transition from 32x2 to 64x4	Macros @sub 1GHz and low bands: transitions from 2x2 to 4x4 MIMO Macros @ med bands: transition from 32x2 to 64x4
Transport network	Backhaul macro + SC: Fibre Aggregation network: Fibre	WAN Network: Fibre Fronthaul: Fibre

		Backhaul: Fibre
MBB requirements (according to use case A2, Enhanced mobile broadband)	Coverage probability: 95% Capacity demand: according to Cisco VNI report	Coverage probability: 95% Capacity demand: according to Cisco VNI report Peak data rate: 10 Gbps
MTC requirements (according to case A.5, Sensor network monitoring and massive nomadic mobile machine type communications)	Coverage probability 99% 200.000 active sensors per sqkm[13]	Coverage probability 99% 200.000 active sensors per sqkm[14]
V2X requirements (according to use case A4, Vehicle communications)	Latency of <10 ms[15], high reliability and coverage of 99%, position accuracy 0.1-1 m, high connection density	Latency of 1-5 ms[16], high reliability and coverage of 99%, position accuracy 0.1-1 m, high connection density
Indoor / Outdoor traffic demand	80% / 20%	80% / 20%
Traffic offload by indoor SC	79% (1% to be served by outdoor macro + SC)	79% (1% to be served by outdoor macro + SC)
Spectrum deployment	Max. 100 MHz per site 4 operators share the whole spectrum	Max. 100 MHz per site 4 operators share the whole spectrum
Architecture Options	4G / LTE-A Pro, LTE-V, NB-IoT	5G NORMA
D-RAN radio stack completely BTS *C-RAN @ central offices*	D-RAN (D2,.2 configuration1) NB-IoT radio @ BTS LTE-V edge cloud @ BTS	D-RAN (D2.2 configuration 1) C-RAN (D2.2 configuration 2) C-RAN density: (tbd) per sqkm,
HetNet:	≤5 SCs per sector @ low and med. frequency bands Indoor SCs[17] >6 GHz incl. M-MIMO	≤5 SCs per sector @ low and med. frequency bands Indoor SCs[18] >6 GHz incl. M-MIMO
Operator scenarios	3 single-service networks realised by legacy technologies owned by 1 MNO	MNO operates multi-service network for 3 independent tenants, slices are completely operated by MNO on behalf of the tenants

[13] Realised by 3GPP NB-IoT

[14] Realised by 5GN MTC slice

[15] Realised by 3GPP LTE-V

[16] Realised by 5GN V2X slice

[17] Indoor SC in residential buildings assumed to be subscriber owned, in public buildings operator owned

[18] Indoor SC in residential buildings assumed to be subscriber owned, in public buildings operator owned

