5G Mobile Network Architecture for diverse services, use cases, and applications in 5G and beyond Deliverable D2.1 Baseline architecture based on 5G-PPP Phase 1 results and gap ana...

Technical Report · October 2017

DOI: 10.13140/RG.2.2.27536.46085

CITATIONS
0

READS
69

1 author:

Sina Khatibi
Nomor Research
24 PUBLICATIONS **87** CITATIONS

SEE PROFILE

H2020-ICT-2016-2
5G-MoNArch
Project No. 761445

5G Mobile Network Architecture

for diverse services, use cases, and applications in 5G and beyond

Deliverable D2.1

Baseline architecture based on 5G-PPP Phase 1 results and gap analysis

Contractual Date of Delivery	2017-10-31
Actual Date of Delivery	2017-11-06
Work Package	WP2 – Flexible and adaptive architecture design
Editor(s)	Marcos Rates Crippa (UNIKL)
Reviewers	Wolfgang Hahn (NOK-DE), Gerhard Kadel (DT)
Dissemination Level	Public
Type	Report
Version	1.0
Total number of pages	84

Abstract: This deliverable delineates a baseline architecture for the 5G-MoNArch project based on the consolidated view coming from the work of the relevant fora, consortia, SDOs (such as, 3GPP and ETSI), 5G PPP Phase 1 projects along with 5G PPP WGs. After defining this baseline architecture, all the 5G system gaps that need to be addressed by the innovations proposed by the project with a particular focus on E2E network slicing are identified. These innovations allow the baseline 5G-MoNArch architecture to support diverse service requirements and enable new business models.

Keywords: 5G Network Architecture, 5G System Gap Analysis, Baseline Architecture, E2E Network Slicing

Executive Summary

The development of the fifth generation (5G) mobile networks is advancing quickly, with research projects and standardisation efforts working on defining the main elements of the 5G architecture. Research projects funded by the European Commission (EC) and running under the auspices of Phase I of the 5G infrastructure Public Private Partnership (5G PPP) have played an important role in this process. There are, however, gaps in the current baseline consensus 5G architecture to be filled by innovations supporting diverse service requirements and enabling new business models.

This deliverable delineates a baseline architecture for the 5G-MoNArch project based on the current relevant state of the art. Relevant state of the art here refers to a consolidated view coming from the work of the most relevant fora, consortia, standards developing organisations (SDOs) such as 3GPP and ETSI, 5G PPP Phase 1 projects along with 5G PPP working groups (WGs).

After defining this baseline architecture, a gap analysis is performed. The gaps that need to be address by the 5G-MoNArch project innovations are identified. 5G-MoNArch proposes multiple key innovations: three enabling innovations contributing to the baseline architecture, and two functional innovations which define characteristics and features of specific network slices. The enabling innovations support the operation of network sliced 5G networks, while the functional innovations are specific functions required when deploying network slices with particular requirements (in this case resilience and security, as well as resource elasticity).

This deliverable, together with 5G-MoNArch deliverable D6.1 ‘Documentation of requirements and KPIs and definition of suitable evaluation criteria’, provides the first baseline architecture and architectural requirements for 5G-MoNArch. The next steps will be defining and extending all architectural elements, concepts and components, aiming at having a 5G-MoNArch initial architecture for Deliverable D2.2 ‘Initial overall architecture and concepts for enabling innovations’.

List of Authors

Partner	Name	E-mail
NOK-DE	Christian Mannweiler	christian.mannweiler@nokia-bell-labs.com
	Diomidis Michalopoulos	diomidis.michalopoulos@nokia-bell-labs.com
	Borislava Gajic	borislava.gajic@nokia-bell-labs.com
UC3M	Albert Banchs	banchs@it.uc3m.es
	Marco Gramaglia	mgramagl@it.uc3m.es
DT	Markus Breitbach	m.breitbach@telekom.de
	Gerd Zimmermann	zimmermanng@telekom.de
NOK-FR	Aravinthan Gopalasingham	gopalasingham.aravinthan@nokia-bell-labs.com
	Bessem Sayadi	bessem.sayadi@nokia-bell-labs.com
HWDU	Ömer Bulakci	oemer.bulakci@huawei.com
	Qing Wei	qing.wei@huawei.com
	Riccardo Trivisonno	riccardo.trivisonno@huawei.com
	Panagiotis Spapis	panagiotis.spapis@huawei.com
	Emmanouil Pateromichelakis	emmanouil.pateromichelakis@huawei.com
TIM	Fabrizio Moggio	fabrizio.moggio@telecomitalia.it
	Andrea Buldorini	andrea.buldorini@telecomitalia.it
SRUK	Mehrdad Shariat	m.shariat@samsung.com
	David Gutierrez Estevez	d.estevez@samsung.com
ATOS	Beatriz Gallego-Nicasio Crespo	beatriz.gallego-nicasio@atos.net
	Jose Enrique González	josee.gonzalez@atos.net
	Joanna Bednarz	joanna.bednarz@atos.net
CEA	Antonio De Domenico	antonio.de-domenico@cea.fr
	Nicola Di Pietro	nicola.dipietro@cea.fr
CERTH	Anastasios Drosou	drosou@iti.gr
	Athanasios Tsakiris	atsakir@iti.gr
MBCS	Dimitris Tsolkas	dtsolkas@mobics.gr
	Odysseas Sekkas	sekkas@mobics.gr
RW	Simon Fletcher	simon.fletcher@realwireless.biz
NOMOR	Kunjan Shah	shah@nomor.de
	Sina Khatibi	khatibi@nomor.de
UNIKL	Marcos Rates Crippa	crippa@eit.uni-kl.de

Revision History

Revision	Date	Issued by	Description
0.1	01.07.2017	5G-MoNArch WP2	Initial draft
1.0	06.11.2017	5G-MoNArch WP2	Final version for delivery

List of Acronyms and Abbreviations

2G	2nd Generation mobile wireless communication system (GSM, GPRS, EDGE)
3G	3rd Generation mobile wireless communication system (UMTS, HSPA)
3GPP	3rd Generation Partnership Project
4G	4th Generation mobile wireless communication system (LTE, LTE-A)
5G	5th Generation mobile wireless communication system
5GS	5G System
5G-PPP	5G infrastructure Public Private Partnership
AAA	Authentication, Authorisation and Accounting
AaSE	AIV agnostic Slice Enabler
AIV	Air Interface Variant
AMF	Access and Mobility management Function
BBU	Base Band Unit
CAPEX	CAPital Expenditure
CCNF	Common Control Network Functions
CN	Core Network
CP	Control Plane
CSC	Communication Service Customer
CSI	Channel State Information
CSMF	Communication Service Management Function
CSP	Communication Service Provider
CU	Central Unit
DC	Data Centre
DCSP	Data Centre Service Provider
DRB	Data Radio Bearer
DU	Distributed Unit
eICIC	enhanced Inter-cell Interference Coordination
eMBB	enhanced Mobile Broadband
feD2D	further enhanced D2D
GHO	Group Handover
gNB	NR NodeB
HARQ	Hybrid Automatic Repeat Request
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
ISRB	Inter-slice Resource Broker
KPI	Key Performance Indicator
LTE	Long Term Evolution
MAC	Medium Access Control
MANO	management and orchestration
MCS	Modulation Coding Scheme
MME	Mobility Management Entity
mMTC	Massive Machine Type Communication
MOCN	Multi-Operator Core Network
MORAN	Mobile Operator Radio Access Network
NAS	Non-Access Stratum
NBI	Northbound Interface
NE	Network Element
NEP	Network Equipment Provider
NF	Network Function
NFV	Network Function Virtualisation

NFVO	Network Function Virtualisation Orchestrator
NGMN	Next Generation Mobile Networks
NOP	Network Operator
NRM	Network Resource Model
NS	Network Service
NSI	Network Slice Instance
NSMF	Network Slice Management Function
NSSAI	Network Slice Selection Assistance Information
NSSF	Network Slice Selection Function
NSSI	Network Slice Subnet Instance
NSSMF	Network Slice Subnet Management Function
NST	Network Slice Template
NWDA	Network Data Analytics
OPEX	OPerational Expenditure
PAN	Personal Area Network
PDCP	Packet Data Convergence Protocol
PGW	Packet Data network Gateway
PHY	Physical Layer
PLMN	Public Land Mobile Network
PNF	Physical Network Function
QoE	Quality of Experience
QoS	Quality of Service
RA	Registration Area
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RLC	Radio Link Control
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio resource Management
RTT	Round Trip Time
SDAP	Data Adaptation Protocol
SDM-O	Software Defined Mobile Network Orchestrator
SDO	Standards Developing Organisation
SDSF	Structured Data Storage network Function
SFC	Service Function Chain
SGW	Serving Gateway
SMF	Session Management Function
TAU	Tracking Area Update
UDSF	Unstructured Data Storage network Function
UE	User Equipment
UP	User Plane
UPF	User Plane Function
VIM	Virtual Infrastructure Manager
VISP	Virtual Infrastructure Service Provider
VNF	Virtual Network Function
VNFM	Virtual Network Function Manager

Table of Contents

1 Introduction

Since the early the fifth generation (5G) research phase starting in 2012, the development of concepts for the 5G system (5GS) has progressed at a rapid pace. Both research projects and standardization efforts have described the main elements of the 5G architecture. Third generation partnership project (3GPP) has already set the completion of the first "non-standalone" release of 5G till the end of 2017. To this end, European Union (EU) funded 5G infrastructure Public Private Partnership (5G PPP) Phase 1 projects[1] have played an important role in establishing consensus and providing various technologies and innovations to the standards developing organizations (SDOs). 5G Architecture Working Group (WG) established by the 5G PPP Phase 1 projects has provided a consolidated output and view on the overall architecture [5GARCH16-WPv2]. Although all these efforts have provided a solid baseline architecture, there are still 5G system gaps that can be filled by innovations to better fulfil the 5G vision of supporting diverse service requirements and enabling new business sectors often referred to as vertical industries. Further, end-to-end (E2E) network slicing spanning over network domains (e.g., core network, CN, and radio access network, RAN) where multiple logical networks corresponding to different business operations are sharing a common infrastructure, is seen as the fundamental pillar of the 5GS. Accordingly, for the 5GS to fulfil its promises, the envisioned innovations shall enable a native E2E network slicing support. This is one of the main goals of the 5G-MoNArch project.

On this basis, this first deliverable for the Work Package (WP) 2 of the 5G-MoNArch project aims to provide a baseline architecture, to be further improved and detailed in future deliverables. To achieve this goal, this document will engage in two main tasks, each one with its specific chapter, as illustrated in Figure 1-1.

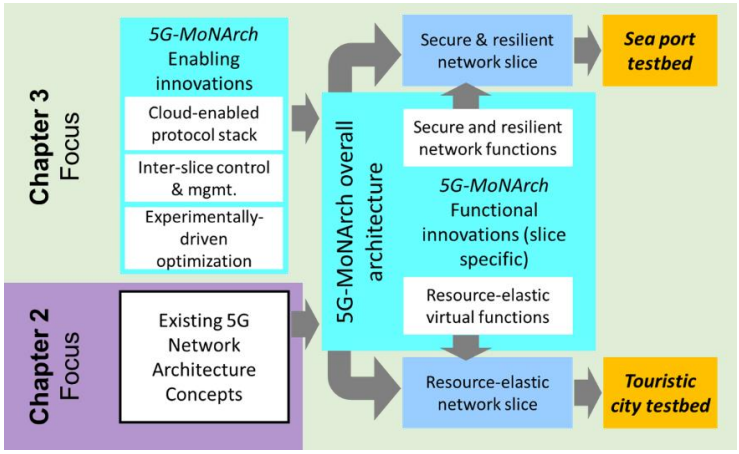


Figure 1-1: 5G-MoNArch approach with building blocks and innovations; Chapter 2 places the focus on the description of the baseline architecture taking into account the most relevant state-of-the-art (SotA) concepts, while Chapter 3 highlights the 5G-MoNArch innovations that will address the identified 5GS gaps along with the brief summary of target testbeds

First, the identification and summary of the key outcomes and architectural commonalities needed to support a virtualised and flexibly managed multi-service, multi-tenancy network, building on key results from all relevant 5G PPP Phase 1 projects (e.g., 5G-NORMA and METIS-II) and other fora like SDOs (e.g., 3GPP RAN/SA and ETSI NFV/MEC) is performed. The most relevant aspects coming from this identification represents the fundamental concepts and components of the 5G-MoNArch architecture as outlined in Chapter 2.

Following that, a 5GS gap analysis is performed around this baseline architecture, identifying where it cannot meet the 5G objectives. Furthermore, 5G-MoNArch innovations are analysed and it is shown how they will address those gaps. This gap analysis associated with the 5G-MoNArch innovation mapping is highlighted in Chapter 3. Some final remarks and future steps are outlined in the conclusion in Chapter 4.

[1] 5G PPP Phase 1 Projects - https://5g-ppp.eu/5g-ppp-phase-1-projects/

2 Fundamental Concepts and Components of 5G-MoNArch Architecture based on State-of-the-Art

In accordance with the motivation given in the Introduction, this chapter outlines the most essential consolidated architecture descriptions coming from the most relevant state-of-the-art (SotA) fora, consortia, and SDOs, such as, 3GPP, ETSI, and 5G PPP Phase 1 projects along with 5G PPP WGs, and highlights the most relevant architectural features that construct the 5G-MoNArch baseline architecture. Chapter 3 then analyses the 5GS gaps in this baseline architecture and puts the 5G-MoNArch innovations forward, which will address these gaps. Accordingly, D2.1 establishes the basis toward 5G-MoNArch initial architecture that will be captured in Deliverable D2.2 [2]. It is worth noting that the 5G-MoNArch innovations and further work can result in modifications and optimisations on the baseline architecture captured herein.

The structure of this chapter is as follows. Section 2.1 provides the high-level 5G-MoNArch preliminary architecture including the main consolidated outcomes from the SotA. In Section 2.2 and Section 2.3, functional architectures of CN and RAN are presented, while the centralised control layer is discussed in Section 2.4. Section 2.5 details the Management and Orchestration (MANO) architecture framework and inter-relates the 5G-MoNArch baseline architecture with the current standardisation progress. Section 2.6 depicts the physical network infrastructures and topologies considering the functional architecture descriptions given in the previous sections. Finally, Section 2.7 presents a concise summary of the inter-relations of technical areas provided in Sections 2.2 - 2.6 w.r.t. the high-level architecture described in Section 2.1.

2.1 High-level architecture

The system architecture for 5G networks shall incorporate the performance and flexibility to support multiple telecommunications services, with heterogeneous key performance indicators (KPIs) and sharing the same infrastructure. Further, 5G shall give operators unique opportunities to address and offer new business models to consumers, enterprises, verticals, and third-party tenants. To this end, Section 2.1.1 presents the architecture requirements and vision from NGMN, which sets the basis for the network slicing framework presented in Section 2.1.2. Section 2.1.3 consolidates the mobile network architecture vision from the 5G PPP Phase 1 projects, whose synergy builds the basis for the preliminary 5G-MoNArch reference functional architecture given in Section 2.1.4. This reference architecture is the starting point for further architectural extensions based on 5G-MoNArch innovations. Section 2.1.5 briefly provides the envisioned communications ecosystem including new business roles and relations.

2.1.1 NGMN 5G architecture requirements and vision

In an early phase of 5G work, [NGMN15] has outlined the basic requirements and principles that the future architecture [NGMN17] shall fulfil from an operator's perspective. For this purpose, a high-level architecture vision has been given that spans over different layers in a vertical view and crosses different domains in a horizontal view. Particularly, the architecture shall enable the deployment of multiple logical networks ("network slices"), which will be discussed in detail in Section 2.1.2 on top of a (mostly) shared infrastructure and the associated resources, cf. Figure 2-1.

Vertical view of the NGMN reference architecture

[NGMN15] envisions network slice to cover **business application layer, business enablement layer, infrastructure resource layer**. The infrastructure resource layer consists of the physical resources of a fixed-mobile converged network, comprising access nodes, cloud nodes (which can be processing or storage resources), 5G devices (in the form of (smart) phones, wearables, Customer Premise Equipment (CPE), machine type modules and others), networking nodes and associated links. The business enablement layer is a library of all functions required within a converged network in the form of modular architecture building blocks, including functions realised by software modules that can be retrieved from the repository to the desired location, and a set of configuration parameters for certain parts of the network, e.g., radio access. The business application layer contains specific applications and services of

[2] 5G-MoNArch Deliverable D2.2 "Initial overall architecture and concepts for enabling innovations," June 2018.

the operator, enterprise, verticals or third parties that utilise the 5G system. The **end-to-end (E2E) management and orchestration entity** comprises the set of functions that operate and maintain the three layers, including the technologies and application programming interfaces (APIs) interconnecting them, e.g., the APIs to translate the use cases and business models into actual network functions and slices.

Horizontal view of the architecture

In a horizontal view, the E2E slice crosses multiple "domains". These domains can be management domain from different network operators/different service providers, different infrastructure segments like RAN, Transport Network (TN) and CN, or control domains for different technologies (e.g., optical/wireless transport, hardware/software based network functions) or control areas.

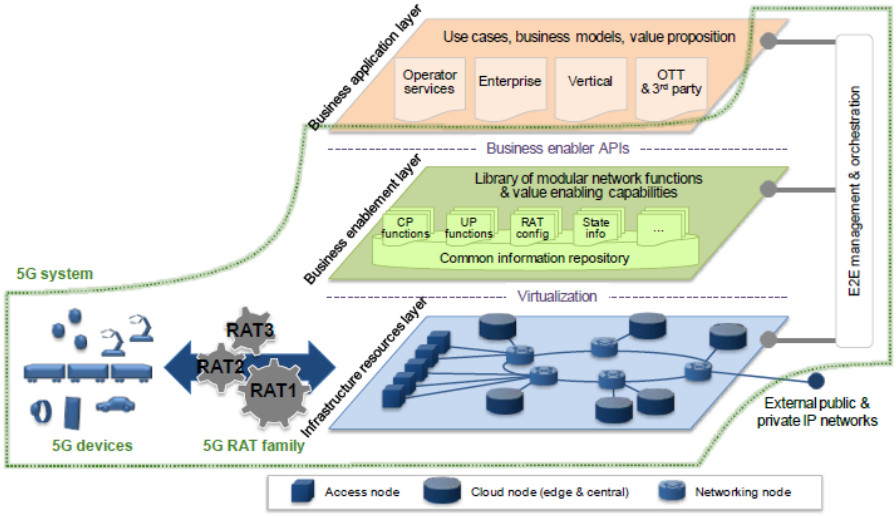


Figure 2-1: 5G architecture [NGMN]

2.1.2 E2E Network Slicing

The need for performance and functional flexibility, as highlighted by NGMN requirements analysis, has surely been the driver for the definition of all 5G embryonal architectural concepts. All in all, and referring to the horizontal view of the architecture as per Section 2.1.1, the majority of proposals feature the ability to enable the dynamic instantiation of tailored Control Plane (CP) and User Plane (UP) functions. In short, 5G (in its mature form) will not have a single network architecture (as it was defined e.g. for 4G systems) but will enable the definition of different logical architectures, built upon a set of basic logical functions, tailored to target requirements of groups of homogeneous use cases.

In parallel, the combination of architecture flexibility with different use cases associated to different business segments has led to the definition of Network Slice. In [NGMN15], a network slice has been defined as a composition of network functions and specific Radio Access Technology (RAT) settings, combined for a specific use case or business model. The slice can span all domains of the network: software modules running on cloud nodes, specific configurations of the transport network, dedicated radio configurations or even a specific RAT, as well as the end devices.

Refining the network slice definition is highly controversial, as it might have significant standardisation, design, and operational impacts. The concept has already been elaborated in some relevant prior art. "Network slicing" applied to RAN has been presented in [Kokku et al], where its introduction mainly aims at enabling spectrum sharing among Mobile Virtual Network Operators (MVNOs) while minimising the impacts on Access Nodes design. Hence, the ultimate goal was optimising the overall radio resource utilisation. Targeting the same goal of efficient sharing of network infrastructure among operators, [Caballero et al] elaborates the slicing concept presenting a multi-tenant slice controller for efficient active RAN sharing. The paper discusses operational aspects relating to slicing. With analogous intentions, but focusing on a different segment of the system, authors of [Nguyen et al] designed a novel approach for CN slicing, to share resources according to traffic demand and to reduce capital

expenditures (CAPEX) and operational expenditures (OPEX). An early concept of network slice is also hidden in 3GPP Décor [3GPP TR 23.707], where dedicated 4G CNs are conceived to meet functional requirements of different set of services. By defining dedicated CN elements (e.g. MME) for different services, Décor implicitly partitions the CN into slices, implemented on dedicated and isolated hardware and handling different services.

Despite of the existence of wide and heterogeneous prior art, within 5G scope network slicing has still several variables to be clarified and set. However, it seems agreed by the majority the concept of network slice will apply "E2E", i.e. a Network Slice shall be defined as the instantiation (over a software-defined or physical infrastructure) of a tailored architecture made by a set of interconnected logical Access and Core Network functions, relating to both CP and UP as well as network management, to support a particular set of use cases. **The rationale behind this "E2E" choice seems to be straightforward:** From 5G use case analysis a fact clearly emerges, namely, functional and performance requirements are defined from an E2E perspective**.** For this reason, the architecture design will benefit from a holistic view that implies the genesis of the **E2E network slicing concept**. In particular, tailored E2E architecture consists of a set of logical functions and related interfaces, by the composition of which tailored CP and UP architectures are defined.

The definition of the E2E slicing concept unravels a set of complex issues to be addressed, including E2E slices **design, instantiation,** and **operation**. Designing a slice for a specific use case requires the definition of CP and UP architecture, procedures and protocols upon the basic set of functions, both on access and core network side. Instantiating a slice deals with mechanisms for its implementation and deployment over the available infrastructure including TN, Data Centre (DC) etc., fulfilling potential isolation requirements. Finally, operating slices requires mechanisms for slice monitoring and reconfiguration.

2.1.3 5G-PPP overall 5G reference architecture

By integrating the architecture view of 5G-PPP Phase 1 projects, [Arch WG WP] the NGMN architecture vision will be further developed by incorporating the view from the industry as well as mapping into real world technology components (e.g., Software-Defined Networking (SDN)/Network Function Virtualisation (NFV) and security framework). Three vertical levels are defined, namely: **Service Level, Network Level, and Resources & Functional Level**, cf. Figure 2-2. Like the NGMN view, there is one **Network and Service Management** entity interacting with all the vertical levels. Different network slices are represented at the **Network Level** with different sets of interconnected network functions (NFs). Such logical representation is further mapped to hardware/software-based infrastructure resources residing in the so-called **Resources & Functional Level** according to E2E service requirements of each individual slice. The 5G-PPP overall architecture provides a converged architecture perspective of the 5G-PPP Phase 1 projects. A special focus has been on network slicing that provides the framework for supporting the demand of vertical industries in a business-driven way. Accordingly, network slicing is in the core of the envisioned overall architecture. Furthermore, the overall architecture natively includes reliable security mechanisms to enable customised network slice instances running on a common infrastructure and to adhere by the requirements of new business models requiring multi-party trust relations.

At the **service level**, the slice provider offers a northbound interface (NBI) to tenants where they can request/modify/monitor/control their slices according to the service-level agreements (SLAs) with the slice provider. The needs of different businesses are captured by SLAs that may necessitate different instantiations of the NFs associated with network slice instances.

At the **network level**, this implies that such on-demand NFs [3GPP TR 23.799] can be tailor-made for different devices based the level of support needed. Particularly, user terminals as part of this E2E chain can play a more focal role by providing location, capabilities, and statistics to enable the network to optimise NFs (e.g., 3GPP SA2 AMF). This is in also line with transitions to further enhance Device-to-Device (D2D) Communications [RP-150441]. The example in Figure 2-2 depicts two example network slice instances tailored for automotive and Internet of Things (IoT) vertical sectors, where 3GPP defined NFs are utilised for illustration. To support the mission-critical and low latency services in the

automotive slice, NFs are rather instantiated at the edge cloud, while NFs can be instantiated in the central cloud for the IoT slice enabling more cost-efficient implementation.

The **network operating system** maps logical network slices to the "resources and functional" level, e.g., using various (programmable) controllers, but also virtualisation techniques.

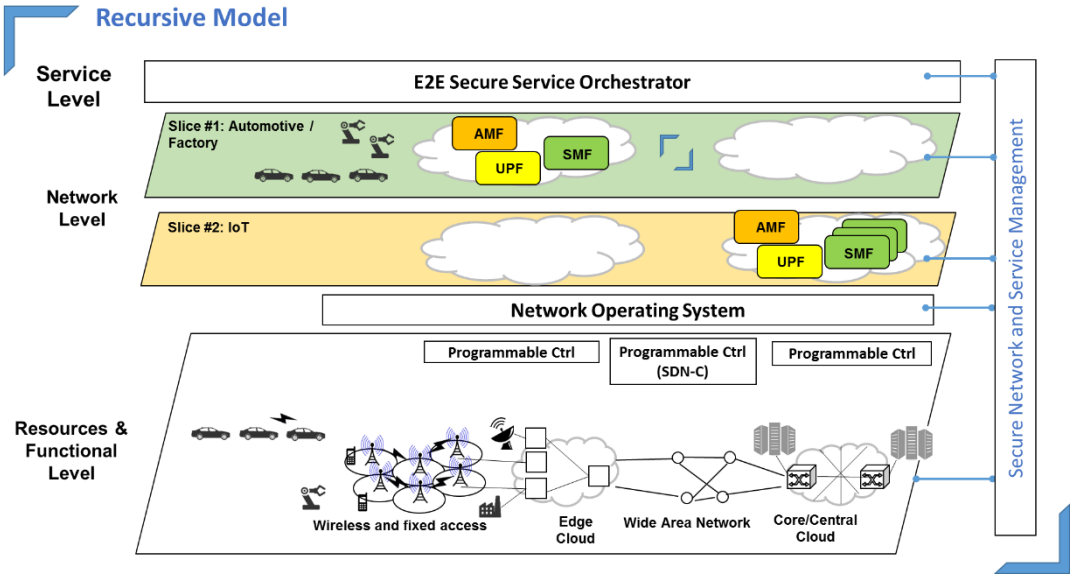


Figure 2-2: Overall 5G architecture identified by 5G-PPP Phase 1 projects [5GARCH16-WPv2]

In addition, the recursive model, as marked in Figure 2-2, implies that the envisioned architecture framework can be applied to different network slices considering their requirements; hence, this illustration itself does not cover all possible implementations.

Generally, it is expected that a multitude of network slices will need to be supported, where customer needs can vary, as well. Therefore, the network management and control framework needs to be highly automated and apply cognitive methods for the entire fault, configuration, accounting, performance, security (FCAPS) and lifecycle management of network slice instances. The preliminary 5G-MoNArch functional architecture is based on this vision, cf. Section 2.1.4.

2.1.4 Preliminary 5G-MoNArch Functional Reference Architecture

Incorporating the results from 5G-PPP phase 1 projects and the 5G requirements initially defined in [NGMN15], the preliminary high-level functional view of the 5G-MoNArch architecture is designed in a modular manner and depicts four layers. For each of these layers, it 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. The high-level functional view of the system architecture is depicted in Figure 2-3. It shows the separation into four layers as well as the differentiation into intra-network-slice and inter-network-slice functions.

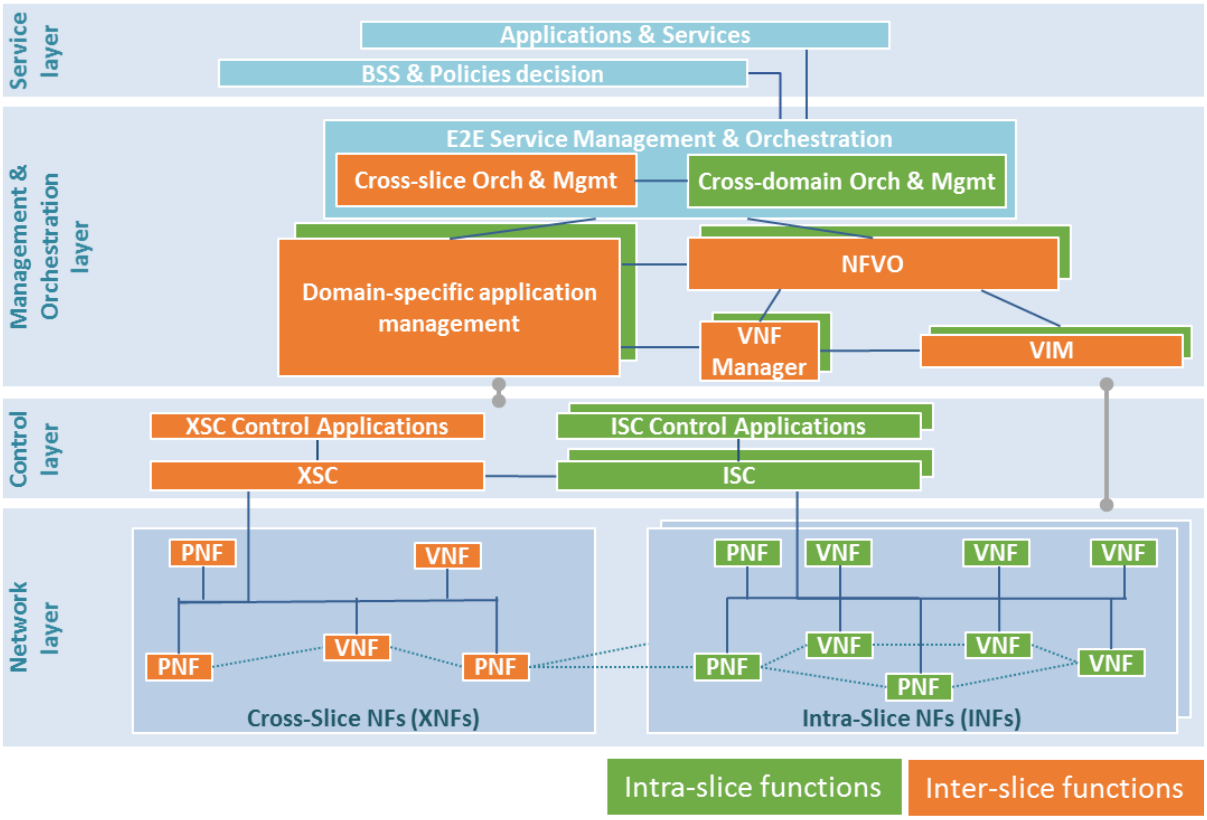


Figure 2-3: Preliminary high-level functional view of overall 5G-MoNArch architecture (adapted from [5GN-D3.3])

The **Service Layer** comprises Business Support Systems and business-level Policy and Decision functions as well as applications and services operated by a tenant or other external entities. These functions of the Service Plane interact with the Management & Orchestration Plane via the Service Management function, see below.

The **Management & Orchestration (MANO) Layer** extends the ETSI NFV management and orchestration architecture towards multi-tenant and multi-service networks. It therefore comprises the Virtual Infrastructure Manager (VIM), the VNF Manager (VNFM) and the NFV Orchestrator. Further, the layer accommodates application management functions from various management domains, e.g., different operator domains, edge/central cloud, (R)AN, CN, and TN. In the case of telecommunications network management, this can comprise Element Managers (EM) and Network Management (NM) functions or their equivalents in 5G systems. Such functions would also implement ETSI NFV MANO reference points to the VNFM and the NFVO. The E2E service Management & Orchestration includes Service Management function and Orchestration functions. The Service Management is an intermediary function between the Service Plane and the Management & Orchestration Plane. It transforms consumer-facing service descriptions into resource-facing service descriptions (and vice versa). The orchestration function includes both Cross-slice orchestration and management (Orch & Mgmt) function and Cross-domain Orch & Mgmt function. The Cross-slice Orch & Mgmt function is responsible for inter-slice management (e.g., common context between different slices/tenants, Inter-Slice Resource Broker (ISRB) which determines and enforces policies for cross-slice resource allocation, particularly in the case of shared network functions, etc.) Cross-domain Orch & Mgmt function is taking care of the coordination/negotiation between different management domains for a single slice.

