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 D2.3

Final overall architecture

Contractual Date of Delivery	2019-04-30
Actual Date of Delivery	2019-04-30
Work Package	WP2 – Flexible and adaptive architecture design
Editor(s)	Ömer Bulakci (HWDU), Qing Wei (HWDU)
Reviewers	Peng Chenghui (HWDU), Muhammad Naseer-Ul-Islam (NOK-DE)
Dissemination Level	Public
Type	Report
Version	1.0
Total number of pages	165

Abstract: This deliverable presents the final 5G-MoNArch "Overall Architecture" as an evolution from D2.2 "Initial Architecture". It presents the final results of the project's *enabling innovations* and elaborates on the concepts for architectural extensibility and customisation capitalising on the *functional innovations*, to enable the realisation of the Smart Sea Port and the Touristic City use cases.

Taking into consideration the results on elastic slice management from WP4, the enablers for resilient and secure slices from WP3, and the 5G system (5GS) gap analysis derived in WP2, the initial architecture has been refined based on three design pillars: (i) split of control and user plane, (ii) unified service-based architecture across all layers and domains of the 5GS, and (iii) end-to-end (E2E) slicing support. The E2E slicing support is provided by the enablers pertaining to inter-slice control and management. The experiment-driven optimisation is highlighted, which paves the way for realistic virtual network function design and novel orchestration algorithms. The telco-cloud-enabled protocol stack focuses on the paradigm shift of flexible RAN network function (NF) implementation, where the inter-NF dependencies are relaxed. The final overall architecture is further detailed, where the functional architecture descriptions for each layer are presented in detail along with the interfaces and specific protocols needed for the 5GS realisation. Finally, the impact of these innovations on the ongoing standards and their benefits for the 5G ecosystem requirements are presented.

Keywords: 5G Architecture, E2E Network Slicing, Enabling Innovations, Architectural Extensibility

Executive Summary

This deliverable presents the final results on the design of the 5G-MoNArch architecture. Specifically, the initial architecture presented in deliverable D2.2 [5GM-D2.2] has been refined towards the "final architecture". This updated architecture, indicated as "Final 5G-MoNArch Overall Architecture", is described in detail, thereby structured into four layers: (i) Service layer, (ii) Management & Orchestration (M&O) layer, (iii) Controller layer, and (iv) Network layer.

The overall 5G-MoNArch architecture is built upon the novel specific design aspects, i.e., end-to-end (E2E) slicing support across four network layers, specification of slice-specific and slice-common functions, multi-tenancy capable M&O, inter-slice resource management, and optional integration of radio access network (RAN) control applications. In addition to these design aspects, in this deliverable, further novel features built upon the initial architecture are highlighted: (i) Service-based characteristics spanning all layers with unified service-based interface (SBI) design; (ii) the ability to collect and analyse per-slice aggregated data, and to aid network optimisation via the novel end-to-end (E2E) integrated data analytics framework comprising domain-specific data analytics functions, i.e., network data analytics function (NWDAF), management data analytics function (MDAF), and radio access network (RAN) data analytics function (RAN-DAF); (iii) the interworking between the aforementioned functions. In particular, considering the 5G-MoNArch Itf-X interface defined in D2.2 [5GM-D2.2] as the basis, the notion of inter-domain and intra-domain interfaces capitalising on the service-based architecture (SBA) principles has been extended. This unified interface description enables the interactions between functions on the M&O layer and functions on the Network layer, in order to achieve enhanced flexibility and orchestration capabilities.

Furthermore provided is the impact of 5G-MoNArch contributions towards standardisation, such as the concept of data duplication for both the data plane and the control plane, to provide resilient RAN operations as well as data analytics both in 5G core network (CN) and management plane. Besides the 3rd Generation Partnership Project (3GPP) contributions, this deliverable highlights the 5G-MoNArch collaboration with the Next Generation Mobile Networks (NGMN) Alliance, GSM Association (GSMA), European Telecommunications Standards Institute (ETSI) Zero touch network and Service Management (ZSM) Industry Specification Group (ISG), and ETSI Experiential Networked Intelligence (ENI) ISG. 5G-MoNArch has been collaborating with GSMA to define the concept of 5G-MoNArch slice blueprint and provide an efficient tool for designing and deploying network slices. Moreover, in the framework of the ETSI ENI, a use case and a proof-of-concept (PoC) related to cross-slice elastic resource management and orchestration based on the 5G-MoNArch innovations have been proposed.

To realise the individual 5G-MoNArch features, a set of novel enablers and innovation elements has been developed that map onto three *enabling innovations*: (i) telco-cloud-enabled protocol stack, (ii) inter-slice control and management, and (iii) experiment-driven optimisation. Herein, final evaluation results are provided in this deliverable along with their impact on the 5G-MoNArch architecture and protocol stack. As the final overall architecture of 5G-MoNArch includes the components emerging from WP3 [5GM-D3.2] and WP4 [5GM-D4.2] *functional innovations*, these are briefly presented in this deliverable as well. The implications of the enabling and functional innovations on the 5G-MoNArch final overall architecture are captured via the 5G-MoNArch novel components that are explicitly highlighted including the functional architecture descriptions of different domains. The specific protocols are described via message sequence charts (MSCs) that are needed for the 5GS realisation.

Finally, the deliverable focuses on the specific instantiations of the 5G-MoNArch architecture with respect to the functional innovations developed in WP3 [5GM-D3.2] and WP4 [5GM-D4.2], and the associated testbed use cases, in order to demonstrate network slicing elasticity, resilience, and security. To achieve this goal, the innovations related to the network slice design, deployment, and lifecycle management are presented. The 5G-MoNArch Network Slice Blueprint concept is the universal means for such service-specific design and operations of network slices. To demonstrate the concept, this deliverable elaborates on how these innovations can be used to create and manage use case-specific network slices in the two 5G-MoNArch testbeds, namely, the Hamburg Smart Sea Port and the Turin Touristic City and provides an analysis on their impact on the evolution of the 5G ecosystem from a business perspective.

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
	Francisco Valera	fvalera@it.uc3m.es
DT	Markus Breitbach	m.breitbach@telekom.de
	Gerd Zimmermann	zimmermanng@telekom.de
	Paul Arnold	paul.arnold@telekom.de
	Michael Einhaus	einhaus@hft-leipzig.de
	Igor Kim	kim@hft-leipzig.de
	Mohamad Buchr Charaf	charaf@hft-leipzig.de
NOK-FR	Aravinthan Gopalasingham	gopalasingham.aravinthan@nokia-bell-labs.com
	Bessem Sayadi	bessem.sayadi@nokia-bell-labs.com
	Fred Aklamanu	fred.aklamanu@nokia-bell-labs.com
HWDU	Ömer Bulakci	oemer.bulakci@huawei.com
	Qing Wei	qing.wei@huawei.com
	Emmanouil Pateromichelakis	emmanouil.pateromichelakis@huawei.com
	Riccardo Trivisonno	riccardo.trivisonno@huawei.com
	Clarissa Marquezan	clarissa.marquezan@huawei.com
	Onurcan Iscan	onurcan.iscan@huawei.com
TIM	Fabrizio Moggio	fabrizio.moggio@telecomitalia.it
	Andrea Buldorini	andrea.buldorini@telecomitalia.it
	Roberto Querio	roberto.querio@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
	Ghina Dandachi	ghina.dandachi@cea.fr
CERTH	Anastasios Drosou	drosou@iti.gr
	Asterios Mpatziakas	ampatziakas@iti.gr
	Stavros Papadopoulos	spap@iti.gr
MBCS	Dimitris Tsolkas	dtsolkas@mobics.gr
	Odysseas Sekkas	sekkas@mobics.gr
RW	Julie Bradford	julie.bradford@real-wireless.com

NOMOR	Sina Khatibi	khatibi@nomor.de
UNIKL	Marcos Rates Crippa	crippa@eit.uni-kl.de
	Bin Han	binhan@eit.uni-kl.de

List of Acronyms and Abbreviations

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
AF	Application Function
AMF	Access and Mobility Management Function
AN	Access Network
AR	Augmented Reality
ARP	Allocation and Retention Priority
B2B	Business-to-Business
BS	Base Station
CAPEX	CAPital EXpenditure
CN	Core Network
CP	Control Plane
CQI	Channel Quality Indicator
CSC	Communication Service Customer
CSCC	Cross-slice Congestion Control
CSMF	Communication Service Management Function
CSP	Communication Service Provider
CU	Central Unit
D-RAN	Distributed RAN
DRB	Data Radio Bearer
DSC	Dynamic Small Cell
DU	Distributed Unit
E2E	End-to-End
eMBB	enhanced Mobile Broadband
eNA	enablers for Network Automation
ENI	Experiential Networked Intelligence
ETSI	European Telecommunications Standards Institute
gNB	NR access node with user plane and control plane
GSM	Global System for Mobile Communications
GSMA	GSM Association
GST	Generic Slice Template
HARQ	Hybrid Automatic Repeat Request
HLS	Higher Layer Split
IAB	Integrated Access Backhaul
IEEE	Institute of Electrical and Electronics Engineers
IOT	Internet of Things
IM	Interference Management
ISCF	Inter Slice Correlation Function
ITS	Intelligent Transport System
KPI	Key Performance Indicator
LTE	Long Term Evolution
MAC	Medium Access Control
M&O	Management and Orchestration layer
MANO	ETSI MANagement and Orchestration
MBB	Mobile BroadBand
MCS	Modulation and Coding Scheme

MF	Management Function
mMTC	Massive Machine Type Communication
MS	Management Service
NAS	Non-Access Stratum
NWDAF	Network Data Analytics function
NBI	NorthBound Interface
NE	Network Element
NEF	Network Exposure Function
NEST	NEtwork Slice Template
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
PER	Packet Error Rate
PDB	Packed Delay Budget
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PHY	Physical Layer
PLMN	Public Land Mobile Network
PNF	Physical Network Function
PoC	Proof of Concept
PRB	Physical Resource Block
QoE	Quality of Experience
QoS	Quality of Service
RA	Registration Area
RAN	Radio Access Network
RAN-DAF	RAN Data Analytics Function
RAT	Radio Access Technology
RCA	RAN Controller Agent
RCM	RAN Configuration Mode
RL	Reinforcement Learning
RLC	Radio Link Control
RM	Resource Management
RRC	Radio Resource Control
RRM	Radio Resource Management
SBA	Service-Based Architecture
SDAP	Service Data Adaptation Protocol
SDO	Standards Developing Organisation
SE	Spectral Efficiency

SINR	Signal-to-Interference-and-Noise Ratio
SLA	Service Level Agreement
SMF	Session Management Function
SMm	Security Monitoring Manager
STZm	Security Trust Zone Manager
TN	Transport Network
TSN	Time-Sensitive Networking
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URI	Uniform Resource Identifier
VIM	Virtual Infrastructure Manager
VNF	Virtual Network Function
VNFM	Virtual Network Function Manager
VNN	Vehicular Nomadic Node
VR	Virtual Reality
V2X	Vehicle to Anything
WG	Working Group
WP	Work Package
ZSM	Zero touch network and Service Management

List of Figures

List of Tables

1 Introduction

Since the early research phase of the fifth generation (5G) of mobile and wireless communications networks starting in 2012, the development of the 5G system (5GS) has progressed at a rapid pace. Within the 5GS, end-to-end (E2E) network slicing, service-based architecture, Software Defined Networking (SDN), and Network Function (NF) Virtualisation (NFV) are seen as the fundamental pillars of the architectural design to support in a cost-efficient way the heterogeneous key performance indicators (KPIs) of the new use cases emerging in 5G. The 5GS, powered by network virtualisation and network slicing, gives mobile network operators unique opportunities to offer new services to consumers, enterprises, verticals, and third-party tenants and to address such heterogeneous KPIs. On this basis, previous 5GPPP Phase I collaborative research projects like 5G-NORMA and METIS-II as well as standardisation bodies have identified the main elements and characteristics of the 5G architecture.

Although all these aforementioned efforts have provided a solid baseline architecture, in our view, there has been still room for 5GS enhancements to better fulfil the 5G vision of supporting diverse service requirements while enabling new business sectors often referred to as vertical industries. In the previous deliverables D2.1 [5GM-D2.1] and D2.2 [5GM-D2.2], we have provided a gap analysis with respect to ongoing 5GS architecture design efforts in the industry and academia. The objectives and design principles of 5G-MoNArch architecture address these gaps. This deliverable extends the "Initial Overall Architecture" of 5G-MoNArch from deliverable D2.2 towards the "Final Overall Architecture". The rest of the deliverable is organised as follows.

Chapter 2 presents in detail the *final overall architecture design*, its components, and the impact on the specifications of the target Standards Developing Organisations (SDOs). 5G-MoNArch has applied a structured approach in building the final overall architecture and dealing with the technological gaps identified in the previous deliverables. Specifically, 5G-MoNArch contributions are based on innovation elements, each one composed of one or more enablers[1] and grouped into three fundamental enabling innovations (see Figure 1-1): *telco-cloud enabled protocol stack, inter-slice control and management, and experiment-driven optimisation* as well as two functional innovations: *resilience & security* [5GM-D3.2] and *resource elasticity* [5GM-D4.2]. By applying a functional decomposition to the 5G-MoNArch enablers, we analyse their impact on the overall architecture in terms of functional extensions and interfaces as well as on protocol implications. As elaborated in D2.2 [5GM-D2.2], these functional extensions are built upon the baseline architecture, which is described in D2.1 [5GM-D2.1]. As also emphasised in Figure 1-1, the final overall architecture is built not only upon enabling innovations provided in this deliverable but also upon the functional innovations provided in D3.2 (resilience & security) [5GM-D3.2] and D4.2 (resource elasticity) [5GM-D4.2] to attain a comprehensive design.

In Chapter 3, this deliverable describes the aforementioned *5G-MoNArch enabling innovations* and provides final results to evaluate their benefits in the overall architecture. Although the main focus has been placed on the higher layer protocol and architecture implications, specific physical layer (PHY) requirements and interactions have been shown wherever applicable. Moreover, the aforementioned protocol implications and architectural extensions are illustrated via message sequence charts (MSCs), which present the needed signalling mechanisms between the architectural modules of the innovation elements/enablers for the system realisation. It is worth noting that these signalling mechanisms constitute standard-relevant aspects and are already captured in various approved 5G-MoNArch technical contributions to target SDOs.

[1] An enabler is a technology component that addresses a specific research problem, e.g., inter-slice radio resource management (RRM) for dynamic time division duplex (TDD) in Section 3.2.1, solves the resource allocation problem considering slice requirements and dynamicity of subframe configurations. Similar enablers are grouped under an innovation element, e.g., enablers dealing with resource allocation and slicing are grouped under inter-slice resource management (RM) in Section 3.2. Subsequently, similar innovation elements form a 5G-MoNArch enabling innovation, e.g., inter-slice RM falls under the inter-slice control & management enabling innovation. This mapping is shown in Chapter 3 in Table 3-1.

In Chapter 4, we further elaborate on the concepts for *architectural extensibility and customisation* for the 5G-MoNArch testbed use cases. First, we present the 5G-MoNArch network slice blueprint concept and design, then we describe how the slice blueprint is used to deploy and orchestrate a network slice starting from the Generic Slice Template (GST)[2]. Second, we detail how the flexible 5G-MoNArch architecture can be extended and customised for specific use cases and requirements as those related to the Hamburg Smart Sea Port and the Turin Touristic City testbeds. Finally, this chapter analyses the impact of the 5G-MoNArch architecture and innovations on the evolution of the 5G ecosystem, from a business perspective.