C.2 Intermediate results

C.2.1 Verification of eMBB performance requirements

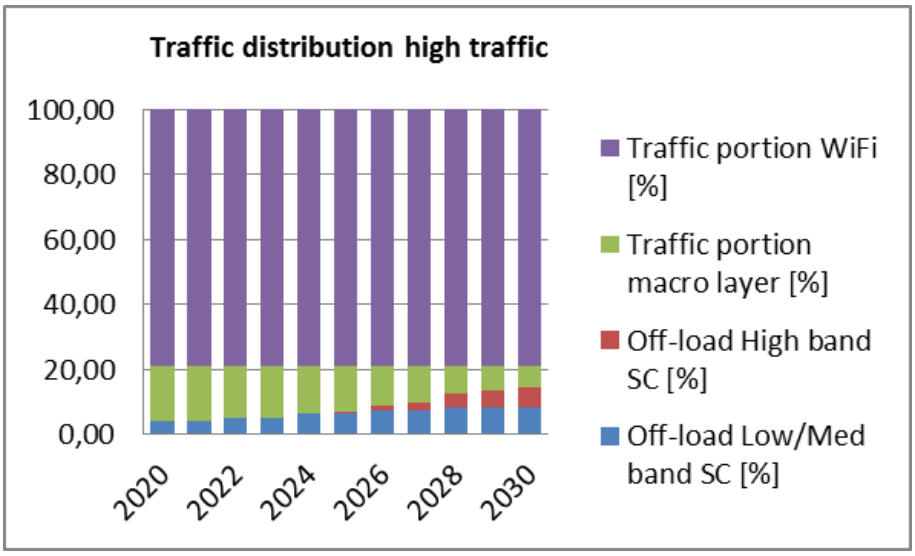


Figure C-8: Traffic distribution 200 m ISD

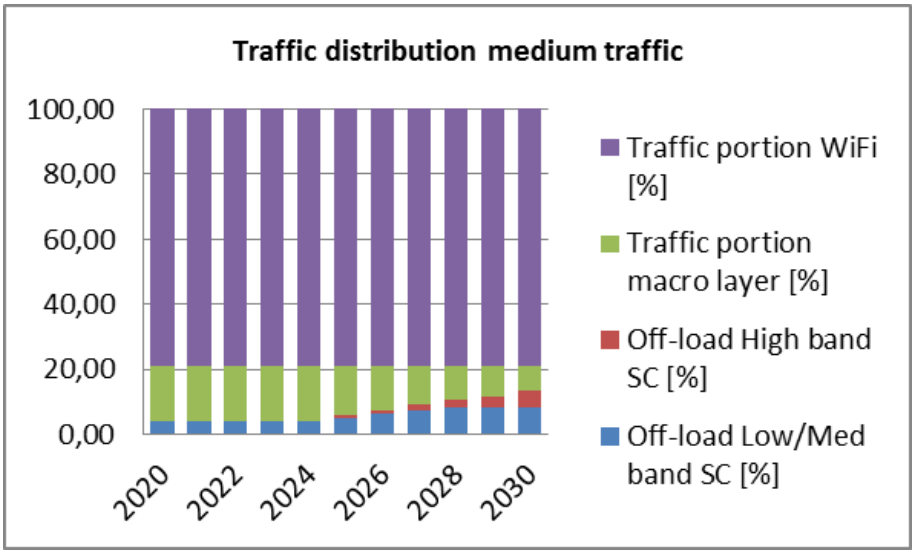


Figure C-9: Traffic distribution 500 m ISD

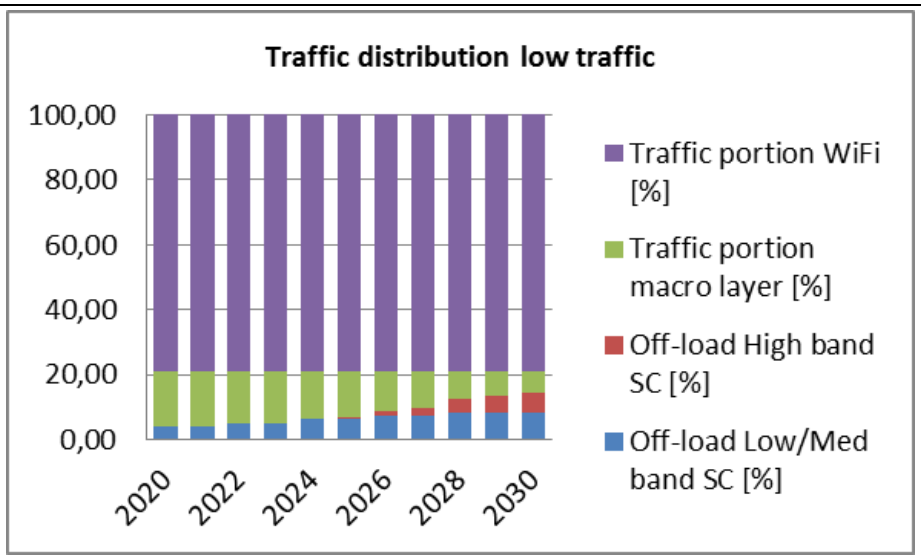


Figure C-10: Traffic distribution 900 m ISD

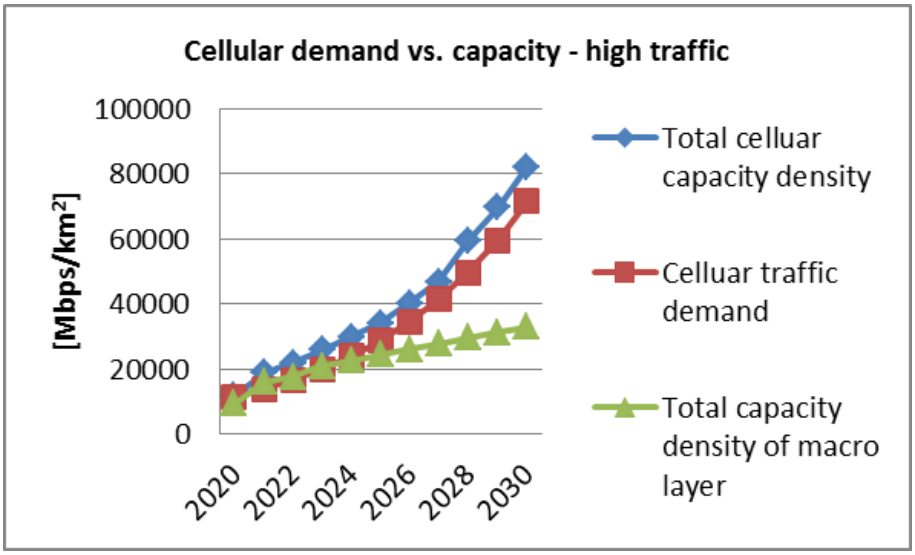


Figure C-11: Throughput demand vs. capacity 200 m ISD

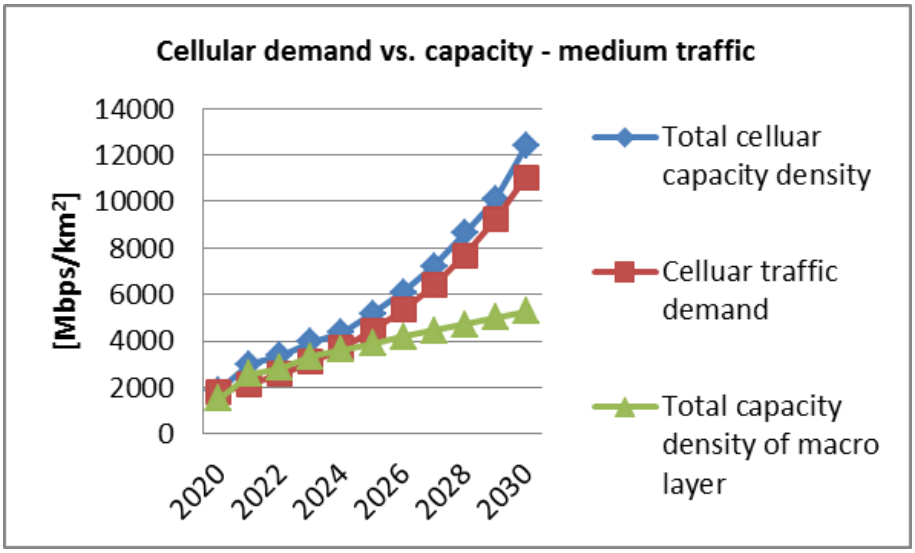


Figure C-12: Throughput demand vs. capacity 500 m ISD

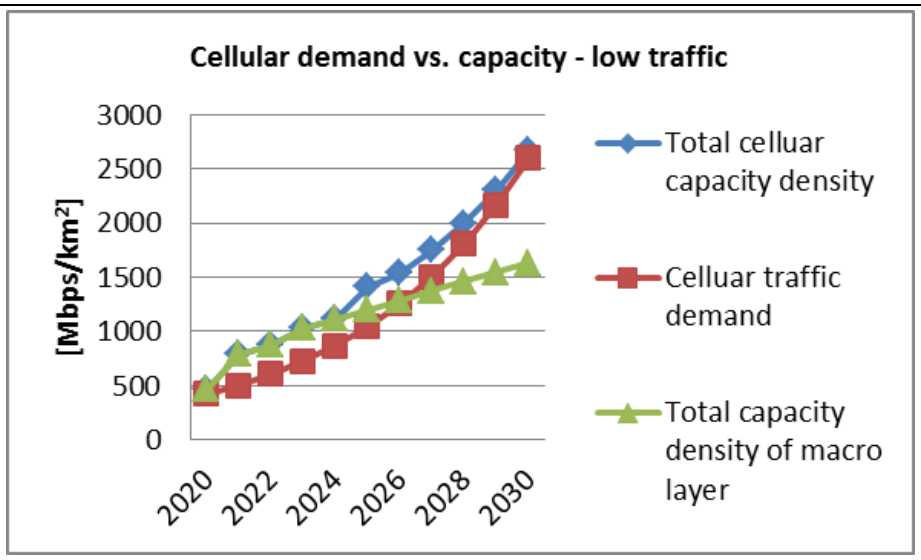


Figure C-13: Throughput demand vs. capacity 900 m ISD