The depicted 5G-MoNArch preliminary architecture focuses primarily on the use of hypervisors (virtual machines or VMs) as an implementation technology for NFV. Nevertheless, the rise of lightweight container [Plauth et al] technologies provided by solutions, such as, Docker, has started to influence the datacentre landscape, with their ease of deployment, lower costs, and shorter development times, as well as faster instantiation and migration. To cope with and benefit from such advancement in NFV technologies, 5G-MoNArch aims to extend the current MANO layer to be container complaint.

The **Control Layer** accommodates the two main controllers: (1) the Cross-slice Controller (XSC) for the control of Cross-slice Network Function (XNFs) that are shared by multiple network slices (depicted in orange) and (2) Intra-slice Controller (ISC) for Intra-slice Network Function (INFs) that are dedicated to individual network slice (depicted in green). Following the SDN principles, XSC and ISC abstract from the technological and implementation-related details of controlled network functions. They translate decisions of the northbound control applications into commands towards southbound Virtualised NFs (VNFs) and Physical NFs (PNFs) in both User and Control Planes.

Finally, the **Network Layer** comprises the VNFs and PNFs needed to carry and process the user data traffic. Such VNFs and PNFs can be either the control plane network functions (e.g., AMF, SMF, MME, and AAA) or user plane network functions (e.g., user plane function - UPF, serving/packet data network gateway - S/PGW, and router). Details of different CP/UP functions are explained in Section 2.2.

It is worth mentioning that the interfaces depicted between different layers will be further defined within 5G-MoNArch future work.

Moreover, Figure 2-3 implicitly illustrates **three fundamental design aspects** that shall be followed in the 5G-MoNArch architecture:

(1) Split of control and user plane

5G-MoNArch applies a consistent split of control plane and user plane throughout all network domains, including RAN, CN, and transport network. Among others, this allows for hosting associated control and user plane functions in different locations and facilitates to aggregate control and user plane functions differently.

(2) Support for E2E network slicing

The architecture allows for different levels of slicing support across MANO, control, and user plane. The first supported option includes slice-specific functions, i.e., each slice incorporates a dedicated and possibly customised function that is not shared with others. The second option includes the possibility to operate functions (or function instances) that are shared by multiple slices and have the capability to address requirements from multiple slices in parallel. Figure 2-3 depicts this split into common or so-called inter-slice functions and dedicated (intra-slice) functions. This split is maintained from the MANO Plane down to the User Plane, i.e., dedicated NFs are controlled and managed by the tenant's own instance of ISC and MANO Plane functions (i.e., ETSI NFV functions as well as domain-specific application management functions). Shared functions are controlled and managed by the XSC as well as the necessary MANO Plane functions, usually operated by the Mobile Network Operator (MNO) or the Mobile Service Provider (MSP). The policies regarding the utilisation of shared functions, particularly the resource allocation to active slices, are determined e.g., by the ISRB, and communicated towards the respective control, management and orchestration functions for further enforcement. Finally, the third option is to not only have slice-dedicated NFs but to additionally assign the associated infrastructure resources (HW), including spectrum, exclusively to a single slice.

(3) Network programmability

The concept software-defined networking (SDN) splits between logic and agent for any functionality in the network. In the context of 5G-MoNArch the applicability of this paradigm with the concepts of elasticity and resiliency will be studied. This means that the network functions are split into the decision logic hosted in the controller application and the NF that executes the decision. The controllers, either ISC or XSC, reside "between" application and NF and abstracts from specific technologies and implementations realised by the NF, thus decoupling the controller applications from the controlled NF. In a more general sense, the network is programmable based on the decision from management plane/control application. This enables the network to flexibly adjust its behaviour according to the requirements of various use cases in high dynamic environment.

According to the above discussion, a summarising table (Table 2-1) of the main concept/components of our architecture is provided.

Table 2-1: Summary of the main concept/components of the 5G-MoNArch architecture

Long Name	Short Name	Description
Intra-slice application	I-APP	Centralised control or management function for one specific slice (Runs on top of ISC NBI)
Cross-slice application	X-APP	Centralised shared control or management function cross multiple slices (Runs on top of XSC NBI)
Intra-slice network function	INF	Network function used by one specific slice
Cross-slice network function	XNF	Network function used by multiple slices
Intra-slice controller	ISC	Software Defined Controller for Intra-slice functions
Cross-slice controller	XSC	Software Defined Controller for Inter-slice functions
Cross-domain Orch & Mgmt function		Orchestration and management function for one slice cross multiple domains
Cross-slice Orch & Mgmt function		Orchestration and management function cross multiple slices

2.1.5 High Level Stakeholder model

Throughout this document consideration of users of the system will necessitate the utilisation of the stakeholder definitions. In today's cellular networks the mobile network operator (MNO) typically employs a vertically integrated model and owns the spectrum, antenna sites and core network sites inclusive of corresponding equipment. They also implement the required functionality at each site to deliver the required service level to either their subscribers and/or a mobile virtual network operator (MVNO). Dependent on his business model the MNO may also own the inter-site transport network (integrated operator) or be leasing the corresponding lines from another operator. Fully utilising the 5G virtualised networks capability as proposed by 5G-MoNArch creates the opportunity to move away from this highly integrated stakeholder model to one with more layers of stakeholders based on a classic horizontal or platforms approach. These extra layers of stakeholders introduce opportunities for new entrants to work with existing ones to provide customised equipment or service implementations wherever and whenever needed. This ability to customise will ideally lead to the seamless integration of new verticals into the mobile ecosystem, opportunities for new revenues streams for mobile service providers, and enable realisation of benefits to society more generally.

One view of the tiered stakeholder model, which is enabled through a flexible 5G network is shown in Figure 2-4. This is largely taken from the 5G-NORMA project [NORMA D3.2] with care taken to align it with the terminology used currently at 3GPP [3GPP TR 28.801]. The definition of the stakeholder roles within this are presented next and followed by how the two testbed scenarios foreseen within 5G-MoNArch might map to this tiered system.

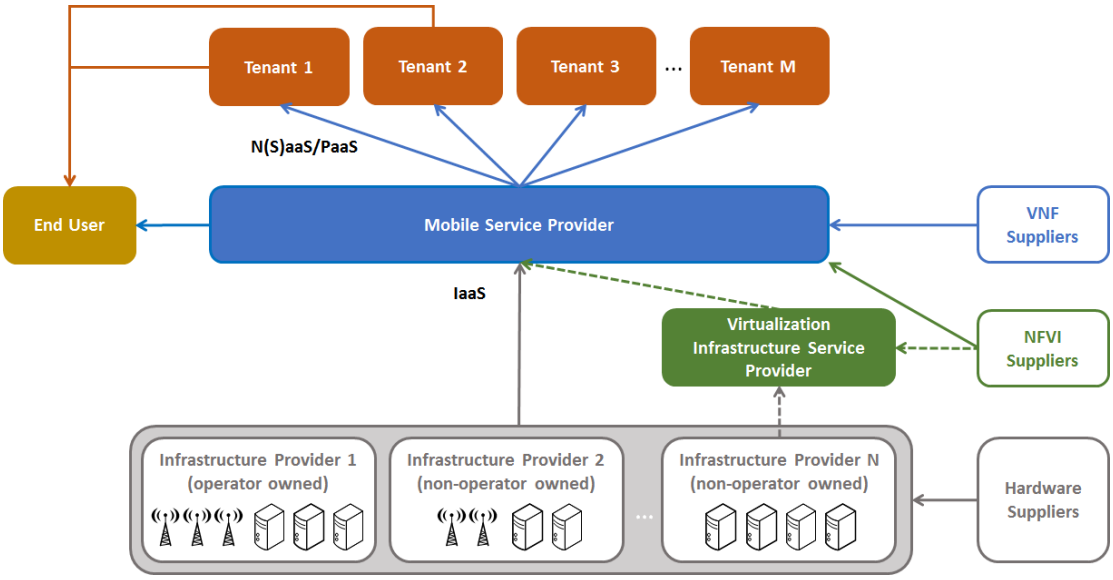


Figure 2-4: Example of the tiered stakeholder model that 5G virtualised networks will enable (modified from [5GN-D32])

Stakeholders are individuals, entities or organisations that affect how the 5G-MoNArch system operates. Where appropriate, and as guided by business model analysis, some stakeholders will be actors in the cost or revenue structure.

A 5G-MoNArch **Mobile Service Provider (MSP)** provides mobile internet connectivity and telecommunication services to either end users directly, i.e. through a business-to-customer (B2C) relationship, or via an intermediate "tenant", i.e. a business-to-business (B2B) or business-to-business-to-anyone (B2B2X) relationship; see next stakeholder description. The dedicated logical mobile network resources offered by an MSP are based on Network Slice Instances (NSIs) realising the relevant NF chains to support the instantiated telecommunication services, e.g., eMBB (enhanced Mobile Broadband) or mMTC (massive Machine Type Communications). In case of intermediate tenants, the MSP's offerings are Network (Slice) As A Service (N(S)aaS) or Platform-As-A-Service (PaaS). An MSP is responsible for design, build and operation of its service offerings.

A **5G-MoNArch tenant**, usually a business entity, buys and leverages a 5G-MoNArch network slice and services provided by the MSP. A tenant can, for example, be equivalent to today's MVNO, an enterprise (e.g., from a vertical industry) or other organisations that require telecommunications services for their internal business operations or for offers to their customers.

A **5G-MoNArch Infrastructure Provider (InP)** is the entity/company that owns and manages parts of, or the complete infrastructure of the network under consideration and offers it to the MSP, i.e., Infrastructure-As-A-Service (IaaS). With respect to the architectural model in 5G-MoNArch, the InP role may be further sub-divided into antenna site infrastructure provider, transport network provider, and data centre service provider (DCSP). The former owns the physical infrastructure such as the antenna sites, the HW equipment for the antennas and Remote Radio Heads (RRHs), monolithic base stations, etc. (i.e., infrastructure related to PNFs). The latter is represented by the collapsed roles of an entity/company that owns and manages local and/or central data centres. Within 5G-MoNArch, there are two types of data centre operators, infrastructure providers acting on small/medium size data centres (in terms of resources to be deployed and geographical presence) and big players (like Amazon) having big data centres deployed world-wide.

In 5G-MoNArch terminology a **Mobile Network Operator (MNO)** is an entity that operates and owns the mobile network, i.e. it vertically integrates into a single entity the roles of MSP and InP.

In practice, there may be also a so-called **Virtualisation Infrastructure Service Provider (VISP)** which designs, builds and operates its virtualisation infrastructure(s) on top of InP services provided by one or more DCSPs. The VISP offers its infrastructure service to the MSP.

Further roles in the stakeholder model to be mentioned are the **HW supplier** offering HW to the InPs (server, antenna, cable …), the **NFV Infrastructure (NFVI) supplier** providing the corresponding NFV

infrastructure to its customers, i.e. to the VISP and/or directly to the MSP, respectively, and finally the **VNF supplier** offering virtualised SW components to the MSP.

This high-level stakeholder model provides the first level of decomposition for architectural analysis. As the 5G-MoNArch architecture is refined, further stakeholder definitions will emerge as necessary to articulate the scope of system and opportunities.

2.2 Core Network

The next generation core network architecture needs to be more flexible to adapt to the requirements of diversified and continuous emerging services, accommodate various types of User Equipment (UE), interconnect different Radio Access Networks (RAN), and scaling with variable traffic demands. On the other hand, the advancing of NFV/SDN technology paves the way of network architecture evolution towards softwarisation. Driven by these factors, the 5G Core Network (5GC) architecture should be based on modular design, C/U separation, "service based" (see below) and support network slicing. [3GPP TR 23.799] defines two architecture options for the 5GC. Option 1 follows the 4G design syntax with the focus on functional aspects and use reference point representation (Figure 2-5). Option 2 proposes a service based architecture addressing especially the flexibility requirements for the 5G era. This document examines on the second architecture option according to the scope and focus of the 5G-MoNArch project. Below are the design principles listed in [3GPP TR 23.799] for Option 2:

- Separate the UP and CP functions
- Allow for a flexible deployment of UP and CP functions, i.e. central location or distributed (remote) location.
- Modularise the function design
- Separated Authentication and Mobility management
- Separated mobility management and session management
- Support a flexible information model with subscription and policy separated from network functions and nodes.
- Minimise access and core network dependencies.
- Procedures (i.e. set of interactions between two NFs) are defined as a service, wherever applicable, so that its re-use is possible.
- The architecture shall support capability exposure.

Based on these principles, 3GPP specified the service based 5G reference architecture as shown in Figure 2-6. This architecture includes both conventional mobile network functions (e.g., CP network functions such as AMF, SMF, AUSF, PCF, AF, UP network function UPF, user data management functions UDM), as well as some special functions introduced to support service based architecture and network slicing (i.e., NSSF, NRF, NEF). The description of these functions is listed as below:

- **AMF (Access and Mobility Management Function)**, including termination of RAN CP interface (N2) and of NAS interface (N1), NAS ciphering and integrity protection, registration/connection /reachability/mobility management, lawful interception, access authentication and authorisation, security anchoring, security context management;
- **SMF (Session Management Function)**, including session management, UE IP address allocation & management, UP functions selection/control, termination of interfaces towards PCF, policy enforcement and QoS, roaming functionality. SMF is connected to UPF via N4 interface;
- **AUSF (Authentication Server Function)**, providing authentication and authorisation functionalities;
- **NEF (Network Exposure Function)**, providing means to collect, store and securely expose the services and capabilities provided by 3GPP network functions (e.g., to third parties or amongst NFs themselves);

- **NRF (Network Repository Function)**, maintaining and providing the deployed NF Instances information when deploying/updating/removing NF instances, supporting service discovery function;
- **PCF (Policy Control Function)**, supporting unified policy framework to govern network behaviours, providing policy rules to control plane function(s) to enforce them;
- **UDM (Unified Data Management)**, supporting Authentication Credential Repository and Processing Function, storing the long-term security credentials and Subscription information;
- **AF (Application Function)**, representing any additional CP function which might be required by specific Network Slices, potentially provided by third parties;
- **UPF (User Plane Function)**, including the following functionalities: anchor point for Intra-/Inter-RAT mobility, External PDU session point of interconnection, packet routing & forwarding, UP QoS handling, packet inspection and Policy rule enforcement, lawful interception, traffic accounting and reporting.
- **NSSF (Network Slice Selection Function)**, selecting the set of network slice instances serving the UE.

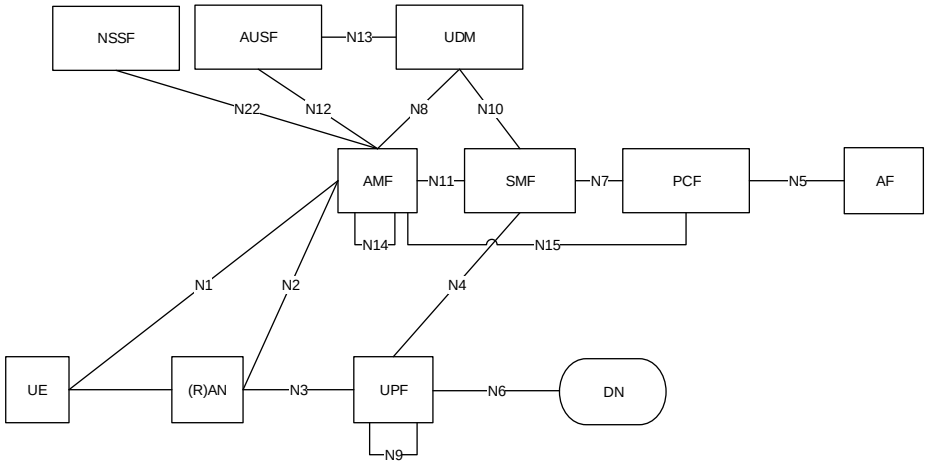


Figure 2-5: Non-Roaming 5G System Architecture in reference point representation [3GPP TS 23.501]

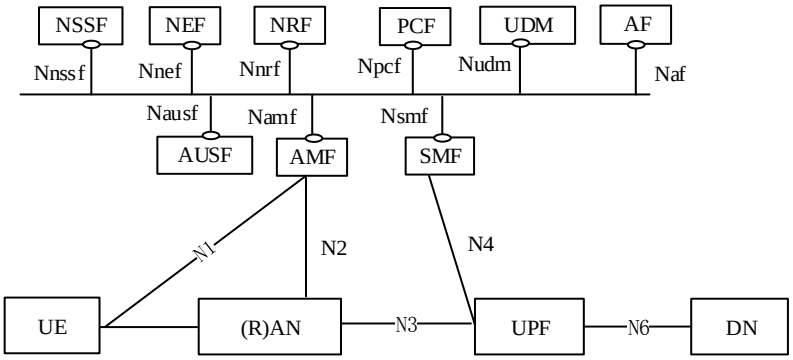


Figure 2-6: 5G System Architecture [3GPP TS 23.501]

5GC supports both the Radio Access Network and Fixed Access Network, i.e. (R)AN. It connects to UE through AMF via N1 interface, connects to (R)AN control plane through AMF via N2 interface, and connects to (R)AN data plane through UPF via N3 interface.

This architecture includes some further features like stateless functions, service based procedures (details in [23.502], concurrent access to local and centralised services, etc.

With the current specification for release 15, 3GPP has built up a flexible 5G Core Network architecture and slice support framework. 5GC can provide differentiated services/network slice for different group of use cases, e.g., by the selection of different network functions (e.g., different types of AMF/SMF/UPF) and network function instances (e.g., at centralised location or local location).

Each network slice is identified by an E2E network slice identifier (i.e., S-NSSAI Single Network Slice Selection Assistance Information). 5GC decides on the S-NSSAI(s) a UE can use (i.e., allowed NSSAI) in a certain area (i.e., Registration Area, RA), which identifies the serving AMF for this UE. The allowed NSSAI can be updated per RA with the registration procedure defined in [23.502]. UE is only allowed to request the slice whose S-NSSAI is in the allowed NSSAI within UE's current RA, which is called requested S-NSSAI. The requested S-NSSAI identifies the SMF to serve this UE and SMF further identifies the UPF to serve the UE. To this end, the core network part of the network slice is uniquely identified. When the UE moves, AMF/SMF/UPF can be reselected due to the Geo coverage of the functions as well as the connectivity to the (R)AN node.

There are still some gaps to be filled considering the complete framework to support the deployment of the E2E network slice in all use cases. For instance, 3GPP SA2 focuses more on the mobility network functionality, slice selection, network service differentiation in 5GC, while the slice deployment/management aspects will need some further study to enable the guarantee of SLAs for an E2E slice. Meanwhile, 3GPP is now working on Rel. 15. Some features will need to be further addressed in Rel. 16, especially considering E2E slicing support. Chapter 3 provides a further analysis regarding the extension of the specified network functions in the current 3GPP architecture.

2.3 RAN

The next generation mobile technology will place unprecedented demands on the efficiency, flexibility and scalability of the radio access network (RAN) to support diverse use cases and their performance requirements without impacting the total cost of ownership. These network demands are forcing a radical re-think of the entire RAN including remote radio head (RRH), the baseband unit (BBU) and the transport connectivity between the two. Softwarisation represents an important enhancement in the process towards 5G RAN design by exploiting the novel features of SDN/NFV paradigms.

The 5G-PPP project METIS-II has identified several RAN design requirements which are necessary to meet the diverse needs of E2E 5G architecture and some important requirements include

- the 5G RAN should be highly scalable with respect to parameters like throughput, the number of devices or the number of connections,
- the RAN need to be re-programmable to enable the overall network to be software-configurable,
- the 5G RAN architecture should enable a tight interworking between LTE-A evolution and novel 5G radio technology on RAN level, and

the 5G RAN design must be future proof, i.e., it should enable an efficient introduction of new features and services and guarantee backward-compatibility of devices in future releases.

However, to achieve the fundamental design aspects of 5G-MoNArch, i.e., the re-programmable and scalable E2E mobile network slicing, the flexibility of the current RAN architecture needs to be studied. In this section, the current 5G RAN (5G NR) protocol architecture, and the existing RAN slicing approach proposed by the 5G-PPP projects [METIS II D2.4] [NORMA D4.1] [5GARCH16-WPv2] is examined.

2.3.1 RAN Control and User Plane NFs

5G New Radio (NR) development is part of continuous mobile broadband evolution process to meet the requirements of 5G as outlined by IMT-2020. 5G New Radio (NR) is expected to expand and support diverse use case scenarios and applications that will continue beyond the current IMT-Advanced standard, for instance, enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communication (URLLC) and massive Machine Type Communication (mMTC).

3GPP has released specification 38.300 V1 on NR and NG-RAN overall description. This standard comes with the detailed descriptions about 5G NR network and Protocol architecture. The User Plane

(UP) and Control Plane (CP) of 3GPP NR protocol stacks are shown respectively in Figure 2-7 and Figure 2-8 [3GPP TS 38.300].

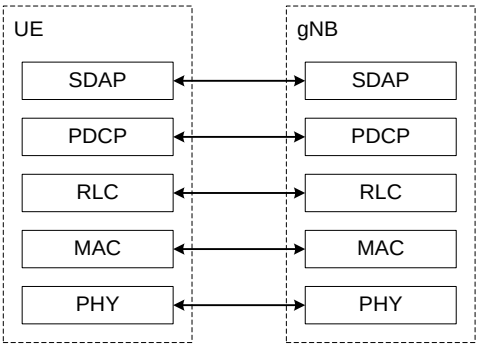


Figure 2-7: (UP) Protocol Stack [3GPP TS 38.300]

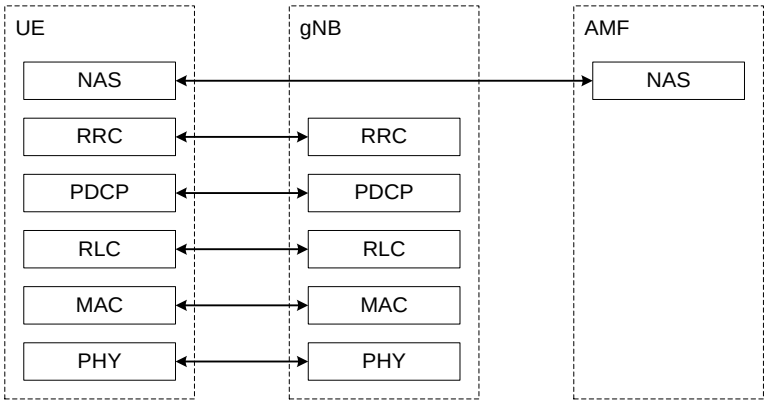


Figure 2-8: (CP) Protocol Stack [3GPP TS 38.300]

At both the User Equipment (UE) and the NR NodeB (gNB), the UP protocol stack is composed by the Physical Layer (PHY), the Medium Access Control (MAC), the Radio Link Control (RLC), the Packet Data Convergence Protocol (PDCP), and the new Service Data Adaptation Protocol (SDAP). The CP protocol stack is composed by the PHY, the MAC, the RLC, the PDCP, and the Radio Resource Control (RRC). The Non-Access Stratum (NAS) is used to convey non-radio signalling between the UE and Access and Mobility Management Function (AMF).

5G-NR UP contains PHY, MAC, RLC, and PDCP same as LTE and has introduced a new layer named as SDAP (Service Data Adaptation Protocol) to handle flow-based Quality of Service (QoS) framework in RAN, such as mapping between QoS flow and a data radio bearer, and QoS flow ID markings.

However, the E2E network slicing, re-programmability and cloudification of mobile network functions in 5G further requires the decomposition of RAN protocol stacks (CP and UP) into cloud enabled software functions called virtual network functions (VNFs) with interfaces for interacting with each other.

There is a consensus on CN/RAN split that enables an independent evolution of CN and RAN functionalities and allows multi-vendor deployments [5GARCH16-WPv2]. A common CN and CN/RAN interface (referred to as S1* [METIS II D2.4] and NG see Section 2.2) for both the novel air interface variants (AIVs) and the evolution of LTE-A. Enhancements are also envisioned for the evolution of the X2 interface (referred to as X2* and Xn interface [3GPP TS 38.300]), which jointly with S1* become interfaces addressing multiple AIVs. Here, it is assumed that the overall air interface (AI) is composed of novel and evolved legacy AIVs [3] [METIS II D2.4]. METIS-II proposes a common protocol architecture for the 5G RAN, illustrated in Figure 2-9 where two AIVs are exemplarily

[3] An AIV is defined as the RAN protocol stack (i.e., PHY/MAC/RLC/PDCP/RRC or 5G equivalents, or subset thereof) and all related functionalities describing the interaction between infrastructure and device, and covering, e.g., a subset of services, bands, cell types that characterise the overall 5G system.

illustrated. Therein, AIV-overarching mechanisms are located at the Access Network – Outer (AN-O) layer while AIV-specific mechanisms are located at the Access Network – Inner (AN-I) layer. It is worth noting that in this implementation AN-O corresponds to a central unit (CU) and AN-I corresponds to a distributed unit (DU). The functional split option is illustrated at PDCP level not to influence the 5G specification with legacy AIV constraints. It is observed that functionalities that are tightly coupled with hardware implementations can be implemented at the DUs, while software-based implementation can be implemented at the CU. In terms of the coupling with the radio frame structure, traditionally slow functionalities (e.g., traffic steering) can be designed to operate on a faster time scale and can still be implemented at a CU. Some of key elements of the common CP are outlined as [METIS II D2.4]:

- **AIV agnostic Slice Enabler (AaSE)** enables performance guaranteeing multi-slice RM with real-time SLA monitoring
- **Multi-AIV Resource Mapping** provides the interface to AN-I to enable fast routing of data flows to the appropriate AIV(s) comprising both novel 5G AIVs and legacy AIVs
- **Real-time Resource Mapping** is a collection of mechanisms which includes flexible multi-service scheduling where different parameters related to the communication using a certain AIV can be adjusted in real time
- **RRC:** includes the RRC state machine handling and the mobility management functions that should be moved to the RAN to optimise Tracking Area Updates (TAU) as well as the way that the UE is configured to perform the measurements for the various AIVs.

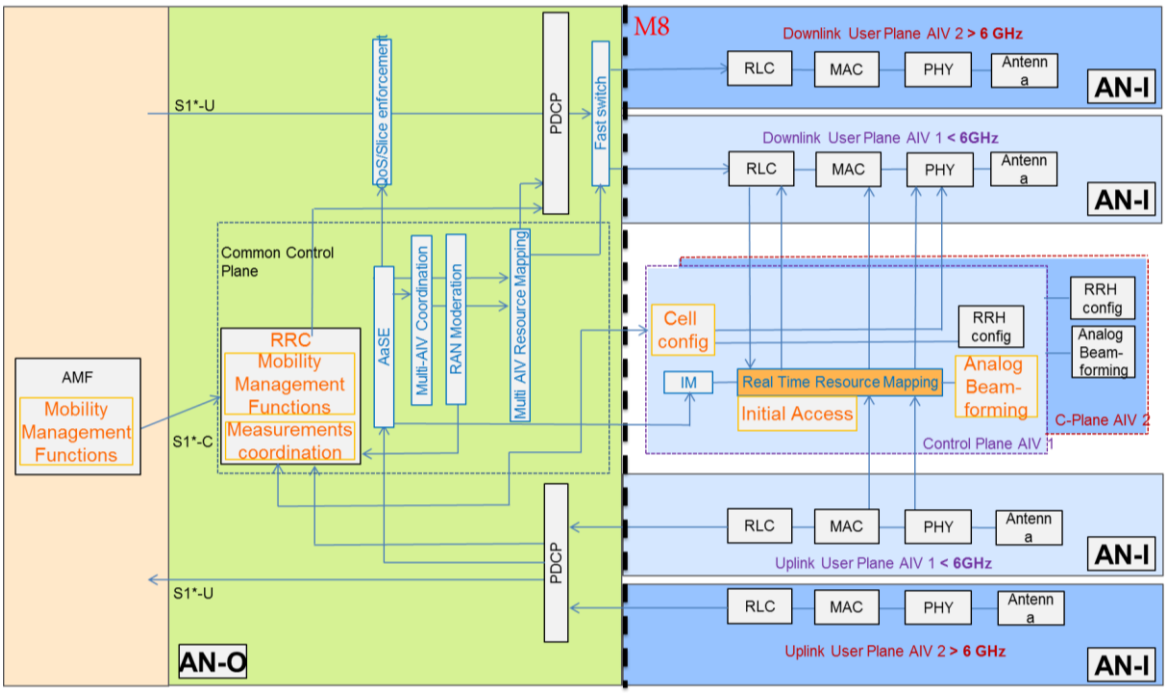


Figure 2-9: Protocol Architecture of common control plane [METIS II D2.4]. Functionalities with blue text font indicate synchronous control functions while the functionalities with orange text font indicate asynchronous control functions

2.3.2 RAN Part of E2E Network Slicing

The RAN infrastructure will typically be the same for all the network slices. There might be cases where the slicing will go down to the physical layer frequency resources, but this will relate to the actual requirements from the use case (UC), e.g., potential UCs that require physical separation of the resources due to regulation reasons as well as the needs of the vertical industries that will utilise the network slices.

In [Silva et al] the requirements for RAN slicing have been summarised as listed below:

- Utilisation of RAN resources should be maximised among multiple slices;

- RAN should be slice-aware via some explicit or implicit identification (e.g., based on an abstraction model);
- RAN should support traffic differentiation mechanism to treat different slices differently and/or different services within the multi-service slices;
- RAN should support protection mechanisms to minimise inter-slice effects (such as the congestion of one slice negatively affecting the other);
- RAN should support efficient management mechanisms e.g. to efficiently set up new slices and to efficiently operate new business/services.

However, the 5G-PPP projects [METIS II D2.4] [NORMA D4.1] [5GARCH16-WPv2] have converged the RAN slicing approach in the following aspects, where an example illustration is provided in Figure 2-10:

- Limited number of configurations in RAN, i.e., RAN configuration modes (RCMs), should exist to cover the different Use cases;
- Certain slices may share all the resources and be differentiated using different QoS classes;
- Each slice can have its own RRC functions and configurations when it comes to particular functions (e.g., discontinuous reception/transmission (DRX/DTX), measurements reporting, TAU periodicity, cell selection strategies, cell configurations / TDD patterns etc.);
- A function should be present for enabling the slice selection. This can be done as initial configuration or via a common functionality (e.g., common RRC part);
- An RRM function should ensure the sharing of the common radio resources and facilitate the slice isolation among the different slices – this can be omitted for full separation of resources. However, each slice can apply its own RRM strategies according to the slice specific characteristics;
- Each slice can apply its own strategies for certain functions (e.g., header compression, ciphering, segmentation, re-ordering, ciphering).

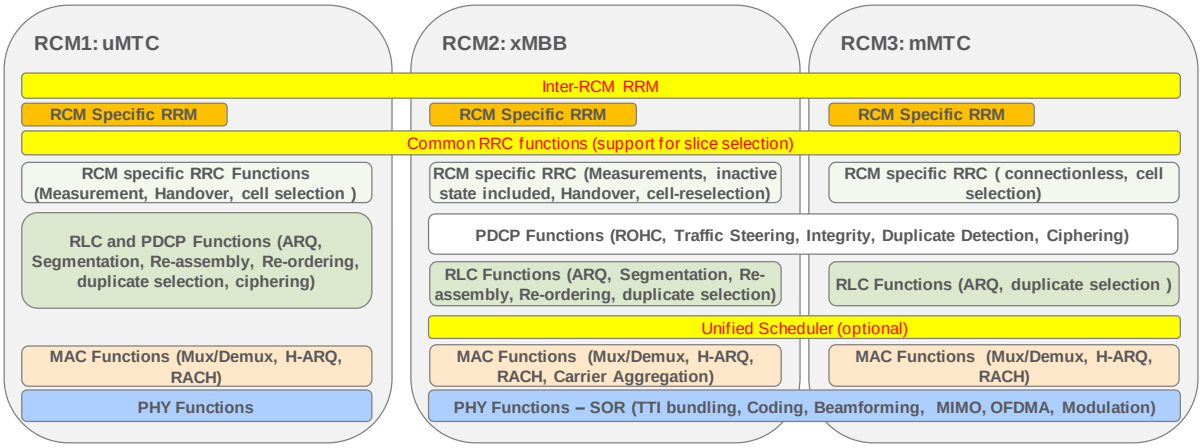


Figure 2-10: Example of RAN support of E2E network slicing with shared and independent functions

Considering the 5G use cases and deployment options, it can be expected that *the depth of network slicing* on the RAN side is also business-driven. The business of a vertical industry may require a complete isolation even in terms of the radio resources. As different vertical industries share a common infrastructure, various slicing implementation options may co-exist. Towards this direction, in 5G NORMA project, three different deployment options have been identified as shown in Figure 2-11:

- In the first option (Option 1) each slice has an individual RAN protocol stack implemented down to the upper part of the physical layer (c.f. PHY-User and PHY-Cell). Only the lower part of the physical layer (c.f. PHY-TP) is shared across slices. Option 1 can be considered as implementing all user-specific functions such as forward error correction encoding, layer mapping and precoding in an individual fashion;

- In the second option (Option 2) each slice uses an individual implementation of service-specific functionality such as PDCP, RLC, and slice-specific RRC, QoS scheduling etc.
- In the third option (Option 3) each slice is differentiated by different QoS classes.