Chapter 5 provides the concluding remarks and highlights how the identified 5GS gaps are addressed by the 5G-MoNArch enablers and architectural extensions.

There are three appendices provided at the end of the deliverable, where Appendix A presents the 5GS gaps briefly, Appendix B provides standard relevance of the 5G-MoNArch innovations and components, and, finally, Appendix C includes further analyses and evaluations on the 5G-MoNArch enablers.

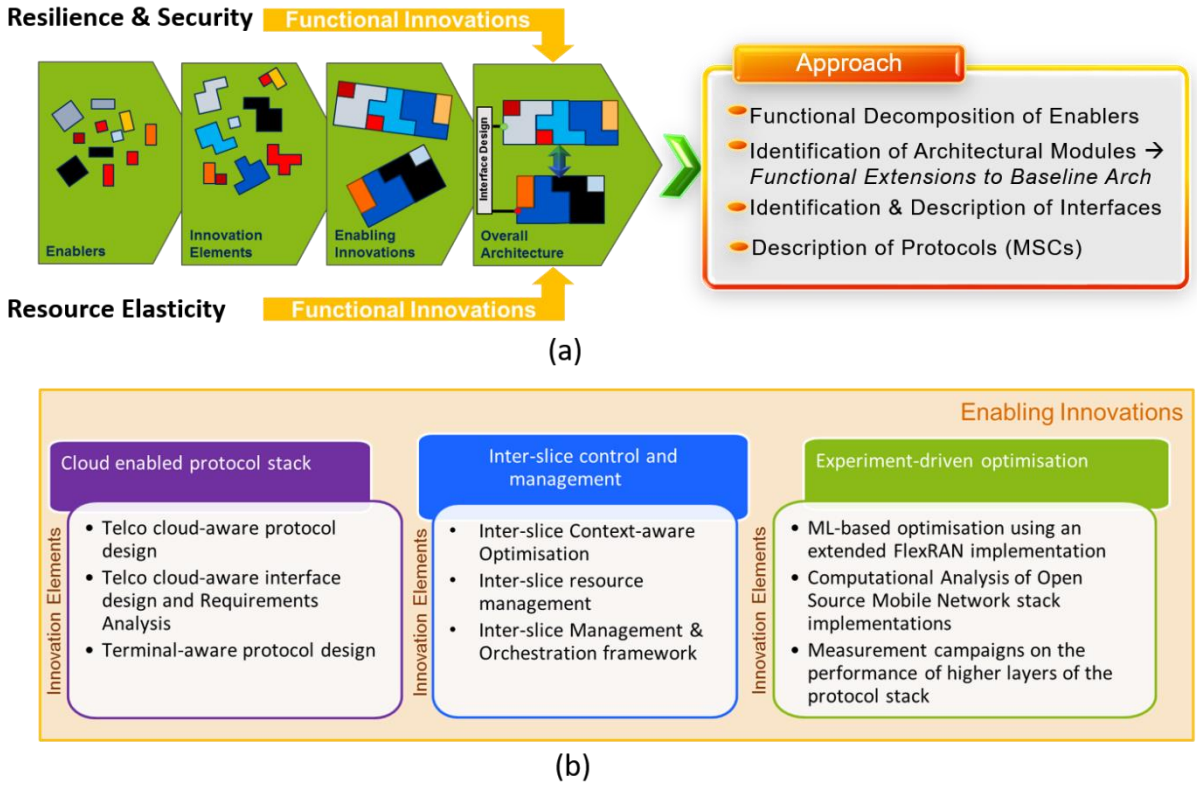


Figure 1-1: (a) The approach of 5G-MoNArch in building the overall architecture based on enabling and functional innovations; (b) High-level 5G-MoNArch enabling innovations

[2] GST has been specified by Global System for Mobile Communications (GSM) association (GSMA). 5G-MoNArch has collaborated with GSMA during the GST specification, where 5G-MoNArch has provided slice realisation vision as well as inputs on 5G-MoNArch testbed use cases and their requirements.

2 5G-MoNArch Final Overall Architecture

This chapter details the 5G-MoNArch final architecture reference model and describes how the following fundamental design objectives are met: (i) E2E slicing support across different technological, network, and administrative domains, (ii) split of control and user plane, and (iii) flexible, per-use-case architecture customisation. For each of the three objectives, a set of design solutions has been developed, which will be shown in the remainder of this chapter. Starting with the overall architecture design which elaborates on the fundamental structuring into network layers and domains, the chapter further depicts where the architecture relies on existing architecture components, e.g., from 3GPP or ETSI NFV. Further, novel NFs for core network (CN) and (R)AN as well as innovative management and orchestration functions introduced by 5G-MoNArch are mapped into the architecture, thus completing the overall picture of the 5G-MoNArch architecture.

2.1 *Overall architecture design – network layers and domains*

The design of the 5G-MoNArch overall functional architecture considers the requirements from the project's use cases and results from 5G-PPP Phase 1 projects (including the White Paper of the 5G-PPP Architecture WG (v2) [5GARCH17-WPv2]), as well as the 5G requirements initially defined in [NGMN15]. Figure 2-1 depicts the fundamental structure of the architecture. It consists of four layers: (1) Service layer, (2) Management & Orchestration (M&O) layer, (3) Controller layer, and (4) Network layer. For each of these layers, there are a set of architectural elements that deliver the system's functionality, including the key functional elements, their responsibilities, the interfaces exposed, and the interactions between them.

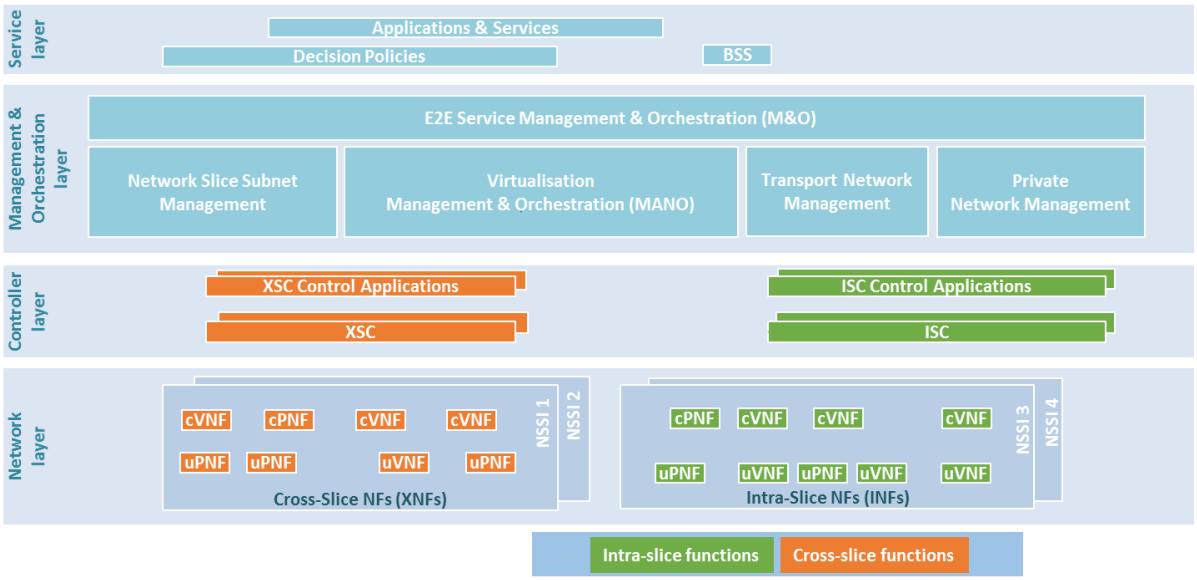


Figure 2-1: 5G-MoNArch high-level structure of the overall functional architecture

The Service layer comprises Business Support Systems (BSS), business-level Policy and Decision functions, and further applications and services operated by a tenant or other external entities.

The M&O layer is divided into an End-to-End (E2E) service M&O sublayer and an additional sublayer containing domain-specific management functions. In [5GM-D2.1] (Section 2.1.2), we have motivated the necessity of an E2E view of a network slice to ensure the satisfaction of service requirements from the customers. The E2E network slice is composed of Network Slice Subnet Instances (NSSIs), typically each from a different network domain, including subnets from radio access, transport, and core network domains, or private (e.g., enterprise) networks.

The Controller layer includes both the cross-slice and the intra-slice controllers (XSC and ISC, respectively). Together with the respective Control Applications, they realise the enhanced network programmability paradigm, which is further elaborated in Section 2.1.2. Typically, each network domain has a dedicated controller that is aware of the technology and implementation characteristics utilised in the domain. Cross-domain coordination would be executed in the M&O layer.

The Network layer hosts Network Slice Subnet Instances (NSSIs). Typically, an NSSI comprises the NFs of a specific network domain, e.g., RAN or CN. An NSSI can be either shared by multiple Network Slice Instances (NSIs) or dedicated to a single NSI. Accordingly, an NSSI consists of cross-slice NFs (XNFs) or intra-slice NFs (INFs), respectively.

Similar to the approach taken by 3GPP for the 5G system architecture representation [3GPP TS23.501], this deliverable also shows the overall architecture in a classical *reference-point representation* as well as a *service-based representation*. While both representations **are equivalent in terms of the offered network functionality**, the service-based representation shows flexible and extensible interactions among NFs, i.e., Service-Based Interfaces (SBI) are provided by service-provisioning functions to service-consuming functions. This is elaborated in more detail in Section 2.1.3.

2.1.1 Key design paradigms of the 5G-MoNArch system architecture

The architecture design of 5G-MoNArch brings several novelties and enhancements compared to prior art network architecture. This section highlights the most crucial aspects. Figure 2-1 through Figure 2-3 implicitly illustrate **three fundamental design aspects** that the 5G-MoNArch architecture design has followed:

(1) **Support for E2E network slicing:** The architecture allows for combining different options of slicing support across M&O and Network layers for each slice instance. The first supported option includes slice-specific functions, i.e., each slice may incorporate dedicated and possibly customised functions that are 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 2-1 depicts this split into common or so-called cross-slice functions and dedicated (intra-slice) functions. This split is maintained in the M&O layer, the Network layer, as well as the Controller layer, i.e., dedicated NFs may be controlled and managed by the tenant's own instance of ISC and M&O layer functions. Shared functions are usually operated by the Mobile Network Operator (MNO) or the Mobile Service Provider (MSP). The MNO (together with potential third-party infrastructure providers) is also in charge of managing the infrastructure. The policies regarding the utilisation of shared functions, particularly the resource allocation to active slices, are determined by the Cross-slice M&O function and communicated towards the respective functions of the Controller and Network layers for further enforcement. Finally, the third option is to not only have slice-dedicated NFs but to additionally assign the associated infrastructure hardware resources (HW), including spectrum, exclusively to a single slice. The slice-specific functions and shared functions in one logical slice are bind together by the network slice identifier at the Network layer. More details on how the Network layer performs network slice selection is described in Sections 2.2 and Appendix B.

(2) **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 TN. Among others, this allows for hosting associated CP and UP NFs in different locations and also facilitates to aggregate CP and UP NFs differently. The split further allows independent scalability and evolution of NFs.

(3) **Flexible architecture customisation:** The customisation of the architecture for a specific network slice instance is triggered by NSMF, which modifies the architecture and functionality used in existing slices. For example, this can include further deployment, management, orchestration, and control instructions for specialised NFs. Hence, these slice-specific modifications and customisations may affect different layers of the architecture.

The following sections further elaborate how these paradigms have been realised.

2.1.2 5G-MoNArch final overall architecture

Figure 2-2 depicts the 5G-MoNArch final overall architecture, including the extensions for security and resilience defined in WP3 and for resource elasticity defined in WP4.

The **Service layer** functions interact with the M&O layer via the Communication Service Management Function (CSMF), see below.

The **Management & Orchestration layer** is composed of the M&O functions from different network, technology, and administration domains, e.g., 3GPP public mobile network management [3GPP TS 32.101], ETSI NF Virtualisation (NFV) Management and Orchestration (MANO) [ETSI NFV13], ETSI Multi-access Edge Computing functions [ETSI MEC16], management functions of TNs and non-public enterprise networks. Further, the M&O layer comprises the E2E M&O sublayer hosting the Network Slice Management Function (NSMF) and CSMF that manage network slices and communications services, respectively, across multiple management domains in a seamless manner. In the *Virtualisation MANO* domain, the ETSI NFV MANO architecture for lifecycle management (LCM) of Virtual Machines (VMs) is extended towards LCM of virtualisation containers (e.g., Docker). Therefore, it comprises, besides the ETSI NFV components, corresponding functions for LCM of containers. Therefore, the Virtualised NF Manager (VNFM) has according components for virtual machine infrastructure (VMI) and container infrastructure (CI). Similarly, the Virtualised Infrastructure Manager (VIM) contains a VMI Management Function (VMIMF) and a CI Management Function (CIMF). NFV Orchestrator (NFVO) provides the network service orchestration capabilities as well as the dispatching functionality for selecting between VM-related or container-related MANO functions. Further, the layer accommodates network slice subnet management domain, e.g., network slice subnet management function (NSMF) or NF Management Function (NFMF) for 3GPP network management. Such functions also implement ETSI NFV MANO reference points to the VNFM and the NFVO. The CSMF transforms consumer-facing service descriptions into resource-facing service descriptions (and vice versa) and therefore works as an intermediary function between the Service layer and the NSMF. The NSMF splits service requirements as received from CSMF and coordinates (negotiates) with multiple management domains for E2E network slice deployment and operation. As a major 5G-MoNArch novelty, NSMF further incorporates a Cross-slice M&O function for inter-slice management (e.g., common context between different slices/tenants, inter-slice resource brokering for cross-slice resource allocation, particularly in the case of shared NFs, etc.). In contrast, the Cross-domain M&O function works on strictly intra-slice level, but across multiple network and technology domains. The M&O layer performs the management tasks on NSIs, which are uniquely identified by an NSI identifier. An NSI may be further associated with one or more Network Slice Subnet Instances (NSSI). The details are further described in Sections 2.2.3 and 4.1.