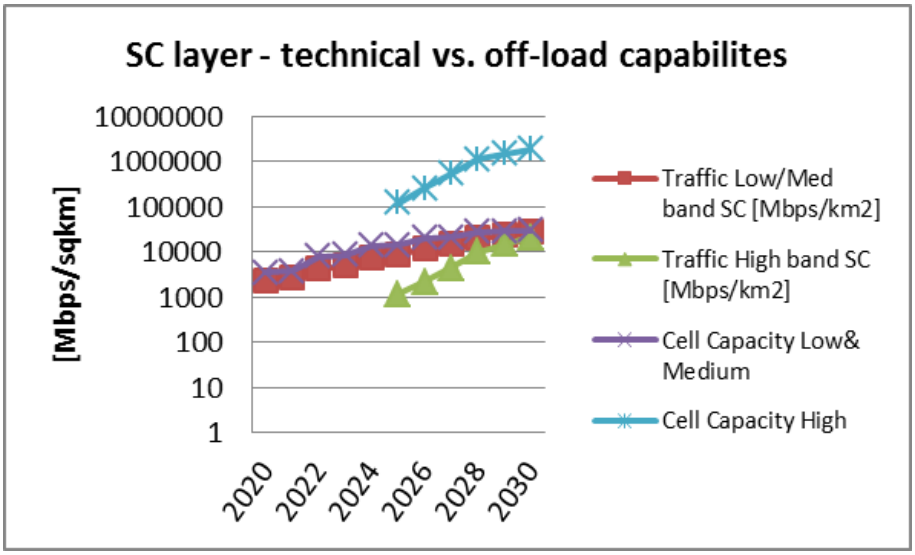


Figure C-14: Technical vs. off-load capabilities 200 m ISD

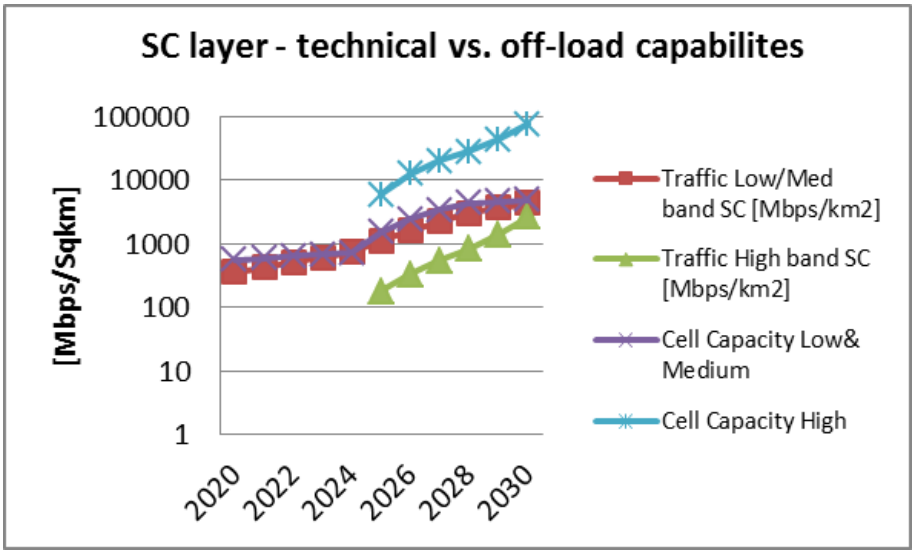


Figure C-15: Technical vs. off-load capabilities 500 m ISD

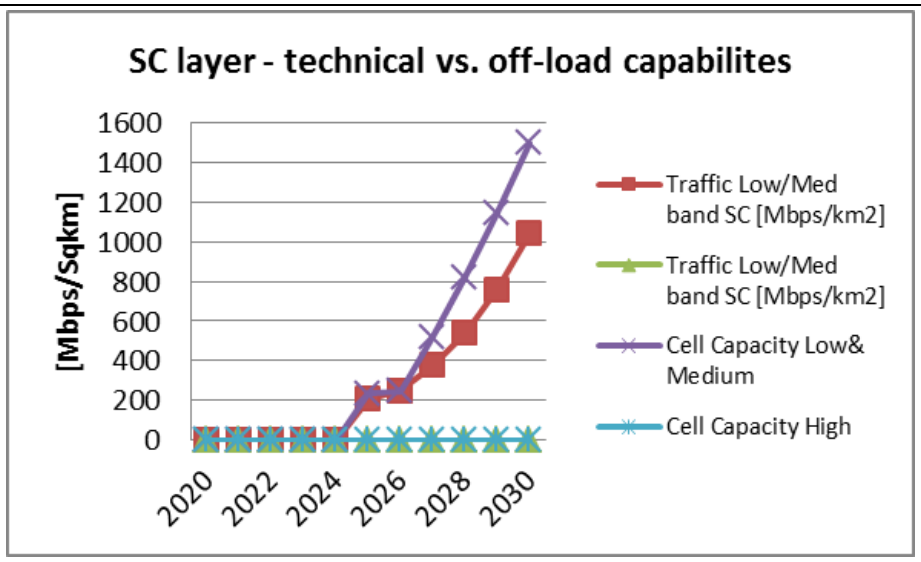


Figure C-16: Technical vs. off-load capabilities 900 m ISD

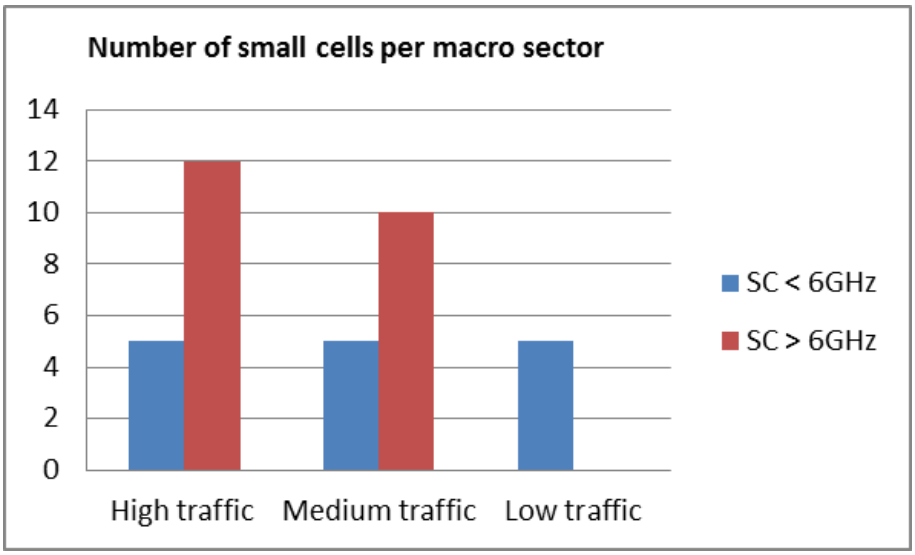


Figure C-17: Need for capacity extension by small cells in the different traffic scenarios

C.2.2 Verification of functional requirements

Preliminary results of functional requirement analysis are compiled in Section 6.2.2. The following tables enclose links to more detailed information.