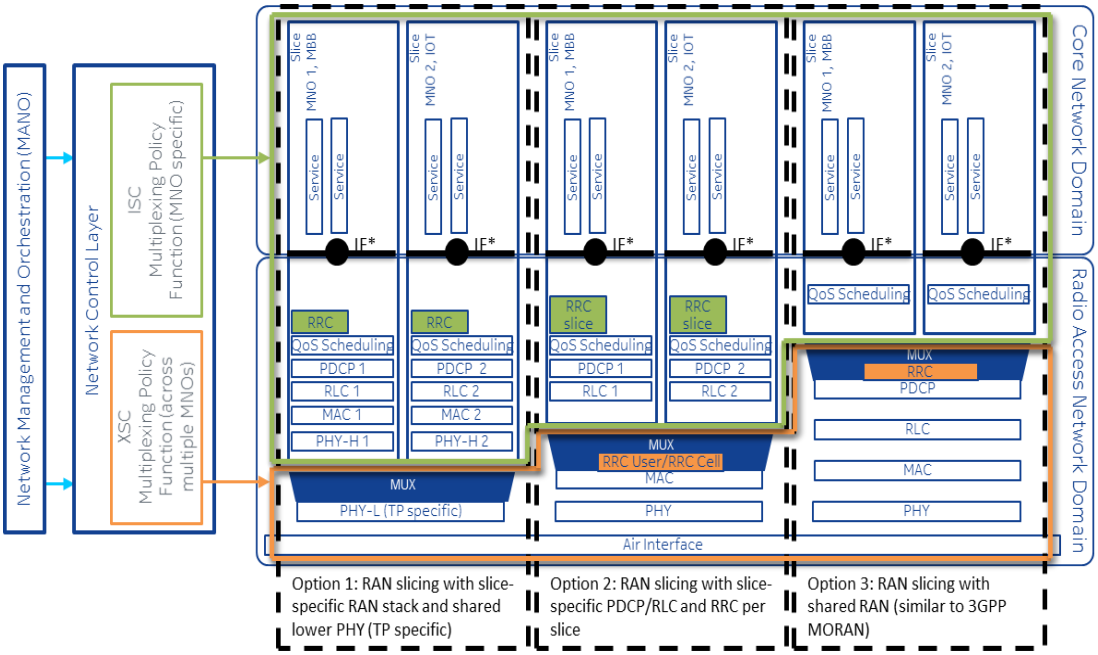


Figure 2-11: Implementation options of RAN part of E2E network slicing [NORMA D4.1]

There are still some gaps to be filled to be able to achieve the fundamental 5G-MoNArch design aspects, for example decomposition of RAN protocol stack into cloud enabled NFs, identification of dedicated and shared functions in the context of slicing and finally the definition of protocols and methods to enable programmability in the RAN.

2.4 Centralised CP Architecture

It is expected that the Software Defined Networking paradigm will be the most prominent way of performing network control in the near future. Driven by the vast amount of available efforts in both research and SDO, the 5G NORMA project proposed a Software Defined control and data layer architecture that will be at the basis of the 5G-MoNArch one.

Software defined network control entails the separation of the formerly monolithic functionality (e.g., an S-GW) into two elements: the function *logic*, running on top of a standardised North Bound Interface of a Controller that, in turn, uses its South Bound Interface to control (V)NFs where the agent is running on the data plane. This approach turned to be successful for e.g., datacentre networks, so it applicability in a broader way was studied. Most notably, the 3GPP SA2 decided to apply this split in the NextGen System (i.e., 5G) in its soon to be finalised Architecture [3GPP TR 23.799]. In there, the core functionality once performed by, among others, S-GW, P-GW and MME, is now split into a User Plane Function (UPF, the agent) controlled by the Session Management Function (SMF, running the logic). Applying this concept to the core network elements is a natural choice, as the fundamental operation that those elements are performing (i.e., packet forwarding) is the same that was initially targeted by the SDN concept. However, 5G NORMA proposed a more ambitious paradigm, and applied the SDN concept also to former RAN elements. Therefore, functions like wireless spectrum management or scheduling are also split into two, well defined, parts: an application that controls the functionality in a possibly centralised fashion and the (V)NF that resides in the data plane, enforcing the rules devised by the application. For this reason, the whole architecture (as described in Section 2.1) and in particular the control and user plane architecture has been designed to natively consider this feature, which has been proven to be efficient when dealing with novel concepts such as multi-tenancy and network slicing.

2.4.1 The controllers

Broadly speaking, a VNF should provide/support the same overall functionality as an equivalent "black box" based function (non-virtual or bare metal). The key difference, though, is that a VNF could be deployed as a software instance capable of running on general purpose servers via virtualisation technologies. To allow for such flexibility, running for example network functions at various network locations (even at the base station), an architecture is briefly detailed below that is in line with the defined overall ETSI MANO ecosystem with the emphasis placed on mobility and QoE support, as described in Section 2.1. This architecture, built to natively support network slicing spanning several network domains, is composed by three main elements MANO, the Intra Slice Controller (ISC), and the Cross-Layer Controller (XSC). Their role, is summarised next.

MANO

Although not directly involved in the control and user plane architecture, the overarching role of the MANO is to provide and maintain a suitable network function chaining to create an E2E service. The proposed orchestration capabilities are in line with (and provide extensions upon) the ETSI logical reference architecture for NFV MANagement and Orchestration (MANO) [ETSI GS NFV-MAN]. Several functional requirements were considered while designing the orchestration side as further described in Section 2.5. The orchestration framework shall tightly interact with the control elements (ISC and XSC), reacting to QoS/QoE-based triggers.

The integration with ISC and XSC is paramount to achieving full QoE/QoS support in a network slice. These modules are instantiated as further VNFs and are the main triggers for QoE/QoS based re-orchestration.

ISC and XSC

The Intra Slice Controller (ISC) controls the network functions belonging to a slice and their associated resources using a Software Defined approach. There is an ISC instance per network slice, which retrieves network requirements through its northbound interface (connected to the MANO layer), and triggers the actions through its southbound interface (connected to both VNFs and Physical NFs), following the SDN paradigm. Such interfaces are used to fulfil slice QoE/QoS constraints. If QoE/QoS targets are not satisfied, the ISC instructs a re-orchestration. The advantages provided by the ISC can be summarised as follows:

- *Flexibility*: Operators would be able to tailor the network to their needs by simply re-programming the controller.
- *Programmability*: It allows third parties to acquire network resources on-demand satisfying their individual Service Level Agreement (SLA) while enhancing the user perceived QoE with customised network resources.
- *Unified control*: Adopting a logically centralised control unifies heterogeneous network platforms and provides a simplified operation of the wireless network. With ISC, network operators only need to control a set of central entities (namely, the controllers) that control the entire network, which possibly includes heterogeneous radio technologies.
- *New services*: New services can be easily introduced by directly modifying the network behaviour by means of applications running on the ISC northbound interface. This would considerably save time in developing, debugging, and deploying new network functions.

While ISC provides isolation between slice's resources/functions, shared components need a dedicated controller to fully exploit the multiplexing gain. This controller is called XSC. Specifically, elements such as transmission points, radio resources, transport and fronthaul capacity are often realised as shared resources. Once they are collected in a common pool, an interaction between XSC and ISC is in place to dynamically use the shared physical resources during operational flows. Physical resources are intended as radio and transmission over other media, processing within areas of computing resources, and storage for user/data plane and control plane information. While resource pooling for storage and processing power may be less demanding due to theoretically large resource pools, the scarcity of radio resources in many cases requires an advanced resource management solution [Sciancalepore et al].

The XSC takes decisions based on the policies provided by MANO layer how to fulfil the demands of several partially competing network slices simultaneously. Slice performance demands might be identified in terms of:

- **throughput**, which requires a management of the radio resources
- **latency**, which requires a management of the placement of the network functions and usable storage entities
- **management of processing / compute and storage resources** in the neighbourhood of access nodes, which may also impact latency and error recovery performance
- **reliability and resilience**, which is also greatly influenced by proper mechanisms for dynamic sharing of all three kinds of resources.

2.4.2 Centralised control layer

As already discussed in Section 2.1, the 5G NORMA architecture that is at the basis of the 5G-MoNArch one, deeply embeds the concepts of network slicing and multi-tenancy into its architectural elements. By categorising resources and the NFs associated to them into dedicated and shared among network slices, 5G NORMA defined two kinds of controllers: the ISC (controlling dedicated network functions) and the XSC (controlling shared network functions). They naturally serve as limit point between the c-layer elements and the d-layer ones, as depicted in Figure 2-12.

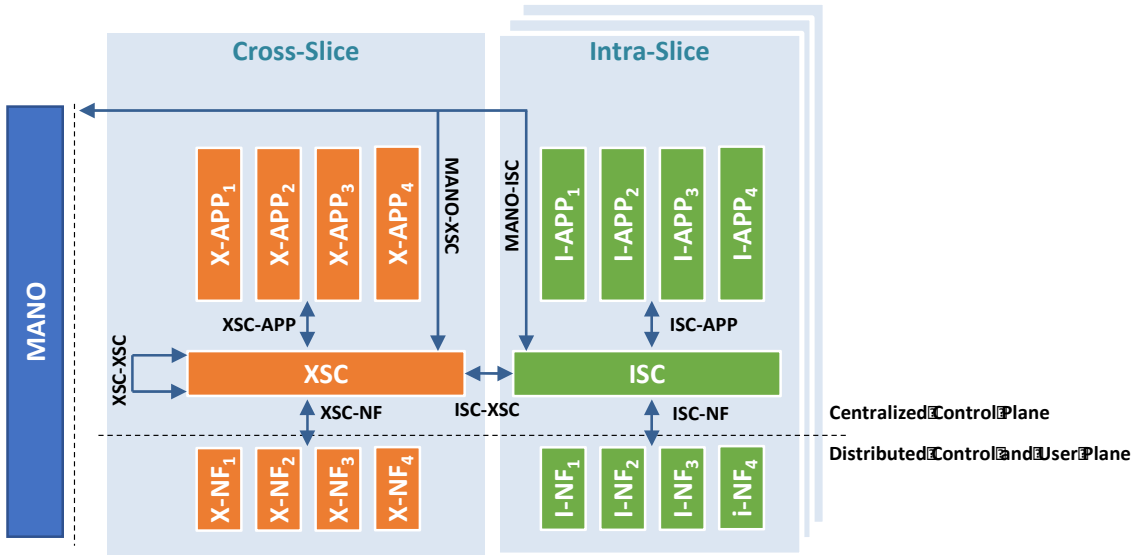


Figure 2-12: 5G NORMA – 5G-MONARCH control and user plane architecture

Former c-plane functions such as, for example, mobility management are performed in this new architecture as an application, running on top of the controllers. According to their nature, applications can be X-APP (if controlling shared network functions) or I-APP (if they work on dedicated ones, I-NF). The main advantages of this approach are given by the centralisation of the controlling application that can then receive feedback from other possibly centralised functions, such as scheduling. So, let us take an exemplary Mobility Management (MM) I-APP running on top of the ISC to explain the whole control framework.

Here, the MM ISC App is controlling/interfacing entities within the NFVI, including former CP functions (e.g., location and paging) and UP functions (for anchoring, forwarding, enforcement, etc.). So, the MM I-APP conveys specific mobility requirements through the ISC to I-NFs being fully under control of one ISC. An additional challenge is, however, the inclusion of I-NFs under ISC control, e.g. radio equipment is included in mobility operations (e.g. lower layer mobility decision). In this case the control operation is performed in conjunction with the XSC. For a more detailed description of this use case, the reader is referred to [Yousaf et al].

2.4.3 Considerations on distributed and hierarchical control

The centralised approach provided by ISC and XSC application is however, just one of the layers of the 5G NORMA control and user plan architecture. Not all the functionality can be split into a logic running on top of an ISC (or XSC) and the agent running in the underlying NFs. That is, for different reasons, some of the control functionality should be managed in a legacy distributed way:

- **Legacy PNFs**: legacy PNFs that should be integrated into a network slice may not support the ISC or XSC approach as envisioned by 5G-MoNArch. Therefore, their behaviour should be integrated through the Southbound Interface of the controllers
- **Data Locality**: some control NFs build on information available locally that should be processed with very low timing constraints. In this case, the performance gains obtained by a centralised approach may not be enough for certain NFs.
- **Scalability**: for some NFs, the overhead introduced by a fully Softwarised and Centralised control may be too much, especially when configured for extreme situations. As an example, a centralised MAC scheduler that controls several base stations may hardly be reconfigurable through an ISC (or XSC) application. Therefore, some of the functionality is necessarily offloaded to distributed control functions that can operate at the fastest time scales needed in each context (e.g., a fast-scheduler that can operate at sub-TTI level)

2.5 Network Management and Orchestration

2.5.1 Overview

The MANO architectural framework has the role to manage the underlying virtualised infrastructure and orchestrate the allocation of resources needed by the Network Services (NSs) and VNFs. Such coordination is necessary as NFV decouples software implementations of NFs from the physical resources they use.

3GPP SA5 management and orchestration concepts in 5G Networks are firstly provided, followed by the MANO architecture defined mainly in 5G NORMA. This is the baseline from which 5G-MoNArch MANO evolves to enable the following key features: flexible location and instantiation of network functions, software defined control and orchestration of the network and joint optimisation of network functions.

2.5.2 3GPP SA5 (Telecom Management) Orchestration of 5G Network

3GPP SA5 normative work focuses on the management of 5G Networks. New management features are required due to the introduction of virtualisation and Network Slicing in 5G. These features transform the network into a flexible aggregation of elements with a lifecycle that reflects the end user needs. To manage a 5G Network, automation and orchestration are mandatory features. 5G-MoNArch key design aspects, such as programmability and Network Slicing (see Section 2.1.2) are therefore also 3GPP SA5 key issues.

In [3GPP TR 28.800] the architecture for the management and orchestration of the next generation network is studied according to use cases of interest.

For the management of the virtualised network elements, the virtualised resource supporting the network element is managed by NFV-MANO. The 3GPP network functions are managed by 3GPP network management functions. For the management of the non-virtualised network elements, both network function and network resource is managed by 3GPP network management functions.

The current line of study is about evolving the LTE management system to fulfil 5G requirements. To do this the following actions must be accomplished:

- New NRMs (Network Resource Models) are needed for the radio access network and core network.
- New measurement specifications are needed for Network Resources features and nodes.
- The existing 3GPP management specifications needs to be checked and eventually enhanced to fulfil 5G requirements.

3GPP study [3GPP TR 28.800] describes the following use cases:

- Management and orchestration of networks containing non-5G NE and 5G NE: potential management options/scenarios for possible network deployments, specifically at the early stages of introducing 5G network elements along with existing non-5G generation 3GPP network(s). Operator wants to manage a network containing both non-5G NE(s) and 5G NE(s).
- Management and orchestration architecture for network slicing: more details in [3GPP TR 28.80]
- Fault and Performance data collection and reporting (single-operator scenario and multi operator scenarios)
- Management and orchestration architecture for management of 5G-RAN and 5GC
- Management of communication services: communication services can be provided by network slice or network without slicing
- Exposure of management interfaces to another operator: Operator's management system exposes suitable APIs to another operator for slice management
- Exposure of management interfaces to communication services provider: a communication services provider should be able to request an operator to host a service management

The current conclusion of [3GPP TR 28.800] is that the management of the next generation networks and services is mainly related to Network Slice management functions. The deepening on management and orchestration architecture for network slicing is in [3GPP TR 28.801]. The 5G-MoNArch architecture has Network Slice as one of its key feature (see Section 2.1.4) so the normative works on it is very important for the project.

3GPP SA5 in [3GPP TR 28.801] defines the following concepts on Network Slices:

- Network slice instance: a set of network functions and the resources for these network functions which are arranged and configured, forming a complete logical network to meet certain network characteristics.
- Network slice subnet template: description of the structure (and contained components) and configuration of the network slice subnet instance
- Network slice subnet instance: a set of network functions and the resources for these network functions which are arranged and configured to form a logical network.
- Network slice template: description of the structure (and contained components) and configuration of a network slice
- Physical resource isolation: physical resource allocated for one network slice cannot be used by other network slices to avoid negative effect between multiple network slice instances.

5G-MoNArch architecture foresees slice specific NFs and common NFs, both controlled by ISC and XSC controllers (described in Section 2.1.4). The model adopted by 3GPP SA5 is coherent with 5G-MoNArch's view. The definition of Network Slice Subnet Instance (NSSI) is intended to enable the management of common NFs.

A Network Slice Instance (NSI) is a managed entity in the operator's network with a lifecycle independent of the lifecycle of the service instance(s). In particular, service instances are not necessarily active through the whole duration of the run-time phase of the supporting NSI. The NSI lifecycle typically includes an instantiation, configuration and activation phase, a run-time phase and a decommissioning phase. During the NSI lifecycle the operator manages the NSI.

The following phases describe the Network Slice lifecycle (see Figure 2-13):

- Preparation phase
- Instantiation, Configuration and Activation phase
- Run-time phase
- Decommissioning phase

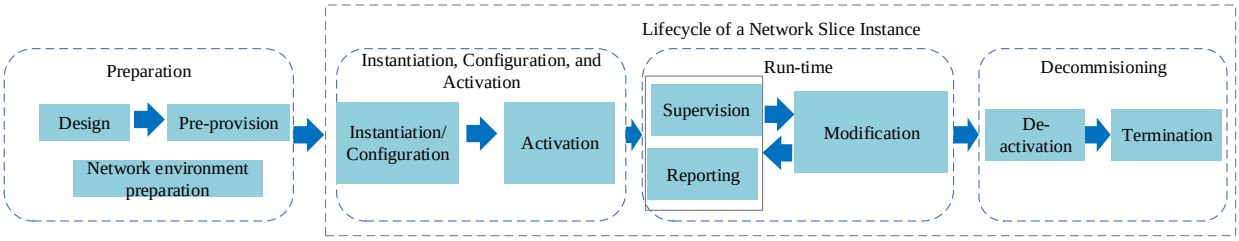


Figure 2-13: Network Slice lifecycle

More details on each phase are available on [3GPP TR 28.801]. 3GPP TR 28.801 defines the following Network slice concepts:

- **Completeness of an NSI**: A NSI is complete in the sense that it includes all functionalities and resources necessary to support certain set of communication services thus serving certain business purpose.
- **Components of an NSI**: The NSI contains NFs (e.g. belonging to AN and CN). If the NFs are interconnected, the 3GPP management system contains the information relevant to connections between these NFs such as topology of connections, individual link requirements (e.g. QoS attributes), etc. For the part of the TN supporting connectivity between the NFs, the 3GPP management system provides link requirements (e.g. topology, QoS attributes) to the management system that handles the part of the TN supporting connectivity between the NFs.
- **Resources used by the NSI**: The NSI is realised via the required physical and logical resources.
- **Network Slice Template**: The Network Slice is described by a Network Slice Template (NST). The NSI is created using the NST and instance-specific information.
- **NSI policies and configurations**: Instance-specific policies and configurations are required when creating an NSI. Network characteristics examples are ultra-low-latency, ultra-reliability, etc. NSI contains Core Network part and Access Network part.
- **Isolation of NSIs**: An NSI may be fully or partly, logically and/or physically, isolated from another NSI.

[3GPP TR 28.801] defines the information model attached to network slices with the following assumptions (see Figure 2-14):

- An NSI may support zero or more communication services
- A communication service may be served by one or more NSIs, possibly with different characteristics.
- An NSI may be composed of network slice subnets of Physical Network Functions and/or Virtualised Network Functions.
- Physical Network Functions and Virtualised Network Functions may belong to one or more network slice subnet(s).
- Virtualised Network Functions are deployed on top of virtualised resources.

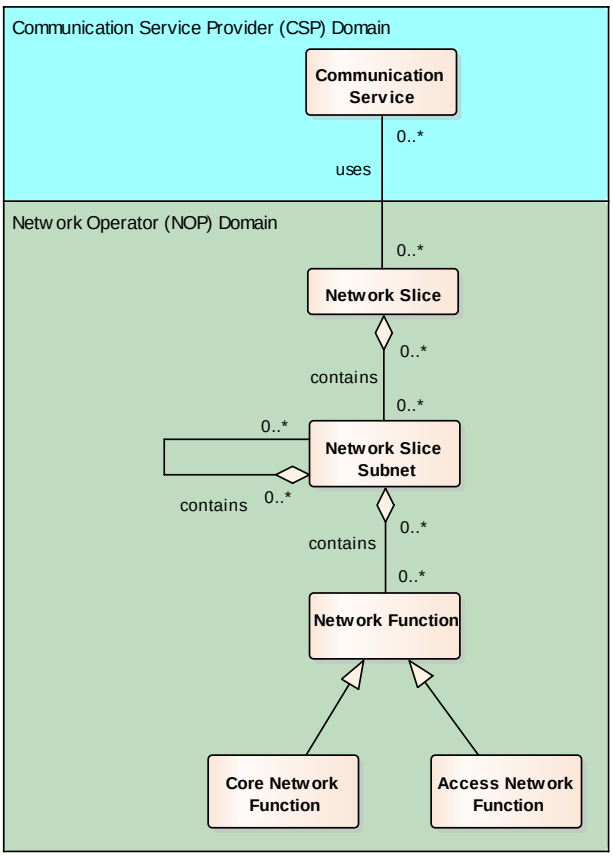


Figure 2-14: NS Information Model

Some of the concepts of a network slice subnet are:

- A NSSI constituent may include NF(s) and other NSSI(s).
- A NSSI may be shared by two or more NSIs, this is called a shared constituent of NSI.
- A NSSI may be shared by two or more NSSI(s), this is also called a shared constituent of NSSI.
- An NSSI that is dedicated to one NSI and is not shared as a constituent by two or more NSSI(s) is called a non-shared NSSI.
- An NSSI may contain CN functions only, AN functions only, or both CN functions and AN functions.
- The resources comprise physical and logical resources. In case of virtualisation, logical resources may be used.

On lifecycle management aspects [3GPP TR 28.801] defines that a communication service can, depending on the communication service requirements, use an existing NSI or be the trigger for the creation of a new NSI. The new NSI may be created just for this communication service or it may be created to support multiple communication services with similar network slice requirements. The lifecycle of a communication service is related, but not dependent on that of a NSI. The NSI may exist before the communication service uses the NSI and may exist after the communication service stopped using the NSI.

An NSI can, depending on the NSI requirements, be created using one or more existing NSSI(s) or initiate the creation of one or more new NSSI(s). The new NSSI(s) may be created just for this NSI or it may be created to support multiple NSIs. The lifecycle of an NSI is related but not dependent on that of an NSSI. The NSSI may exist before the NSI is created and may exist after the NSI is not needed anymore.

To improve the operational sustainability in 5G, SON (Self Organising Network) concepts, as key enablers introduced in 4G, may be reused for 5G. Evolution to 5G may bring increased network scale

and complexity, especially considering the multiple services/devices/tenants in 5G networks. In this context, operators may want to use concepts of SON as key feature to leverage 5G network slicing management.

The use of SON concepts for 5G management is coherent with the programmability key feature of 5G-MoNArch architecture. SON works on centralised and/or distributed algorithms programmed to automate actions on the network.

The main aspects of SON concepts that can potentially be used to leverage for network slicing management include:

- **NSI Automated Configuration**. Automated configuration of an NSI can happen during the Instantiation, Configuration, and Activation phase, a newly created NSI can be automatically configured with appropriate parameters before it is activated.
- **NSI Automated Reconfiguration**. Automated reconfiguration happens during Run-time phase, an activated NSI can be reconfigured automatically because of change of service requirements.
- **NSI Automated Optimisation**. The NSIs can be modified automatically to avoid degradation of services in case of network function overload, dynamic topology change, etc. The status of the target NSIs is monitored, including the status of network functions and services.
- **NSI Automated Healing**. For the running NSIs, SON algorithms could identify the failures of NSIs and apply some corrective actions. The network functions which compose the NSI support fast failure recovery and healing mechanisms, thus enabling automatic convergence of the affected network functions to a stable desired state. The results of the Self-Healing needs to be notified to the operator.

Examples of SON concepts applied to Active Antenna Systems are in [3GPP TR 38.865].

In the context of next generation networks responsibilities regarding operations must be clearly defined and assigned to roles. SA5 defines the following high-level business roles (see Figure 2-15):

- **Communication Service Customer (CSC)**: Uses communication services.
- **Communication Service Provider (CSP)**: Provides communication services. Designs, builds and operates its communication services.
- **Network Operator (NOP)**: Provides network services. Designs, builds and operates its networks to offer such services.
- **Virtualisation Infrastructure Service Provider (VISP)**: Provides virtualised infrastructure services. Designs, builds and operates its virtualisation infrastructure(s). Virtualisation Infrastructure Service Providers may also offer their virtualised infrastructure services to other types of customers including to Communication Service Providers directly, i.e. without going through the Network Operator.
- **Data Centre Service Provider (DCSP)**: Provides data centre services. Designs, builds and operates its data centres.
- **Network Equipment Provider (NEP)**: Supplies network equipment. For sake of simplicity, VNF Supplier is considered here as a type of Network Equipment Provider.
- **NFVI Supplier**: Supplies network function virtualisation infrastructure to its customers.
- **Hardware Supplier**: Supplies hardware.

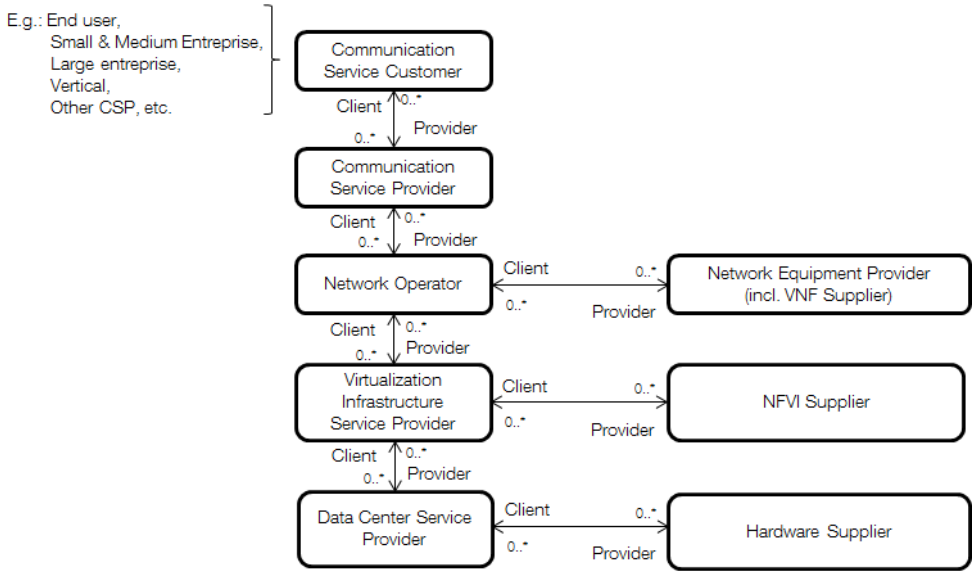


Figure 2-15: Functional model of business roles

The business roles identified by 3GPP can be related to 5G-MoNArch Stakeholder Model that is described in Section 2.1.5. The identified roles are mainly the same (Table 2-2). The Mobile Service Provider is, for 5G-MoNArch, one role that can be mapped with two 3GPP business roles. The Mobile Service Provider is both a Communication Service Provider and a Network Operator.

Table 2-2: Relationship between the 5G-MoNArch Stakeholder Model and the 3GPP Stakeholder Model

5G-MoNArch Stakeholder	**3GPP Stakeholder**
End User	Communication Service Customer
Tenant	Communication Service Provider
Mobile Service Provider	Communication Service Provider Network Operator
VNF Supplier	Network Equipment Provider
Virtualisation Infrastructure Service Provider	Virtualisation Infrastructure Service Provider
NFVI Supplier	NFVI Supplier
Infrastructure Provider	Data Centre Service Provider
Hardware Supplier	Hardware Supplier

Network slice related management functions: the following management functions are needed to manage the NSIs to support communication services:

- **Communication Service Management Function (CSMF)**: responsible for translating the communication service related requirement to network slice related requirements. Communicate with Network Slice Management Function (NSMF).
- **Network Slice Management Function (NSMF)**: responsible for management and orchestration of NSI. Derives network slice subnet related requirements from network slice related requirements. Communicates with the Network Slice Subnet Management Function (NSSMF) and Communication Service Management Function.
- **Network Slice Subnet Management Function (NSSMF)**: responsible for management and orchestration of NSSI. Communicates with the NSMF.

2.5.3 Baseline 5G-MoNArch MANO Layer

5G-MoNArch MANO layer takes 5G NORMA [NORMA D3.2] as the starting point, which will be described in the following. The baseline MANO layer defined in 5G-MoNArch is derived from the 5G-PPP Phase 1 projects, following the ETSI MANO Framework [ETSI GS NFV-MAN] and 3GPP SA5.

5G-NORMA has extended the ETSI NFV MANO architecture to support multi-service and multi-tenancy systems. Figure 2-16 shows the different blocks in the 5G-NORMA MANO layer.