The **Controller layer** realises the software-defined networking concepts [ONF14], extends them to mobile networks, and therefore accommodates two controller types and their respective control applications:

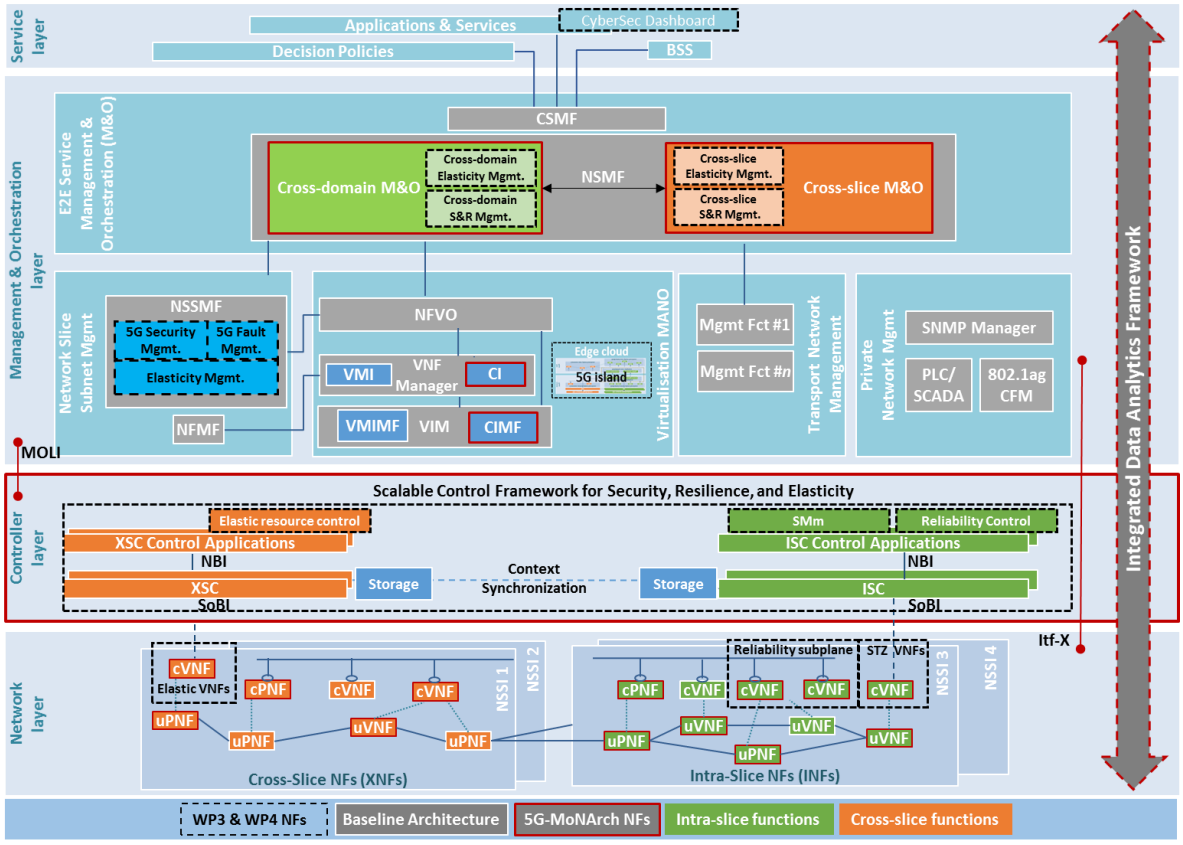
(1) The Cross-Slice Controller (XSC), e.g., a RAN controller for the control of Cross-slice NFs (XNFs) shared by multiple network slices. One or multiple applications on top of the controller host the control logic that shall be applied to the controlled NFs.

(2) The Intra-Slice Controller (ISC), e.g., a CN controller for Intra-Slice NF (INFs) within a dedicated CN-NSSI. One or more slice-specific applications per slice (but no cross-slice applications) use the northbound interface to communicate with the controllers in order to convey the control logic residing in each of the independent applications. In principal, these applications could even be sourced from different vendors.

These controllers expose a northbound interface (NBI) towards control applications and a southbound interface (SoBI) towards virtualised network functions (VNFs) and physical network functions (PNFs) in the Network layer. Interfaces towards the M&O layer are provided via the so-called Management & Orchestration Layer Interface (*MOLI)* reference point, see Section 2.1.3 for more details.

The Controller layer facilitates the concept of **mobile network programmability.** Generally, software-defined networking (SDN) splits between *logic* and *agent* for any functionality in the network. This means that the NFs are split into the decision logic hosted in a control application and the actual NF in the Network layer (usually a uPNF or uVNF) that executes the decision. In other words, for the given uVNF or uPNF, the corresponding cVNF or cPNF disappears since the functionality is provided by the Controller layer functions. The controller resides "between" application and NF and abstract from specific technologies and implementations realised by the NF, thus decoupling the control application from the controlled NF, cf. Figure 2-1. 5G-MoNArch has investigated the applicability of this paradigm, focusing on the concepts for elasticity (WP4) and resiliency (WP3). If no such split between control

logic and agent is applied, i.e., conventional control plane functions (cPNF or cVNF) are utilised, the Controller layer disappears. In this sense, it is an optional layer of the 5G-MoNArch architecture.



BBS	Business Support System	NSMF	Network Slice Management Function
CFM	Connectivity Fault Management	NSSMF	Network Slice Subnet Management Function
CI	Container Infrastructure	PLC	Programmable Logic Controller
CIMF	CI Management Function	SNMP	Simple Network Management Protocol
cPNF	Control plane Physical Network Function	SCADA	Supervisory Control and Data Acquisition
CSMF	Communication Service Management Function	SoBI	South-Bound Interface
E2E	End-to-End	STZ(m)	Security Trust Zone (management)
INF	Intra-slice Network Function	S&R	Security and Reliability
ISC	Intra-Slice Controller	uVNF	User plane Virtualised Network Function
Itf-X	Interface X	VIM	Virtualised Infrastructure Manager
Mgmt	Management	VMI	Virtual Machine Infrastructure
M&O	Management & Orchestration	VMIMF	VMI Management Function
NBI	North-Bound Interface	XNF	Cross-slice Network Functions
NFMF	Network Function Management Function	XSC	Cross-Slice Controller
NFVO	Network Function Virtualisation Orchestrator		

Figure 2-2: 5G-MoNArch final overall architecture

The **Network layer** comprises the VNFs and PNFs of both control plane (i.e., cVNF and cPNF) and user plane (i.e., uVNF and uPNF). NFs can include, for example, 3GPP CP functions (AMF, SMF, AUSF, RRC, etc.) and user plane (UP) functions (e.g., UPF, PDCP, etc.) or novel NFs developed in the project, e.g. for resource elasticity, resilience, and security. Generally, the 5G-MoNArch Network layer can comprise different CP/UP architectures. For example, also a 4G mobile network with EUTRAN and EPC functions could constitute an instance of the Network layer, nonetheless exhibiting the associated limitations in supporting 5G-MoNArch features (e.g., no support for network slicing in 4G). In the reference point representation, interfaces towards the M&O layer are provided via the *Itf-X* reference point. It is an evolution of the 3GPP *Itf-S* interface for facilitating fault, configuration, accounting, performance, and security (FCAPS) management as well as domain-agnostic LCM procedures. For associating a UE to the correct NSI, the Network layer uses the Single Network Slice Selection Assistance Information (S-NSSAI), which is provided by the UE. Moreover, the CN part of the CP in

the network layer is realised as a service-based architecture (SBA) [3GPP TS 23.501]. Further details of CN functionality, slice identification, and SBA are explained in Sections 2.2.2 and Appendix B; details on the 5G-MoNArch RAN architecture are shown in Sections 2.2.1 and Appendix B.

Moreover, Figure 2-2 also shows NFs that specifically implement the solutions developed for reliability and security (WP3) and for resource elasticity (WP4). They can be found in all layers. On the Service layer, Security Dashboard enables efficient security management by human operators or via APIs as well as intuitive security status monitoring.

On the M&O layer, Cross-slice M&O may incorporate WP3 **Cross-slice Security and Resilience (S&R) Management** functions for addressing jointly increased security and resilience requirements. Furthermore, to enable flexible orchestration across elastic slices, Cross-slice M&O can integrate WP4 Slice-aware and Orchestration-driven **Cross-slice Elasticity Management** functions. Cross-domain M&O can include WP3 **Cross-domain S&R Management** that manages security and resilience issues within a single slice, but between different management domains. In addition, Cross-domain M&O can integrate WP4 Computational and Orchestration-driven **Cross-domain Elasticity Management** functions if intra-slice flexible resource management is required. The M&O layer also hosts the Big Data Analytics module defined in WP4 (as part of the **Integrated Data Analytics Framework**), which can support the elasticity functions located in the Cross-slice M&O and Cross-domain M&O. Further, the WP3 **5G Fault Management (FM)** and **5G Security Management** and the WP4 **Elasticity Management** functions address network alarms, security procedures, and elasticity performance, respectively, on the granularity level of NSSIs. Edge Cloud infrastructure nodes can host so-called "5G Islands", i.e., mobile networks with minimal functionality that can operate autonomously and do not require connectivity to central data centres.

The Network layer can be enhanced with WP3 **Reliability sub-plane** functions for multi-connectivity (data duplications) and network coding for improved RAN resilience. For further customisation, the Network layer may host WP3 **Security Trust Zone (STZ)** VNFs, which enable to guarantee a required level of security and trust in a dedicated area of the infrastructure, and VNFs with enhanced robustness and elasticity characteristics in case of unforeseen fluctuations of resource availability.

The Controller layer can host the WP3 **Scalable Control Framework for Security, Resilience, and Elasticity**, which automatically scales the controller nodes with respect to the underlying traffic in the network in order to enhance network scalability but to ensure high availability of controllers. For slicing elasticity, dedicated **Intra-slice Controller (ISC) and Cross-slice Controller (XSC)** have been developed in WP4 (see Section 4.3.2). These controllers can provide an 'inner-loop' control of NFs, to enforce elasticity at fast time scale. Exemplary controller applications include Reliability Control and Security Monitoring manager (SMm) from WP3 or Elastic resource control from WP4. Further details on specific WP3 and WP4 functionality and their impact on the customisation of the architecture can be found in Section 4.3.

Finally, in 5G-MoNArch an **Integrated Data Analytics Framework** has been developed to allow the exchange of data and analytical services across all layers of the 5G-MoNArch architecture. The details can be found in Section 2.2.4.

2.1.3 Reference-point and service-based system architecture representations

Figure 2-2 depicts the reference point representation of the 5G-MoNArch overall system architecture. With the exception of the service-based interfaces between control plane functions in the Network layer, it shows dedicated reference points between NFs and between NFs and functions from the M&O layer. Besides dedicated intra-layer reference points between NFs, the reference point architecture representation incorporates the inter-layer *Itf-X*, *MOLI*, and *SoBI* reference points.

The **5G-MoNArch *Itf-X* reference point** subsumes all interfaces between functions on the M&O layer and functions on the Network layer. Each interface depends on (1) the management domain in charge of managing the NF and (2) the implementation details of the NF, e.g., whether it is a PNF or VNF or which features of the NF are exposed to the M&O system. For example, *Itf-X* subsumes *Itf-S* interface as defined for pre-Rel. 15 3GPP network management systems [3GPP TS 32.101] or *Ve-vnfm-vnf* reference point between the VNF Manager and the VNF as specified in [ETSI NFV SOL002]. Another example comprises NETCONF- or SNMP-based interfaces for managing transport NFs, such as Time-

Sensitive Networking (TSN) bridges. In a fully service-based architecture representation, a dedicated *Itf-X* reference point becomes obsolete.

The interaction between M&O and Controller layer is supported by **Management and Orchestration Layer Interface (MOLI)**. In general, MOLI supports two functionalities 1) Configuration of various parameters in NFs (e.g. MAC & PHY layer configuration, Computation resource allocation) decides by the Orchestrator for the slice under deployment, 2) Re-orchestration request to M&O layer triggered by the controller upon monitoring of VNFs (e.g. due to scarcity of computational resources, security vulnerability detection in the platform, failure or malfunction of deployed VNFs) [5GM-D4.2][5GM-D3.2]. For the interaction with the NFs of the Network layer, i.e., for executing control actions, the controllers implement the **Southbound Interface (SoBI)**.

Figure 2-3 depicts the service-based representation of 5G-MoNArch overall architecture. The service-based interaction between NFs provides a set of features and associated advantages. Among others, NFs can be realised in a stateless manner since such state-related data (e.g., session data) are shared via an SBA bus, sometimes referred to as data bus. Moreover, SBA brings along several characteristics such as agility, flexibility in deployment, testability, scalability, performance, and simplicity [5GPPPW2C]. SBA also facilitates the design of modularised NFs, uniform interaction procedures between NFs (e.g., NFs can offer their functionality as a service to other NFs), unified authentication framework between NFs, and concurrent access to services. Besides the Network layer, the SBA approach is also adopted in the M&O layer.

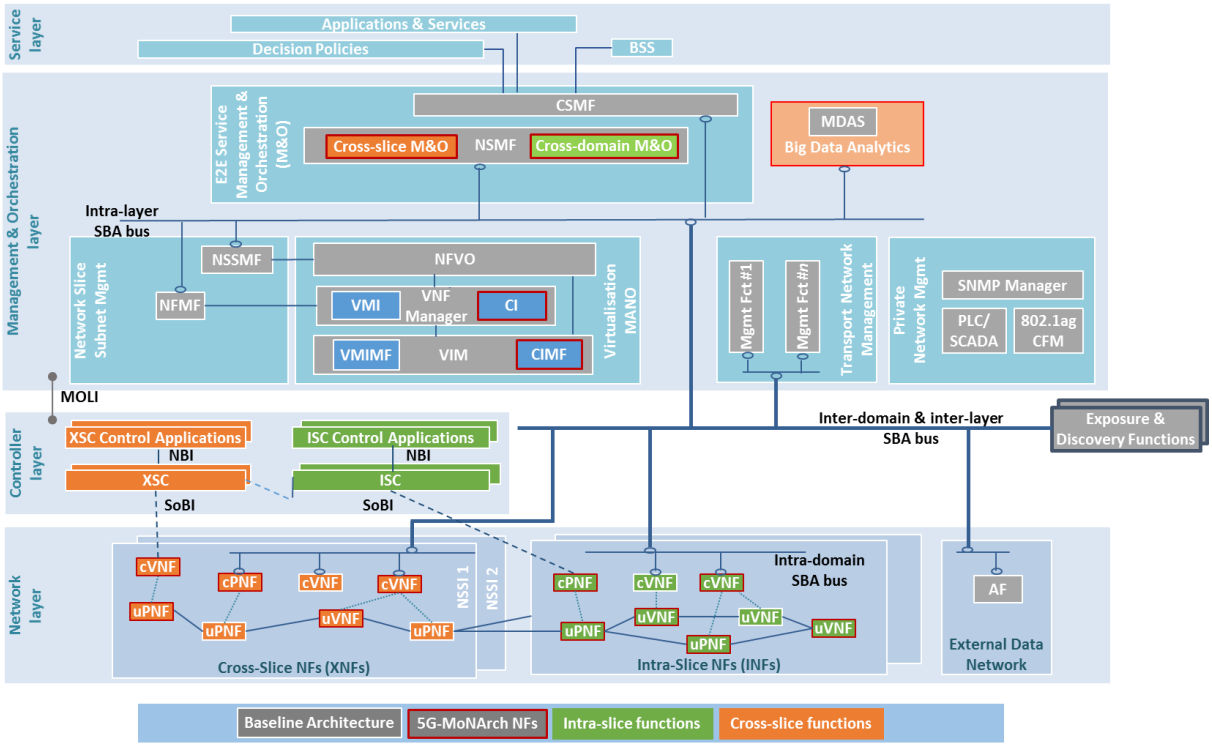


Figure 2-3: Service-based representation of 5G-MoNArch overall architecture

In order to achieve service-based interaction across layers, across management domains, and across network domains (RAN, CN, etc.), the architecture incorporates an *inter-domain & inter-layer SBA bus* which interconnects the individual intra-layer and intra-domain SBA buses. For example, the Network layer contains several *intra-domain SBA buses*, including the one for the external Data Network (DN) domain. The *intra-layer SBA bus* on the M&O layer connects services of different management domains and exposes (a subset of) them to the inter-domain an inter-layer SBA bus. In order to fully support such service-based operation, each SBA bus incorporates *Exposure & Discovery Functions*, which include but are not limited to:

- Service repository: a database that catalogues consumable services and includes fundamental information about the service, e.g., provisioning layer/domain or where and how to consume it;
- Service (de)registration: functionality providing the capability to service providers to announce their service(s) in order to add it to the service repository. Similarly, service providers can deregister services;
- Service discovery: functionality to match service requests from consumers with available services of the repository;
- Enablers for efficient service consumption:
    - Routing, forwarding, and load balancing of service requests and data streams related to actual service provisioning;
    - Functionality for service consumers to subscribe to available services;
    - Support of event-based or periodic service provisioning;
    - Data persistence, processing and delivery enablers (e.g., means for data storage, data streaming, or batch data processing);
- Service exposure governance and access control: functionality to configure and enforce rules for authenticating and authorising service consumers, including restrictions for service availability across layers and domains.