C.2.2.1 eMBB link table

Requirement	**Details to be found in…**
Application awareness (H – RG#8)	[5GN16-D41], *RAN support for advanced QoS control, page 23* [5GN16-D41], *application aware scheduling, page 106* [5GN16-D41] *QoS-aware 5G PDCP or Uu application protocol, page 107* [5GN16-D51], *QoE aware eICIC, page 114* [5GN16-D51], IETF Distributed mobility management, *page 160*

Multi-layer and multi-RAT connectivity (for TP and coverage) (M – RG#8)	[5GN16-D41, *Multi-Connectivity Functional Architectures, page 49* [5GN16-D41, *Inter-RAT multi-connectivity, page 54*
Efficient backhaul (H – RG#9)	[5GN-D32], *Interface 5GNORMA-SDMC-SDN, Section 4.4*
User privacy and security is required at least at the level provided in LTE, and should be enhanced by options for even better protection (e.g. "IMSI-catching" protection). While security is important for mobile broadband, it is not in the main focus of this use case. (M – RG#11)	[5GN-D32], *Applicability of LTE Security Concepts, Section 5.3*
Capacity for uplink and downlink can be flexibly allocated and optimised on cell and sector level based on just-in-time user requirements and used applications (H – RG#5)	[5GN16-D41], *Centralised RRM for the Virtual Cells, page 153*