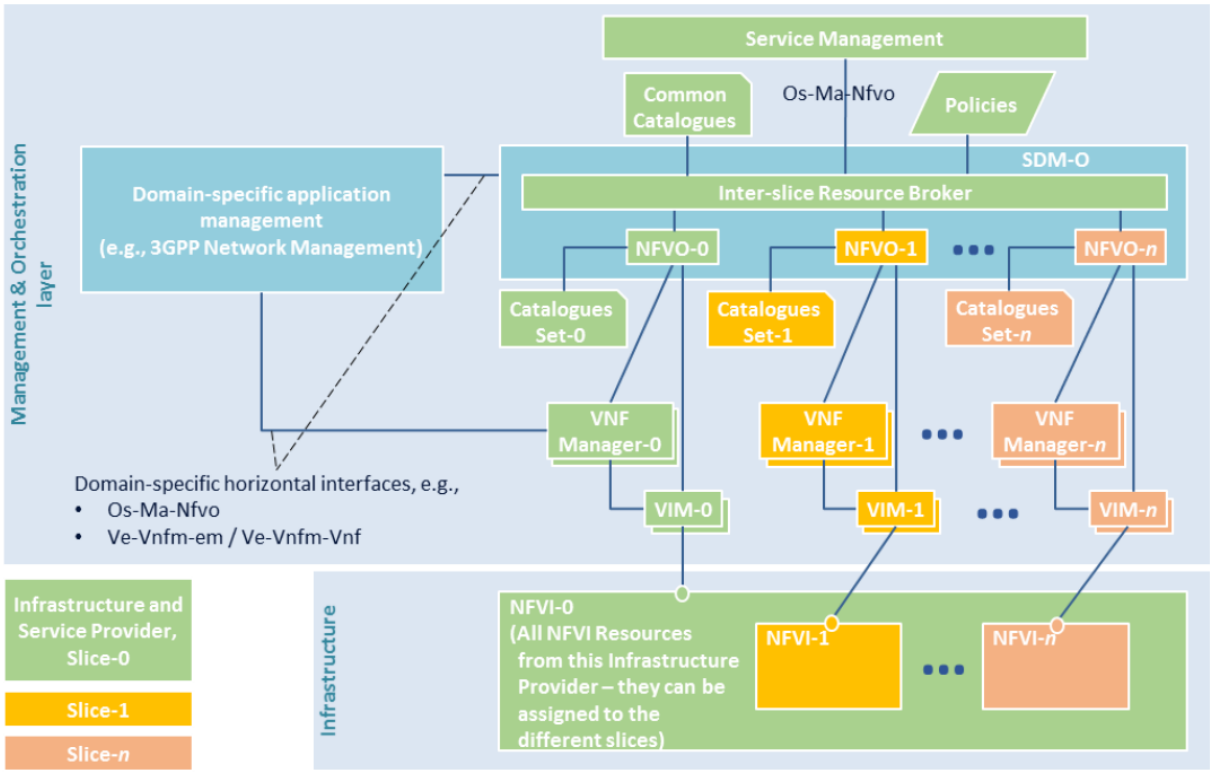


Figure 2-16: 5G NORMA Main Management and Orchestration Blocks

VIM, VNF Manager and NFVO modules are the same blocks as those defined by the ETSI NFV MANO specification, including functionality and reference points:

- VIM is responsible for controlling and managing the virtualised infrastructure (compute, storage and network resources), usually under one operator's infrastructure domain. Infrastructure domain refers to the infrastructure layer or part of it owned by a dedicated infrastructure provider (see Section 2.1.5 for the description of that stakeholder role model). Among the set of functions performed by the VIM are the following:
  - Control and manage the NFVI resources
  - Collect performance measurements and events
  - 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).
- The VNF Manager is responsible for the lifecycle management of VNF instances. Each VNF instance is assumed to have an associated VNF Manager. A VNF manager may be assigned to the management of a single VNF instance, or the management of multiple VNF instances of the same type or of different types. The functions performed by VNFM include:
  - Instantiation & termination of VNFs (lifecycle management).
  - VNFs scaling (up & down, in & out)
  - Updating or upgrading VNFs.
  - Configuration and event reporting.
- The NFVO is in charge of the network wide orchestration and management of NFV (infrastructure and software) resources, and realising NFV service topology on the NFVI. The NFVO manages and automates the distributed NFV Infrastructure. The NFVO has control and visibility of all VNFs

running inside the NFVI. The NFVO provides Graphical User Interface (GUI) and external NFV-Interfaces to the outside world (i.e. Business System Support (BSS) / Operations System Support (OSS)). 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 authorisation of NFVI resource requests.

The novelty provided by 5G-NORMA project consists on a new functional block called ISRB that has specifically been designed to manage and orchestrate resources allocation for network services and functions across different slices and multiple tenants. It is the main component to perform the multitenant and multiservice paradigms. It 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.

This ISRB, together with the above described NFVO, forms the SDM-O (Software Defined Mobile Network Orchestrator) functional block.

Apart from the described functional blocks, each NFV MANO stack instance can either work on a common set of catalogues for network services and NFs as provided and operated by the Infrastructure Provider/Mobile Service Provider or on a dedicated catalogue set as on-boarded by the tenant and certified by the provider.

Finally, the Service Management maps the service requirements as provided by the tenant to the appropriate network slice template. As a result of this mapping process, Service Management provides a network slice descriptor to the ISRB. As shown in Figure 2-16, the 5G-NORMA architecture provides the possibility to commission multiple NFV MANO stack instances, e.g., dedicated to a tenant or a network slice. The Service Management receives performance, fault, and configuration data about commissioned network slices from the ISRB. These data are used for performance reporting as well as accounting and charging towards the tenant.

5G-NORMA architecture is the baseline for 5G-MoNArch. As described in Section 2.1.4, Figure 2-3, 5G-MoNArch Management and Orchestration functions have been enhanced to support multi-tenant and multi-service networks by adding Cross-domain and Cross-slice Orchestration and Management functions.

Based on the depicted state of the art, 5G-MoNArch requires extending the orchestration and management algorithms to support:

- Cross-domain management & orchestration of mobile networks that will be able to orchestrate all kind of resources across the different domains, see Section 3.1.2. It is needed to define the east/west interfaces between ISRB of different domains, allowing the management and orchestration of E2E slices with resources belonging to more than one administrative domain.
- Efficient NF resilience and security mechanisms and network elasticity described in Sections 3.2.1 and 3.2.2 respectively.

2.5.4 Multi-tenancy and multi-service in the 5G-MoNArch MANO Layer

The multi-tenant- and multi-service-aware reference point in the 5G-MoNArch (MANO) plane is 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 or 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, cf. Figure 2-17.

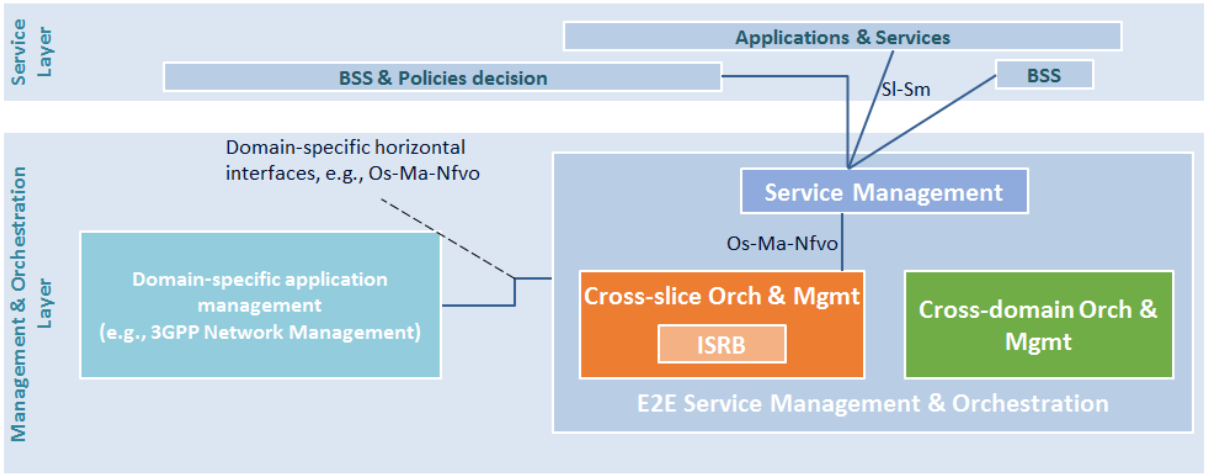


Figure 2-17: Interaction between MANO Plane and Service Plane

2.5.4.1 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 Sp-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 2-16 (in Section 2.5.3), the 5G-MoNArch 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-MoNArch network slice descriptor does not only contain information on control and UPFs, but also on MANO plane 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 [3GPP TR 28.801] and detailed in Section 2.5.2, lifecycle management is composed of 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.

In a first step, the ISRB 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 ISRB uses a combination of the resource commitment models as outlined in Section 2.5.4.2. Commissioning of the network slice control and UPFs is triggered by the ISRB 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, Figure 2-16 (in Section 2.5.3). 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., ISRB to Service Management), the ISRB 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 Plane. 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);

2.5.4.2 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 ISRB. 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.

To summarise, the approach described in this chapter is in line with the three fundamental concepts for 5G-MoNArch described in Section 2.1.4, namely split of control and user plane - described in Section 2.4, Support for E2E network slicing and Programmability. Management and orchestration provide the necessary management functions to create E2E slices and interact with the control plane as well as allow network programmability.

2.6 Physical Network Infrastructure and Topology

Whereas the 5G system considered for the 5G-MoNArch architecture design has been described from a functional perspective in Sections 2.2 to 2.5, latest considerations on physical network infrastructures and topologies as well as options for flexible orchestration of network functions (NFs) onto this infrastructure layer, are presented in the following.

Figure 2-18 provides a high-level view on the physical infrastructure layer expected for 5G mobile networks[4]. Please note that 5G is intended to cover also fixed-mobile convergence (FMC) [NGMN15], but as FMC is not in primary focus of 5G-MoNArch dedicated fixed network parts (especially in the access area) are not explicitly shown in the figure.

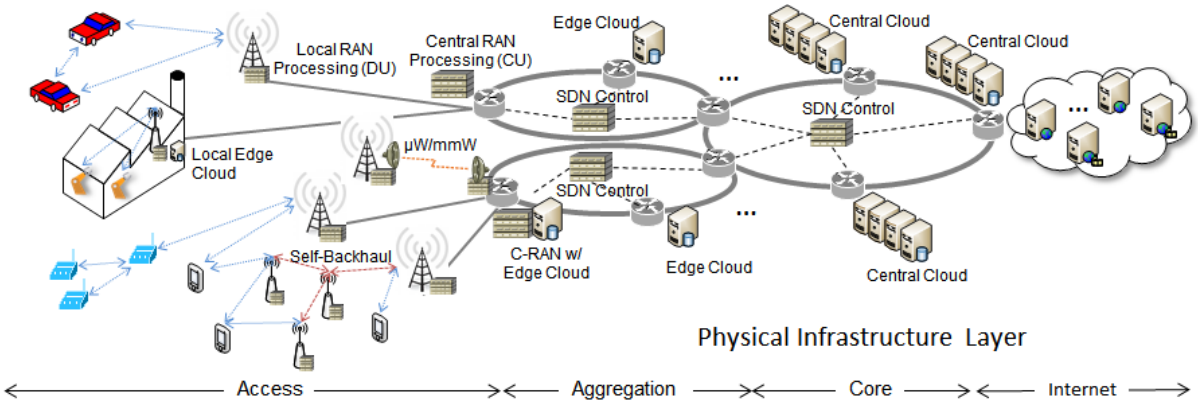


Figure 2-18: High-level view on physical infrastructure of 5G system considered within 5G-MoNArch (modified from [METIS II D2.4])

In contrast to former generations, the 5G infrastructure layer will increasingly provide computing power and data storage capabilities across its network components taking care of the feasibility of technical concepts like software defined radio (SDR), software defined networking (SDN) and network function virtualisation (NFV) [5GPPP16]. This will result in improved flexibility and programmability of network components with the help of mostly virtualised network functions (NFs) (especially in the CN) which can be flexibly placed according to requirements of different 5G use cases [3GPP TS 22.185] [3GPP TS 22.186] [3GPP TS 22.261] [3GPP TR 22.891]. One example is the orchestration of processing and application functions (incl. CN NFs) as near as possible to the antenna sites by using edge clouds to guarantee e.g. low service latencies. This may happen also in conjunction with cloud-based radio access network (C-RAN)) components which may serve for the centralisation of distinct radio access related NFs. C-RAN is also noted as V-RAN (virtualised RAN) in case of using highly virtualised RAN NFs in central RAN components. Edge clouds und C-RANs can be part of the operator-owned access network, but also part of a local, operator independent infrastructure (e.g., in a factory hall used by a vertical industry player) which again may be integrated into a larger wide area (logical) network (WAN) by a mobile service provider (MSP) offering network slices to its customers.

SDN and NFV approaches have a strong impact on future base station (BS) implementation in 5G. 4G eNBs can still be seen as monolithic blocks (except of well-known IQ sample based physical layer split into baseband unit (BBU) and remote radio head (RRH) components which are connected via fibre-based CPRI [CPRI15] or ORI interfaces [ETSI14-ORI]). 5G BSs (aka gNB in 3GPP terminology w.r.t. 5G New Radio (NR) air interface) will be able to be split into a so-called centralised unit (CU) and one or more distributed units (DU) [3GPP TS 38.300] [3GPP TR 38.801] (see Figure 2-19). Like "classical" C-RAN in 4G, NFs related to higher layer RAN protocol stack can be placed in the CU, whereas lower layer NFs are located in the DUs. With that so-called horizontal split gains from centralisation can be achieved, e.g. through common resource management (RM) and flow control [METIS II D2.4]. But that split also allows NFs to be flexibly placed into CU and DUs according to performance criteria like latency as well as to adapt the placement to the characteristics of the underlying x-haul (front-/mid-

[4] The figure shows only a simplified sketch and includes only a single infrastructure domain. Network slices, e.g., for continental-wide or global applications, may cover several domains and therefore may require a hierarchical model for the interrelations between a slice provider and the infrastructure domain owners.

/backhaul) transport network (TN) between CU and DUs [5GC] [5GX] (aka known as F1 interface in 3GPP NR terminology [3GPP TS 38.300]).

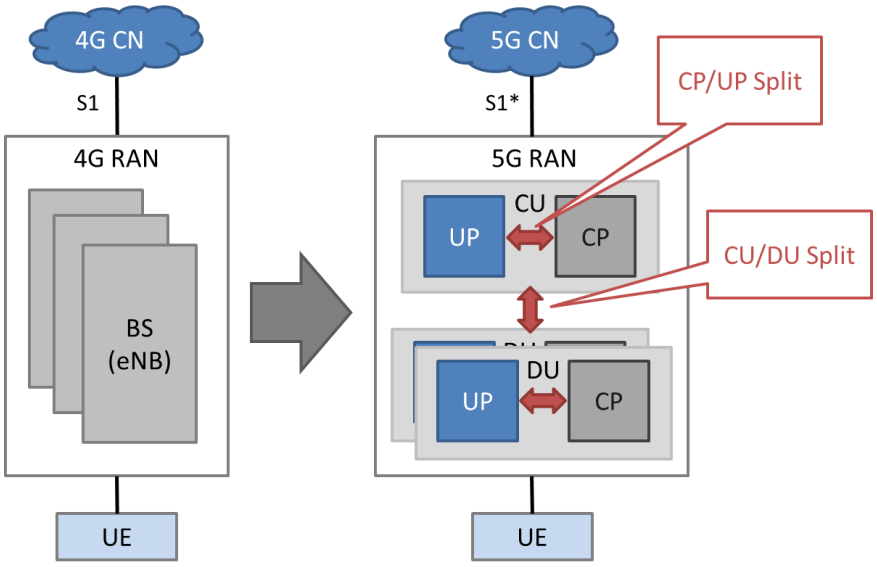


Figure 2-19: High-level view on architectural evolution from 4G to 5G RAN considering two-dimensional split in control/user plane (CP/UP) and central/distributed units (CU/DUs) [METIS II D2.4]

The impact of horizontal split options for the LTE protocol stack has been already considered in the EU FP7 project iJOIN [iJOIN D5.3]. 3GPP started with a similar approach during their initial study item on 5G NR. The different split options were classified according to Figure 2-20. Table 2-3 provides a high-level summary of the characteristics of each split option.

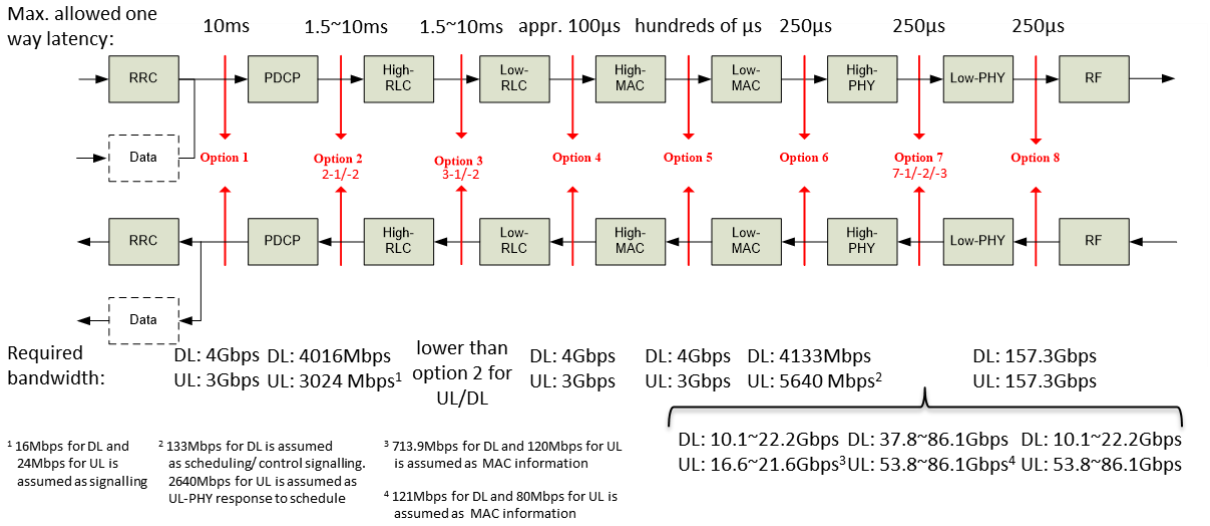


Figure 2-20: Horizontal functional CU-DU split options for the 3GPP radio protocol stack and their impact on latency and data throughput of the x-haul interface [3GPP TR 38.801]

Table 2-3: High level summary on characteristics of different horizontal CU-DU split option [3GPP TR 38.801]

	Opt. 1	**Opt. 2**	**Opt. 3-2**	**Opt. 3-1**	**Opt. 5**	**Opt. 6**	**Opt. 7-3 (only DL)**	**Opt. 7-2**	**Opt. 7-1**	**Opt. 8**
Baseline available	No	Yes (LTE DC)	No							Yes (CPRI)

Traffic aggregation	No	Yes							
ARQ location	DU			CU May be more robust under non-ideal transport conditions					
Resource pooling in CU	Lowest	In between (higher on the right)							Highest
	RRC only	RRC + L2 (partial)			RRC + L2	RRC + L2 + PHY (partial)			RRC + L2 + PHY
Transport NW latency requirement	Loose			FFS	Tight				
Transport NW peak BW requirement	N/A	Lowest	In between (higher on the right)						Highest
	No UP req.	Baseband bits					Quantised IQ (f)		Quant. IQ (t)
	-	Scales with MIMO layers						Scales with antenna ports	
Multi-cell/freq. coordination	Multiple schedulers (independent per DU)			Centralised scheduler (can be common per CU)					
UL Adv. Rx	FFS					N/A	FFS	Yes	
Remarks	Note 1				Note 2/3	Note 2	Note 2	Note 2		

- Note 1: Beneficial for URLLC/MEC (FFS)
- Note 2: Complexity due to separation of Scheduler & PHY processing
- Note 3: Complexity due to separation of Scheduler & HARQ

Figure 2-20 includes also some exemplary values demonstrating the requirements to the x-haul interface with respect to latency and bandwidth (throughput) for data transfer between CU and DU. Basic assumptions for the computation of those bandwidth values are given in Table 2-4.

Table 2-4: Underlying assumptions for data throughput values in Figure 2-20 [3GPP TR 38.801]

Items	**Assumption**	**Applicability**
Channel Bandwidth	100MHz (DL/UL)	All options
Modulation	256QAM (DL/UL)	
Number of MIMO layer	8 (DL/UL)	
IQ bit width	2*(7~16)bit (DL), 2*(10~16)bit (UL)	Option 7-1 Option 7-2 Option 7-3
	2*16bit (DL/UL)	Option 8
Number of antenna ports	32 (DL/UL)	Option 7-2 Option 7-3 (UL) Option 8

As it can be seen from Figure 2-20, centralisation of lower layer NFs generally increases the x-haul requirements in terms of throughput and latency as known from today's CPRI interface implementation. With 5G, those requirements may be further tightened because of, e.g., shortened transmission time intervals (TTIs), wider channel bandwidths and strongly increased number of antenna ports with Full Dimension (FD) or Massive MIMO [Björnson et al], as already assumed in Table 2-4. This is especially true for frequency bands above 6 GHz where CPRI-like interfaces would counteract wide-area centralisation due to hundreds of gigabits per second per carrier to be transferred between CU and DU

(see e.g., [METIS II D4.2] [5GPPP17]). Therefore, it is expected that there will be a co-existence of C-RAN (CU/DU) deployments with the classical fully distributed approach (D-RAN).

In addition to the horizontal split also a vertical split may be considered to allow separation of CP and UP NFs according to SDN principles (see e.g. [XRAN16]). Open standardised interfaces between CP and UP would enable operators to have consistent control over components and NFs from different vendors and allow to change or upgrade CP NFs without the need to replace in addition UP NFs often tightly coupled to CP in today's RAN implementations (resulting in significant cost savings). As CP-UP split is a rather straightforward approach for CN [3GPP TS 23.501] (see Section 2.2) and TN (x-haul, aggregation) [5GC] [5GX], it is much harder to be implemented in the RAN due to mentioned tight integration of CP and UP NFs in the protocol stack. Especially the required time synchronicity with air interface TTI framing in lower layers limits the feasibility of a fully centralised CP (i.e., of non-collocated processing) [NORMA D4.2] [METIS II D2.4].

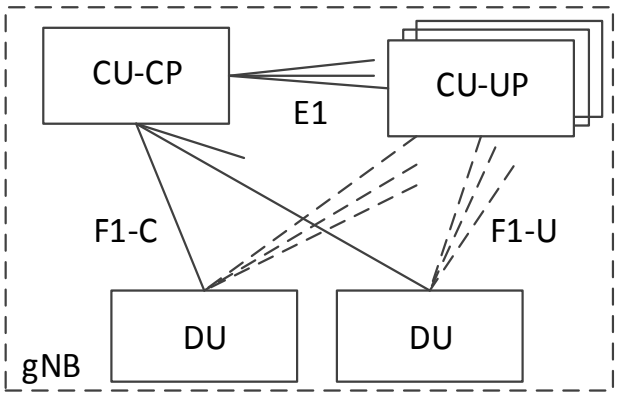


Figure 2-21: 3GPP gNB architecture w.r.t. CP-UP split under discussion [3GPP TR 38.806]

Figure 2-21 describes a possible architecture for a NR gNB as under discussion in 3GPP with following characteristics:

- A gNB may consist of a CU-CP, multiple CU-UPs and multiple DUs;
- The CU-CP is connected to the DU through the F1-C interface;
- The CU-UP is connected to the DU through the F1-U interface;
- The CU-UP is connected to the CU-CP through the E1 interface;
- One DU is connected to only one CU-CP.

Due to theoretically high degree of freedom in combinations of vertical and horizontal split options and the high effort to define open interfaces 3GPP is focusing in the initial standardisation phase of NR (Rel-15) only to a single higher layer horizontal split (HLS), equal to Option 2 below PDCP as shown in Figure 2-20 without explicitly considering also a CP-UP split. This Option 2 directly supports the dual-connectivity (DC) feature of LTE for combining LTE macro cells with NR macro or small cells in initial, so-called non-standalone (NSA) deployments [3GPP TS 38.300] [3GPP TS 38.401]. Centralised CP NFs in the CU will be related to e.g. RRC, RAN mobility, admission control, and high level inter-site and/or air interface resource coordination like eICIC (enhanced Inter-cell Interference Coordination), AaSE (AIV agnostic Slice Enabler), or dynamic traffic steering (see also Figure 2-10 in Section 2.3.2).

Other split options in NR might come with new study and work items for Rel-15 and beyond (see e.g. [3GPP-LLS] with respect to a study item for lower layer split (LLS)), but from a practical perspective it is expected that only a limited number will be finally standardised as open interfaces. In principle, RAN vendors are still in a position to implement additional proprietary interfaces to be usable in single vendor deployments, but this would highly limit the achievable flexibility for operators.

With respect to the x-haul (F1) interface there is some activity ongoing at IEEE on specification of so-called next generation fronthaul interface (NGFI) [NGFI15] covering user data, management, and control traffic. Two projects are active, one (P1914.1) working on a standard for packet-based fronthaul

transport networks and the other (P1914.3) on a standard for radio over Ethernet encapsulations and mappings (see [NGFI] for more details).

In addition, the vendor-based CPRI initiative has published a new so-called eCPRI interface description [eCPRI17] that will apply new split options within the PHY layer inclusive of flexible scalability to the UP traffic, so targeting a reduction of the transport bandwidth by a factor of 10. As shown in Figure 2-22 the eCPRI specification focuses on three different reference splits, two splits in downlink (DL) and one split in uplink (UL) (noted as I_D, II_D and I_U) which are addressed by 3GPP under Options 7-1/7-2/7-3 (see Figure 2-20). Any combination of the different DL/UL splits is possible. eCPRI will be – similar to NGFI – Ethernet- and IP-enabled.

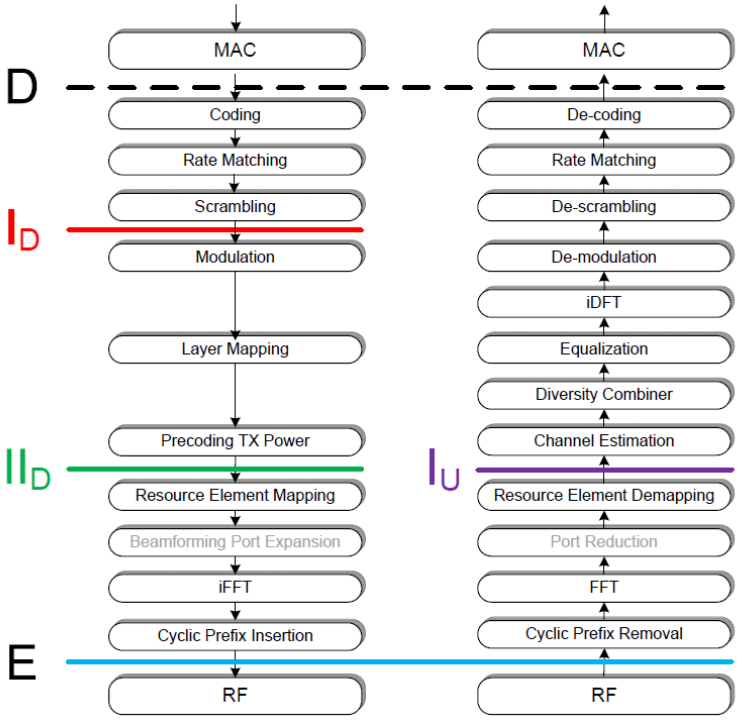


Figure 2-22: Horizontal splits supported by eCPRI [eCPRI17]

To support diverging requirements of 5G services and network slice types, respectively, all x-haul interfaces (in combination with CU/DU hardware/software) must provide the flexibility with respect to NFs the requirements entail. That means:

- Simultaneous support of different splits by the same (sub)network;
- Flexible allocation (in space) due to physical network infrastructure;
- Flexible allocation (in time) based on service requirements;
- Simultaneous support of different splits (per-UE, per-bearer, per-slice).

In addition, such split allocations are not static, but may also vary in time due to orchestration of new slices with corresponding NFs or re-orchestration of existing ones. A still open issue is the handling of required MANO layer functions in case of flexible F1 interfaces.

Due to cost reasons, not all antenna sites in a 5G network are expected to be connected via fibre (especially in rural areas), i.e., micro-wave (µW) and millimetre-wave (mmW) point-to-point (PtP) or point-to multipoint (PtMP) systems will still have their merits. Especially in mmW frequency range, new bands above 90 GHz beyond already used V- and E-Bands are under consideration (see e.g. [ETSI-mWT]). Within ONF's [ONF] Wireless Transport Project there were already several successful proof of concepts (PoC) demonstrating that SDN-based approaches are also applicable to wireless backhaul. With respect to small cell concepts in mmW bands (covering also moving cells like nomadic nodes (NNs) [METIS II D2.4]), it is intended that the 5G air interface specification should support self-backhauling, i.e., the wireless access to UEs and the wireless x-haul to infrastructure gateways (e.g., a transmission/reception point (TRxPs) of a 5G BS), will share common resources. mmW-based solutions

are especially of big interest in the fixed access area as alternative to fibre-to-the-home (FTTH), i.e. the so-called fixed wireless access (FWA). From an economical perspective, the applicability of FWA solutions will be strongly dependent on the considered environment (urban, rural, etc.) and the already existing fixed infrastructure situation.

In 5G access networks UEs will not only be directly connected to BS TRxPs, but also indirect communication will be possible via other devices. That means, a device may act as a relay to a TRxP (aka as cluster head, e.g., for sensor networks [METIS II D2.4]) or alternatively, devices will also directly communicate with each other without or with only limited interaction with the RAN infrastructure (device-to-device communication (D2D)). The latter one is especially addressed in vehicle-to-anything (V2X) communication supporting future autonomous driving [3GPP-22186].

With respect to relevant site locations in the physical infrastructure, Figure 2-18 provides only a rough high-level view as any detailed deployment will be strongly dependent on the already existing infrastructure topology of an operator and his migration strategy to a highly virtualised and softwarised 5G network. There may be some hierarchical approaches, e.g. for the aggregation sites where the traffic from different macro and small cell TRxPs will be aggregated (e.g. at central offices (COs) of operators). Those sites are especially interesting for the placement of C-RANs (CUs). Dependent on the environment and the resulting distance to the TRxPs (relevant w.r.t. latency aspects) a C-RAN site may cover between 10 to 100 cells. In countries like France or Germany typically several hundreds of COs per operator are existing, which may then be used for C-RAN implementation, probably in combination with dedicated edge cloud installations, whereas only less than about 10 central cloud sites would be part of a country-wide network.

C-RAN sites may be connected to each other via ring or meshed aggregation TNs on fibre basis, whereas tree structures are typically used for the links between the aggregation site and antenna/TRxP/DU sites. This has to be considered w.r.t. finally achievable reliability for 5G services, as such a link is a single point of failure, at least for customers served in the coverage area of the linked TRxP. To achieve high reliability values, different approaches have to be considered, e.g., multi-connectivity between different locations combined with overlapping coverage areas. In the FMC market combinations of fixed access and LTE in home gateways are already available (so-called Hybrid Access) which are also simply extendable to LTE and WLAN integration up to the UE, e.g. via multipath TCP (MPTCP) approaches (non-optimum OTT solution). TNs in aggregation and core already provide a high availability due to the implemented ring and meshed structures based on optical DWDM technologies, so no additional measure is expected to be needed to cope with 5G requirements.

Following the advanced cloudification and softwarisation in data centres, the RAN NF processing is in principle already feasible via pure SW functions. Due to performance and energy efficiency reasons, there may still be some so-called physical NFs (PNFs) that are coupled with underlying HW, e.g., DSPs or FPGAs (especially related to physical layer processing in the RAN protocol stack), so not all NFs in a system will be fully virtualised. Nevertheless, the other NFs may run on general purpose processor (GPP) HW and therefore can be dynamically deployed on different servers under the limitations set by service and other requirements (e.g., from a security perspective). Therefore, both edge cloud and central cloud data centres are in principle feasible to handle RAN NFs in virtual machines or containers in addition to already foreseen CN NFs and application functions. W.r.t. the HW coupling mentioned before it is to be noted that also in virtualised data centre environments HW acceleration features may be implemented [ETSI15-NFV]. Small edge clouds (aka cloudlets) carrying dedicated CN NFs and application functions may be also be implemented at antenna sites to serve certain low latency service requirements, but up to now there are no 5G service descriptions that would require such a costly approach. One exception may be the implementation in a local network domain, e.g., for a vertical industry player as mentioned before.

In summary, the infrastructure layer acting as a reference for 5G-MoNArch is consisting of three main types of deployment nodes (see also [NORMA D2.2]):

- Bare metal node:
  - Execution of PNFs with tight coupling between HW and SW platforms (in many cases SW is even highly embedded in HW).

- Typically used at antenna sites for processing of lower layer radio protocol stack (e.g. realising a DU or RRH).

- Edge cloud node:
  - Comprises a small, locally placed (i.e., within the access network close to or at antenna sites) collection of processing, storage, networking, and other fundamental computing resources.
  - Typically, the number of edge cloud nodes is at least one order of magnitude higher than the number of central cloud instances.
  - Deployment particularly in rather densely populated metropolitan, urban, and sub-urban areas.
- Central cloud node:
  - Hosting a significantly large collection of processing, storage, networking, and other fundamental computing resources.
  - Typically, only a few of them are found in a nationwide operator network.

The following features are to be considered for both cloud node types:

- Ability to deploy and run arbitrary SW (incl. operating systems and applications; limiting factor is the available processing, storage and/or networking power).
- Provisioning of virtualised resources based on NFV principles to execute VNFs and MANO functions.
- Edge clouds are generally exhibiting greater heterogeneity than central cloud in terms of utilised HW and hypervisors, geographical deployment, and topological structure.

It is to be noted that 5G-MoNArch has initially no fixed position w.r.t. applicable 5G network infrastructures and implementation options. Different alternatives will be considered during the techno-economic evaluations to be performed in WP6 of the project as part of the architecture verification and validation process [NORMA D6.1].