One of the examples where 5G-MoNArch further elaborates on the advantages of the SBA representation comprises the **Integrated Data Analytics Framework**. The framework enables the customised composition of data analytics capabilities from the M&O layer, from the Network layer (both RAN and CN network domains), from 3rd party domains (e.g., Application Function (AF), DN), or from the UE. For instance, the Big Data Analytics module in the M&O layer (cf. Figure 2-3) can host Management Data analytics Services (MDAS) as defined in [3GPP TS 28.533], network domain-specific services., e.g., RAN performance management data analytics, NWDAF, NFVI data analytics services, etc.) for both cross-domain and cross-slice optimisations. Details of the Integrated Data Analytics Framework using SBI are described in Section 2.2.4.

2.1.4 Key technology areas in SDOs impacted by 5G-MoNArch architecture

Overall, 5G-MoNArch novelties in network architecture design have also impacted the study-items and work-items in several standardisation working groups, in particular in 3GPP and ETSI. Table 2-1 lists contributions to standards that are based on the project's results. They are grouped into five Key Technology Areas (KTAs):

- KTA 1: Service-based Architecture Enhancements
- KTA 2: Integrated Data Analytics Framework
- KTA 3: RAN support of Network Slicing & RAN Enhancements
- KTA 4: Network Slicing Enhancements
- KTA 5: 5G M&O Enhancements

Further details on the relationship with standards and standardisation roadmaps can be found in Appendix B.

Table 2-1: 5G-MoNArch contributions to the standards

Standard Body	Title	Reference	5G-MoNArch KTA Mapping
3GPP SA2	New Key Issue: Network Slicing for eV2X	S2-180147	KTA 1, KTA 4
3GPP SA2	23.726: FS_ETSUN (Enhancing Topology of SMF and UPF) / 23.726 Scope	S2-181046	KTA 1
3GPP SA2	Clarification on Key Issue: Network Slicing for eV2X	S2-183735	KTA 4

3GPP SA2	NWDA-assisting E2E QoS Assurance	S2-183634	KTA 1, KTA2
3GPP SA2	Solution for AF Data Exposure to/from NWDAF	S2-183637	KTA 1, KTA2
3GPP SA2	Use Case on UE-driven analytics sharing	S2-185816 (S2-185290)	KTA 1, KTA2
3GPP SA2	Update to the general framework for 5G network automation (TR 23.791)	S2-186271	KTA 1, KTA2
3GPP SA2	TR 23.742: Solution for NF reliability	S2-186151	KTA 1
3GPP SA2	Updated SID: Study on Enhancement of Network Slicing	S2-186185	KTA 4
3GPP SA2	New SID on Enhanced support of Vertical and LAN Services	S2-186182	KTA 1
NGMN NWMO	Cross-slice user stories	NA	KTA 5
3GPP RAN2	Support for SRB duplication with CA	R2-1803233	KTA 3
3GPP SA5	Add Data Analytics Management Service for Network Slice and Network Slice Subnet	S5-183560	KTA 2, KTA 4, KTA 5
3GPP SA5	Add example of functional management architecture	S5-183409	KTA 5
ETSI ZSM	Proposal on the overview and architecture of ZSM framework	ZSM(18)000236r2	KTA 1, KTA 5
3GPP SA2	Key Issue: UE-driven analytics sharing mechanisms to 5GC	S2-187264 (S2-186919)	KTA 1, KTA 2
3GPP SA2	LS from FS-eNA to SA5/RAN3	S2-186667	KTA 2, KTA 3
3GPP SA2	UC and KI for KI4 Interactions with OAM for Analytics Exposure	S2-186668	KTA1, KTA 2, KTA 5
3GPP SA2	New Solution to Key Issue #3: Data Collection by subscription to NFs/AFs	S2-186346	KTA 1, KTA 2
ETSI ZSM	Proposed ZSM Architecture Diagram Changes	ZSM(18)000325r2	KTA 1, KTA 5
3GPP SA2	Solution: UE-driven analytics sharing	S2-188512 (S2-187903)	KTA 1, KTA2
3GPP SA2	Discussion paper on V2X slicing KI	S2-188307	KTA 4
3GPP SA2	Solution for Data Collection from OAM using Existing SA5 Services	S2-188263	KTA 1, KTA2, KTA 5
ETSI ZSM	Automated discovery of services offered by a management domain	ZSM(18)000364r2	KTA 5
ETSI ZSM	Definition of integration fabric	ZSM(18)000378r1	KTA 5
3GPP SA2	Updates to Impacts and Evaluation of Solution 12	S2-1810696	KTA2
3GPP SA2	Updates to Solution 1 for Network Data Analytics Feedback	S2-1860695	KTA 2
3GPP SA2	Solution for KI#2 on Analytics Exposure to AF	S2-1810694	KTA 2

3GPP SA2	Solution to NWDAF assisting traffic routing using MEC information	S2-1810334	KTA 2
ETSI ZSM	Management service related to network service orchestration	ZSM(18)000445	KTA 1, KTA 5
ETSI ZSM	Management service related to service performance assurance	ZSM(18)000446	KTA 1, KTA 5
ETSI ZSM	Add domain performance report service	ZSM(18)000450	KTA 1, KTA 5
ETSI ZSM	Architecture Diagram Changes	ZSM(18)000501	KTA 1, KTA 5
ETSI ZSM	Clarify capability of domain orchestration and some clarifications	ZSM(18)000442	KTA 1, KTA 5
3GPP SA5	YANG definitions for network slicing NRM	S5-185532	KTA 4, KTA 5
3GPP SA2	Integration of the 5G System in the TSN network	S2-188459	KTA 1
3GPP SA2	Update to SID: Study of enablers for network automation for 5G	S2-189047	KTA 1
3GPP SA5	Update the UC and requirements for performance data streaming	S5-186429	KTA 5
3GPP SA2	Updates to Solution 19	S2-1812173	KTA 2
3GPP SA2	Overall Conclusion for Key Issue 4	S2-1812175	KTA 2
3GPP SA2	Updates to Solution 12	S2-1812174	KTA 2
3GPP SA2	Updates to Solution 24	S2-1812172	KTA 2
ETSI ISG ENI	Use case on "Elastic resource management and orchestration"	ENI(18)000162r1	KTA 3, KTA 5
ETSI ISG ENI	Proof of concept on "Elastic network slice management"	ENI(18)000175r4	KTA 3, KTA 5
3GPP RAN3	Slice support of IAB nodes	R3-186014	KTA 3, KTA 4
3GPP SA5	Update NRM IRP Solution Set to support slice priority	S5-187439	KTA 4, KTA 5
3GPP SA5	Update NRM root IOCs to support slice priority	S5-187370	KTA 4, KTA 5
3GPP SA5	Solution for performance data streaming	S5-187372	KTA 4, KTA 5
ETSI ZSM	Add capabilities to Analytics Service	ZSM(18)000596r2	KTA 2, KTA 5
ETSI ZSM	Add E2E SLA Management	ZSM(18)000601r2	KTA 4, KTA 5
3GPP SA2	Adding reference to new TS 23.288 in TS 23.502	S2-1901040	KTA 4, KTA 5
3GPP SA2	TS 23.288 skeleton for 5G analytics framework	S2-1901041	KTA 2
3GPP SA2	CR for TS 23.501 based on conclusion of eNA TR 23.791	S2-1901042	KTA 2

3GPP SA2	Adding Selected Solutions #12 from eNA to TS 23.288	S2-1900949	KTA 2
3GPP SA2	Adding Selected Solutions #24 from eNA to TS 23.288	S2-1901024	KTA 2
ETSI ZSM	Policy management service for E2E	ZSM(19)00021	KTA 5
3GPP SA2	TS 23.288: Update to Data Collection from OAM	S2-1902400	KTA 2, KTA 5
3GPP SA2	TS 23.501 CR0987: CR for TS 23.501 Clarifications NWDAF Discovery and Selection	S2-1902521	KTA 1, KTA 2
3GPP SA2	TS 23.288: Remove the FFS for AF registration during Data Collection procedure	S2-1902398	KTA 1, KTA 2
3GPP SA2	TS 23.288: Analytics exposure to AF via NEF	S2-1902395	KTA 1, KTA 2
3GPP SA2	TS 23.502 CR1060: NEF service for NWDAF analytics	S2-1902524	KTA 1, KTA 2
3GPP SA2	TS 23.501 CR0964: NEF service for NWDAF analytics	S2-1902397	KTA 1, KTA 2
ETSI ZSM	Update of the analytics service	ZSM(19)000121	KTA 2, KTA 5
ETSI ZSM	Discussion on mapping the ZSM002 list of services	ZSM(19)000122	KTA 1, KTA 5
ETSI ZSM	Update mapping ZSM002 to SA5	ZSM(19)000192	KTA 5
3GPP SA2	23.501 CR1258: Clarifications NWDAF Discovery and Selection	S2-1903964	KTA 1
3GPP SA2	23.502 CR1298: Extensions to NRF Services	S2-1903965	KTA 1, KTA 2
3GPP SA2	P-CR TS 23.288: Setup of Network Map for Data Collection	S2-1903814	KTA 2
3GPP SA2	P-CR TS 23.288: Clarification of FFS on Analytics Exposure to AFs via NEF	S2-1903966	KTA 1, KTA 2
3GPP SA2	P-CR TS 23.288: Clarifying Flexible AF Registration	S2-1904011	KTA 1, KTA 2
3GPP SA2	TS 23.501 CR1299: Extending Exposure Capability to support Analytics Framework	S2-1903968	KTA 1, KTA 2
3GPP SA2	TS 23.502 CR1300: Updating NEF and NRF Services to Support AF Available Data Registration	S2-1903999	KTA 1, KTA 2
3GPP SA2	Update to NF Load Analytics Procedures	S2-1903917	KTA 2, KTA 5
3GPP SA2	Update to Network Performance Analytics Procedures	S2-1903939	KTA 2, KTA 5
3GPP SA5	pCR 28.861 Add Multi-dimensional Resource Optimisation	S5-193221	KTA 4, KTA 5
3GPP SA5	Update NRM requirement to support SBA management	S5-193396	KTA 1, KTA 5
ETSI ZSM	ZSM002 update of service feasibility check	ZSM(19)000195r2	KTA 4, KTA 5

ETSI ZSM	ZSM002 Management communication service to solve pub-sub debate	ZSM(19)000032r3	KTA 5
ETSI ZSM	Informative examples on ZSM deployment architectures	ZSM(19)000203r2	KTA 1, KTA 5

2.2 *Novel components and interfaces of the 5G-MoNArch architecture*

This section introduces the novel NFs and interfaces that 5G-MoNArch has introduced beyond state-of-the-art mobile network architectures. In Table 2-2, a brief overview of the essential 5G-MoNArch novel components is provided along with their key features within the 5G-MoNArch architecture. Also, a mapping to 5G-MoNArch D2.3 sections and other deliverables is tabulated, where the details on these components can be found in the respective sections.

Table 2-2: Overview of 5G-MoNArch novel components

Architecture Domain / Layer	**Name of the 5G-MoNArch Novel Components and Key Features**	**Mapping onto 5G-MoNArch WPs and Deliverables**
RAN	*RCA*: Main interface to the controller layer, RAN exposure function	WP2 D2.3 Section 2.2.1
	RAN-DAF: Data analytics function in the RAN CU	WP2 D2.3 Section 2.2.4
	MM: Slice-aware mobility function in the RAN RRC	WP2 D2.3 Section 2.2.1 & Section 3.3.3
	Dynamic RAN Control Unit: Control unit handling dynamic small cell (DSC) operation as well as slice-aware functional operations	WP2 D2.3 Section 2.2.1 & Section 3.3.2
	Inter-slice RRM: RRM among network slices to ensure slice-requirements including slice isolation in the RAN CU	WP2 D2.3 Section 2.2.1 & Section 3.3
	IM: Slice-aware interference management in the RAN CU	WP2 D2.3 Section 2.2.1 & Section 3.3.1
	Packet Duplication: PDCP-level packet duplication to increase reliability	WP2 D2.3 Section 2.2.1 & WP3 D3.2 [5GM-D3.2]
	Unified Scheduler: MAC-level resource allocation factoring in slice requirements as well as reliability- and elasticity- tailored improvements	WP2 D2.3 Section 2.2.1 & WP3 D3.2 [5GM-D3.2] & WP4 D4.2 [5GM-D4.2]
	Group Coordination & Group Communication as part of *D2D group mobility*: UE-centric mobility enhancements in case of group mobility	WP2 D2.3 Section 2.2.1 & Section 3.1.3
Controller Layer	*Reliability Control and SMm*: Applications tailored for the functional innovation *resilience and security*	WP2 D2.3 Section 2.2.1 & WP3 D3.2 [5GM-D3.2]
	Elasticity Control: Applications tailored for the functional innovation *resource elasticity*	WP2 D2.3 Section 2.2.1 & WP4 D4.2 [5GM-D4.2]
	Slow inter-slice RRM: Long-term resource allocation optimisation and support for real-time RRM	WP2 D2.3 Section 2.2.1
	Slice-aware RAT selection: Support of MM in the RAN via slice-awareness	WP2 D2.3 Section 2.2.1 & Section 3.3.3

CN	*ISCF* with enhancements on *PCF, NWDAF, NSSF:* Enabling multi-slice coordination to realise new use cases, such as multi-slice UE services	WP2 D2.3 Section 2.2.2 & Section 3.2.2
	NWDAF enhancements: Data analytics function in CN with novel enhancements on UE-centric data analytics and inter-slice context sharing	WP2 D2.3 Section 2.2.2 & Section 3.2.1 & Section 3.2.3
	Data analytics framework: *NWDAF* and AF/OAM interaction regarding data collection and CN analytics exposure, which enables per slice cross domain optimisation.	WP2 D2.3 Section 2.2.2 & Section 2.2.4 & Section 3.2.1
M&O	*CSMF* comprising *Requirements Translation, Requirements Update, Service Allocation*, *Service Activation*, and *Service Analytics*: Service allocation and management, and translation of service requirements into network requirements along with view on the service status and performance	WP2 D2.3 Section 2.2.3
	NSMF Cross-domain M&O comprising *Slice Blueprint, NSSI decomposition, Cross-domain S&R Management*, and *Cross-domain Elasticity Management*: Management of a single network slice across the different management domains along with extensions by functional innovations *resilience and security*, and *resource elasticity*	WP2 D2.3 Section 2.2.3 & Section 4.1 & WP3 D3.2 [5GM-D3.2] & WP4 D4.2 [5GM-D4.2]
	NSMF Cross-slice M&O comprising *Cross-slice Requirements Verification, Cross-subnet Requirements Verification, Cross-slice Elasticity Management*, and *Cross-slice S&R Management*: Management of the interaction and resource sharing among the deployed network slices along with extensions by functional innovations *resilience and security*, and *resource elasticity*	WP2 D2.3 Section 2.2.3 & WP3 D3.2 [5GM-D3.2] & WP4 D4.2 [5GM-D4.2]
	Big Data Module comprising *MDAS*: Data Analytics function in the M&O layer	WP2 D2.3 Section 2.2.3 & Section 2.24 & Section 3.3.5 & WP4 D4.2 [5GM-D4.2]
	NSSMF comprising *NSD Creation, 5G Security Management, 5G Fault Management*, and *Elasticity Management*: Management of the network slice subnets that are constituents of a network slice along with extensions by functional innovations *resilience and security*, and *resource elasticity*	WP2 D2.3 Section 2.2.3 & Section 3.3.4, WP3 D3.2 [5GM-D3.2] & WP4 D4.2 [5GM-D4.2]
	Virtualisation MANO: ETSI NFV MANO architecture with extensions towards LCM of virtualisation containers (e.g., Docker).	WP2 D2.3 Section 2.1.2 & WP3 D3.2 [5GM-D3.2] & WP4 D4.2 [5GM-D4.2]

2.2.1 Radio access network components

The 5G-MoNArch RAN architecture takes the baseline architecture presented in [5GM-D2.1], where the baseline architecture covers 5GPPP Phase 1 consensus and the 3GPP status from the publication time, i.e., the latest 3GPP Release specification on NG-RAN [3GPP TS 38.300] [3GPP TS 38.401], e.g., addition of Service Data Adaptation Protocol (SDAP) layer and F1 interface with CU-DU split. On this basis, the 5G-MoNArch RAN architecture enhances the baseline architecture by functional models emerging from the 5G-MoNArch innovations as outlined in Chapter 3 and Chapter 4.