C.2.2.2 mMTC link table

Requirement	**Result**
The protocol stack (access/core) should allow the management of a massive number of devices w.r.t. ID management and addressing. (H – RG#7)	[5GN16-D51] *Centralised mobility management @ SDMC, page 34* [5GN16-D41], *User-centric connection area, page 174*
The access network should handle the network resources in C-Plane and U-Plane in a highly efficient way, it should especially only require a low signalling overhead. (H – RG#7, #8)	[5GN16-D41], *User-centric connection area, page 174*
The system should support the use of sensors-type devices with very low cost and long battery lifetime. (H – RG#7, #8)	[5GN-D32], *not in scope of the project, Section 6.2.2.2*
Depending on device type the network access should be applicable via dedicated RATs and frequency bands or in a flexible way. (M – RG#4, #5)	[5GN16-D41], *RAT/Link Selection, page 84*
The mobility management should support stationary, nomadic, and highly mobile devices and should consider also roaming across network boundaries. (H – R#1, #4, #6, #7, #8)	[5GN16-D51] *Selection of Mobility Management scheme, page 130*

Connectivity to a radio network shall be provided directly in case of local radio networks or via relay of other devices in surrounding (gateways, cluster heads). (H – R#3, #4)	[5GN-D32], *not in scope of the project, Section 6.2.2.2*
The network should support direct device-to-device (D2D) connectivity between sensors as well as connection of a sensor to intermediate gateways (cluster heads or aggregators). The local links should be manageable and controlled by the network. (H – RG#1, #4, #6, #7, #8)	[5GN-D32], *not in scope of the project, Section 6.2.2.2*
The system should support both unidirectional as well as bidirectional communication between sensors and other radio nodes. (M – RG#5, #6, #7)	[5GN-D32], *not in scope of the project, Section 6.2.2.2*
The network should provide flexible security and authentication procedures for mMTC as well as means for easy security credential provisioning for massive number of devices. Such security aspects are of high importance for the use case. (H – RG#7, #11)	[5GN-D32], *Multi-service support by tailored radio interface security algorithms, Section 5.4.4.3*

C.2.2.3 V2X link table

Requirement	**Result**
The network should be able to create ad-hoc subnetworks, linking specific nodes and allowing for local access. Ad-hoc networks may support geographical addressing and geographical routing between network elements (i.e., GeoNetworking). (H – RG#1, #3, #4, #5, #6)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*
The system should guarantee the coexistence of safety and non-safety vehicular applications operating over the same scenario. (H – RG#4, #5, #8)	[5GN16-D51] *Network Slice Brokering, page 51*
The system should provide the fast, targeted dissemination of safety messages. (H – RG#4, #5, #7, #8)	[5GN16-D41, *Multi-Connectivity Functional Architectures, page 49*
The system should allow the retrieval of network information to be processed by external applications (e.g., for traffic levels estimation). (M – RG#8)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*

The system should be able to keep track of devices' precise location without incurring in signaling overloads. (H – RG#8)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*
The system should be able to discover the topology of V2V networks established, even if links have been established using non 5G links (e.g., Bluetooth, 802.11p). (L – RG#3, #6)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*
The system should be able to support content discovery for certain types of information (e.g., traffic conditions in the roads). (L – RG#8)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*
The system should enable optimizations for control plane and data plane functions such as optimal routing and handover minimization in the ad-hoc, vehicle supporting subnetworks. (M – RG#3, #6, #8)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*
The system should be able to predict its own reliability against changing traffic conditions or other factors. (H – RG#8)	[5GN-D32], *not in scope of the project, Section 6.2.2.3*
Very high network availability and therefore superior robustness against attacks, in particular DoS attacks, is required. This includes strong authentication between devices and network in order to prevent unauthorised communication. Moreover, integrity protection and encryption is required for the signalling traffic and – unless the applications build on application layer security mechanisms – also for the user plane. Security mechanisms must be robust against loss of network nodes; security mechanisms must be available also in RAN parts that are isolated from central components. Security aspects are of high importance for the use case. (H – RG#11)	[5GN-D32] *Security requirements, Section 6.2.4*
The network should be able to provide intrinsic security mechanisms and protection against attacks. To increase security the network should be able to combine data from different sources for redundancy and consistency checks. (H – RG#11)	[5GN-D32], *not in scope of the project, Section 6.2.4*
The network should be able to cooperate with other existing Vehicular	[5GN-D32], *not in scope of the project, Section 6.2.2.3*

Ad Hoc NETworks (VANETs), e.g., those based on ITSG5/IEEE 802.11p	
The system should be able to expose its reliability prediction to third parties through and open API	[5GN-D32], *not in scope of the project, Section 6.2.2.3*