There are other ongoing activities evaluating deployment alternatives suitable for 5G. For example, the M-CORD (Mobile CORD) project as part of the CORD (Central Office Re-architected as a Data centre) initiative [CORD] considers the implementation of C-/V-RAN in operators' COs. M-CORD combines NFV, SDN, and the elasticity of commodity clouds to bring data centre economics and cloud agility to the telco CO. The reference implementation of CORD will be built from commodity servers, white-box switches, disaggregated access technologies (e.g., vOLT, vBBU, vDOCSIS), and open source software (e.g., OpenStack, ONOS, XOS). In Figure 2-23 the intended open reference implementation for 5G is shown. In the meantime, there is also a cooperation between M-CORD and xRAN w.r.t. the integration of the xRAN controller [XRAN16] into the reference platform.

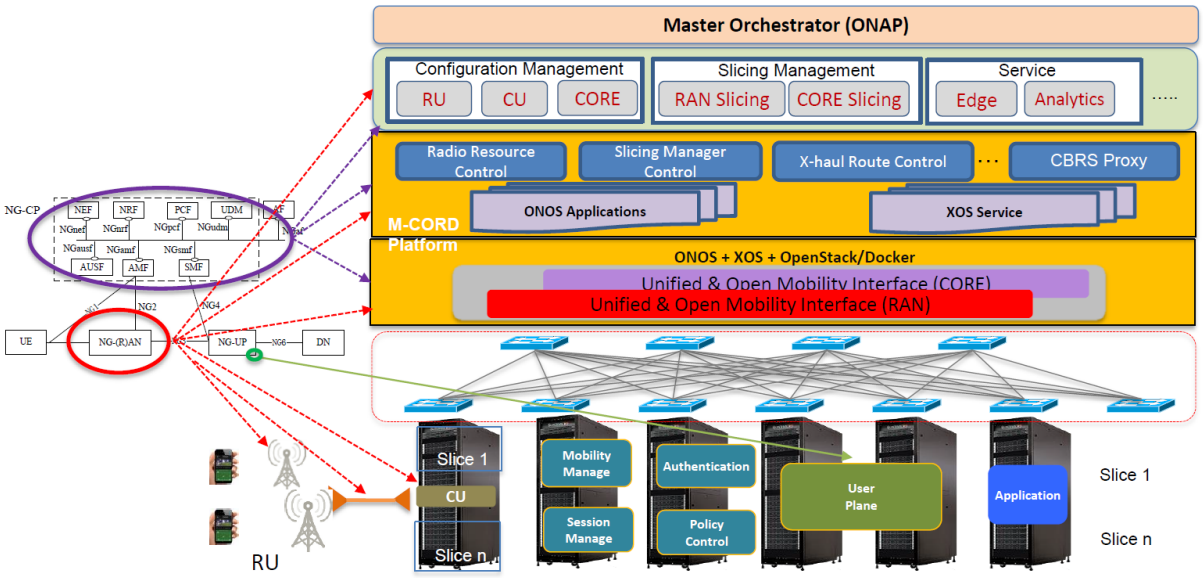


Figure 2-23: Open reference platform implementation of M-CORD for 5G [CORD]

The increasing traffic demand with 5G requires operators to find cost-efficient solutions towards the network deployment. Especially the high capacity radio access applying high frequency bands above 3 GHz will result in dramatically more sites. In addition, also improved network coverage and reliability will be required for certain 5G URLLC use cases. To avoid significantly increasing operational and infrastructure costs, network sharing between operators, especially for the RAN part, might be a measure to combat with the cost challenge. It will also allow viable solutions based on network slice provisioning across the infrastructure domains of the involved operators or using a neutral 3rd party host that may build and run the network and the operators act as tenants. Well-known legacy concepts are passive site sharing, active RAN sharing based on MORAN (Mobile Operator RAN) or MOCN (Multi-Operator CN), or roaming based approaches. With softwarisation in 5G and considering the RAN split architecture there will be new sharing options going beyond the traditional ones. Examples are C-RAN infrastructure sharing (where each operator uses its own CU VNFs), sharing of DU PNFs/VNFs only (but using separate CU VNFs), sharing of CU VNFs only (but not of DU PNFs/VNFs), sharing the network infrastructure based on the network slicing approach, taking care of novel stakeholder models as described in Section 2.1.5. The physical infrastructure shown in Figure 2-18 must support these new sharing models.

2.7 Summary and Positioning of Technical Domains within 5G-MoNArch Preliminary Reference Architecture

Chapter 2 has identified the most essential architectural concepts and components of the baseline 5G-MoNArch architecture. Especially, the architectural components for network slicing support are examined. Chapter 3 then aims at identifying the 5GS gaps in the baseline architecture that will be addressed by the 5G-MoNArch innovations. Accordingly, the innovations and future work can necessitate changes and extensions on the baseline architecture.

In this section, a brief summary of the technical areas is given, and their positioning within the 5G-MoNArch preliminary reference architecture is highlighted. The three fundamental design aspects of 5G-MoNArch architecture outlined in Section 2.1.4 are inter-related to the technical areas. An overview of such mappings of the technical areas to the three 5G-MoNArch fundamental design aspects is provided in Table 2-5.

- CN takes the latest specification from 3GPP SA2 as the basis and provides the descriptions of the most essential NFs and interfaces among NFs as well as UE and RAN. Further, CN architecture takes the modularisation and CP/UP split and provides network slicing support. A network slice selection framework is included, where slice-specific NFs can be determined. This framework enables slicing support in case of roaming, as well, and defines expected network behaviours in terms of features and services, which can also be tenant specific.
- RAN takes the latest specification from 3GPP RAN WGs as well as the most relevant consolidated outcomes from the 5G-PPP Phase 1 projects. This includes a new protocol stack sublayer, namely, SDAP on the UP, which enables a dynamic QoS framework on RAN level. In addition, the most relevant CP NFs are outlined, and RAN support for E2E network slicing is described, where slice specific and inter-slice NFs are exemplified. It is further highlighted that depending on the business needs of the tenants (e.g., different degrees of network slice isolation), different implementation options can co-exist.
- Centralised CP architecture describes the ISC and XSC of the centralised control layer and how these controllers interact with the MANO. ISC provides the mechanisms for controlling slice-dedicated NFs, while XSC provides mechanisms for controlling shared NFs and achieving the multiplexing gains.
- Network management and orchestration presents the latest status from 3GPP SA5 and describes the baseline 5G-MoNArch MANO, which takes the most relevant aspects from 5G-PPP Phase 1 projects, ETSI MANO and SA5. The baseline 5G-MoNArch MANO aims at multi-tenancy and multi-service support with a focus on automation and orchestration.
- Physical network infrastructure and topology depicts how the functional architectures described by the above-mentioned technical domains can be implemented onto the physical infrastructure. In this

regard, different functional splits options along with their specification needs are discussed, and the network features, which shall enable these splits, associated with the performance needs are highlighted. In addition, different deployment options are outlined considering the cloudification and softwarisation trends.

Table 2-5: Mapping of Technical Areas to 5G-MoNArch Fundamental Design Aspects

Technical Areas	**CP/UP separation**	**E2E Slicing support**	**Network Programmability**
CN	Separated CP/UP NFs	• Dedicated slice support functions • Modular NF design with Service Based Interfaces • E2E slice identifier	Modular design of NFs facilitates the management plane to configure the network according to the requirements of different use cases
RAN	Common CP/UP interfaces	Implementation options of RAN part of E2E network slicing	CP/UP separation enables the programmability of RAN by aligning with the concept of SDN
Centralised CP architecture	Controllers are the responsible of performing CP/UP separations	Two controllers are defined according to their involvement in the intra or cross slice operation	I-APP and X-APP are indeed different software components that implement the programmability.
Network management and orchestration	Orchestration framework interacts with CP elements ISC and XSC.	• MANO layer provides inter-slice functions and dedicated (intra-slice) functions that support E2E network slicing. • Network slice lifecycle management	• Resource allocation and communication with the controllers. • SON algorithms applied to network slices
Physical network infrastructure and topology	Infrastructure layer supporting CP-UP split (vertical split) in RAN, CN, and TN	Infrastructure layer supporting programmability and flexible placement of CP/UP VNFs as well as configurability of PNFs to realise network service chains for dedicated slices	5G infrastructure layer with increased computing power and data storage capabilities across its network components taking care of the feasibility of technical concepts like SDR, SDN, and NFV

3 Overview of 5G-MoNArch Innovations and 5GS Gap Analysis

Chapter 2 established a basis for the architecture of 5G-MoNArch, by highlighting the most relevant aspects coming from the state of the art and indicating the fundamental concepts and components of the 5G-MoNArch. In this chapter, the architectural gaps with respect to 5G objectives are identified. Besides that, how these gaps will be addressed by the 5G-MoNArch innovations will also be detailed here.

5G-MoNArch will extend this basis with five key innovations: three enabling innovations contributing to the baseline architecture (cloud-enable protocol stack, inter-slice control and management, and experiment-driven optimisation), and two functional innovations which correspond to specific network slices (secure and resilient network functions, and resource-elastic virtual functions). The enabling innovations support the operation of network sliced 5G networks, while the functional innovations are more specific functions required when deploying network slices with particular requirements (resilience and security, and resource elasticity).

To validate the feasibility of the whole architecture, two architectural instantiations will be deployed as testbeds. One testbed will reside in the Hamburg sea port, while the other will run in a touristic city.

This chapter is structured as follows:

- Section 3.1 discusses the enabling innovations of the 5G-MoNArch architecture
- Section 3.2 follows with the functional innovations
- Section 3.3 provides a summary of the gap analysis and the relationship between those gaps and the innovations
- Section 3.4 discusses the planned architectural instantiations of the 5G-MoNArch architecture

3.1 Enabling Innovations

3.1.1 Cloud-enabled Protocol Stack

One of the key concepts of the 5G-MoNArch architecture is the flexible function decomposition and allocation [Sabella et al]. This concept builds on the work carried out by the 5G-PPP Phase 1 projects for the 5G mobile network architecture which, relying on *orchestration and virtualisation* technologies, decouples mobile network functions (NFs) from the underlying hardware infrastructure and enables their flexible placement within the different nodes that conform the physical network. The entire network, comprising of edge and core nodes in different locations, thus becomes a large "*telco cloud*," where NFs can be appropriately located depending on the requirements of the associated service. Building on this concept of Phase 1 projects, 5G-MoNArch will implement a flexible execution platform that builds on MANO and NFV as enablers for flexible function allocation.

While 5G-PPP Phase 1 projects have defined the concept that enables the flexible allocation of NFs, they have applied this concept to the LTE RAN protocol stack that is not necessarily optimised for this purpose (see Section 2.3).

As a consequence of the LTE design of the protocol stack, there are *inter-dependencies between the NFs co-located in the same node* (**GAP #1**), which must be considered to not harm the performance of the protocol stack. In fact, a protocol stack layer is composed by several NFs, which interact and depend from each other and exchange signalling with NFs that belong to other layers. Indeed, "traditional" protocol stacks have been designed under the assumption that certain functions reside in the same (fixed) location and, while they work close to optimality as long as such NFs are co-located in the same node, they do not account for the possibility of placing these NFs in different nodes.

To deal with these challenge, 5G-MoNArch aims to design a new RAN protocol stack tailored to fully exploit virtualisation and orchestration techniques. This new protocol stack is referred to as "*cloud-enabled protocol stack".*

The 5G-MoNArch cloud-enabled protocol stack will be based on two innovation elements, which are presented in the following, the **Telco cloud-aware protocol design** and **Terminal-aware protocol design**.

In 5G-MoNArch, the mapping of NFs to nodes follows a different paradigm: as discussed in the following sections, the 5G-MoNArch architecture provides the flexibility to shift NFs to the nodes that better fit the specific requirements of each service (**GAP #2**). As a result of this, NFs that were traditionally co-located in the same node (**GAP #1**) may now be placed on different nodes. More specifically, the aim is to design a flexible network that simultaneously supports different protocol splits (for instance per-UE, per-bearer, and per-slice). In addition, the NF deployment can be adapted in time and space depending on the momentary availability of network resources and service requirements.

The problem is that traditional protocol stacks are not well adapted to allow such a flexible placement of NFs. Indeed, placing certain NFs with heavy inter-dependencies in different nodes may incur very high overheads or may simply not be possible. This poses significant constraints on the flexibility of placing NFs, which compromises the overall gains obtained from the flexible function allocation.

One example of logical dependencies within the stack is the recursive dependencies between Modulation Coding Scheme (MCS), Segmentation, Scheduling, and RRC. These functions depend on each other and require close synchronisation among each other for their operation.

To overcome the problems related to the inter-dependencies between NFs in the protocol stack, one of the key innovations of 5G-MoNArch is the redesign of the RAN protocol stack with the goal of leveraging the benefits of the flexible function decomposition and allocation. Specifically, 5G-MoNArch will focus on the design of **Telco cloud-aware protocol design** adapted to its execution in a cloud environment. With the **Telco cloud-aware protocol design,** the aim is to relax and (as much as possible) remove the logical and temporal dependencies between NFs, with the goal of providing a higher flexibility in their placement.

In addition to the logical dependencies, traditional protocol stacks also impose stringent temporal dependencies, such as e.g. for Hybrid Automatic Repeat Request (HARQ). This is a lower layer protocol that requires the receiver to send immediate feedback informing of the decoding success of a packet transmission. In the current LTE stack, the time between the reception of a packet and the indication of the successful decoding is 4 ms, which forces the decoding function to be located very close to the radio interface and thus limits the possibility of centralising such functionality.

From a research perspective, there has been substantial work devoted to the re-design of radio access functions to increase the flexibility in their placement within the network cloud (see Section 2.6). For instance, [Rost and Pravad] proposes a novel HARQ scheme, called Opportunistic HARQ, which decouples the decoding from the sending of acknowledgements, which provides a high degree of flexibility in the placement of decoding and other related functions. Similarly, [Fritzsche et al] proposes a robust adaptation scheme that can cope with imperfect channel state information (CSI) and thus allows for centralising this functionality in deployments where long backhaul latencies cause CSI imperfections. While there have been other additional proposals focusing on the re-design of specific functions of the protocol stack to enable their flexible allocation, to the best of our knowledge 5G-MoNArch is the first to completely re-design the entire radio access protocol stack to enable the flexible allocation of all functions and facilitate their orchestration.

To enable a fully flexible functional split within a cloud-enabled protocol stack, the role of user terminal needs to be revisited. Such terminal-aware protocol design enables highly distributed deployments that can provide a faster adaptation and reconfiguration of the NFs.

As outlined in Section 2.1, user terminal has transitioned into more prominent roles in line with new developments in standards. This has enabled D2D relaying solutions since Release 13 [RP-150441] and is currently being further enhanced (feD2D) in Release 15 specifications [3GPP TR 36746] via emerging concepts like Group Handover (GHO) to improve remote UE reachability, to support efficient traffic differentiation, signalling, and service continuity at a controlled level of device complexity and power consumption on linked UEs. In this manner, the contexts of remote and relay UEs can remain collocated in the network. This will increase the time available for handover execution, reduce the risk of handover failure, and result in more accurate resource allocation at the target cell.

The above D2D framework combined with modular NFs provides a platform to further push costly NFs beyond the network edge to *save energy or to off-load resource demanding tasks* (**GAP #4**). The key is to find optimal balance between centralised NFs and distributed ones towards the edge (based on

Terminal-aware protocol design as above). This, in turn, may enable faster adaptation and reconfiguration of certain NFs (including mobility) in agreement with cyber foraging concept [Yang et al] already established in the literature in the area of pervasive computing (**GAP #5**). In this manner, relay nodes (as network surrogates) act as the last stand of "local" computing beyond the edge. This leads to a truly organic network with self-sufficient sub-networks connected to the umbrella network via single anchor UEs.

3.1.2 Inter-slice Control and Management

Network slicing is one of the key aspect of the 5G-MoNArch project. Network slicing has a strong impact on the RAN design, both in the UP and CP. As outlined in Section 2.1, 3GPP SA2 assumes the logical isolation of the slices and has defined per-slice policy management to complement the QoS management framework from 4G system. Although the fundamental blocks for inter-slice coordination are identified for 5GS (see, e.g., Section 2.3.2), *E2E cross-slice optimisation is not fully supported* (**GAP #6**). Indeed, a new design paradigm is needed to allow the simultaneous operation of multiple network slices, each with tailored access functions and functional placements to meet their target KPIs. While each network slice is considered as a fully operational (logical) network on its own, multiple network slices are operated on a common physical / virtualised infrastructure, which requires specific inter-slice control and management functions. Since 5G-MoNArch is targeting to design and develop network control functions to achieve a flexible and programmable inter-slice control and management framework that can be used to realise multiplexing gains across slices while guaranteeing slice-specific SLAs, this section examines in detail the existing proposals and gaps in current standards and, scientific literatures.

Inter-slice control and resource management

Inter-slice resource management, aka multi-slice resource management, is very important for improving the system efficiency, especially on shared infrastructure resources, which is a means for cross-slice optimisation. The inter-slice resource management thus factors in the slice SLAs, e.g., to adapt the instantaneous radio resource allocation. In addition to the slice-adaptive radio resource allocation, slice awareness can be extended to the so-called hard network resources, namely, wireless access nodes, particularly self-backhauled dynamic small cells. That is, the slice support may not only include the conventional radio resources like time and frequency resources, but it can also include the adaptation of the network topology considering the dynamic small cells available in a certain region. Accordingly, the slice-adaptive resource control shall also consider the changing radio topology including different access node types, e.g., micro-cells, pico-cells, relays, and vehicular nomadic nodes (VNNs).

A VNN is a low-power node integrated into vehicles (e.g., as part of a car sharing or taxi fleet), which can take the form of unplanned small cells [METIS II D5.2] [Bulakci et al]. Within the framework of the 5G dynamic radio topology as illustrated in Figure 3-1, VNNs can be activated and deactivated based on the traffic demand. Moreover, the dynamic radio topology can also include unplanned small cell deployments, which, e.g., can perform self-backhauling. In addition, the needed flexibility on the wireless backhaul link can be reached by employing a relaying functionality. A fixed relay can be typically deployed as fixed radio frequency (RF) amplify & forward (AF) /repeater or layer 3 (L3) decode & forward-(DF) node [3GPP TS 36.300]. In this context, *the functional operation of small cell networks is fixed* (**GAP #3**) and does not change relative to service requirements or the location of the small cell. That is, the functional operation and the associated operation mode of the small cells based on the pre-determined functional operation remain fixed. This can also incur higher operational expenditure (OPEX), when the network is planned for the highest or peak service requirements.

Furthermore, based on the assumption (as stated in Section 2.3.1) that each slice can have its own RRC functions and configurations, the UL/DL configuration of the operation of small cells (e.g. TDD pattern) may depend on the slice requirements (e.g. KPIs, traffic demand and characteristics) and the deployment of small cells. In this direction, in [SSS+16] the development of dynamic adaptation of TDD patterns was proposed; however, *this focused mainly on fixed deployments and pre-defined allocation of resources to slices* (related to **GAP #12**).

With respect to **GAP #3** and **GAP #6**, one main disadvantage of fixed small cells is, thus, the aforementioned lack of flexibility which is essential in 5G systems, where slice-awareness and 5G tight key performance indicators (KPIs) can necessitate on-demand flexible small cell operation.

Furthermore, fixed functional operation and fixed small cell cannot flexibly adapt to changing service (and traffic) requirements. Moreover, VNNs (as particular case of small cells) can be positioned at different parts of the cells; thus, the optimum functional operation in terms of performance changes based on the location and the associated channel link qualities. In addition, slice awareness can necessitate the dynamic configuration of small cells (e.g. TDD patterns) to meet slice-specific KPIs on demand. Consequently, **slice-aware functional operation is needed addressing GAP #3 and GAP #6** so that 5G RAN can adapt to network changes while fulfilling the requirements of different network slices**.** Accordingly, **inter-slice resource management addressing GAP #12** can be considered as one component of slice-aware functional operation, e.g., to allocate radio resources to meet slice-specific SLAs.

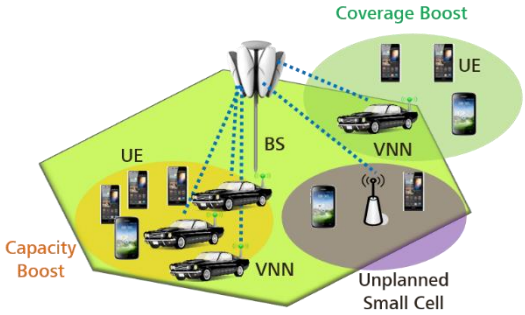


Figure 3-1: Flexible network deployment example based on dynamic small cells

The flexibility supported by the network architecture enables the network slicing: Ability to (re)configure the network instances. Applying this principle in the core requires only to consider the infrastructure resources. However, slicing the RAN is more challenging since it requires to handle both the infrastructure and radio resources. For that slicing in RAN is highly dependent on the way the RAN stack is split and configured.

The previous configurability requires the NFs to be parameterisable and in various cases to have multiple functions implemented. In particular, depending on the actual slice instantiation, a different instance of a NF may be used (e.g., implementation of various coding mechanisms, where depending on the characteristics of the slice, the most proper coding mechanism is utilised). In certain extreme cases, some functions can be totally omitted (e.g., ciphering in header compression can be omitted in mMTC and/uMTC scenarios) [Silva et al].

However, according to [3GPP TR 38.801] "the Xn interface shall be future proof to fulfil different new requirements, support new services and new functions". The latter implies that the RAN should:

- be able to introduce new functionalities for meeting new requirements coming from new UCs
- not be vendor specific for enabling the future proof nature of the 5G system. This statement stands also for the NG-C, and NG-U.

To support the previous aspects, new mechanisms should be developed within the **slice-aware functional operation** framework for:

- configuring the RAN protocols, and
- introducing new functionalities in the RAN.

Inter-slice context management

Isolation of the network slices is one of the essential requirements. As discussed in Section 2.3.1, the depth of the slicing influences the isolation means; the network slices may be isolated physically (e.g., slice-specific access node elements in case of public safety and slice-specific spectrum allocation) or logically (e.g., slice-specific NFs). Even in case of slice isolation, there can be some context information that can be shared among network slices. For instance, the UE-related context in case the UE is connected to multiple slices, such as UE location/mobility pattern, can be utilised by all the network slices associated with that UE. Such context can include information about the shared infrastructure, such as, the load/failure of common network functions and network connections. The exchange of such

information between different network slices can be helpful for the optimisation the performance of the 5G system (5GS).

Currently, with regards to **GAP #6**, 3GPP SA2 only defines the context processing/sharing between the NFs in the same network slice, i.e., NetWork Data Analytics (NWDA) [3GPP TS 23.501]. The utilisation of common context information among network slices depends on the isolation level and the customised NFs per slice. In case of high logical separation (e.g., high customisation level per network slice via substantial use of slice-specific NFs), parallel operation of control functionalities in these network slices associated to a UE can increase signalling cost along with CP latency. Accordingly, **inter-slice context sharing and management shall be enabled** where the context management mechanisms cross network slices shall aim:

- To identify context information that can be commonly utilised by network slice instances,
- To utilise context information available in one network slice instance to improve the functionalities in another, and
- To minimise the signalling cost and control plane latency due to parallel-running control functionalities in multiple network slice instances.
- To improve or optimise slice-specific KPIs;
- To improve or optimise UE-specific KPIs;
- To aid network slices reconfiguration, management, troubleshooting;
- To aid slice-specific and inter-slice features, e.g. traffic splitting, traffic steering, traffic switching, load balancing, and network slice scaling IN/OUT.

In many use cases, multiple slices may be used. For instance, in AR/VR cases, control signalling and some sensor data like UE position/header direction are time critical and need to be transported via URLLC slice. While the perceived data by the user are transported via eMBB slices. Since multiple network slices work together for one application, correlations exist between the needed performances of different network slices. In this sense, inter slice performance management and coordination using context sharing becomes especially important for the success of related business use cases.

Some context information could be common for logically isolated slices. For example, the UE related context in case of one UE is connected to multiple slices, such as UE location/mobility pattern, etc. E.g., the context from the shared infrastructure such as the load/failure of common network functions/network connections, etc. The exchange of such information between different network slices is helpful for optimisation the performance of the 5GS. While current SA2 only defines the context processing/sharing between the network functions in the same slice, i.e., NWDA NetWork Data Analytics [23.501], further enhancement is needed for context sharing cross network slices and maybe also from the infrastructure.

Following the conventional research roadmap of mobile networks, the research domain is split into (R)AN and CN. However, with the advancing of NFV and network cloudification in the 5G era, the boundary between (R)AN and CN becomes blur. Some network functions are moved to the edge (e.g., C-RAN cloud) to reduce the E2E latency of the applications. Some conventional CN mobile network management functions can be implemented by RAN (e.g., RAN based UE reachability/paging). In these cases, the context sharing between RAN and CN becomes extremely important. However, 4G mobile network provides only general RAN/CN context exchange over S1 interface which is not slice specific. Since the mapping of RAN configurations and CN slice can be complicated (e.g., one E2E slice with the same CN slice can be deployed at different site with different RAN configurations, one RAN configurations can apply to multiple E2E slices which are mapped to different CN slice.), the per slice context exchange between RAN and CN is not straight forward.

Terminal analytics-driven slice selection / control

As outlined earlier, NextGen protocol and reference points are defined for each NF. Such NFs can be implemented either as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, or as a virtualised function instantiated on an appropriate platform. On the other hand, separation between control and user planes guarantees each plane resources to be scaled independently. This allows User plan functions (UPFs) to be deployed separately from control plane

functions in a distributed fashion. UPFs may be deployed very close to UEs to shorten the Round Trip Time (RTT) between UEs and data network for some applications requiring low latency.

From slicing perspective, above CP and UP separation enables one common control network functions (CCNF) with multiple slice-specific CP/ UP NFs per UE [3GPP TR 23.799].

Here, UE intelligence is needed in slice selection and control. In effect, the UE may be pre-configured with Network Slice Selection Assistance Information (NSSAI) as also outlined in 3GPP SA2 studies. The NSSAI can be standardised and shared across Public Land Mobile Networks (PLMNs), or it can be specific per PLMN. The Configured NSSAI is a NSSAI configured by default in a UE to be used in a PLMN before any interaction with the PLMN ever took place. If the UE doesn't store any NSSAI for the ID of the PLMN that the UE accesses, the UE provides no NSSAI in RRC and the RAN sends the signalling to a default CCNF.

During the initial attachment, the NSSAI is used by RAN as input to select the CCNF. Then, Network Slice Selection Function (NSSF) in CCNF selects the network slice instance based on NSSAI, UE subscription data and other information available (e.g. UE capabilities, SLA information or local configuration). The Accepted NSSAI is the NSSAI used by the UE after the PLMN has accepted an "Attach Request" from the UE. The "Attach Accept" message includes the Accepted NSSAI.

The UE may cause the network to change the set of network slices it is using by submitting the value of a new NSSAI in a mobility management procedure. However, the final decision is up to the network. This will result in termination of on-going PDU sessions with the original set of network slices. Change of set of slices used by a UE (whether UE or Network initiated), may lead to CCNF change subject to operator policy.

From QoS framework perspective, NextGen will support flow-based QoS for better flexibility. QoS Flow is the finest granularity for QoS treatment in the NG System. User plane traffic with the same Next Generation Interface 3 (NG3) marking value within a PDU session corresponds to a QoS flow. This enables flexible mapping of QoS flows to the Data Radio Bearers (DRBs) in the RAN. Here, IP flows are mapped to QoS flows (at UPF in the core or UE) and QoS flows are mapped to DRBs at AS layer (either within RAN or UE).

Another important development is related to Reflective QoS where UE creates a new derived QoS rule for UL when it receives a DL packet for which Reflective QoS is indicated by the network. UE can perform UL rate limitation on PDU Session basis for non-GBR traffic and on QoS Flow basis for GBR traffic.

Moving beyond the current developments in NexGen specifications, UE may have more prominent role (to address **GAP #6**) via **terminal data analytics- driven slice selection or QoS control**. This is also in line with utilisation of Network Data Analytics (NWDA) as introduced in recent standard developments to optimise mobility decisions.

Inter-slice management and orchestration

In 5G-MoNArch as described in Figure 2-3, there are two set of functionalities located in the control layer (ISC and XSC) and MANO plane (cross-slice and cross-domain Orch & Mgmt) to implement the management and orchestration of E2E network slices while respecting their service specific KPIs. Since most of those management and orchestration technologies are inherited from IT world, adapting such technologies in the telco domain without key performance degradation is the greatest challenge (to address **GAP #5**). The main focus of 5G-MoNArch with respect to **inter-slice management and orchestration framework** involves: (i) **the identification of NFs that can be shared between slices,** (ii) **identification of functionalities and decisions that can be deployed in the (inter slice) control layer (ISC) to monitor and re-configure shared VNFs to maximise the overall resource utilisation, and improving QoS/QoE addressing GAP #2 and GAP #12 within inter-slice resource management,** (iii) **analysing the current SotA SDN/NFV management frameworks.**

Regarding the inter-slice control, although there are variety of SDN-C frameworks and approaches, none of them is designed with the focus on manging next generation mobile networks [IETF SDN] [Salman et al] [SDxCentral] [ONOS]. For example, though one of the main stream SDN-C frameworks such as ONOS [ONOS] provides distributed and clustering operation (cf. Figure 3-2) for network scalability, fault-tolerance and resilience, it is developed especially with the focus on management of fixed transport

networks., The adaptability of such solutions into mobile network infrastructure requires further study, especially on the extension of functions, protocols and algorithms for performance improvement (to address **GAP #5**). Since the state-of-the-art orchestration frameworks and software are designed and developed with more focus on deployment and life cycle management of VNFs in the traditional cloud computing domain (non-telco cloud) [Mirantis OpenCloud], there are limitations in such frameworks to be adapted directly into the telco domain to satisfy the required SLAs for various use cases (to address **GAP #5**).

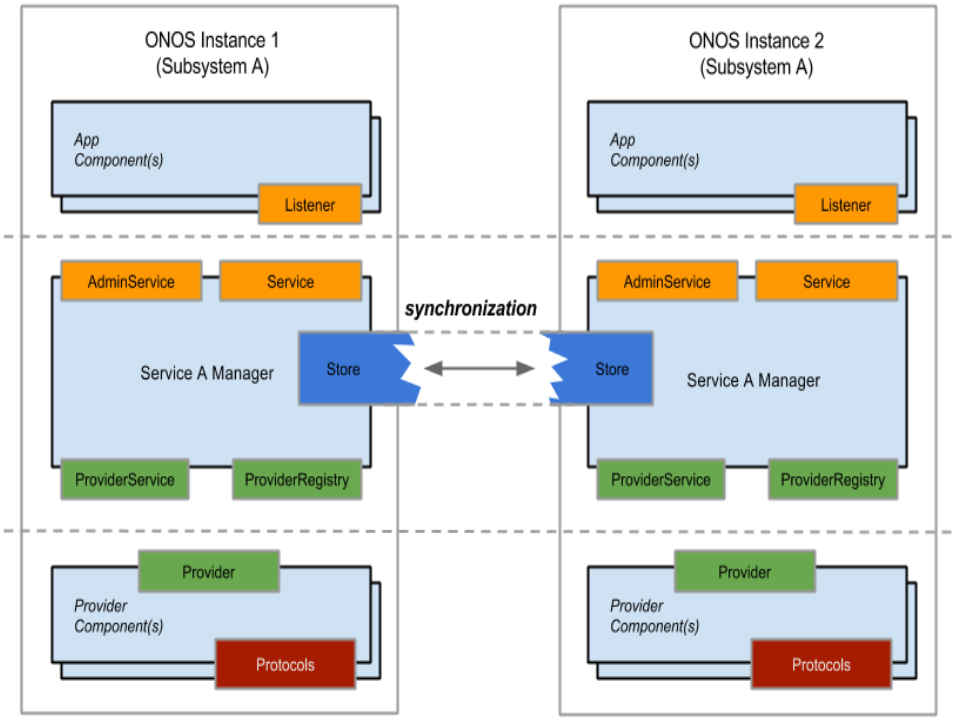


Figure 3-2: Distributed Clustering Architecture in ONOS framework

The 5G-MoNArch architecture defines the elements that take care of inter domain automation and E2E management and orchestration (Figure 2-3). SA5 has completed the specification phase for automation of networks including VNF and started the specification phase to define the management and orchestration architecture for 5G networks.