The 5G-MoNArch extensions not only include the new functional enhancements on the CU and DU but also the F1 interface implications (see Chapter 3) as well as the novel Controller Layer described herein for RAN. It is worth noting that, in 5G-MoNArch, the Controller Layer is envisioned only for RAN, which provides means to introduce RAN control functions as specific application implementations. It is worth noting that such flexibility is already available for the CN thanks to the application functions (AFs) as part of the SBA (see Section 2.2.2 and Section 2.2.3). A high-level illustration of the 5G-MoNArch RAN architecture is given in Figure 2-4. Therein, the Controller layer is identified by XSC and ISC along with the corresponding applications (APPs) running on the NBI. The control commands and interactions with the gNBs take place via the SoBI.

Building upon the high-level RAN architecture in Figure 2-4, a detailed illustration of the 5G-MoNArch RAN protocol architecture is given in Figure 2-5. The protocol architecture includes both the CP and UP functions at the Controller layer, CU, DUs and UEs. The extensions introduced by 5G-MoNArch innovations are highlighted and the associated descriptions are provided in the following paragraphs. The interface implications are captured by Message Sequence Charts (MSCs) which are provided in Chapter 4 in accordance with the 5G-MoNArch innovations.

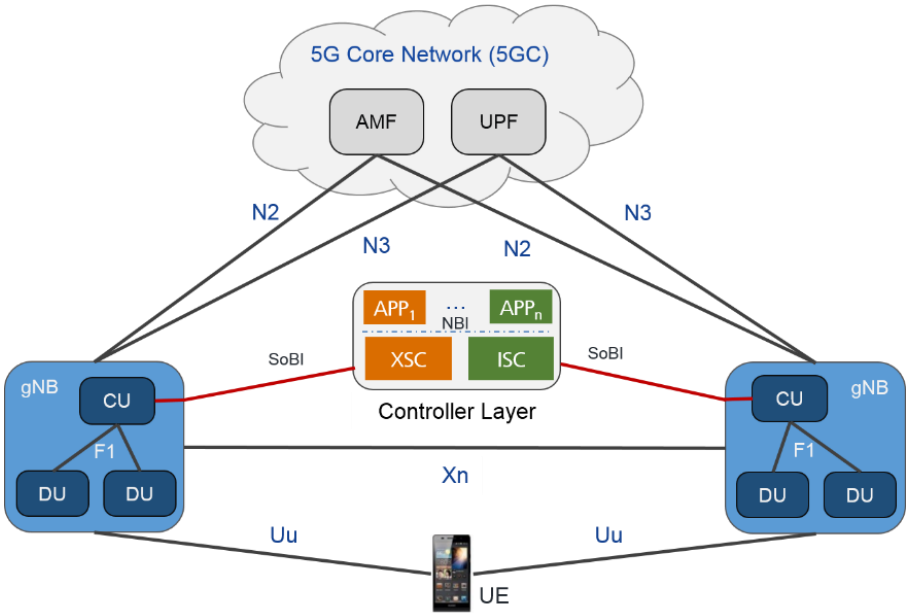


Figure 2-4: High-level 5G-MoNArch RAN architecture

As shown in Figure 2-5, the **RAN Controller Agent (RCA),** one of the novel components defined in 5G-MonArch is introduced in the CU to interface distributed and centralised VNFs to the logically centralised controller. In general, **RCA** acts a middle ware between controller and network layer with a local data-store capable to store most recent monitoring information from the network layer. In this regard, RCA can be considered as one of the exposure and discovery functions, as shown in Figure 2-3. The amount of the data that can be exposed to the controller layer can thus be controlled by the RCA. The SoBI is the unified interface between **RCA** and controller layer for monitoring and re-configuration of VNFs. Each VNFs in DU and CU follows application-based approach to interact with RCA for control applications deployed in the controller. The **RCA** is interfacing the novel RAN data analytics function (RAN-DAF), which is responsible for collecting monitoring information related to both UEs and RAN, such as Channel Quality Indicator (CQI), Power Level, Path Loss, Radio Link Quality, Radio Resource Usage, Modulation and Coding Scheme (MCS), Radio Link Control (RLC) buffer state information, etc. The information obtained from RAN-DAF can be sent by RCA to controllers in the form of NBI applications (Slow Inter-slice RRM, Slice Aware RAT Selection, Elastic Resource Control, etc., as detailed in the following sections) for further optimisation. Together with RAN-DAF, **RCA** is also responsible for routing re-configuration information from controller to the respecting VNFs in the CU and DU of RAN. Further details on the RAN-DAF are provided in Section 2.2.4.

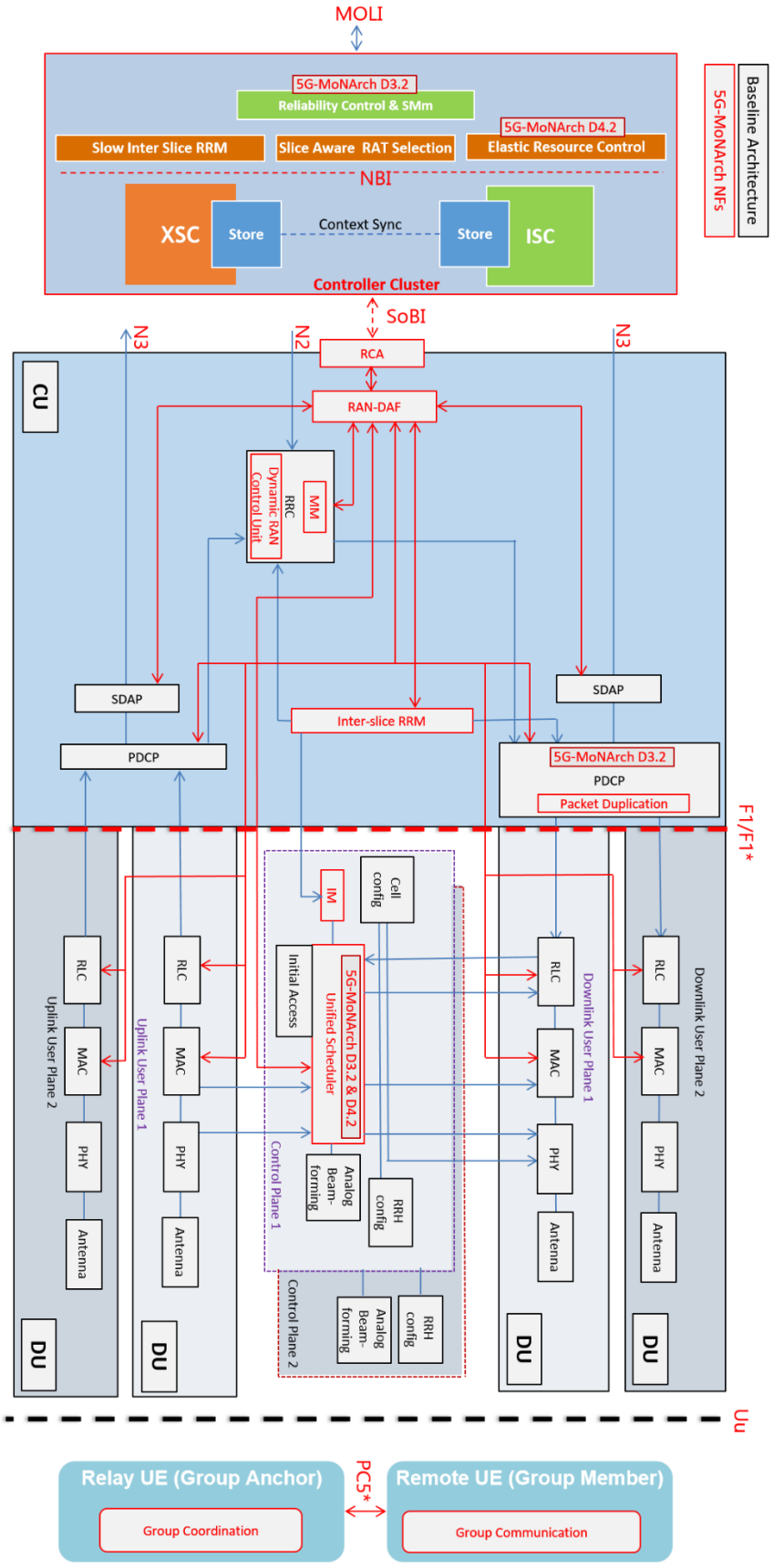


Figure 2-5: 5G-MoNArch RAN protocol architecture: the extensions from WP3 and WP4 are marked by 5G-MoNArch deliverables D3.2 [5GM-D3.2] and D4.2 [5GM-D4.2], respectively; the interfaces impacted by the 5G-MoNArch novel components are highlighted with red colour

2.2.1.1 Slice-aware RRM and RRC

The overall 5G-MoNArch architecture supports the isolation of NSIs, including resource isolation, OAM isolation, and security isolation. Resource isolation enables specialised customisation and avoids one slice affecting another slice. For instance, RAN needs to provide and enforce differentiation, and maintain isolation between slices where resources are constrained including RF resource, backhaul transport resource and computing resource. Each slice may be assigned with either shared or dedicated radio resource depending on RRM implementation and SLA. The amount of allocated resources can be scaled up or down for higher utilisation efficiency depending on the traffic load of each NSI. This section details inter-slice resource allocation approaches followed in 5G-MoNArch to efficiently share and manage resources between slices. Namely, first radio resource allocation schemes are described, which is followed by the extension toward new types of resources introduced in 5G (i.e., extended notion of a resource). Further extensions on the slice-aware RAT selection and group mobility are provided next.

Inter-slice RRM

The network slice-awareness in 5G RAN will strongly affect the RAN design and particularly the CP design, where multiple slices, with different optimisation targets, will require tailored access functions and functional placements to meet their target KPIs. To this end, RRM is one of the key aspects which will be affected. Here to mention that the operation and placement of RRM will be strongly affected by the aforementioned slice realisation variants which correspond to the slice isolation at RAN level. In Slice-aware RAN, in order to offer the flexibility that multiple slices can meet diverse KPIs (e.g., data rate, latency, and reliability), some RRM functionalities will be required to be tailored for different slice requirements.

On the other hand, the RAN deployment may provide some limitations on the efficiency of RRM due to the wireless channel, traffic load, and resource availability constraints, which may affect the overall performance (assuming numerous slices re-using the same RAN deployment). In particular, in dense urban heterogeneous scenarios, the signalling and complexity of RRM will be higher due to more signalling exchanges needed for passing RRM information to different entities. Moreover, the distribution of RRM functions in different radio nodes will provide new dependencies between RRM functions, which should be taken care of in order to optimise performance. In addition, in case of HetNet RAN deployments, non-ideal backhaul between access nodes (macro and small cells) will put some limitations on the RRM decisions and placement options to meet certain KPIs.

In slice-aware RAN, the CP can be categorised in the following groups of functionalities based on the RAN Configuration Modes (RCM) framework[3]. RAN Slice or RCM has been proposed in literature; and is a composition of RAN NFs, specific function settings and associated resources (HW /SW, and network resources). These RCMs will multiplex the traffic to/from core network slices to ensure optimisation across slices. To ensure meeting the E2E slice requirements, assuming limited RCMs, which may be mapped to numerous slices, a CP functionality framework is introduced, which is required to allow for slice-tailored optimisation in RAN. In particular:

- **Intra-RCM RRM:** For slice specific resource management and isolation among slices, utilising the same RAN is an open topic which is currently investigated. In literature [YT16], the conventional management of dedicated resources can be seen as intra-slice RRM, which can be tailored and optimised based on slice specific KPIs.
- **Inter-RCM RRM/RRC**: On top of Intra-RCM RRM, Inter-RCM RRM/RRC (which includes also Inter-slice RRM and slice-aware Topology RRM for wireless self-backhauling) can be defined as the set of RRM policies that allow for sharing/isolation of radio resources among slices or slice types. This can be used to optimise the resource efficiency and utilisation, by flexibly orthogonalising them in coarse time scales. Inter-RCM RRM can be defined as an "umbrella" functional block which dictates the RAN sharing and level of isolation / prioritisation among network slices or slice types. In this direction, an Inter-RCM RRM mechanism is proposed in [PP17], where slice-aware RAN clustering, scheduler dimensioning and adaptive placement of Intra-slice RRM functions is discussed in order to optimise

[3] Further details on RCMs have been captured in 5G-MoNArch Deliverable 2.1 [5GM-D2.1].

performance in a dense heterogeneous RAN. Given the requirement of new access functions which can be tailored for different network slices, the distribution of RRM functionalities in different nodes will be a key RAN design driver which can allow for multi-objective optimisation in a multi-layer dense RAN. The adaptive allocation of such functions is also envisioned as key feature to cope with the dynamic changes in traffic load, slice requirements and the availability of backhaul/access resources.