SA5 has specified the management architecture related to automated VNF lifecycle management, defining the interaction between MANO and the Network management as described in Figure 3-3.

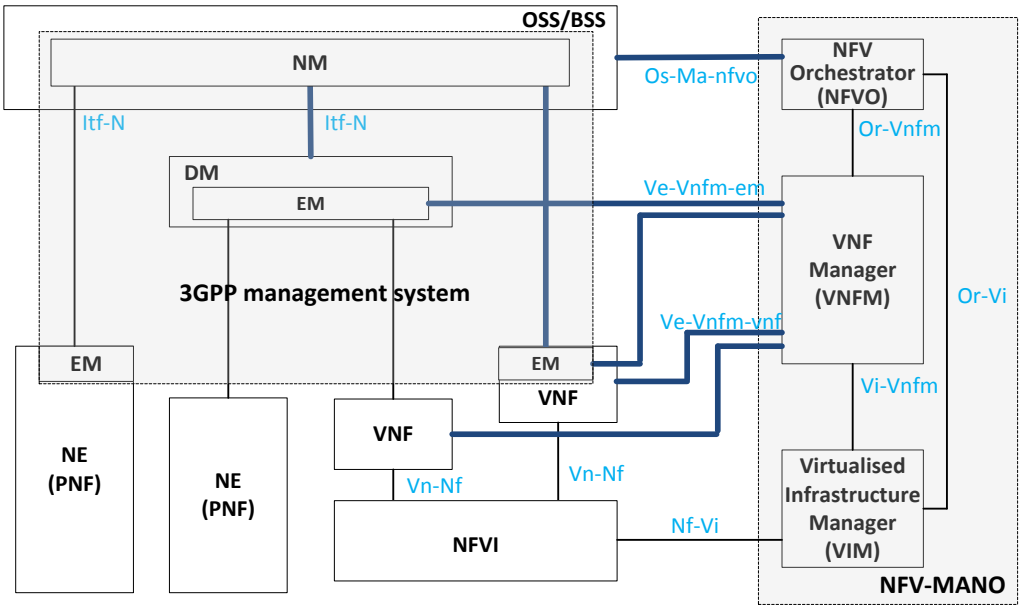


Figure 3-3: 3GPP Management System and MANO relationship

This representation is similar to what 5G-MoNArch foresees for a specific domain. 5G-MoNArch architecture for Management and Orchestration defines an E2E Service Management & Orchestration layer that takes care of both service and network requirements. The service requirements are translated into network slice requirements that are managed by the Cross-slice and Cross-domain Orchestration and Management functions. 3GPP SA5 has not yet defined an architecture for 5G networks orchestration. The study done on this topic foresees new management functions for Network Slice and Network Slice subnet management and a new network function for slice selection (NSSF: Network Slice Selection Function). 3GPP SA5 does not split these new functions into Cross-slice and Cross-domain. In 3GPP SA5 view the Network Slice, because it is an E2E concept, is by default cross-domain. The Network Slice Management Function takes care of every kind of slice, in different domains and with or without shared Network Slice subnet (cross slice). In 3GPP SA5 view, the Customer Service Management Function takes care of service requirement and translates them into network requirements managed by the Network Slice Management Function and by the Network Slice Subnet Management Function.

According to [3GPP TR 28.801] NFs can be shared among network slices, this is one of the main reason of network slices subnets definition. Being the NS an E2E entity that fulfils a customer service, a NS cannot be used as a component to build up a different service. To optimise network deployment and management, sharing NFs is an important requirement that has been implemented by 3GPP management architecture using the network slice subnet entity. A network slice subnet is not E2E, it aggregates NFs and it can be shared among network slices as describes in Figure 2-14 (NS Information Model) and in Figure 3-4.

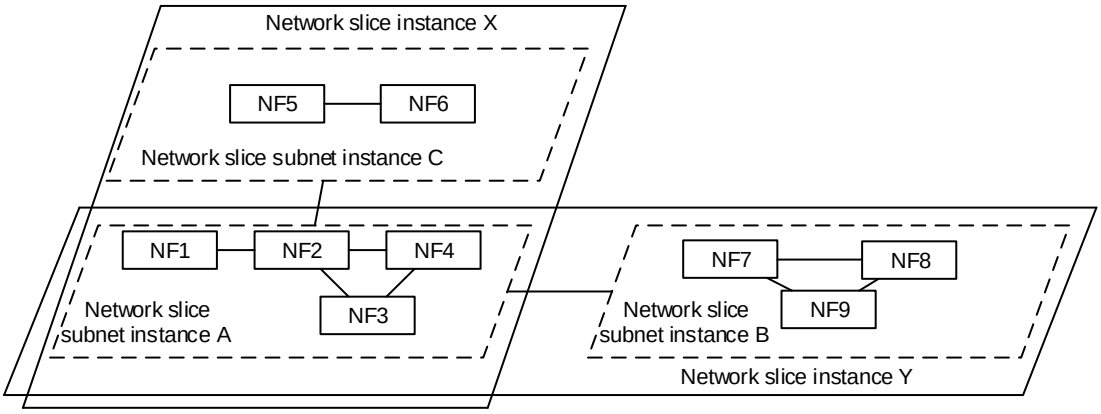


Figure 3-4: NSI X and Y composed by NSSI A, B and C

Orchestrating NSIs that have shared NFs, according e.g. to elasticity requirements, requires the automated management to rightfully take care of all the requirements of the involved NSIs that shares that NFs. As an example, a scenario is represented by two NSIs with a shared VNF. If the service requirements of one NSI change, implying some scaling of shared VNF, this scaling must be performed coherently with the service requirements the other NSI. Based on both service requirements the scaling could be performed or, if it is not compatible, a new VNF must be deployed and activated according to the new requirements.

The modelling of network resources must be done considering the requirements on NFs sharing. The 5G-MoNArch SDM-O needs to be developed, according to 3GPP, with the new NSMF and NSSMF as defined in Section 2.3.1. The SDM-O must be designed with the capability to be aware of all the involved NSI network requirements when managing a NF that serves different NSIs (to address **GAP #2**).

3.1.3 Experiment-driven Optimisation

Experimental optimisation is one of the key elements in the designing and implementation of the next generation of mobile networks. Having different functionalities virtualised, the cloud infrastructure providers must develop an experimental procedure to be able to meet the QoS requirements of each VNF optimally. Scaling and elasticity decisions (either vertical or horizontal) cannot be made without having a practical experimental optimisation approach. Experiment-driven optimisation is enabled through measurement campaigns (i.e., a monitoring process). The measurements from these campaigns feed a modelling procedure, which models the VNF behaviour regarding their computational, storage and networking resource demands. The resulted models may facilitate the overall resource management of the cloud infrastructure. Algorithms and functions that apply upon the 5G protocol stack can improve their performance by exploiting experiment-driven insights and, thus, taking more intelligent decisions. In contrast to importance of this issue, it was not the focus of many studies so far.

In the following, insights regarding critical issues for applying experimental-driven modelling and optimisation in a cloud-enabled infrastructure are provided, leading to gap identification compared to existing approaches.

Current state of play and gaps identification regarding the experiment-driven approach

In the literature, many projects have led to results that can feed the experiment-driven modelling and optimisation approach (for more details see Appendix A). In these projects, the Open Air Interface (OAI) [OpenAirInterface] has been recognised as the major tool for reliable measurement campaigns. The main experimental results so far are related to the C-RAN architecture. C-RAN demands very high capacity fronthaul solutions (~50 times higher than backhaul) [Checko et al] due to the digital transmission of the I/Q samples between RRH (Radio Remote Head) and BBU (Baseband Unite) pool. The most commonly used fronthaul solutions for experiments and real implementations are the CPRI (Common Public Radio Interface), the OBSAI (Open Base Station Architecture Initiative) [Fujitsu], and the recently emerged Next Generation Fronthaul Interface (NGFI). Additionally, the Hybrid Automatic Repeat Request (HARQ) imposes the most critical processing requirement to C-RAN architecture. In LTE, the sent packets must be ACK/NACK within next eight subframe (i.e., within the next 8 ms). This 8 ms delay budget is required for encoding/decoding of UE, propagation over air interface, fronthaul propagation, and eNodeB processing. In case of NACK, retransmission occurs, that packet must be retransmitted. It is estimated that BBU has approximately 3 ms timeframe to decode one subframe [Alyafawi et al]. The delay budget calculation is depicted in Figure 3-5.

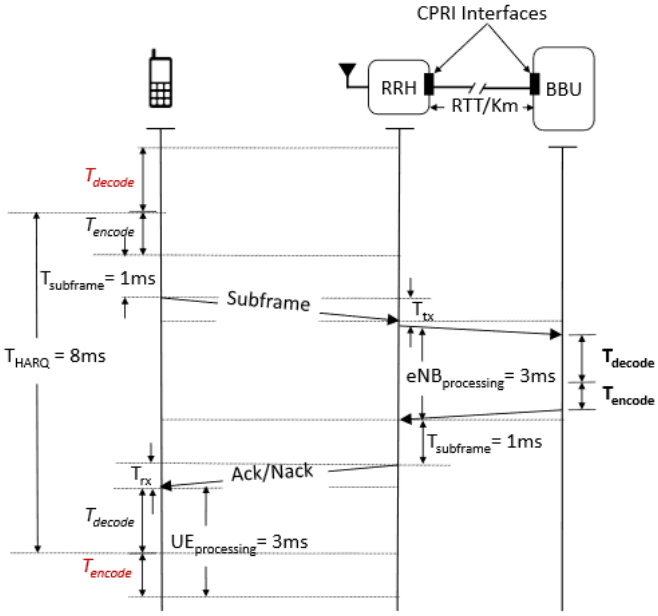


Figure 3-5: LTE delay budget constraint (extracted from [Alyafawi et al])

Moreover, from current C-RAN approaches it is revealed that the use of the dynamic allocation/deallocation of computational resources can minimise operating cost for the 5G network operators [Nikaein]. Network function should be designed in a way that there is an optimal trade-off between cost and reliability. Dynamic allocation/deallocation is expected to be studied in detail during the next phases of 5G-MoNArch. Additionally, one of the least addressed issue regarding the realisation of the C-RAN is, Channel State Information (CSI) accuracy. Due to the fronthaul introduction in C-RAN architecture, CSI information becomes outdated once it reaches to BBU. Due to Inaccurate CSI, radio resource management can´t be performed optimally. Such inaccuracy leads degradation of overall system performance [Cai et al].

Considering the above analysis, current efforts towards realising 5GS, lack experiment-based E2E resource management of VNFs that takes advantage of E2E software implementations on commodity hardware and utilises, in a dynamic manner, behavioural patterns of NFs' resource demands (**GAP #7**).

Experiment-driven modelling and optimisation in a cloud enabled network

As mentioned above, experiment-driven modelling and optimisation is a key innovation enabler for the 5G-MoNArch project. The key innovation element of this enabler is the **E2E management of computational, storage and networking resources consumed by VNFs.** It is worth noting that although this innovation element primarily places the focus on **GAP #7**, it is expected that all 5G-MoNArch innovations can benefit from the experiment-driven modelling and optimisation; therefore, this innovation element can be inter-related to all other identified 5GS gaps as well. This innovation will be realised through measurement campaigns, i.e., targeted monitoring of the performance of VNFs and Network Slices, with the major goal to provide implementation and evaluation results that bridge the gap identified in the previous paragraph. From a more general perspective, the innovation element mentioned above brings a new paradigm in network management and orchestration by feeding with experiment-based inputs the two other enablers of the project (cloud-enabled protocol stack and inter-slice control) as well as the functional innovations of the project (mainly, the approaches towards resource elasticity).

One illustrative example of such innovation, is the experiment-based orchestration of VNFs within the telco cloud. Indeed, since some of the nodes in the telco cloud (particularly at the edge) may be equipped with limited resources, the placement of VNFs in nodes needs to consider the availability of computational, storage and networking resources in addition to other criteria such as service requirements, slice awareness, and functional split options. To better clarify this example, assume that the target is to perform the placement of VNFs of a slice based on the availability of the computational

resources. Traditional approaches assume that a fixed amount of computational resources is required for each network function. However, this model is very coarse and clearly insufficient to understand the performance of a real environment in which the computational and traffic load fluctuates significantly over time. To overcome this, accurate models of the computational behaviour are required to be able to determine the performance impact resulting from a given VNF allocation. In particular, new and precise models are needed which characterise the available resources at each node and the utilisation profile of VNFs. A potential modelling will consider, among others, the following aspects:

- Available resources at a node: When evaluating the processing and memory resources available at a node, not only the resources in the node are to be considered, but also the overhead imposed by the platform itself including the handing of virtual machines (VMs) or containers.
- Resources consumed by network functions: It is also essential to model the resources required by each of the NFs, including the time-variant behaviour rather than just average values, as well as the statistical correlation resulting from logical dependencies between NFs. This necessarily requires implementing and executing these functions to observe their behaviour.

Figure 3-6 illustrates the different tasks that need to be performed to address this innovation. Our framework builds on two models, which capture the behaviour of the NFs and the infrastructure, respectively (the two boxes inside the Model repository container in Figure 3-6).

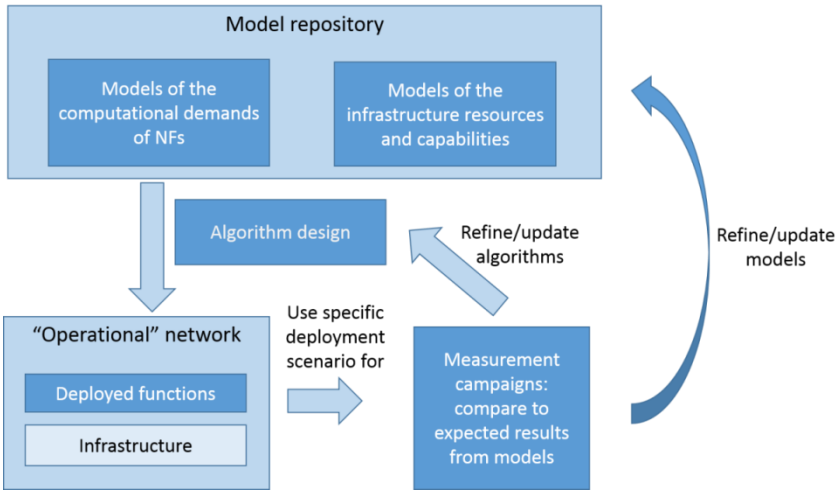


Figure 3-6: Experiment-driven modelling and optimisation

To take advantage of the experiment-driven modelling and optimisation in a cloud enabled network, new challenges arise. First, it is required the conduction of exhaustive measurement campaigns per VNF and per network slice, that will focus on consumption of computational, storage and networking resources and considering cost-effectiveness and the special characteristics and peculiarities due to the use of commodity hardware (the key choice for the cloud-enabled networking). Among others open issues are: (i) the characterisation of temporal behaviour, i.e., occurrence of peaks of resource consumption and periods of lower load, (ii) the evaluation of the (non-negligible) overhead incurred by computational resources used to run system management software, and (iii) the impact of the communication environment as well as the logical dependencies between VNFs, which introduce statistical dependencies in the computational demands of such functions.

3.2 Functional Innovations

3.2.1 Secure and Resilient Network Functions

5G-MoNArch project puts particular attention to the security and resilience aspects of the network to ensure the network robustness to different kinds of unexpected events. The security aspects aim at preventing and, when not possible, minimising the effects of unexpected events originated by a human (attacks). Such man-made network disruptions either compromise fundamental security properties e.g., integrity, confidentiality, and availability in the network or entail any other deliberate misuse of the network that can turn into a security threat with major consequences.

Apart from the problems in the network operation that are caused deliberately by the human factor, i.e. security threats, the problems can be related to other network aspects e.g. software, infrastructure, the actual implementation, deployment and configuration of the network functions etc. Such potential problems will be addressed by 5G-MoNArch through the investigation on network resilience. The actual deployment of 5G network can include the network functions running on virtualised infrastructure (telco cloud) as well as on the specialised physical hardware instances (RAN), with potentially different resilience issues and mechanisms suitable for achieving the resilience. The 5G-MoNArch project distinguishes between RAN and telco cloud in order to address the resilience issues in a domain-specific manner. Therefore, the network resilience in 5G-MoNArch will be treated through the concepts of RAN reliability and telco cloud resilience.

New security requirements

The 5G-PPP Security Working Group recently released a white paper [5GPPP Phase1 Security] describing the 5G-PPP Security Landscape of Phase 1 projects. This white paper lays out design principles and is a first step toward a common 5G security framework, but still requires further discussion on several implementation-related topics and cross-domain orchestration. The paper raises awareness on major security risks and identifies new requirements introduced by the 5G context, which are outlined in Table 3-1:

Table 3-1: Security Requirements

Security Requirement	Description
SR1 - Security Level	*5G must provide a security and privacy level higher or at least equal to the security and privacy level in 4G.*
SR2 – Security Automation	*5G infrastructures' heterogeneity and complexity require security to be dealt at multiple levels and across domains, and thus automation to handle this successfully is vital.*
SR3 - Security Monitoring	*5G systems must support security monitoring capable of detecting advanced cyber security threats and support coordinated monitoring between different domains and systems (e.g. mobile and satellite).*
SR4 – Security Management	*E2E security management and orchestration should be put in place considering correlation and coherence / consistency between data exchanged/shared at Security Architecture Inter-domain interfaces.*
SR5 – Security Liability Schemes	*New responsibility schemes should be proposed, in coherence with existing Regulation, regarding the distribution and allocation of responsibilities and obligations in a multi-tenant softwarised telecom infrastructure, and in particular for potential delegation of regulation obligation to non-regulated third parties.*
SR6 - Inter-tenant/Slice Isolation	*Infrastructure sharing by multiple network operators will require strict isolation at multiple levels to ensure the expected security level.*
SR7 - 5G Liability	*The chain of Trust and liability of multi-tenants should be managed and auditable for each service, component supplier, operator and customer.*
SR8 - Enabling Value Added Services with E2E encryption	*To comply with privacy regulations and protection of user data, traffic encryption is expected to be generalised across 5G networks.*
SR9 - 5G regulation conformity	*5G technologies should be developed in compliance with applicable legislation/regulation.*

5G reference security architecture

To address these requirements, the gaps left by the existing 3G and 4G security architectures [3GPP TS 23.101] [3GPP TS 33.102] [3GPP TS 33.401] need to be filled in. The reference 5G security architecture described in the 5G-PPP Security WG white paper, proposes some design principles, such as: being a logical rather than physical architecture; following a distributed, hierarchical and recursive approach; providing a multi-domain and vertical support; bringing security management into scope; fostering flexibility and extensibility; ensuring support for massive and critical Machine-type communication (MTC); and implementing regulatory compliance.

Building upon existing security architectures for 3G and 4G, a summary of the architectural extensions introduced by 5G context is provided below, see also Figure 3-7:

- **Strata**: One extension introduces a new stratum (i.e. "grouping of protocols and functions related to one aspect of the network services") focused on management, in addition to the existing ones: access, transport, serving, home, and application stratum.
- **Security feature groups**: Associated to the managements stratum, a new feature group (i.e. "groups of security mechanisms and protocols, each one related to a stratum") "Security management" is added, comprising e.g. securing orchestration, key management etc. Additionally, the "Visibility & configurability (V)" feature group in 4G has been renamed to "Trustworthiness (V)" and extended in terms of scope, to cover other typical 5G facets.
- **Domains**: support for multiple domains need to be introduced in 5G context, which is paramount to implement trust between actors, devices, services, etc. On the one hand, it must be distinguished between *physical domains* or Infrastructure Provider Domains (IP Domains) and the *logical/functional domains* called Tenant Domains. On the other hand, transversal "domains-across-domains" must be model as *slice domains*. Similarly, to feature groups and strata, *management domain* is added to model the corresponding management functionalities. A *third-party domain* captures third party services to deal with authentication and identity management, e.g. device-specific providers. The *IP service domain* model external IP networks. The final extension captures so-called direct-mode, UE-to-UE communication by adding the (Additional) UE Domain (which internally has the same de-composition as the UE domain).

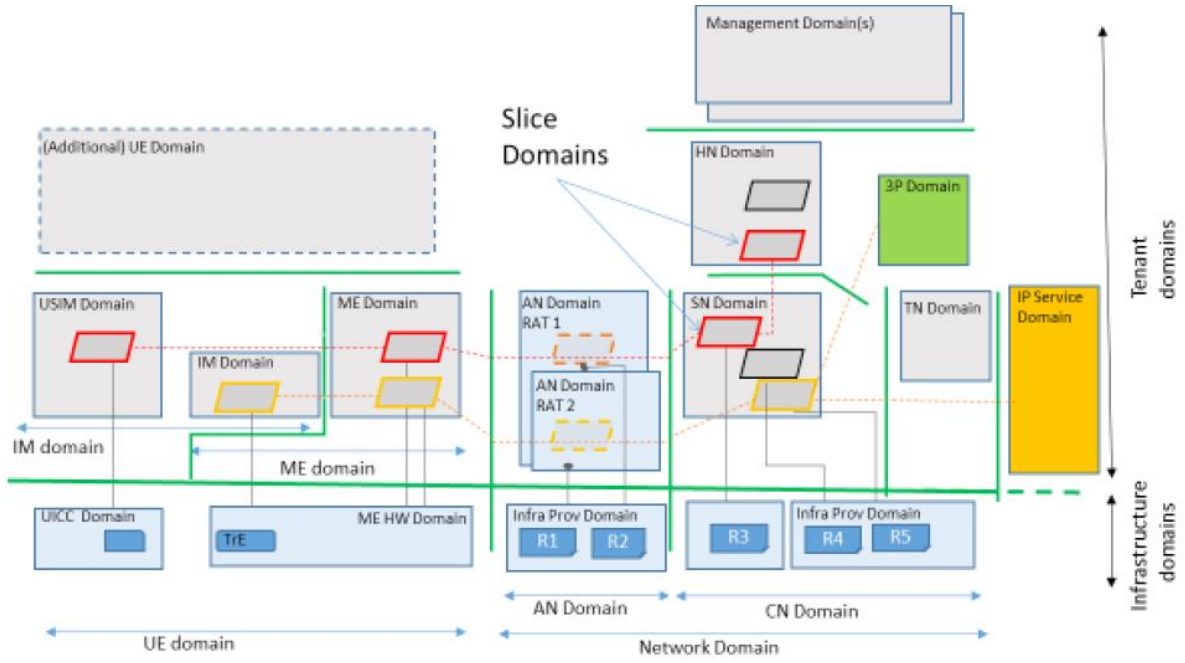


Figure 3-7: Proposed domains for the 5G security architecture (Source: 5G-PPP Security WG [5GPPP Phase1 Security])

There is a lack of concrete deployments in industrially-relevant pilot activities that prove the existing security technologies indeed address the new 5G requirements, and most importantly, that evaluate critical non-functional aspects (e.g., performance) in the specific 5G context. In particular, it is essential to evaluate the trade-off between security and resilience, to incorporate a mechanism in the overall

architecture design that governs this relation from a service level perspective, with an impact in the underlying architecture layers (see Section 2.1.3).

Although the 5G-PPP Phase 1 projects already address major security concerns in the 5G context, a series of critical gaps related to orchestration & management, accountability, compliance & liability, as well as performance and resilience remain fully open. Therefore, **a more complete and refined design of a 5G security architecture is still missing (GAP #8)**.

Self-adaptive and slice-aware deployment model

The concept of network slices, already introduced in Section 2.1.2, brings many benefits towards achieving a more flexible, reactive and self-adapting network security management. But also, slices permit high performance deployments, which are critical for real-time security monitoring (i.e. analytics), as well as deploying tenant-scoped security functions that are needed to efficiently manage a cyber-attack reaction. But, to maximise the benefits, slices should be fully isolated and come with minimal but key security functions, i.e. core security functions, such as guaranteed E2E isolation, communications confidentiality & integrity, and AAA/traceability.

Moreover, to make 5G network cyber resilient, i.e. resistant to cybersecurity threats, it is paramount to firstly, include security monitoring features in the design of the 5G reference architecture, and secondly, to ensure these features are continuously activated during the 5G network infrastructure operation. This way, security threats and attacks will not only be efficiently detected and countered, but most importantly, will be prevented in a dynamic and context-aware manner. The use of security analytics that continuously process a large number of logs produced at the network and application layers provide essential inputs to learning processes (based, for example, on Machine Learning or Artificial Intelligence approaches). Moreover, auditing capabilities produce the necessary evidence to prove conformity to applicable regulation and support for the implementation of liability schemes, fostering trust among stakeholders and ensuring a wider adoption.

The green lines in Figure 3-7 reflect interfaces/reference points between domains. These are the points where security monitoring features need to be deployed to manage security from and end-to end perspective. However, several challenges outlined in [5GPPP Phase1 Security] arise with regards to this topic, such as the way to combine the needs for E2E security monitoring with the need for strong isolation between slices (at Core and Access level) and how to prevent security shortcuts via a monitoring and management system; how to adapt real-time E2E security monitoring system to evolving infrastructure topologies; selecting appropriate machine learning algorithm and its learning styles for an accurate and efficient prediction for a given security problem and conditions; are just a few examples.

Security monitoring is an essential feature that contributes significantly towards ensuring a continuous and long-term robust 5G network operation and infrastructure. It is fundamental to effectively meet several of the new security requirements (as listed in Table 3-1). Besides helping to address SR1 and SR2, by supporting the correct triggering of preventive and reactive measures in an automatic (or at least semi-automatic) manner, it also produces evidence to achieve accountability, and thus, supports meeting SR5, SR7 and SR9 requirements too. Although there are technologies to implement security monitoring effectively (e.g. detection tools tailored to both specific infrastructure layers and specialised in certain attack families, advanced frameworks for security analytics) with very good results in distributed and multi-layered architectures (e.g. Cloud or IoT environments), these still need to adapt to the particularities of 5G, in terms of infrastructure and threats, and must be validated in industrially relevant deployments.

There are plenty of room for research and improvement in terms of cross-layer orchestration, secure information exchange and specially, to align with the E2E network slicing concept. These concepts have implications in the way security management is traditionally addressed (centralised vs de-centralised) by existing security solutions but most importantly, in the way these are deployed and interact with the underlying infrastructure, to ensure proper isolation, preservation of privacy, integrity and cyber-resilience. Also, traditional security zoning, monitoring and risk mitigation strategies need to be revisited considering the advances achieved in the areas of machine learning and active security monitoring. Thus, **a slice-aware deployment model for security controls with self-adaptive capabilities need to be defined (GAP #9).**

Resilient Network Functions

5G-PPP Phase 1 projects have been focused on development of the key concepts and building blocks of the 5G architecture. Therefore, the development of more specific network functions that are required for implementing a slice with particular requirements has not been addressed to a large extent. One of such technology gaps that has not been addressed in 5G-PPP Phase 1 projects is enabling of resilience in 5G networks.

Resilience is the ability of the network to continue operating correctly during and after a natural or man-made disturbance, such as the loss of mains power [NORMA D6.1]. In other words, the resilient network needs to be able to recover after an unexpected event and to resume its normal operation. This capability is of paramount importance for network reliability and providing a service with satisfying performance, especially for critical communication type services such as envisioned in URLLC slice. However, the 5G-PPP Phase 1 projects did not explicitly or to a large extent target this aspect in their architectures. The requirements on resilience has been implicitly addresses by the management and control entities and mechanisms that are designed in a way to promptly react on unexpected events. E.g. in 5G NORMA after a violation of QoS requirements is detected on centralised controllers (SDM-C/X) the problem mitigation is attempted through network reconfigurations by controllers. In the case that this was not sufficient to overcome the problem, a trigger is sent to MANO blocks of 5G NORMA architecture, i.e., SDM-O to perform needed action for problem mitigation e.g. scale out the resources of network functions, deploy new functions etc. [NORMA D5.2]. Although the architecture developed in 5G NORMA is capable of reacting to unexpected traffic/network events and mitigate their negative influence to a certain extent, the architecture and mitigation mechanisms are not build with resilience in mind and are not optimised for such specific use case. Therefore, the aforementioned problem mitigation actions and processes can be suboptimal and cannot meet high reliability requirement in an efficient way. To fill this gap between 5G-PPP Phase 1 projects architectural design and resilience needs of particular network slice types, e.g. URLLC slices or industrial enterprise slice, a special attention in design of 5G-MoNArch architecture is put on enabling and integrating resilience and reliability aspects. Rather than being an "afterthought" in 5G-MoNArch the resilience is one of the main objectives with which the architecture will be built.

Whereas the support for reliability and resilience is required on E2E level, i.e. at the overall network and service level, different mechanisms for enabling such features can be applied to different network domains, elements and parts of the protocol stack, e.g. different mechanisms can be applicable to physical network components, lower RAN protocol stack, virtualised network functions, infrastructure components, etc. With this in mind the approaches for enabling resilience can be separated into two main categories targeting two environments: RAN and telco cloud.

RAN reliability

As defined in [MONARCH D6.1] the reliability is a percentage (%) of the amount of sent network layer packets successfully delivered to a given system node (incl. the UE) within the time constraint required by the targeted service, divided by the total number of sent network layer packets. A relatively novel approach which can be used to increase the reliability at the RAN is multi-connectivity [Ravanshid et al], [Koudouridis et al], [Michalopoulos et al]. The main idea of multi-connectivity is to utilise the simultaneous connection of the terminals to multiple access points. This enables exploitation of a larger set of available resources, and thus increases the connection reliability. The origins of multi-connectivity are visible already in LTE technology, first in the form of carrier aggregation [3GPP TR 36.808], and later through the form of dual connectivity [3GPP TR 36.842], with the main difference in the aggregating the data at different layers of the RAN protocol stack. The target of the carrier aggregation and dual connectivity is the throughput increase, suitable for applications with high data rate requirements. However, the 5G-MoNArch aims at utilising the multi-connectivity concept not as a technique for throughput increase but as the technique for *reliability increase*, instead.

Utilising multi-connectivity as a means to increase reliability instead of throughput entails amendments with regards to its functionality. In particular, for reducing the probability that the user terminal receives an erroneous version of the transmitted packets, data needs to be *duplicated* and transmitted to the terminal in the form of *redundancy*. That is, the communication setup is designed in a novel way, where the terminals receive multiple replicas of the same message from the corresponding access points. As

such, RAN reliability comes as an outcome of the inherent *diversity* of telecommunication setups that involve simultaneous transmission/reception to/from multiple access points.

Furthermore, it is worth mentioning that although traditional dual connectivity is deployed over non-virtualised architecture, within the 5G-MoNArch framework multi-connectivity in the context of virtualised networks is studied, where additional benefits are expected in terms of flexible resource deployment and re-usability.

Apart from multi-connectivity, the network coding technique can be used as a mean for increasing the reliability. The basic idea is that network nodes transmit composite messages i.e. two or more messages together. At the destination nodes, the composite messages are inferred rather than directly decoded. In the context of network coding the work in [Nazer and Gastpar] illustrates a method which uses the interference from neighbouring nodes to generate a set of linear equations that can be solved at the destination nodes which increases the throughput. Furthermore, the network coding can be used for increasing the reliability as discussed in [Ghaderi et al] and aimed in 5G-MoNArch context. 5G-MoNArch will focus on leveraging existing network coding techniques and adapting it for usage in virtualised environment. In this regard, the study will involve investigations of implementing network coding as a special network function block, where the optimal location of such block into the 5G-MoNArch flexible architecture framework will be pursued.

The implementation of multi-connectivity for the purpose of reliability increase is anticipated to take place at using the CU / DU architecture split discussed in Section 2.3.1. In particular, multi-connectivity used for high reliability implies the use of special functionalities, such as data duplication and network coding. Such special functionalities need to run in a centralised location (i.e., the CU) such that the coordination of the multiple physical links involved is facilitated. At the same time, the lower-layer functions associated with high reliability (such as, for instance, duplicated packets and network coded -related scheduling) are carried out at the DU, due to the physical constraints involved.