- **Topology RRM**: This can be seen as another category of Inter-RCM RRM, mainly for distributed RAN (D-RAN), where the resource allocation of wireless self-backhauling is essential to allow for joint backhaul/access optimisation [LPL+17]. Thus, Topology RRM can be tailored for different slices [PSW+17] in order to allocate backhaul resources among RCMs in a slice-tailored manner in order to avoid backhaul bottlenecks.
- **Unified scheduler**: An overarching Medium Access Control (MAC) Scheduler, where different slice types share the same resources and dynamic resource allocation and slice multiplexing is required on top of RCM-specific MAC. In addition, it needs to enable computational elasticity, which means in principle to consider available computational resources in the scheduling decision. Therefore, the unified scheduler has been considered as a possible VNF with novel elastic schemes for adapted MCS, PRB, and rank allocation have been proposed in [5GM-D4.2]. Similarly, the unified scheduler can also allow supporting the network coding procedures, such as enabling/disabling network coding and the choice of the parameters for the presented approaches in [5GM-D3.2].

Based on this categorisation, an interesting aspect which may define the CP functionality requirements and the interface / signalling requirements between the CP functions is the functional split which is dependent on the CU - DU split options. It is to mention that CU and DU split commonly refers to the split of the 5G base station (gNB, ng-eNB) protocol stack; however, it may also refer to functional splits involving cloud entities (e.g. central cloud - edge cloud split) when part of the RAN is virtualised.

Currently, in 3GPP, one split has been specified (from a set of introduced split options), namely Higher Layer Split (HLS) which is the splitting below PDCP-level. For the HLS split Figure 2-6 presents the possible placement of Inter-RCM and Intra-RCM RRM and RRC functionalities. Depending on the placement the interface requirements might be different due to the time/resource granularity of the CP functionalities and their possible interconnections. The interfacing between upper layers (RRC, PDCP) to lower layers for the configuration of the lower layer functionalities as well as the UP forwarding is specified in 3GPP as part of F1 and E1 interfaces.

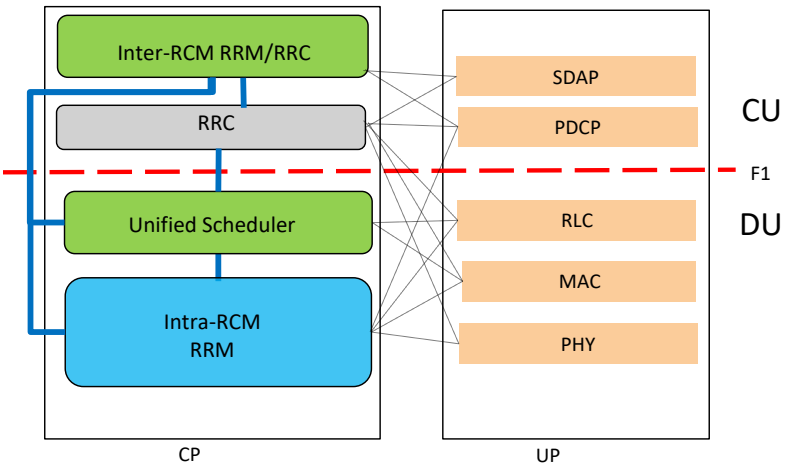


Figure 2-6: Exemplary slice-aware split for CU-DU (functional deployment and interactions)

Slow inter-slice RRM based on the Controller layer

The advantages of SDN such as centralised network abstraction and re-programmability can be used as an alternate solution to implement slice aware RRM. SDN controller instances along with centralised

RRM function (as NBI application) can be deployed as a service on demand managing radio resources between slices. Slices in turn consist of chained VNFs and PNFs to implement the E2E network and can be deployed in the same or physically separated cloud infrastructure. The major objective of using SDN framework for inter slice RRM is to achieve cross layer optimisation by using various network parameters such as radio resource status, current data-rate, buffer status, network latency, priority of slices, etc.

Though the radio resources usage can be better optimised by using centralised and joint optimisation techniques via SDN framework, there is an inherent latency added due to back and forth communication between RAN functions and the controller. The centralisation of RRM i.e., to have a radio resource allocation decision from the controller and send it to the RAN, every scheduling period (~1ms) is almost impossible due to various communication latencies (e.g., between controller and NB application, NB application and controller, controller and Scheduler). So, in order to better realise this centralised approach, 5G-MoNArch introduced the two-level approach based Inter-slice RRM, i.e., slow Inter-slice RRM from the controller (>1ms period) and native Inter-slice RRM from the CU (~1ms period). As introduced in Section 2.2.1, the **RCA** facilitates such approach by interacting with the native or fast Inter-slice RRM in the network layer and the slow Inter-Slice RRM in the controller layer.

Slice-aware functional operation

As introduced in D2.1 [5GM-D2.1][BRZ+15][BP19], slice-aware functional operation can comprise multi-slice resource management on shared infrastructure resources as well as hard/physical network resources, namely, wireless access nodes. 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 (DSCs) available in a certain region. A DSC can comprise access nodes that are not bound to a fixed topology. Slice-aware functional operation thus considers network topology and slice requirements to determine the functional operation (see Section 3.3.2).

The functional operation can be determined by a **dynamic RAN control unit** (see Figure 2-5 and Figure 2-7). Depending on the functional operations of the DSCs and how frequently the functional operations are configured by the RAN control unit, the RAN control unit may reside at different NEs, as exemplified in Figure 2-7. Two example realisations are provided in the following, while more details are given in Section 3.3.2.

- In case of L1 functional operation (e.g., physical layer, PHY), L2 functional operation (e.g., {PHY, MAC}, or {PHY, MAC, RLC}, or {PHY, MAC, RLC, PDCP}), and amplify-and-forward modes, the DSC cell is part of the surrounding macro cell and the UE may be connected both to the DSC and macro cell. Then the configuration of the DSCs can be dynamically or semi-dynamically changed by a RAN control unit connected to and possibly residing at the RRC of the macro cell Base Station (BS)[4], as depicted in Figure 2-7 (a). This would necessitate additional signalling on the self-backhaul (e.g., Un*) interface between macro BS and DSC. This signalling can configure the reception and transmission modes of the DSC based on the network slice customer served by the DSC. The dynamicity of such configuration updates depends on the backhaul link quality and slice requirements. In particular for DSCs like vehicular nomadic nodes (VNNs) [5GM-D2.1][BRZ+15] [BP19], the backhaul link quality will depend on the position and, thus, at each location change, a new configuration for the functional operation can be needed. The frequency of such updates can range from minutes to hours. Accordingly, having the RAN control unit at the macro cell BS enables more dynamic functional operation updates in case of dynamic radio topologies.
- In case of L3 functional operation (e.g., {PHY, MAC, RLC, PDCP, SDAP and RRC}), the DSC may have its own cell, e.g., with a physical cell ID (PCI), and the configuration of the functional operation may take place at a slow time scale. In this case, as shown in Figure 2-7 (b), in one implementation, the RAN control unit may have a **Cross-slice M&O functionality** that resides

[4] The set-up of the BS may determine the interface that may be impacted. In case of monolithic BS, the impacted interface would the wireless backhaul interface, while in case of disaggregated BS (i.e., CU-DU split), the impacted interface would be F1 [3GPP TS38.470].

at the NSMF. In another implementation, RAN control unit may reside at the RAN, e.g., at RRC, and may communicate with the Cross-slice M&O functionality for configuring L3 DSC functional operation, where part of the configuration parameters (such as transmit power) may be obtained from this Cross-slice M&O functionality.

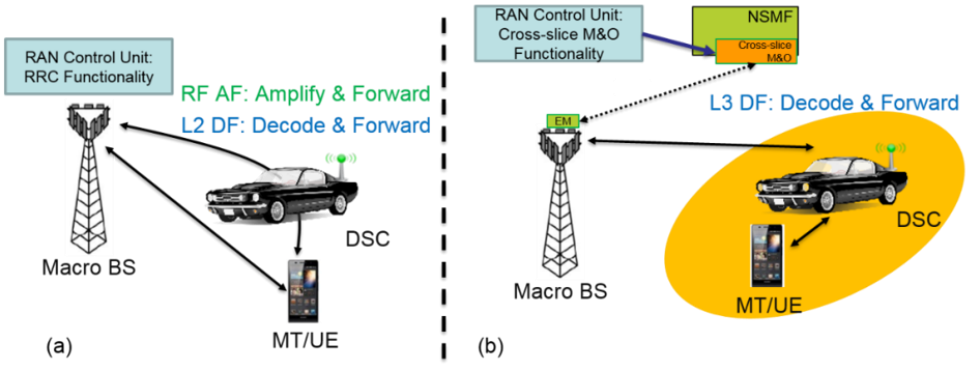


Figure 2-7: Example illustration of the slice-aware functional operation with three different modes (RF-L1, L2, and L3), which are determined and configured by a RAN control unit; amplify-and-forward and decode-and-forward are marked as AF and DF, respectively

Slice-aware RAT selection

As already mentioned, the efficient control of the available radio resources and technologies to satisfy the heterogeneous requirements of different slices is currently an important 5G research challenge. In the 5G RAN architecture, the 5G-MoNArch concept foresees that in the central unit a Mobility Management (MM) module is in charge to optimise the access of the users to a specific RAT, provided by a 5G distributed unit, according to the slice to which the user is associated. In contrast to classical association paradigms, this approach will enable to consider slice specific KPIs, like the reliability of the access technology, see Section 3.3.3 for more details.

D2D group mobility

Device to Device (D2D) communications facilitates an enabling innovation for further support of service continuity and smooth mobility (beyond the network edge). This can be realised via offloading some signalling at the RAN level (from direct signalling to the gNB to indirect signalling between anchor and remote UEs) as shown in Figure 2-5. As will be detailed in Section 3.1.3, a novel group mobility paradigm is established with floating mobility anchor as a Group Coordinator, not necessarily "pinned" to single Relay UEs. The above solution can be confined to RAN domain where the mobility management is handled by gNB. However, gNB needs to be aware of anchor assignment and associated remote UEs (i.e., Group Members) per group as will be outlined later.

2.2.1.2 Elastic resource control

Current trends on big data and its pervasiveness open an opportunity to exploit data analytics to improve the operation of different aspects of the mobile networks by means of data analytics. Although also previous generations of mobile networks incorporated monitoring data for the basic network management, the possible extensions of the new available data and the heterogeneity of the management decisions that shall be taken (e.g., radio and cloud resource assignment, per slice) bring this aspect to another level. Moreover, the increasing availability of machine learning / artificial intelligence algorithms make the data analytics-based network management even more appealing. One of the innovations of 5G-MoNArch, referred to as **elastic resource control** and illustrated in Figure 2-5, is to incorporate such data analytics into the architecture to improve the operation of certain functions, such as radio resource optimisation and slice selection. Such data analytics are fed with the data that can be gathered from the network as a whole (i.e., VNFs and PNFs composing a certain network slice and the attached UEs). Generally speaking, the more relevant the data is, the more accurate are the features that may be extracted from this data, so their applicability to the overall optimisation.

Indeed, data gathered from the RAN NFs is probably the most important kind of data that has to be gathered from a resource assignment perspective. The selected MCS highly depends on the Signal-to-Interference-and-Noise Ratio (SINR) margin of a certain user and have a big impact on the computational effort induced by the decoding / encoding functions of the RAN.

Data gathering in the RAN could be achieved by means of a monitoring application running at the XSC that, e.g., gets from the gNB DU information about the used Physical Resource Blocks (PRBs) by each tenant or the SINR of each user. Novel possible example information elements to enable big data also for computational elasticity are defined in Section 3.1.2. Other probes can be placed directly on tenants controlled VNFs (cVNFs) (e.g. AMF, to keep track of registered users) or directly in the Orchestration.

Resource assignment to slices for radio purposes entails taking optimal decision on both the spectrum assignment to slices, but also on the computational capabilities needed by each slice to encode / decode data of UE using that spectrum. Therefore, by gathering such data, a Big Data analytics component such as the one defined by ETSI Experiential Networked Intelligence [ETSI ENI] could provide algorithms for the following enablers defined by 5G-MoNArch (see more details in [5GM-D4.1] and [5GM-D4.2]):

- Characterisation of network slices load, in terms of used bandwidth, both in the spatial and in the temporal component. This can be leveraged for the correct dimensioning of the data centres, the slice admission control algorithms and intra-slice orchestration algorithms.
- Composability of Network Slices: in order to exploit multiplexing gains of elastic resources assignment, assessing the degree of complementarity on both the spatial and temporal dimensions will be leveraged by proactive cross-orchestration algorithms. That is, different network slices may provide high gains in statistical multiplexing, allowing thus for a higher efficiency in the resource assignment.

Possible examples of how this data can be leveraged for this purpose are available in [5GM-D4.1] and [5GM-D4.2]. In there, the complementarity of mobile network services is investigated, showing the level of complementarity on both temporal and spatial dimensions.

While predominantly the data coming from the (core and access) VNFs has been addressed, UEs can have a more prominent role for data preparation for the network based on past profile of intra-slice vs. cross-slice information they have gathered. As outlined earlier, network analytics can play a focal role in load balancing and radio resource optimisation at intra-slice and cross-slice levels. In order to establish UE interfaces to the analytics engines, new signalling procedures have to be devised between some cross-slice core associated network NFs related to MM and network slice selection and the RAN. For instance, mobility management data can be communicated e.g. via RRC messages at connection establishment and / or mobility procedures.

2.2.1.3 Reliability control and security monitoring management

In order to achieve the RAN reliability levels needed for URLLC services, 5G-MoNArch extends the architecture by a **reliability control and a security monitoring management (SMm) application** in the controller layer, as well as a reliability and a security trust zone VNF sub-plane in the network layer, as further detailed in Section 4.3.1 herein and in [5GM-D3.1] as well as in [5GM-D3.2]. In this regard, RAN reliability can be improved by either **data duplication at the PDCP layer** or **network coding functions as part of unified scheduler**, as shown in Figure 2-5. In addition to deploying standalone implementations of data duplication and network coding, a combination of such approaches is possible by means of a hybrid data duplication and network coding scheme, as is detailed in [5GM-D3.2].

The reliability control and management procedure pertaining to the data duplication technique is as follows. In order to support the duplication reliability function, the introduction of PDCP acknowledgments is envisioned. The PDCP acknowledgments operation is a new approach that was not included in LTE standards. In LTE standards, the packet acknowledgment feedback (ACK) sent from the receiver to the transmitter in order to indicate whether the transmission was correctly received is carried out in two layers: At the MAC layer by means of hybrid automatic repeat request (HARQ), and at the RLC layer by means of outer ARQ. Given that the RLC layers of the two links involved in **data duplication** procedure do not process the exact same packet sequence, in the **data duplication** case feedback should be sent with the PDCP packet numbering. As a result, duplicate packets can be

coordinated via specially designed mechanisms at the PDCP layer, ensuring thus that lost packets are recovered within a limited time interval [5GM-D3.1][5GM-D3.2].

2.2.2 Core network components

This section focuses on the 5G-MoNArch enhancements for the core NFs in the network layer. As outlined in Appendix B, technical specification [3GPP TS 23.501] of the 3GPP – Working Group SA2 (as standardisation body dealing with the 5G System Architecture) defines the service-based NFs and interfaces as illustrated in Figure 2-8. Here, also the NFs have been highlighted where enhancements to the CN beyond current developments in 3GPP (based on 5G-MoNArch studies) are envisioned, namely **Inter Slice Correlation Function (ISCF)**, and enhancements on **PCF**, **NSSF**, and **NWDAF**. The figure also shows an Application Function AF: it represents a general (unspecified) NF, possibly hosted and managed by 3rd parties, which can interact with other NFs directly or via network exposure function (NEF) using SBI and allows customisation of Network Slices as well as tighter interaction between 5GS and 3rd parties (e.g. Network Slice tenants, 5GS users).