The 5G-PPP phase 1 projects did not target the RAN reliability as a built-in solution/element of the fundamental architecture, but the RAN reliability requirements have been mostly implicitly addressed by the management and control entities and mechanisms. There is a clear need for enhanced and inherent support for RAN reliability (**GAP #10**). The 5G-MoNArch aims at addressing this gap by considering the RAN reliability intrinsically in the architecture, by **applying the mechanisms such as multi-connectivity and network coding.**

Telco cloud resilience

In addition to RAN reliability, 5G-MoNArch will focus on resilience in virtualised part of the architecture, i.e. telco cloud. In this context 5G-MoNArch will focus on three main topics which are strongly interrelated, namely:

- Improving the resilience of individual network elements and functions
- Network fault isolation
- Developing the mechanism for failsafe operation

Some network functions can have higher importance in overall network functionality, thus to achieve required resilience level for such functions special techniques need to be applied e.g. for centralised SDN controllers or safety critical network functions in URLLC network slice. 5G-MoNArch will investigate such techniques, e.g. additional redundancy, built in resilience mechanisms as well as trade-offs in applying such techniques, e.g. in terms of overprovisioning and resource reservation.

Furthermore, 5G-MoNArch will elaborate on fault management as mean for fault isolation in telco cloud. Fault management is responsible for providing the information regarding the current network state and to react against problems in network operation that cause the performance degradation. The fault management includes processes such as: the collection of data with respect to network performance; detection of network anomalies, degradation and faults; root cause analysis and fault isolation as well as recovery actions for degradation mitigation. The aim of 5G-MoNArch project is to enable suitable fault management techniques for supporting network slice resilience even for slices with very stringent resilience requirements. Special attention in the context of fault management will be at improving the fault isolation, e.g. preventing the situation where a fault in one part of the network affects

normal functionality of other parts of the network. Network faults can be related to different parts of the network (storage, network node or communication network [Wu et al]), or can originate from different layers and planes e.g. being physical issue, a control plane or a user plane issue [Zhou et al]. Some of the common techniques for mitigating the network faults are self-healing Self-Organising Network solutions [Hämäläinen et al] comprising the techniques for coping with the outages on the level of individual network cells, including outage detection, root cause analysis and fault mitigation. Additionally, techniques such as re-routing [Yu et al] [Xu et al] can be used. 5G-MoNArch will utilise the existing fault management techniques, adapt and extend them to 5G network slicing context.

The failsafe network operation is the ultimate target in the resilience context, and all aforementioned techniques that will be addressed by 5G-MoNArch contribute to this target. Additionally, 5G-MoNArch will approach the topic of failsafe network operation by discussing alternative ways in achieving this target. In this context 5G-MoNArch will study the potential of dimensioning and configuring edge cloud resources for creating "5G islands" that are envisioned to operate in an autonomous way and provide basic network services without being continuously connected to the central cloud. Such technique will improve the resilience of network especially in case of unexpected outages of links between central and edge cloud.

The 5G-PPP phase 1 projects did not consider the telco cloud resilience in a systematic and detailed way. Telco cloud resilience was supported mainly in indirect and rudimentary way through management and control mechanisms (**GAP #11**). The 5G-MoNArch aims at filling this gap by addressing the telco cloud resilience in a rather structured way considering different aspects that can contribute/impact the telco cloud resilience, e.g. **improving the resilience of individual network elements/functions and telco cloud components, improving the fault management and failsafe mechanisms.**

3.2.2 Resource-elastic Virtual Functions

Resource elasticity comprises the second main functional innovation of 5G-MoNArch. It addresses the need for assigning and scaling computational, storage, and communication resources where and when they are needed. For instance, in the case of a typical urban downtown scenario, the required services may range from augmented reality to video chats and instant messaging, each imposing different requirements over a certain period of time at a specific location. To holistically address the problem of spatial and temporal traffic fluctuations in a cost-efficient manner, the mobile network must be able to assign, scale and cluster resources to those parts of the networks where they are needed; and the network functions need to be elastic enough to adapt to the available resources without impacting performance significantly. The ability to gracefully scale down the network operation by means of efficiently scaling those resources according to the demand by when insufficient resources are available is here defined as *resource elasticity*.

The concept of elasticity is well known in the cloud computing community [Coutinho et al, Herbst et al]. However, solutions from cloud frameworks will need to be enhanced within 5G-MoNArch as (i) timescales involved in RAN functions are usually much smaller than those considered in cloud solutions, hence leading to possible outages, and (ii) cloud resources are typically limited and sparser at the edge, sometimes preventing centralised solutions to exploit multiplexing gains. 5G-MoNArch will introduce the concept of elasticity at both edge and central clouds considering the associated constraints of the cloud infrastructure and the mobile network. Furthermore, the elasticity framework will need to consider the fact that cloud resources are shared by different slices and their availability may change according to the dynamic request of tenants.

5G-MoNArch will develop elasticity techniques that take into consideration not only the availability of communication resources, as in the state-of-the-art, but also computational and storage resources. Elasticity principles will be also applied to orchestrate the deployment of NFs, moving functions between central and edge clouds to minimise computational outage.

An elastic system should be able to be optimally dimensioned such that, to support the same services, less communication resources are required compared to a non-elastic system. In highly loaded scenarios, an elastic system efficiently exploits the multiplexing gain and provides large resource utilisation efficiency and high quality of service, by deploying a high number of VNFs over the same physical infrastructure. In addition, in lightly loaded scenarios the elastic system should avoid the usage of

unnecessary resources and reduce the energy consumption (thus limiting the operation expenditure). Moreover, the design of the NFs will be re-visited with the goal of adapting their operation to the available resources and thus minimise the impact of outages on their performance, when such outages occur.

To meet the above described requirements and objectives, three main research areas in the context of resource elasticity are explored in 5G-MoNArch, namely (i) computational elasticity, (ii) orchestration-driven elasticity, and (iii) slice-aware resource elasticity. These three research areas account for an E2E elastic operation of the network, including the RAN. The different dimensions of elasticity, as described below, will require new features in the architecture that are not currently envisioned by the state of the art. Here, the discussion is limited to the architecture described throughout the document, as this is considered being the bleeding edge in the field.

Computational elasticity

One of the most appealing advantages of a cloudified network is the possibility of reducing costs by adapting and re-distributing resources following (and even anticipating) temporal and spatial traffic variations in a centralised manner. However, it is expectable that the cloud resource assignment is occasionally exceeded by the induced burden. This is a particularly true for C-RAN deployments which are known to be highly variable [Checko et al]. In this scenario, allocating resources based on peak requirements would be highly inefficient. VNFs, instead, shall efficiently use the resources they are assigned, and become computationally elastic, i.e., adapt their operation when temporal changes in the load and hence resources available occur.

In the context of wireless communications, the concept of elasticity usually refers to a graceful degradation in performance, for example when the spectrum becomes insufficient to serve all users. However, in the context of a cloudified operation of mobile networks, when addressing elasticity to resource shortages, other kinds of resources are also considered that are native to the cloud environment, such as computational and storage resources.

Elasticity has also been considered by non-VNFs cloud operators, but as mentioned earlier our concept deviates very much from theirs: the time scales involved in RAN functions are significantly more stringent than the ones required by e.g., a Big Data platform or a web server back-end. Another key difference is that resources are way sparser (e.g., they are distributed across the "edge clouds"), which reduces the possibility of damping peaks by aggregating resources.

To overcome such computational outages, NFs will be designed that can gracefully adjust the amount of computational resources consumed while keeping the highest possible level of performance. RAN functions have been typically designed to be robust against shortages on communication resources; hence, the goal of 5G-MoNArch is to make RAN functions also robust to computational shortages, by adapting their operation to the available computational resources. The design of such computationally elastic NFs will be investigated in its horizontal and vertical dimensions, i.e., the ability to scale either the number of virtual machines or containers executing the functions (horizontal) or the resource capabilities of the allocated virtual machines or containers (vertical); it will furthermore be studied how these approaches impact NF performance.

In addition to the obvious gaps in the NF design that are orthogonal to the architectural construction, some elements will need additional components to cope with elasticity at network function level. Obviously, I-APP and X-APP (see Section 2.4) will have to deal with resource availability, which will be one of the parameters passed through the NBI to the agent I-NF (or X-NF). Also, the amount of resources available at any time should be constantly monitored by the monitoring module, which should provide useful information to both the controllers and the orchestration. **Hence, novel elastic functions will be designed as well as mechanisms for NF scaling, addressing GAP #1 and GAP #5.**

Orchestration-driven elasticity

This area addresses the ability to re-allocate NFs within the edge and the central cloud depending on the available resources, considering service requirements and the current network state, and implementing preventive measures to avoid bottlenecks. This may imply scaling the edge cloud based on the available resources (e.g., releasing unneeded resources), clustering and joining of resources from different locations, shifting of the operating point of the network depending on the requirements, and/or adding

or removing of edge nodes [O+14]. Furthermore, edge cloud resources may also be required to provide MEC features, which may have higher priority than specific NFs depending on the service requirements.

Therefore, this objective aims at investigating solutions that enhance the elasticity of the cloud infrastructure in 5G networks by leveraging the orchestration of functions in different locations. The proposed solutions will cope with the shortage of computational resources by moving some of the functions to other locations. Special attention will be paid to (1) the trade-off between central and edge clouds and the impact of choosing one location for a given function, and (2) the coexistence of MEC and RAN functions in the edge cloud. The orchestration solution designed will be aligned with the experiment-driven orchestration techniques also developed within the project.

Elastic VNFs will increase the system resilience per-se, by performing graceful degradation in case of shortages. However, the QoE/QoS perceived by the users depends on the network slice they are attached to, and hence, by the service function chaining building the services. Therefore, the MANO (and specifically the NFV-O) should have a global view of the elasticity achieved by the concatenation of several elastic VNFs (and by different subsets) to perform operation such as resource provisioning and VNF location. Finally, there should be an additional interface between the MANO and the controllers, to provide them with the information about the assigned resources. **5G-MoNArch will design elastic orchestration mechanisms and enablers for the MANO, hence addressing GAP #2 and GAP #4.**

Slice-aware resource elasticity

Finally, this area addresses the ability to accommodate multiple slices within the same physical resources while optimising the network scaling and resource consumption. This facilitates the reduction of CAPEX and OPEX by exploiting statistical multiplexing gains. Indeed, due to load fluctuations that characterise each slice, the same set of physical resources can be used to simultaneously serve multiple slices, which yields large resource utilisation efficiency and high gains in network deployment investments (as long as resource orchestration is optimally realised). 5G-MoNArch will thus devise elastic mechanisms that improve the utilisation efficiency of the computational and radio resources by taking advantage of statistical multiplexing gains across different network slices. The resulting framework will aim at optimising the system scaling (thus limiting CAPEX), the slice performance, and the consumption of computational resources.

In a similar way to the previous item, the Inter Slice Resource Broker should also have a global view of the elasticity of the different functions. To achieve the most from the resources available, the ISRB must rely on the information provided by the different NFV-O to perform the following operations: (i) admit (or reject) new slices to the system and (ii) consistently assign resources to them. Also, the ISRB should have a "Data Analytics" module, to learn from the past usage of the different network slices using e.g., machine learning techniques, and proactively react (and event anticipate) to resources demands. **Hence, an elastic ISRB will be designed for handling elastic slices, by which GAP #6 and GAP #12 will be addressed.**

3.3 Summary of the Gap Analysis and 5G-MoNArch Innovations

After the overview of all enabling and functional innovation of 5G-MoNArch, this section will provide a summary of all the observed 5GS gaps gathered so far, as listed in Table 3-2. In Table 3-3 and Table 3-4, each innovation is broken down into innovation elements, which are then mapped onto different identified 5GS gaps. Besides that, relevant fora, consortia, and SDOs for those 5GS gaps along with the innovation elements are marked[5].

Table 3-2: List of observed gaps

Gap	Description
GAP #1	Inter-dependencies between Network Functions co-located in the same node
GAP #2	Orchestration-driven elasticity not supported
GAP #3	Fixed functional operation of small cells
GAP #4	Need for support for computational offloading

[5] A direct mapping is marked by "X" and a possible extension to cover a 5GS gap is marked by "{X}".

GAP #5	Need for support for telco grade performance (e.g. low latency, high performance, scalability)
GAP #6	E2E cross-slice optimisation not fully supported
GAP #7	Lack of experiment-based E2E resource management for VNFs
GAP #8	Lack of a refined 5G security architecture design
GAP #9	Lack of a self-adaptive and slice-aware model for security
GAP #10	Need for enhanced and inherent support for RAN reliability
GAP #11	Indirect and rudimentary support of telco cloud resilience mainly through management and control mechanisms
GAP #12	Need for (radio) resource sharing strategy for network slices

Table 3-3: 5GS Gap Analysis and how they will be covered by 5G-MoNArch Enabling Innovations

		5GS Gaps											
5G-MoNArch Innovations	**Innovation Elements**	GAP #1	GAP #2	GAP #3	GAP #4	GAP #5	GAP #6	GAP #7	GAP #8	GAP #9	GAP #10	GAP #11	GAP #12
Cloud enabled protocol stack	Telco cloud-aware protocol design	X (3GPP RAN2)	X (ETSI NFV/ENI)	X (3GPP RAN2/3)									
	Terminal-aware protocol design				X (ETSI MEC)	X (3GPP RAN2, SA2)							
Inter-slice control and management	Inter-slice Context-aware Optimisation						X (3GPP RAN3, SA2)						
	Slice-aware Functional Operation			X (3GPP RAN2/3)			X (3GPP RAN2/3)						{X} (3GPP RAN2/3)
	Inter-slice resource management		X (ETSI NFV/ENI)										{X} (3GPP RAN2/3)
	Terminal analytics driven slice selection / control						X (3GPP SA2)						
	Inter-slice Management & Orchestration framework					X (Open Source SDN-C, 3GPP SA5)							
Experiment-driven optimisation	E2E management of computational, storage and networking resources consumed by VNFs	X All 5G-MoNArch innovations can benefit from experiment-driven optimisation (GAP #7: Open source emulation platforms, i.e., OpenAirInterface Software Alliance and srsLTE)											

Table 3-4: 5GS Gap Analysis and how they will be covered by 5G-MoNArch Functional Innovations

		5GS Gaps											
5G-MoNArch Innovations	**Innovation Elements**	GAP #1	GAP #2	GAP #3	GAP #4	GAP #5	GAP #6	GAP #7	GAP #8	GAP #9	GAP #10	GAP #11	GAP #12
Secure and resilient network functions	Multi-connectivity and network coding for improving the RAN reliability										X (3GPP RAN2)		
	Enhancements in telco cloud resilience through improved failsafe mechanisms and fault management											X (3GPP SA5, ETSI NFV)	
	Flexible security monitoring and detection algorithms								X (5G-PPP Security WG, 3GPP SA3)	X (5G-PPP Security WG, 3GPP SA3)			
	Inter/Intra slice security management								X (5G-PPP Security WG, 3GPP SA3)	X (5G-PPP Security WG, 3GPP SA3)			
	Secure exchange of threat intelligence								X (5G-PPP Security WG, 3GPP SA3)	X (5G-PPP Security WG, 3GPP SA3)			
	Self-adaptive slice-aware deployment model								X (5G-PPP Security WG,	X (5G-PPP Security WG,			

									3GPP SA3)	3GPP SA3)			
Resource-elastic virtual functions	Elastic function redesign	X (ETSI NFV)											
	Elastic NF scaling mechanisms					X (ETSI NFV)							
	MANO elastic orchestration mechanisms		X (ETSI MANO/NFV)		X (ETSI MANO/NFV)								
	ISRB for handling elastic network slices.						X (3GPP SA2/SA5)						X (3GPP SA2/SA5)

3.4 Architectural Instantiation of two use cases (5G-MoNArch Testbeds)

5G-MoNArch will not only propose and develop the architecture and innovations described above, but it will also deploy and validate their feasibility and performance in testbeds. These testbeds will cover a variety of requirements, involve multiple technologies, implement a plethora of applications and target various KPIs. The two testbeds are described below.

3.4.1 Sea Port

General description

A sea port is a typical example for a larger environment operated by a vertical industry player for different end customer groups, e.g. shipping companies (both passenger and cargo), logistic companies, railway companies, retailers. Sea ports manage the traffic and trade of goods, aiming at maximising its throughput. This requires a well-designed and reliable ICT infrastructure that take into consideration the following aspects:

- The nature of the network infrastructure and network services within the area of the sea port is quite diverse, consisting of critical applications (e.g., control of water gates), massive broadband applications (e.g., offering Internet to passengers of large cruise ships) and massive sensor applications.
- Failure in a sea port's ICT infrastructure can have international impact, e.g., a stoppage at a central hub such as Hamburg may have an impact on goods transport in all central Europe.
- As crucial part of a country's economy, the ICT infrastructure of a sea port will be subject to cyber-attacks on data and infrastructure.

Based on these aspects, the main requirements here are resilience (guaranteed availability, even in the case of failures), security and support for service diversity. To show how the proposed architecture can satisfy these requirements, 5G-MoNArch will deploy a testbed located at the Hamburg sea port, operated by Hamburg Port Authority (HPA).

Technologies involved

The testbed will implement and demonstrate the following technologies to be developed in the project:

- ***Resilience***: specific functions will be implemented that provide a very high level of resilience, incorporating multi-connectivity and network coding techniques on the radio link. The goal is to achieve resilient behaviour in the presence of radio or other types of impairments.
- ***Security***: other specific functions focus on security, including lightweight encryption mechanisms for low power devices (such as the sensors) as well as the 5G Island concept and security across slices.
- ***Network slicing***: multiple slices will be implemented in the sea port testbed, accounting for the diversity of services (enhanced mobile broadband (eMBB) service, ultra-reliable and/or low latency communications (URLLC), massive machine type communications (mMTC)). Each slice instance must fulfil its service's requirements.
- ***Inter-slice control***: since this testbed will have multiple slices, an inter-slice control functionality will be deployed to dynamically and efficiently share the resources between the slices.

Applications implemented

A multitude of applications are operational in the Hamburg sea port, such as environmental monitoring, ship location monitoring, level metering and traffic management (including traffic light control and parking space management). A major part of these applications is presently implemented through (partly analogue) fixed line networks, lacking the scalability and flexibility required by future applications which involve a substantially larger number of sensors.

The challenges arising from this current scenario range from pure infrastructure problems (i.e., connecting devices) to application-layer problems (i.e., interworking of protocols). These issues will be addressed through the following three applications which will be implemented in the testbed:

(1) Traffic light control (*URLLC*): The testbed will connect traffic lights through wireless links with the following requirements: (i) connections must be reliable and resilient; (ii) traffic lights may be added/removed over time; and (iii) security and data integrity are very important to guarantee proper operation.

(2) Video surveillance (*eMBB*): Video surveillance is needed to control entrance to areas and their general monitoring, imposing the following requirements: (i) no side effects to URLLC services are allowed; (ii) reliable connections are needed; and (iii) data privacy and security are important due to regulations.

(3) Sensor measurements (*mMTC*): Sensor measurements on barges (small sized boats) must be connected through wireless terminals, representing a scenario with many terminals with uplink traffic of varying amount and mobility requirements (i.e., intelligent mobility concepts are necessary).

Figure 3-8: Initial setup of Hamburg sea port testbed

The sea port setup is illustrated in Figure 3-8. In the areas indicated in red, there is the intention to place traffic lights and co-located video surveillance cameras. Here the proposed architecture will be deployed using prototypes and providing different quality of service (QoS) requirements. A second area will be used to operate sensor reading on barges in locations with currently low wireless coverage (indicated in yellow). Please note that this is only a draft initial sketch and there might be changes in the final definition of the testbed area according to upcoming specifications to be created by WP5 of 5G-MoNArch.

3.4.2 Touristic City

General description

This testbed represents a typical case of future advanced multimedia services, in this case in a touristic city environment. One future service envisioned is the provisioning of interactive Augmented Reality (AR) / Virtual Reality (VR) content to end-users, see Figure 3-9 and Figure 3-10. Such applications require dedicated slices with high speed and low delay, while also having the necessary elasticity to adjust to the available computing and network resources.

5G-MoNArch will support the deployment of network slices tailored to specific requirements. In the scenario described here, these slices will include a network slice responsible for the media content transfer, another latency-efficient slice dedicated to the interaction with the VR world, e.g., through haptic communication, and finally, slices providing other services, such as MBB.

In addition to the above, this testbed will also be used to evaluate and demonstrate the following features of the 5G-MoNArch architecture: (i) flexible creation of a specifically localised MBB slice; (ii) solutions related to the dynamic placement of functions; (iii) resource elasticity when overall network conditions

change; (iv) assessment of the system's KPIs in terms of latency and throughput; and (v) provisioning of new functions such as localisation of traffic to meet the desired KPIs.

The location for this testbed will be a touristic site in Turin, Italy. The specific deployment location is currently under evaluation.

Technologies involved

To show that 5G-MoNArch technologies not only meet the requirements of future applications in an environment with resource outages, but also efficiently uses the spectrum and computational resources, the testbed will implement the following technologies:

- ***Computationally-elastic network functions***: NFs will scale down their operation in case of computational resource shortages, while aiming for only a small degradation in performance.
- ***Cloud-enabled protocol stack***: the protocol stack will be enhanced to better perform in a cloud environment, allowing protocol stack functions to be flexibly placed either in the edge or central cloud; this will support different allocations for different slices.
- ***Network slicing and slice-aware elasticity***: three slices will be instantiated: one slice for MBB service, another for a low-latency interactive service and a third one for the AR/VR content. Slice-aware elasticity will be implemented in the context of these slices.
- ***Network orchestration and orchestration-driven elasticity***: network orchestration will provide the low-latency slice with a function allocation that places the delay critical functions close to the end-user. Orchestration will follow the experiment-driven optimisation approach, and consider resource elasticity.

Applications implemented

Visitors will experience the building's interactive AR/VR touristic content, enhancing their visit. Cloisters and some indoor areas are open to the public, while large indoor and outdoor areas are available for demonstrations. This situation allows for multiple scenarios to be assessed.

This testbed covers several use cases:

(1) *On-site Live Event Experience*: Live events in the cloisters transmitted using 360 degree videos with superimposed AR/VR overlay information to enhance the live experience.

(2) *Immersive and Integrated Media*: Remote visitors can tour the Future Centre building con-currently with people on-site in real-time as well as virtual characters.

(3) *Cooperative Media Production*: Virtual visitors will record real time 360 degree AR/VR experiences including exhibits, virtual characters, VR/AR overlaid information, people that are visiting the building, as well as other concurrent users that are participating in the same VR experience.

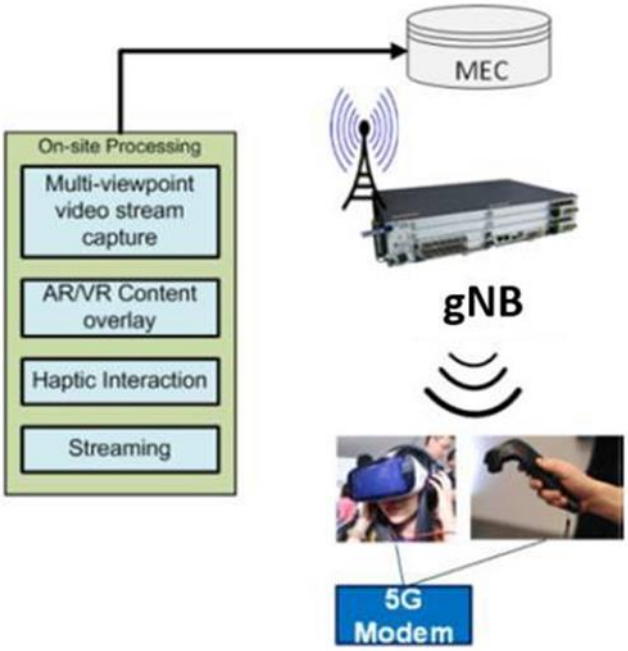


Figure 3-9: Touristic city testbed setup

Figure 3-10: Mock-up illustration depicting the envisioned VR application in-situ

4 Conclusions and Outlook

In this deliverable, the essential architectural concepts and components of the baseline 5G-MoNArch architecture are described considering the most relevant SotA work. Being one of the fundamental pillars of the 5G-MoNArch vision, the native support for network slicing covering E2E technical domains is examined. Baseline concepts for the CN, RAN, MANO, centralised CP architecture, and physical network infrastructure are established. On this baseline, a 5GS gap analysis is performed, where it is detailed how the 5G-MoNArch innovations address these gaps. A brief overview of the two planned architectural instantiations into testbeds are also given.

The work presented herein provides a solid starting point for establishing the 5G-MoNArch vision, and it is also expected that the 5GS gap analysis provided can be a reference for other 5G PPP Phase 2[6] projects as well as other research consortia.

The work presented here, together with the work in deliverable D6.1 (Documentation of Requirements and KPIs and Definition of Suitable Evaluation Criteria), defines the first baseline architecture and architectural requirements. Future work will be further defining and extending architectural elements, concepts and components, aiming at having a first full iteration for deliverable D2.2 (Initial overall architecture and concepts for enabling innovations). The timeline of the work is illustrated in Figure 4-1 along with the various milestones and the upcoming WP2 deliverables.

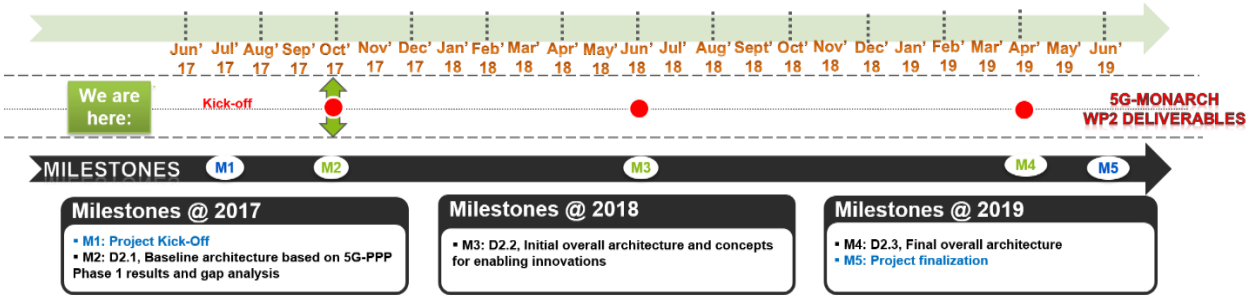


Figure 4-1: Timeline including upcoming WP2 deliverables

[6] 5G PPP Phase 2 Projects - https://5g-ppp.eu/5g-ppp-phase-2-projects/

5 References

[3GPP TR 22.891]	3GPP TR 22.891, "Feasibility Study on New Services and Markets Technology Enablers; Stage 1 (Release 14)", V14.2.0, September 2016
[3GPP TR 23.707]	3GPP TR 23.707 V13.0.0 (2014-12), "Technical Specification Group Services and System Aspects; Architecture Enhancements for Dedicated Core Networks; Stage 2 (Release 13)"
[3GPP TR 23.799]	3GPP TR 23.799 V1.1.0 (2016-10), "Technical Specification Group Services and System Aspects; Study on Architecture for Next Generation System (Release 14)"
[3GPP TR 23.799]	3GPP TR 23.799, "Study on Architecture for Next Generation System"
[3GPP TR 28.800]	3GPP TR 28.800, "Study on management and orchestration architecture of next generation networks and services (Release 14)"
[3GPP TR 28.801]	3GPP TR 28.801, "Study on management and orchestration of network slicing for next generation network (Release 15)"
[3GPP TR 28.802]	3GPP TR 28.802, "Study on management aspects of next generation network architecture and features (Release 14)"
[3GPP TR 36.808]	3GPP TR 36.808, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Carrier Aggregation; Base Station (BS) radio transmission and reception (Release 10)", 2013
[3GPP TR 36.842]	3GPP TR 36.842, "Study on Small Cell Enhancements for E-UTRA and E-UTRAN Higher layer aspects (Release 12)", 2014
[3GPP TR 36746]	TR 36.746, "Study on further enhancements to LTE Device to Device (D2D), User Equipment (UE) to network relays for Internet of Things (IoT) and wearables," Release 15
[3GPP TR 38.801]	3GPP TR 38.801, "Study on new radio access technology: Radio access architecture and interfaces (Release 14)," V14.0.0, March 2017
[3GPP TR 38.806]	3GPP TR 38.806, "Study of separation of NR Control Plane (CP) and User Plane (UP) for split option 2," (Release-15), v0.1.0
[3GPP TS 22.185]	3GPP TS 22.185, "Service requirements for V2X services; Stage 1 (Release 14)," v14.3.0, March 2017
[3GPP TS 22.186]	3GPP TS 22.186, "Enhancement of 3GPP Support for V2X Scenarios; Stage 1 (Release 15)," v15.1.0, June 2017
[3GPP TS 22.261]	3GPP TS 22.261, "Service requirements for the 5G system; Stage 1 (Release 16)," v16.0.0, June 2017
[3GPP TS 23.101]	3GPP TS 23.101, "General Universal Mobile Telecommunications System (UMTS) architecture (Release 13)"
[3GPP TS 23.501]	3GPP TS. 23.501 "System Architecture for the 5G System, Stage 2", Rel. 15 v1.1
[3GPP TS 23.501]	3GPP TS 23.501, "System Architecture for the 5G System; Stage 2 (Release 15)," v1.3.0, September 2017
[3GPP TS 23.501]	3GPP TS 23.501 "System Architecture for the 5G System; Stage 2 (Release 15)," June 2017.
[3GPP TS 23.502]	3GPP TS 23.502, "Procedures for the 5G System, Stage 2," v0.4, Rel. 15
[3GPP TS 33.102]	3GPP TS 33.102, "Technical Specification Group Services and System Aspects; 3G Security; Security architecture"
[3GPP TS 33.401]	3GPP TS 33.401, "Technical Specification Group Services and System Aspects; 3GPP System Architecture Evolution (SAE); Security architecture"
[3GPP TS 36.300]	3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," v13.3.0, April 2016.

[3GPP TS 38.300]	3GPP TS 38.300, "NG Radio Access Network; Overall Description; Stage 2 (Release 15)", v1.0.0, September 2017
[3GPP TS 38.401]	3GPP TS 38.401, "NG-RAN; Architecture description," Release 15, v0.1.0, May. 2017.
[3GPP-LLS]	3GPP RAN TDoc RP-170818, "Study on CU-DU lower layer split for New Radio," March 2017
[5GPPP Phase1 Security]	The 5G Infrastructure Public Private Partnership Security Work Group, "5G-PPP Phase1 Security Landscape", June 2017, available online: https://t.co/XUXHVpEq5C.
[5GARCH16-WPv2]	5G-PPP, Architecture White Paper v2.0, September 2017.
[5GC]	5G-PPP project 5G-Crosshaul, http://5g-crosshaul.eu/
[5GPPP14]	5G-PPP White Paper: "Vision on Software Networks and 5G," Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP_SoftNets_WG_whitepaper_v20.pdf
[5GPPP16]	5G-PPP Architecture Working Group: White Paper "View on 5G Architecture," v1.0, July 2016
[5GPPP17]	5G-PPP White Paper: "View on 5G Architecture (version 2.0)," Available: https://5g-ppp.eu/wp-content/uploads/2017/07/5G-PPP-5G-Architecture-White-Paper-2-Summer-2017_For-Public-Consultation.pdf
[5GX]	5G-PPP project 5G-XHaul, http://www.5g-xhaul-project.eu/
[Alyafawi et al]	I. Alyafawi, E. Schiller, T. Braun, D. Dimitrova, A. Gomes and N. Nikaein, "Critical Issues of Centralized and Cloudified LTE-FDD Radio Access Networks"
[Björnson et al]	E. Björnson, E. Larsson, T. Marzetta, "Massive MIMO: ten myths and one critical question," IEEE Communications Magazine, February 2016
[Bulakci et al]	Ö. Bulakci, Zhe Ren, Chan Zhou, et al, "Towards Flexible Network Deployment in 5G: Nomadic Node Enhancement to Het Net," June 2015.
[Caballero et al]	Caballero Garces, P.; Costa Perez, X.; Samdanis, K.; Banchs, A., "RMSC: A Cell Slicing Controller for Virtualized Multi-Tenant Mobile Networks," in proc. 81st Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st, pp.1-6, 11-14 May 2015.
[Cai et al]	Y. Cai, F. R. Yu and S. and Bu, "Dynamic operations of cloud radio access networks (C-RAN) for mobile cloud computing systems," IEEE Transactions on Vehicular Technology.
[Checko et al]	Checko, A., Christiansen, H. L., Yan, Y., Scolari, L., Kardaras, G., Berger, M. S., & Dittmann, L., "Cloud RAN for Mobile Networks: A Technology Overview," IEEE Communications surveys & tutorials, 2015.
[Checko]	A. Checko, "Cloud Radio Access Network architecture. Towards 5G mobile networks," 2016.
[CORD]	Central Office Re-architected as a Datacentre (CORD), http://opencord.org/
[Coutinho et al]	Coutinho, E. F., de Carvalho Sousa, F. R., Rego, P. A. L., Gomes, D. G., and de Souza, J. N, "Elasticity in cloud computing: a survey," Annals of Telecommunications, 2015.
[CPRI15]	CPRI, "Common Public Radio Interface (CPRI); Interface Specification," v7.0, October 2015
[eCPRI17]	CPRI, "Common Public Radio Interface; eCPRI Interface Specification," v1.0, August 2017
[ETSI GS NFV-MAN]	ETSI GS NFV-MAN 001, "Network functions virtualisation (NFV); management and orchestration," v1.1.1, December 2014
[ETSI14-ORI]	ETSI GS ORI 002-1/2, "Open Radio equipment Interface (ORI); ORI interface Specification; Part 1: Low Layers / Part 2: Control and Management," v4.1.1, October 2014

[ETSI15-NFV]	ETSI GS NFV-IFA 001, "Network Functions Virtualisation (NFV); Acceleration Technologies; Report on Acceleration Technologies & Use Cases," v1.1.1, December 2015
[ETSI-mWT]	European Telecommunications Standards Institute Industry Specification Group Millimetre Wave Transmission (ETSI ISG mWT), http://www.etsi.org/technologies-clusters/technologies/millimetre-wave-transmission
[Fritzsche et al]	R. Fritzsche, P. Rost, and G. P. Fettweis, "Robust Rate Adaptation and Proportional Fair Scheduling With Imperfect CSI," IEEE Transactions on Wireless Communications, vol. 14, No. 8, August 2015
[Fujitsu]	Fujitsu White Paper, "The Benefits of Cloud-RAN Architecture in Mobile Network Expansion"
[Ghaderi et al]	M. Ghaderi, D. Towsley and J. Kurose, "Reliability Gain of Network Coding in Lossy Wireless Networks," INFOCOM 2008. The 27th IEEE Conference on Computer Communications, Phoenix, AZ, 2008
[Hämäläinen et al]	S. Hämäläinen, H. Sanneck, C. Sartori, "LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency," John Wiley and Sons, Dec. 2011.
[Herbst et al]	Herbst, N. R., Kounev, S., and Reussner, R. H., "Elasticity in Cloud Computing: What It Is, and What It Is Not," in IEEE ICAC, 2013.
[IETF SDN]	IETF Proceedings, "SDN Controller Performance Evaluation," Available online: https://www.ietf.org/proceedings/96/slides/slides-96-sdnrg-11.pdf
[iJOIN D5.3]	iJOIN Deliverable D5.3, "Final definition of iJOIN architecture", Apr. 2015
[Kerttula et al]	J. Kerttula, N. Malm, K. Ruttik, R. Jäntti and O. Tirkkonen, "Implementing TD-LTE as Software Defined Radio in general purpose processor"
[Kokku et al]	Kokku, R.; Mahindra, R.; Honghai Zhang; Rangarajan, S., "CellSlice: Cellular wireless resource slicing for active RAN sharing," in proc. 5th conference on Communication Systems and Networks (COMSNETS), pp.1-10, 7-10 Jan. 2013.
[Koudouridis et al]	G P Koudouridis, P Soldati and G Karlsson, "Multiple Connectivity and Spectrum Access Utilisation in Heterogeneous Small Cell Networks", International Journal of Wireless Information Networks, March 2016, Volume 23.
[LTEsim]	G. Piro et al., "Simulating LTE Cellular Systems: An Open-Source Framework," IEEE Transactions on Vehicular Technology, 60(2):498–513, Feb 2011.
[Makris et al]	N. Makris, P. Basaras, T. Korakis, N. Nikaein and L. Tassiulas, "Experimental evaluation of functional splits for 5G cloud-RANs," 2017 IEEE International Conference on Communications (ICC), Paris, 2017.
[MCN D4.3]	MCN, "D4.3: Algorithms and Mechanisms for the Mobile Network Cloud," 2014.
[MCN factsheet]	MCN factsheet, "Mobile Cloud Networking: Mobile Network, Compute, and Storage as One Service On-Demand," Available: http://cordis.europa.eu/fp7/ict/future-networks/documents/call8-projects/mobilecloudnetworking-factsheet.pdf
[MCN]	Mobile Cloud Networking website, http://www.mobile-cloud-networking.eu/site/ [accessed August 2017].
[MEC]	Juniper, "Mobile Edge Computing Use Cases & Deployment," Available: https://www.juniper.net/assets/us/en/local/pdf/whitepapers/2000642-en.pdf
[METIS II D2.4]	METIS-II, Deliverable D2.4, "Final Overall 5G RAN Design," June 2017
[METIS II D4.2]	METIS-II, Deliverable D4.2, "Final air interface harmonization and user plane design," April 2017

[METIS II D5.2]	METIS-II, Deliverable D5.2, "Final Considerations on Synchronous Control Functions and Agile Resource Management for 5G," March 2017.
[Michalopoulos et al]	D. S. Michalopoulos, I. Viering and L. Du, "User-plane multi-connectivity aspects in 5G," 23rd Int. Conference on Telecommunications (ICT), 2016
[Mirantis OpenCloud]	Mirantis OpenCloud Digest "What is the best NFV Orchestration Platform?," Available: https://www.mirantis.com/blog/which-nfv-orchestration-platform-best-review-osm-open-o-cord-cloudify/
[Nazer and Gastpar]	B. Nazer and M. Gastpar, "Compute-and-Forward: Harnessing Interference through Structured Codes," IEEE Transactions on Information Theory, Vol. 57, Oct 2011, pp. 6463-6486.
[NGFI]	IEEE Next Generation Fronthaul Interface (1914) Working Group, https://standards.ieee.org/develop/wg/NGFI.html http://sites.ieee.org/sagroups-1914/
[NGFI15]	China Mobile, et al., "White Paper of Next Generation Fronthaul Interface," October 2015
[NGMN 5G security]	NGMN Alliance, "5G security recommendations, Package #2: Network Slicing, V1.0," April 2016, Available: https://www.ngmn.org/uploads/media/160429_NGMN_5G_Security_Network_Slicing_v1_0.pdf
[NGMN15]	NGMN Alliance, "5G white paper," White Paper, v1.0, February 2015.
[NGMN17]	NGMN Alliance, "5G End-to-End Architecture Framework by NGMN Alliance," 11-May-2017.
[Nguyen et al]	Van Giang Nguyen; Young Han Kim, "Slicing the next mobile packet core network," 11th International Symposium in Wireless Communications Systems (ISWCS), pp.901-904, 26-29 Aug. 2014.
[Nikaein]	N. Nikaein, "Processing Radio Access Network Functions in the Cloud: Critical Issues and Modeling"
[NORMA D2.2]	5G-NORMA Deliverable D2.2, "Evaluation methodology for architecture validation, use case business models and services, initial socio-economic results," August 2016
[NORMA D3.2]	5G NORMA Deliverable D3.2, "5G NORMA network architecture – Intermediate report," January 2017
[NORMA D4.1]	5G NORMA Deliverable D4.1, "RAN architecture components – preliminary concepts," November 2016
[NORMA D4.2]	5G NORMA Deliverable D4.2, "RAN architecture components – final report," June 2017
[NORMA D5.2]	5G NORMA Deliverable D5.2, "Definition and specification of connectivity and QoE/QoS management mechanisms – final report," June 2017
[NORMA D6.1]	5G NORMA Deliverable D6.1, "Demonstrator design, implementation and initial set of experiments," October 2016
[ns3]	ns3, http://networks.cttc.es/mobile-networks/software-tools/lena/
[ONF]	Open Networking Foundation (ONF), https://www.opennetworking.org/
[ONOS]	ONOS, "Introducing ONOS - a SDN network operating system for service providers," ON.LAB, 11 2014. Available: http://onosproject.org/wp-content/uploads/2014/11/Whitepaper-ONOS-final.pdf
[OpenAirInterface]	OpenAirInterface, Available: http://www.openairinterface.org/
[Ravanshid et al]	A. Ravanshid et al., "Multi-connectivity functional architectures in 5G," 2016 IEEE International Conference on Communications Workshops (ICC), Kuala Lumpur, 2016, pp. 187-192.

[Rost and Pravad]	P. Rost and A. Prasad, "Opportunistic Hybrid ARQ – enabler of centralized-RAN over non-ideal backhaul," IEEE Wireless Communications Letters, vol. 3, no. 5, October 2014.
[RP-150441]	RP-150441, "Revised WI: Enhanced LTE Device to Device Proximity Services," Release 13
[Sabella et al]	D. Sabella et al., "Benefits and challenges of cloud technologies for 5g architecture," in Proceedings of 1st International Workshop on 5G Architecture (5GArch 2015), Glasgow, UK, May 2015.
[Salman et al]	Salman, Ola & Elhajj, Imad & Kayssi, Ayman & Chehab, Ali. (2016), "SDN controllers: A comparative study," 1-6. 10.1109/MELCON.2016.7495430
[SCF 159]	Small Cell Forum, "Small Cell Virtualisation Functional Splits and Use Cases," Document 159.07.02, Jan. 2016.
[Sciancalepore et al]	V. Sciancalepore, K. Samdanis, X. Costa, D. Bega, M. Gramaglia, A. Banchs, "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization," in IEEE INFOCOM 2017
[SDxCentral]	SDxCentral, "Market Report; The Future of Network Virtualization and SDN Controllers," 2016.
[SF]	SUPERFLUIDITY, https://5g-ppp.eu/superfluidity/ [accessed Aug, 2017].
[Silva et al]	Icaro da Silva, Gunnar Mildh, Alexandros Kaloxylos, Panagiotis Spapis, Enrico Buracchini, Alessandro Trogolo, Gerd Zimmermann, Nico Bayer, "Impact of network slicing on 5G Radio Access Networks," EuCNC 2016: pp.153-157
[simuLTE]	simulLTE, http://simulte.com/
[srsLTE]	srsLTE, http://www.softwareradiosystems.com/products/#srslte
[Tran et al]	T. X. Tran, A. Younis and D. Pompili, "Understanding the Computational Requirements of Virtualized Baseband Units using a Programmable Cloud Radio Access Network Testbed"
[Wu et al]	L. Wu, H. Luo and Z. Ai, "Proactive fault management in large scale computing systems," Journal of Huazhong University of Science and Technology, vol. 38, no. S1, pp. 20-24, 2010
[XRAN16]	xRAN.org, White Paper "The mobile access network, beyond connectivity," October
[Xu et al]	M. Xu, Y. Yang and Q. Li, "Selecting shorter alternate paths for tunnel-based IP fast reroute," Computer Networks, vol. 56, no. 2, pp. 845-857, 2012
[Yang et al]	Yang Cao, Shiyong Yang, Tao Jiang and Daiming Qu, "Performance Optimization for Cyber Foraging Network via Dynamic Spectrum Allocation," IEEE INFOCOM 2011 WS on Cloud Computing, Apr. 2011
[Yousaf et al]	F. Z. Yousaf, M. Gramaglia, V. Friderikos, B. Gajic, P. Arnold, D. v. Hugo, V. Sciancalepore, M. R. Crippa, O. Holland, I. Labrador, "Resource Sharing for a 5G Multi-tenant and Multi-service Architecture," European Wireless 2017, Workshop on COmpetitive and COoperative Approaches for 5G networks (COCOA)
[Yu et al]	T. Yu, S. Chen, Z. Qin, "A rerouting scheme using connected dominating set in large-scale disaster scenario," Chinese High Technology Letters, vol. 18, no. 1, pp. 11-15, 2008
[Zhou et al]	S. Zhou, Z. Ou and Y. Yuan, "The technology of on-line diagnosis, fault isolation and dynamic rebuilding for network," Application Research of Computers, vol. 20, no. 1, pp. 92-93, 2003.

6 Appendix: Detailed State-of-the-Art for Experiment-driven Optimisation

The 5G-MoNArch project moves one step beyond the current state of play and it not only validates an improved 5G architecture, but also device and implement system-level and NF-level algorithms that build on insights gained from experiments. In the literature, many projects have led to results that can feed the experiment-driven modelling and optimisation approach. For example, the Mobile Cloud Networking (MCN) project provides some key baseline approaches for telco cloud deployment, focusing on [MCN]: (i) Radio Access Network (RAN) Virtualisation, (ii) infrastructure-as-a-Service (IaaS), (iii) investigation, implementation and evaluation of Cloud RAN, which is On-Demand, Elastic, and pay-as-you-go. Additionally, in an experiment performed by a group from Rutgers University, the computational requirement for the implementation of small-scale C-RAN has been studied in [Tran et al]. In SUPERFLUIDITY project another important side aspect of experimental based approach was revealed. As shown in the project, a significant volume of complex work is required to identify, install and put in operation the different virtualised and cloud-enabled components that define the network.

The 5G-NORMA project has also contributed with a set of experimental based results through a series of demos. The most relevant to 5G-MoNArch project are the 5G-NORMA Service-aware QoE/QoS Control Demo and the 5G-NORMA Secured Multi-Tenancy Virtual Network Resources Provisioning [NORMA D6.1].

In these projects, the Open Air Interface (OAI) [OpenAirInterface] has been recognised as the major tool for reliable measurement campaigns. OAI is an open source complete protocol stack software consisting PHY, MAC, RLC, PDCP and RRC layers provided by Open air interface (OAI) Software Alliance (OSA). OAI fills the gap between dedicated hardware and software based network functions.

Other simulation, emulation and testbed tools, include srsLTE [srsLTE], NS3 LENA module [ns3], LTE-Sim [LTEsim], and SimuLTE [simuLTE]. All of them provide a good basis for simulations that will extend and put into the cloud the LTE protocol stack. However, they lack a real execution environment that respects frame timing constraints, integrated all the network components in an E2E and cloud-enabled basis, and have slow evolution process regarding the 5G features. Regarding the emulation profiles of the OAI, i.e., the "DLsim" and "ULsim" tools further study can be found in [MCN-D4.3], while an interesting evaluation of the reliability of the OAI emulated results, compared to the OAI testbed ones (realistic measurement) can be found in [Makris et al].

In the following, more details are given bellow on relevant experiment-based approaches that are adopted in research and innovation project relevant to 5G-MoNArch.

Mobile Cloud Networking

The Mobile Cloud Networking (MCN) was a EU FP7 Large–scale Integrating Project funded by the European Commission started in November 2012 with the duration of 36 months. The motivation behind this project was to deploy telecommunication network to telco cloud. Despite being the research period of the project half a decade ago, it provides some key baseline approaches for telco cloud deployment [MCN].

Key research and innovation issues handled by MCN are as followings [MCN factsheet]:

- Radio Access Network(RAN) Virtualisation,
- Infrastructure-as-a-Service(IaaS),
- Investigate, implement and evaluate Cloud RAN, which is On-Demand, Elastic, and pay-as-you-go.

Profiling approach

In the setup prepared by MCN, C-RAN has been studied, in which signal processing has been performed in the cloud environment on virtual machines deployed on General Purpose Processors (GPP). Open Air Interface (OAI) Release 8 is used as a protocol stack software [OpenAirInterface]. As mentioned above, OAI provides two profiling tools namely "dlsim" and "ulsim". "dlsim" emulates PDSCH of eNodeB and UE while "ulsim" emulates PUSCH for the both. For each process, number of clock cycle has been measured. Time taken by each process has been measured by dividing number of clock cycle by CPU frequency. Processing time taken by LTE PHY layer has been observed given different number of

assigned Physical Resource Blocks(PRBs), different platform environment (GPP, KVM and cloud (ZHAW open stack bed)) and different Modulation and Coding Scheme (MCS).

Profiling Results

Profiling results of MCN are summarised below. Dlsim emulates PDSCH downlink channel and ulsim emulates PUSCH uplink channel. In Figure 6-1, processing time taken by OAI dlsim to encode one subframe has been plotted with several MCS (0, 2, 9, 10, 16, 17, and 27), system bandwidth (5,10, and 20 MHz), and machine environment (dedicated GPP, KVM and cloud). Machines have 2 GB of RAM and CPU with clock frequency 2.4 GHz. In Figure 6-2, processing time required by OAI ulsim to decode one subframe have been plotted.

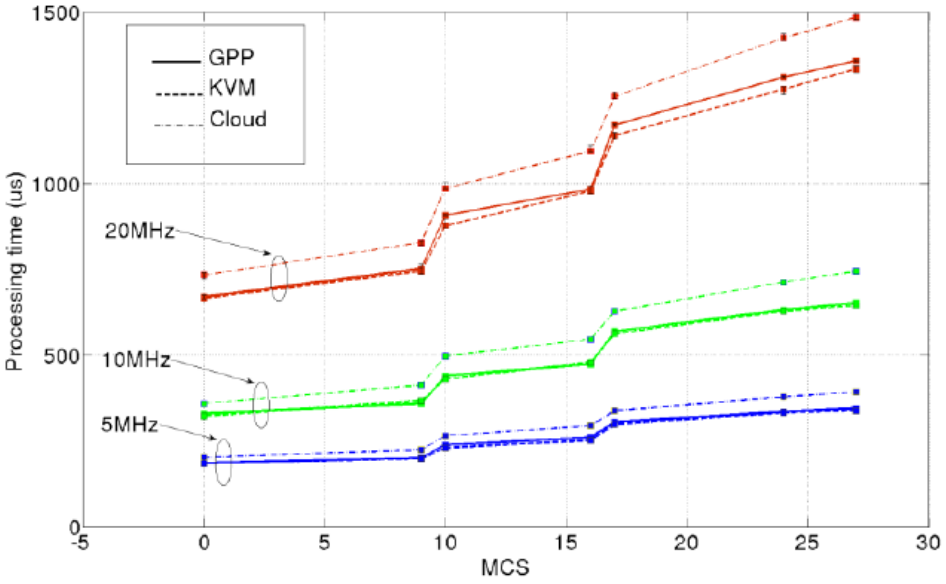


Figure 6-1: Processing time taken by dlsim to encode one subframe (extracted from [MCN-D4.3])

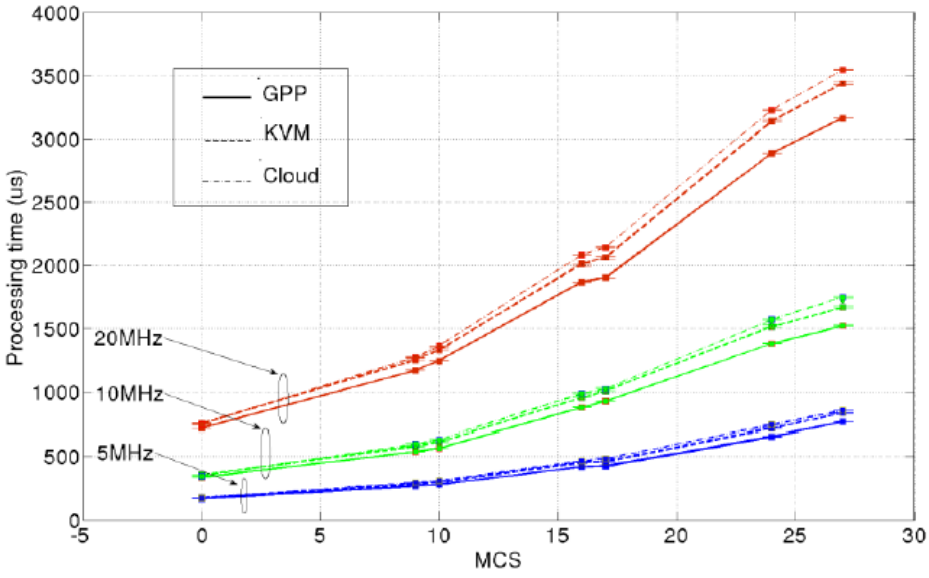


Figure 6-2: Processing time taken by OAI-ulsim (extracted from [MCN-D4.3])

From the graphs shown above, it has been observed that computational resource requirement is directly proportional to modulation and coding scheme (MCS) and number of PRBs allocated to UE. A worst case scenario can be modelled with UE transmitting PUSCH with MCS 27 [Kerttula et al].

After proposing BBU-RRH split, further research has been done to find maximum BBU-RRH distance possible. Existing fronthaul solutions limit the maximum BBU-RRH distance to 60 km.

C-RAN Testbed Rutgers University

The characterising the computational requirement has been studied in [Tran et al] based on implementation of small-scale C-RAN. This Experiment performed by a group from Rutgers University, also used OAI eNodeB as the virtualised RAN. Figure 6-3 illustrates the logical architecture of the C-RAN testbed. Unlike the previous project where only physical layer has been studied, in this experiment the higher layers are included as well. RRH front heads are implemented using SDR (Software Defined Radio) USRP B210. BBU computation has been performed on Intel Xeon-E5 1650 CPU which consists of 12 cores operating at 3.5 GHz and 32 GB RAM. UE runs on Intel Core i7 with operating frequency 3.6 GHz.

Performance of virtualised BBU has been measured in terms of packet delay, GPP processing time and utilisation under various MCS and PRB configuration. Outcomes of the project are briefly provided below [Tran et al]

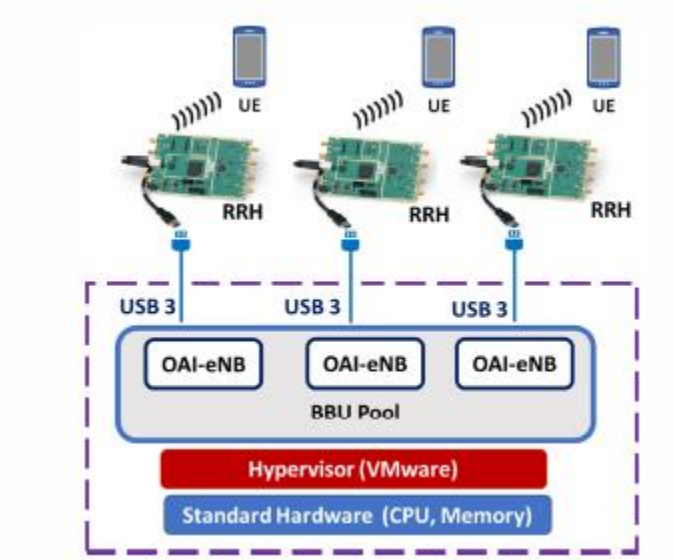


Figure 6-3: The C-RAN testbed logical architecture used in [extracted from [Tran et al]]

Based on the experimental results, it is concluded that the latency requirements can be satisfied only if the frequency used GPP be greater than or equal to 2.5 GHz. In addition, the paper proposed a model for the subframe processing time, given by:

$$t_{sub[\mu s]} = \frac{\alpha_{PRB}}{f_{[Hz]}} + \beta_{MCS} + 2.508$$

where:

- t_{sub}: the subframe processing time,
- f: the frequency of used GPP,
- α_{PRB}: Coefficient based on number of PRBs
- β_{MCS}: Coefficient based on the used MCS (Modulation and Coding Scheme)

Table 6-1: Coefficient values for PRB and MSC (extracted from [Tran et al])

PRB	25	50	100				
$\alpha_{PRB}[\mu s]$	900	940	970				
MCS	0	9	10	16	17	24	27
$\beta_{MCS}[\mu s]$	0	9.7	11.8	37.5	39.7	64.8	75

The CPU utilisation as function of downlink throughput is given by:

$$U_{CPU[\%]} = 0.6237 R_{b[\text{Mbps}]}^{DL} + 21.3544$$

where:

- U_{CPU}: percentage of CPU usage,
- R_b^{DL}: downlink throughput.

5G-NORMA Service-aware QoE/QoS Control testbed

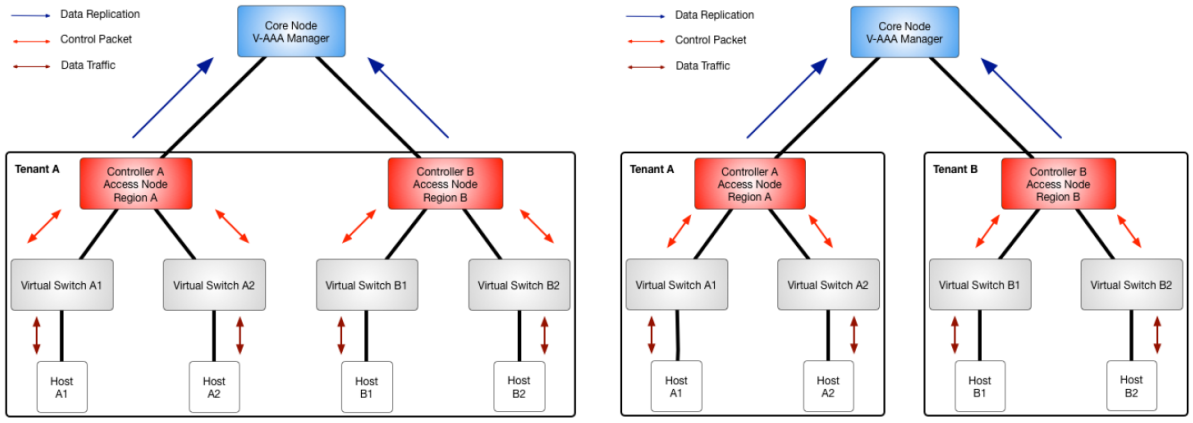
The main innovation demonstrated by 5G-NORMA Service-aware QoE/QoS Control testbed [NORMA D6.1] is network slicing up to the RAN functions. Lower RAN functions (up to PDCP) are considered as common network functions for the two slices. Higher RAN and EPC are deployed as dedicated NFs, with different instances per network slice. This is aligned with the 5G NORMA architecture.

Other functionalities demonstrated are service aware orchestrator and enhanced mobility. Orchestration of dedicated network functions provides different outcomes according to the targeted network slices. For the LL slice, the data-plane is located in the edge cloud, while MBB gets the data-plane located in the central cloud. Although not achieved through an on-line algorithm, the outcome is still a valuable result for the 5G NORMA innovations. Enhanced mobility mechanisms are applied to the LL slice such that the local breakout routing optimisation avoids unnecessary usage of the Packet Data Network (PDN) [MEC]. The demo HW/SW platform is based on the Universal Software Radio Peripheral Software Defined Radio (USRP SDR) and the OpenAirInterface software implementation of the RAN and EPC stack.

5G-NORMA Secured Multi-Tenancy Virtual Network Resources Provisioning via V-AAA

The KCL Virtual Authentication Authorisation Accounting (V-AAA) testbed is a complementary demonstration in 5G NORMA and provides infrastructure for conducting small-scale repeatable experiments of the 5G NORMA architecture, especially experiments that involve secure multi-tenancy and multi-tenant data isolation on the access network (edge cloud). The testbed is based on 1) commodity hardware, i.e., Raspberry Pi, home Wi-Fi router and switch and 2) open source software, i.e., Ryu controller, Open Authentication protocol version 2, couchDB, openVSwitch, that has been configured and extended to provide a hierarchical and distributed database cluster for Tenant isolation and replication of the Tenant data (e.g. billing data, Tenant service logs).

Two showcases were developed based on the commodity hardware testbed that is presented in Figure 6-4. These two showcases have been divided into two scenarios, i.e., in a single-tenant scenario and a multi-tenant scenario. More specifically, the multi-tenant scenario has been divided into two parts: direct request of network resources and indirect request of network resources. Mainly, these scenarios are differentiated by the network resource provisioning in deploying a network resource (e.g. virtual switch port) and obtaining the information for billing purpose. Once the network resources have been deployed, the billing information writes into the database and replicates this information to the hierarchical database at the core node.



a) single-tenant scenario b) multi-tenant scenario

Figure 6-4: A logical entities representation of data flow in single-tenant and multi-tenant scenarios

SUPERFLUIDITY

SUPERFLUIDITY is a European Union's Horizon 2020 research and innovation programme which aims at achieving superfluidity in the Internet: the ability to instantiate services on-the-fly, run them anywhere in the network (core, aggregation, edge) and shift them transparently to different locations. The experimentation approach of the SUPERFLUIDITY project is based on a sequence of use cases called scenes. The initial scene, represents the first step that the infrastructure and the management/orchestration/design tools are deployed. A significant volume of complex work is required for this initial preparation, to identify, install and put in operation a large number of complex components that inter-operate, providing support to the integrated execution of Superfluidity components. Some of the tools and software (e.g. OpenStack, Kubernetes, Grafana, InfluxDB) are not specific results of the project but their configuration and modification have been important to realise the demonstrator. Additionally, an end-user perspective is provided, to better understand the scene objective and the role of the end-users and other primary users in the scene execution. Table 6-2 summarises aspects regarding the technical and end-user perspective of the use cased that are demonstrated in SUPERFLUIDITY project.

Table 6-2: Summary of use cases/scenes demonstrated in SUPERFLUIDITY project

Scene	Summary	End-User perspective/experience	Technical perspective
0	Infrastructure setup	No end-user involvement. Infrastructure provider is involved at this stage.	Establish the infrastructure (hardware & NFVI)
1.a	Superfluidity system design	No end-user involvement. The system designer, telecommunications operator and the network administrators would be involved at this stage.	Edge and cloud, network and service platforms design and deployment artefacts generation.
1.b	Initial components deployment (CRAN and MEC)	No end-user involvement. The system administrator and telecommunications operator would monitor and operate at this point.	Network and service platforms components deployment at edge and central clouds. CRAN and MEC are defined and deployed programmatically.
2.a	Workloads offline characterisation	No end-user involvement. To understand the likely impact of a workload on the deployed infrastructure it is important to model and classify the workloads. This is done by the tools and modelling experts.	Characterisation of workloads on the deployment environment, for scaling mechanisms support.
2.b	Workload (video streaming) / service deployment	No end-user involvement. The service provider will set up the services that will be exposed to the end-users.	Initial deployment of the workload based on a combination of model profiles and the service providers' inputs.
3.a	Central cloud services automatic scaling	An end-user streams a video from the network. The demand at the core grows or there are noisy-neighbours triggering scaling.	Other users are simulated using a simulated load profile and this load automatically triggers the server scaling based on predefined actions.

3.b	Services relocation to edge cloud	An end-user, now connected to the CRAN, also streams a video from the network.	An operator policy triggers a service component to be instantiated at the attachment edge and the video content is then streamed via the MEC.
3.c	Container based services deployment at the edge cloud	Due to the end-user reaching a certain level of content watched, a user-specific advertisement is displayed.	Using a service at the edge that combines video stream content and user-specific content into an advert stream.
3.d	Unikernel based services deployment at the edge cloud	The end-user continues to stream the video, meanwhile there is a DDoS attack at the edge cloud.	A DDoS attack is detected through the OPP running in the MEC and triggers a set of SDN rule modifications combined with xFSM.
3.e	Services' optimisation at the edge cloud	The end-user changes to utilise an encrypted variant of the video stream.	The ADC is realised through a deployment of Citrix NetScaler at the Edge cloud.
4.a	Alternative load balancing function at the central cloud	The end-user continues to stream the video. The network operator though decides to replace the open LBaaS with a commercial variant that promises better performance.	The VNF of the LBaaS is replaced with minimal service disruption to the video streaming service.
4.b	Alternative edge offloading function	The end-user continues to watch the video stream. The operator decides to switch the traffic off-loader to use fast-click for performance reasons.	The TOF used in 3.b is replaced with a Fast-Click based alternative as a plug-in replacement.
4.c	Advanced network control and management	Currently it is not possible to keep the end-user service running through this disruptive change.	Once modified the system operator can modify the SDN rules programmatically and change the end-user experience based on configurations and desired SLAs.