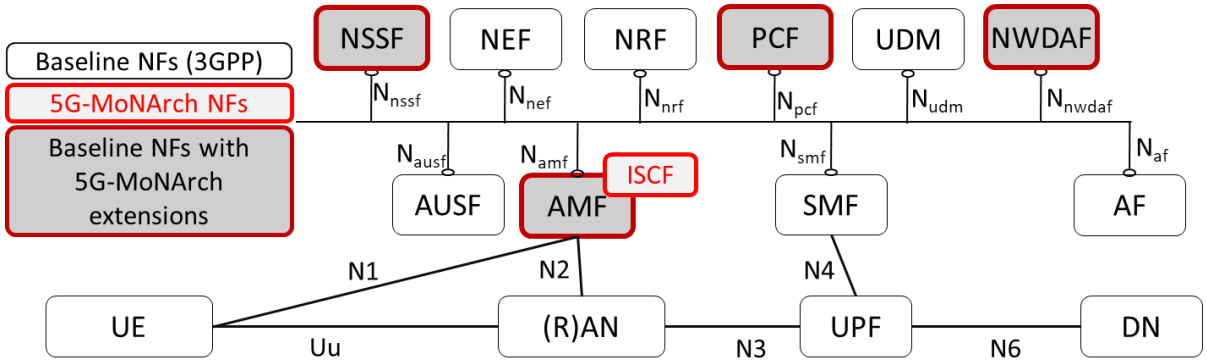


Figure 2-8: 3GPP 5G architecture and functional enhancements in the core network

The cross-slice optimisation includes two aspects: one aspect is related to the interaction between different layers or parts of the network, the other aspect is related to the interaction between different slice at the same layer or part of the network.

To fulfil targets on inter-slice context aware optimisation (see details in Section 3.2.1) on the first aspect, the following enhancements are envisioned:

- Enhancement of NWDAF to collect per slice/cross slice information and provide data analytics to the NFs.
- Optimisation for slicing and M&O layer based on context awareness:
  - Enable the Control Plane (CP) as well as the M&O layer to close the decision-making loop between the CP and M&O layer entities using context awareness in order to optimise the Mobile Core Networks (CN) operation.
  - Enhancement of NWDAF to collect information from M&O layer and maybe also provide feedback to M&O layer per slice/cross slice.
  - Enhancement of NWDAF, NFs, and M&O layer to coordinate the execution of changes in the 5G system based on the feedback provided by NWDAF in case of the CP / M&O layer joint optimisation cases.
  - Enhancement of NWDAF and/or NSSF to collect/ process terminal-driven analytics in order to improve slice selection and control.

A key NF here is the NWDAF. 5G-MoNArch envisions that this NF should be capable of collecting the data from the Controller layer (include the 5GC NFs, AFs) via SBI. It should also be able to collect information from and to provide analytics data to the management layer, RAN and UE where each layer may have its own implementation of a data analytics engine, which should be able to interact with NWDAF at 5GC to support end to end per slice service assurance.

For the second aspect, on Inter-slice coordination, the CP CN architecture solution needs the following enhancements (see further elaborations in Section 3.2.2):

(1) Enhancement of NF to provide per service traffic flow binding.
(2) Enhancement of NFs to distribute the service traffic flow binding
(3) Enhancement of PCF to treat per service correlated QoS profile
(4) Enhancement of network performance and capabilities exposure towards verticals
(5) Enhancement of NFs to perform cross slice optimisation.

The introduction of enhancements (3) and (4), starting from 3GPP Release 15 as baseline, may be achieved via alternative solutions. A simple option (quite straightforward to standardise) is via the extension of services exposed by PCF, as well as of the mechanisms for such services to be accessed by NFs, in particular by AF (possibly from 3rd parties, i.e. Verticals). As far as concerning enhancement (4), the services to be extended would of course depend on the requirements of the specific verticals considered. The AF represents the contact point between 5GS and the Vertical, and the extended services will be tailored to the specific vertical requirements.

The PCF service extensions are conceived considering 3GPP Release 15 services as baseline, summarised in Table 2-3 below. Details can be found in TS23.502 [3GPP TS 23.502], Section 5.2.5. The service exposure to 3rd parties' mechanisms is extended upon the 3GPP Release 15 NEF services as baseline, summarised in Table 2-4. Details can be found in TS23.502 [3GPP TS 23.502], Section 5.2.6.

Table 2-3: PCF Services (3GPP Rel15 baseline)

Service Name	**Service Operations**	**Operation Semantics**	**Example Consumer (s)**
Npcf_AMPolicyControl	Create	Request/Response	AMF
	Update	Request/Response	AMF
	UpdateNotify	Subscribe/Notify	AMF
	Delete	Request/Response	AMF
Npcf_Policy Authorisation	Create	Request/Response	AF, NEF
	Update	Request/Response	AF, NEF
	Delete	Request/Response	AF, NEF
	Notify	Subscribe/Notify	AF, NEF
	Subscribe		AF, NEF
	Unsubscribe		AF, NEF
Npcf_SMPolicyControl	Create	Request/Response	SMF
	UpdateNotify	Subscribe/Notify	SMF
	Update	Request/Response	SMF
	Delete	Request/Response	SMF
Npcf_BDTPolicyControl	Create	Request/Response	NEF
	Update	Request/Response	NEF
Npcf_UEPolicyControl	Create	Request/Response	AMF, V-PCF
	Update	Request/Response	AMF, V-PCF
	UpdateNotify	Subscribe/Notify	V-PCF
	Delete	Request/Response	AMF, V-PCF

Table 2-4: NEF Services (3GPP Rel15 baseline)

Service Name	**Service Operations**	**Operation Semantics**	**Example Consumer(s)**
Nnef_EventExposure	Subscribe	Subscribe/Notify	AF
	Unsubscribe		AF
	Notify		AF

Nnef_PFDManagement	Fetch	Request/Response	SMF
	Subscribe	Subscribe/Notify	SMF
	Notify		SMF
	Unsubscribe		SMF
	Create	Request/Response	AF
	Update	Request/Response	AF
	Delete	Request/Response	AF
Nnef_ParameterProvision	Update	Request/Response	AF
Nnef_Trigger	Delivery	Request/Response	AF
	DeliveryNotify	Subscribe/Notify	AF
Nnef_BDTPNegotiation	Create	Request/Response	AF
	Update	Request/Response	AF
Nnef_TrafficInfluence	Create	Request/Response	AF
	Update	Request/Response	AF
	Delete	Request/Response	AF
Nnef_ChargeableParty	Create	Request/Response	AF
	Update	Request/Response	AF
	Notify	Request/Response	AF
Nnef_AFsessionWithQoS	Create	Request/Response	AF
	Notify	Request/Response	AF

Furthermore, [TR 23.786] has defined a set of key issues that are directly addressed by 5G-MoNArch solutions (e.g., those detailed in Sections 3.2.1, 3.2.2, and 3.2.3), particularly "Key Issue #3: QoS Support for eV2X over Uu interface", "Key Issue #7: Network Slicing for eV2X Services" and "Key Issue #15: Enhancements to assist Application Adjustment"

In SA2#127-bis, a new use case "UE-driven analytics sharing" proposed based on 5G-MoNArch studies in 3GPP SA2 and agreed that may require enhancement of NWDAF and/or other NFs. In SA2#128, some key issues proposed and agreed for this use case such as:

- How the NWDAF collects the UE's information;
- How the NWDAF uses the data provided by the UE to do analytics and provides the analytics information to other NFs.

We will detail some solutions for the above within Section 3.2.3.

2.2.3 Management and orchestration components

This section describes the novel Management and Orchestration (M&O) components introduced by the 5G-MoNArch architecture. In particular a functional split of the CSM/NSMF/NSSMF M&O entities is defined presenting several new components with respect to the SotA. The section further details the internal architecture and functions of the management entities that compose the overall M&O architecture. The proposed functional split is an enhancement with respect of what is currently standardised in 3GPP SA5, adding new functions according to 5G-MoNArch requirements.

This functional decomposition is intended to highlight 5G-MoNArch novelties that have been defined working on the MANO related enablers (see Section 3.4). In an SBA approach each function offers its functionality as a service to any authorised consumer anyway, in the practical implementation, some services are used locally and other are exposed to other management domains (cf. Figure 2-9). It is important to notice that the model adopted by 5G-MoNArch for its novelties, thanks to the adoption of the SBA approach, leads to a very flexible management layer. The interaction among the management function can be different according to the specific management needs. The orchestration process can be tailored and can use the 5G-MoNArch novel APIs not in a predefined and static way, but it can compose the modules as needed.

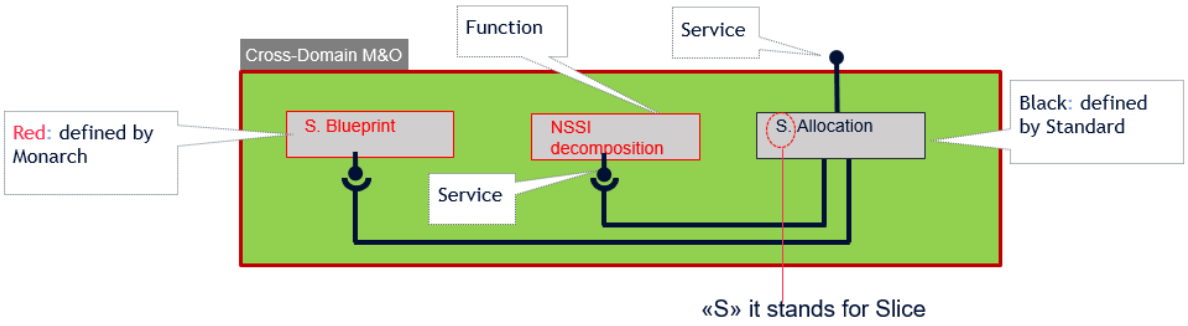


Figure 2-9: M&O functional split principle

The E2E Service Management and Orchestration sublayer, detailed with internal functions, is depicted in Figure 2-10. The black boxes represent the functions already defined in 3GPP SA5, while the red boxes represent the novelty functions defined by 5G-MoNArch.

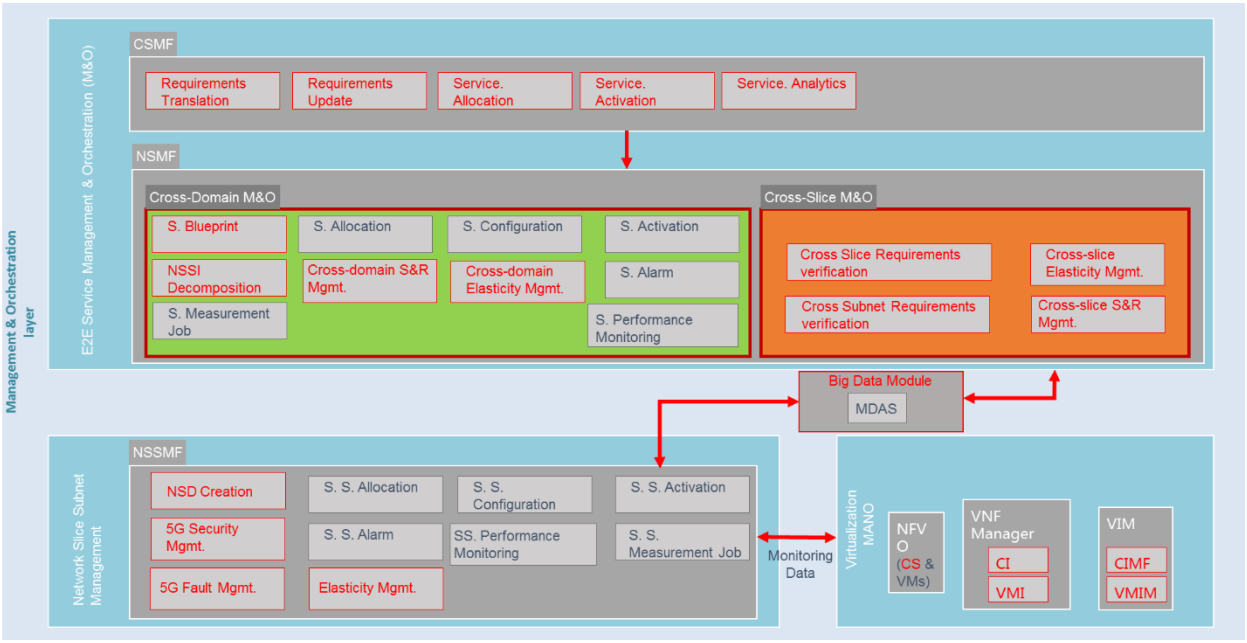


Figure 2-10: Breakdown of the E2E Service Management & Orchestration sublayer

In the following sections, each of the three management functions (CSMF, NSMF, and NSSMF) will be presented with its functional decomposition. The Big Data Module is explained in depth in Section 3.3.5.

2.2.3.1 5G-MoNArch communication service management

The **Communication Service Management Function (CSMF)** takes care of service allocation and management, also translating the service requirements into network requirements and offering a view on the service status and performance. Please note, that all the functions defined for the CSMF are new with respect to SotA defined in 3GPP SA5. The functions defined for the CSMF are described below:

- **Communication Service Allocation**: this function exposes a service to the tenant to request the allocation of a communication service. This request from the customer triggers a request to the NSMF for the allocation of an NSI. This function receives as an input the service requirement. Before asking for an NSI the service requirement has to be translated to network requirements that are the input for the request to the NSMF. The Service Allocation function also exposes a service for the update of the requirements. When a communication service has been requested, the tenant can modify it updating the service requirements. This function takes care of

requirement management consuming the service exposed by the Service Requirement Translation.

- **Communication Service Requirements Translation**: this function translates service requirements into network requirements. This function is consumed by the Service Allocation function.
- **Communication Service Activation**: the allocation of a service is intended to setup all the required infrastructure to build up the service without exposing it to the end users. This function takes care of the actual service activation to have it exposed to the customers.
- **Communication Service Analytics**: performs network data analytics in order to obtain analytics at service level.

2.2.3.2 5G-MoNArch network slice management

With reference to the 5G-MoNArch M&O layer, the **NSMF** is divided into the **cross-domain M&O** and the **cross-slice M&O**.

The **cross-domain M&O** takes care of the management of a single network slice across the different management domains. The function split for this management entity is as described below:

- **Slice Allocation:** the slice allocation function takes as an input the network requirements provided by the CSMF and, also consuming the services exposed by the cross-slice M&O management entity, reuse an existing NSI or create a new NSI to satisfy the allocation request. To create a new NSI the network requirements has to be used to create the actual structure of the NSI in terms of NFs, topology, connectivity and configuration. These data are managed in the Network Slice Blueprint.
- **Slice Blueprint:** the function produces the slice constituents/attributes/configuration starting from the network requirements maybe with the support of predefined templates for specific well known or standardised slices. The 5G-MoNArch Slice Blueprint, that defines the Network Slice Instance in terms of NFs, their interconnection and configuration according to a specific service request, will be presented in Section 4.1.
- **NSSI Decomposition:** the function decomposes the network slice into slice subnets producing the network slice subnet blueprint for each required NSSI
- **Cross-domain S&R Mgmt.:** Slice life cycle optimisations as well as slice configuration and performance enhancements, such as (re-)configuration and troubleshooting (self-healing), enabling resilience/ security features in the Management & Orchestration layer, using **inter-slice context-aware optimisation** and **inter-slice resource management** enablers described in Section 3.2 and Section 3.3, respectively. In addition, it integrates WP3 **intra-slice security and resilience management** functions presented in Section 4.3.1.
- **Cross-domain Elasticity Mgmt.:** Slice life cycle optimisations as well as slice configuration and performance enhancements, such as resource scaling and NF dynamic deployment, enabling elasticity features in the Management & Orchestration layer, using **inter-slice context-aware optimisation** and **inter-slice resource management** enablers described in Section 3.2 and Section 3.3, respectively. In addition, it integrates WP4 **computational and orchestration-elasticity** functions described in Section 4.3.2.
- **Slice Configuration:** once the NSI is deployed, this function takes care of the configuration of slice. As an example, it configures (through the appropriate domains controllers) the connectivity among the NSSIs.
- **Slice Activation:** the allocation of a network slice is intended to setup all the required resources to build up the slice without having it up and running. This function takes care of the actual slice activation to have it exposed to the customers to support a communication service.
- **Slice Alarm:** alarms management at slice level, filtering and aggregating the alarms coming from the different slice subnets to provide alarm information and management for a specific network slice.

- **Slice Performance Monitoring:** performance data management at slice level, filtering and aggregating the performance data coming from the different slice subnets to provide performance data information and management for a specific network slice.
- **Slice Measurement Job:** measurement job management at slice level. This function transforms a request of measurement job for a slice into the appropriate measurement jobs for NSSIs that compose the NSI.

The **cross-slice M&O** takes care of the interaction and resource sharing among the deployed NSI. The function split for this management entity is as described below:

- **Cross slice requirements verification**: this function supports the allocation of a network slice evaluating if an existing NSI can also support the new requested communication service in terms of requirements.
- **Cross subnet requirements verification:** this function supports the creation of a new network slice evaluating if existing NSSIs can also support the requirement for the slice subnets that are constituents of the new network slice.
- **Cross-slice Elasticity Mgmt.:** function dedicated to the Cross-Slice algorithms hosting **inter-slice resource management** and **inter-slice Management & Orchestration** enablers described in Section 3.3 and Section 3.4, respectively. In addition, it integrates WP4 **slice-aware and orchestration-elasticity** functions described in Section 4.3.2.
- **Cross-slice S&R Mgmt.:** function dedicated to the Cross-Slice algorithms hosting **inter-slice resource management** and **inter-slice Management & Orchestration** enablers described in Section 3.3 and Section 3.4, respectively. In addition, it integrates WP3 **cross-slice security and resilience management** functions presented in Section 4.3.1.

The Slice Blueprint, the NSSI decomposition, the Cross-slice Elasticity/S&R Mgmt and all the functions defined inside the cross-slice M&O represent new elements with respect to the SotA defined in 3GPP SA5.

2.2.3.3 5G-MoNArch network slice subnet management

The **NSSMF** manages the network slice subnets that are constituents of a network slice. 5G-MoNArch architecture foresees multiple NSSMFs e.g. for different domains or technologies. Each NSSMF takes care of a groups of NFs collected into a management entity named sub network slice that can also be shared among NSIs for resource optimisation.

The functions defined for the network slice subnet are described below:

- **Slice Subnet Allocation:** the slice subnet allocation function takes as an input the network slice subnet requirements provided by the NSMF and, also consuming the services exposed by the cross-slice M&O management entity, reuse existing NSSIs or create new NSSIs to satisfy the allocation request.
- **Slice Subnet Configuration**: once the NSSI is deployed, this function performs the configuration of the NSSI. As an example, it activates the configuration of the application part of the VNFs through the network management domain.
- **Slice Subnet Activation:** the allocation of a network slice subnet is intended to setup all the required resources to build up the slice subnet without having it up and running. This function takes care of the actual slice subnet activation to provide the requested network service.
- **NSD Creation:** creates the Network Service Descriptor (NSD) for MANO. The NSD is a template file, whose parameters are following the ETSI MANO specification, used by the NFV Orchestrator (NFVO) for deploying network services (as combination of multiple VNFs). This file carries e.g. the configuration parameters, from the virtualisation prospective, of the VNFs that compose a Network Service and the links, in terms of transport network (physical or virtualised), among the VNFs.
- **5G Security Mgmt.:** function dedicated to the Security Mgmt algorithms for slice subnets.
- **5G Fault Mgmt.:** function dedicated to the Fault Mgmt algorithms for slice subnets.
- **Elasticity Mgmt.:** function dedicated to the Elasticity Mgmt algorithms for slice subnets.

- **Slice Subnet alarm:** alarms management at slice subnet level, filtering and aggregating the alarms coming from the different NFs to provide alarm information and management for a specific network slice subnet.
- **Slice Subnet Performance Monitoring:** performance data management at slice subnet level, filtering and aggregating the performance data coming from the different NFs to provide performance data information and management for a specific network slice.
- **Slice Subnet Measurement Job:** measurement job management at slice subnet level. This function transforms a request of measurement job for a slice subnet into the appropriate measurement jobs for the NFs that compose the NSSI.

The NSD Creation and the Elasticity, Security, and Fault Mgmt. functions represent new elements with respect to the SotA defined in 3GPP SA5.

2.2.4 Integrated data analytics framework

5G-MoNArch architecture has an integrated data analytics framework as shown in Figure 2-2 and Figure 2-3, respectively. The framework considers the data analytics capability at the Management & Orchestration layer, the network layer as well as the application layer or UE. Each logical data analytic module is implemented as multiple instances for different use case, and purposes. For instance, the Big Data Analytic Module in the Management & Orchestration layer could be implemented as multiple instances per domains (e.g., RAN M&O data analytics, VNF data analytics, etc.) at different levels (e.g., cross/intra domain). Such framework allows for dedicated data analytic module design at different layers, also enabling cross-layer optimisation. Details of the data analytics module in the NW layer (i.e., NWDAF) and M&O layer (i.e., MDAF) are described in Appendix B.

These individual data analytics modules serve at different parts and layers of the architecture. While each module may need to collect data from other layers for generating data analytics. The analytics generated by each individual module could also be a data source to be collected by other data analytics modules. Meanwhile, in some use cases, the data analytics generated by different modules also needs to be coordinated to avoid conflict actions trigged at different part of the architecture, or even for joint optimisation. All these needs to have efficient way of communication between these data analytics modules. Following 5G-MoNArch design paradigms, we propose to interconnect different data analytics modules with SBI. Below is a list of example implementation of interfaces as shown in Figure 2-11.

- Interface 1: NWDAF interact with AF (via NEF) using NW layer SBI
- Interface 2: N1/N2 interface
- Interface 3: O&M layer configure the NF profile in the NRF, and NWDAF collect the NF capacity information from the NRF
- Interface 4: MDAF interact with application/tenant using NBI
- Interface 5: MDAF interact with RAN DAF using O&M layer SBI
- Interface 6: NWDAF consumes the services provided by MDAF using cross layer SBI
- Interface 7: MDAF consumes the services provided by MWDAF using cross layer SBI
- Interface 8: MDAF collects data from NW layer via trace file/monitoring services

Part of this integrated data analytics framework (i.e., the interaction between the NW layer and M&O layer for data collection and analytics sharing) has already been successful brought into 3GPP Rel. 16 [3GPP TS23.288]. Some parts (e.g., UE DAF) are postponed to the later release in 3GPP SA2. And some parts (e.g., MDAF, RAN DAF) are recognised to be important for 5G system in 3GPP. The individual work has already been trigged in different 3GPP WGs (i.e., RAN2/3, SA5) and the interaction between them will be defined with the progressing of the work in different WGs.

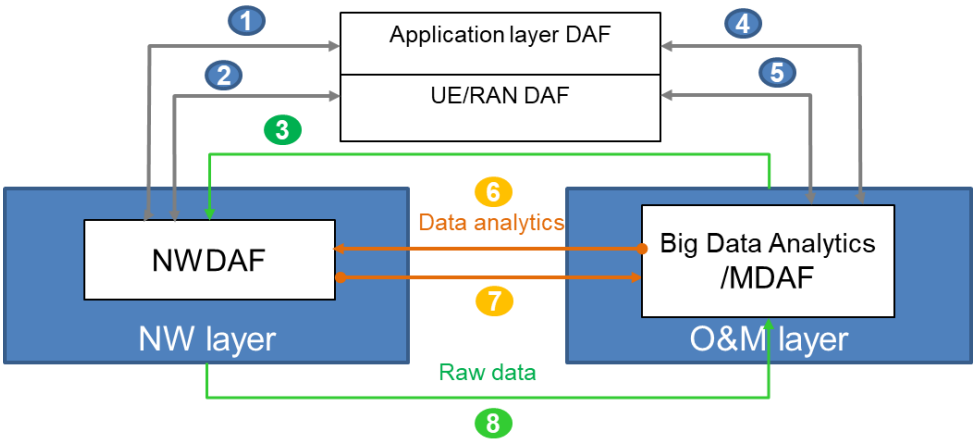


Figure 2-11: Data analytics framework in 5G-MoNArch

2.2.4.1 Data Analytics Characterisation

Firstly, we decompose the analytics functionalities in different levels, based on the monitored parameter (in case of predictive analytics, this should be the expected parameter). This can involve a UE session, or the resource load/situation in a particular domain, or the application/service operation.

- **UE/Session-related parameters**: These parameters may include the prediction of the UE context/behaviour to enable the network to better provision the resources. One example can be the mobility of the user or group of users, which can be used for handover management, or the prediction of interference that the UE will suffer from/cause in a particular area. One further example is the prediction of QoS for one or more UEs in a given area.
- **Network-related parameters:** Here, these parameters can be grouped based on the domain they apply to. In **RAN**, parameters can include the ones regarding the radio resource conditions and availability (e.g., average channel quality, load, and interference) as well as the traffic (e.g., user density) and other factors in real-time or non-real time. In **transport / backhaul**, the parameters that can be estimated concern resource conditions, backhaul/fronthaul (BH/FH) type, topology, availability, dynamicity, etc. Finally, for **CN,** some parameters that can be monitored are based on the processing load and availability of CN functions.
- **Service-related parameters:** This category includes the analytics which can be performed at the application domain (e.g., at terminal or at the application function) and may be used by the 5G network to improve the service operation. One example which is specific for V2X slicing case, is the prediction of UE trajectory/route, traffic conditions, or expected Level of Automation (LoA) for a particular area.
- **Management-related parameters:** This category includes Performance Management (PM) and Fault Management (FM) analytics as introduced in 3GPP SA5. This set of parameters may consider the current slice/subnet performance and statistics on, e.g., radio failures and will provide analytics to MDAF.
- **Cloud-related parameters:** This includes the cloud processing parameters, e.g., the load and availability of computational resources, which may affect the decision for virtualisation of NFs to cloud platforms. In a distributed cloud-based architecture, the above categories of parameters may be deployed on demand in edge or core cloud platforms. Given the tight latency and reliability requirements of some virtualised NFs (e.g., in RAN domain), performing analytics on the estimated computational resource load/conditions is of key importance for performing actions, like offloading the processing load to other cloud processing units.

Granularity of analytics

Real-time: The analytics can be performed in real-time operations (e.g., channel prediction in ms time scale); however, this is more challenging task due to the fact that additional processing might be required, and the overhead may affect the performance. One possible enabler for real-time analytics is

the automation of RAN functionalities which correspond to user sessions, as well as resource optimisation. Analytics on the performance fluctuations in RAN (e.g. SINR distribution over a given time window) can allow for pro-active scheduling which may be essential for URLLC type of traffic. Table 2-5 adds some example parameters that can be monitored/predicted in non-real time.

Near-real time / Non-Real time: In this case, the analytics, are performed in sec/min/hour time scale and may apply to certain types of prediction (e.g., load distribution in a geographical area). In O-RAN [ORAN], near-real time operations have been defined to capture operations like QoS management, traffic steering, mobility management which may be semi-dynamic (e.g., 100s of ms timescale). Table 2-5 adds some example parameters that can be monitored/predicted in non-real time.

Type of analytics

There are different types of analytics that can be useful for the network according to the Gartner's Graph on stages of data analysis [G12]:

- Descriptive Analytics – Explaining what is happening now based on incoming data.
- Diagnostic Analytics – Examining past performance to determine what happened and why.
- Predictive Analytics – An analysis of likely scenarios of what might happen.
- Prescriptive Analytics – This type of analysis reveals what actions should be taken.

2.2.4.2 Integrated Analytics Architecture

In order to allow this overall concept to have some real impact in Mobile Networks, 5G Monarch devoted efforts to promote this concept in 3GPP standardisation. The process of pushing the overall integrated analytics architecture concept into standard had to be broken down in steps. In a first moment the focus of the contributions was to introduce in 5G networks (from Rel. 16) the capabilities of:

- AFs and the NWDAF (a CN function) to interact for analytics exposure from NWDAF to AFs and for data collection from AFs to NWDAF. This interaction is performed via an intermediary CN Function called NEF (Network Exposure Function) as AFs from external parties are not allowed to directly interact with any CN Function except the NEF.
- NWDAF and OAM to interact for allowing direct data collection from OAM by NWDAF.

A series of contributions to 3GPP SA on enablers for Network Automation (eNA) Study Item (documented in 3GPP TR 23.791 [3GPP TR 23.791]) and most recently in the normative phase of eNA (documented in 3GPP TS 23.288 [3GPP TS 23.288]) have been successfully delivered and accepted by 3GPP as listed in Table 2-1. This is the first step towards the architectural enhancements to allow integrated analytics architecture among AF, CN/AN, and OAM.

To this end, in this section, we investigate the novel proposed Integrated Analytics Architecture in detail.

Apart from NWDAF and MDAF, the functionalities which can be defined as necessary parts of the E2E analytics design framework, as shown in Figure 2-12, are the following:

AF-DAF and DN-DAF: Outside the 3GPP 5GS, there are two relevant domains where additional data analytics functions can be deployed. In the DN, the network operator or the vertical can place functions that provide data related to service or performance of non-3GPP networks (e.g., metropolitan wide area networks, WANs) to other DAFs within 5GS or OAM domain. AFs or dedicated AF-DAFs can interact with CN-domain NWDAF, either via 3GPP Network Exposure Function (NEF) or via an inter-domain message bus as depicted in Figure 2-12. AF-DAFs enable the operator to deploy on demand new functionality customised for AF-domain requirements, or the vertical to perform analytics that can support the E2E service operation. This can prove highly beneficial for vertical industries like IIoT and V2X, where the vertical requires exposure of selected data from 3GPP network operation, a higher level of control of the network, as well as flexibility of deployment.

RAN-DAF: Real-time analytics are required for improving RAN NFs, like radio resource management. Since the RAN need to provide fast decisions, the analytics based on the processing of real-time measurements may need to stay local for optimising performance dynamically. Also, there are the business aspects, which may involve different stakeholder among RAN, CN, and OAM. So, the storage and analysis of radio-related measurement may not be available at CN or OAM. An example deployment of such functionality is shown in [PMM+19], where more complex RAN deployments with CU-DU

splits, better motivate for such functionality. Here, different options for performing RAN analytics may be examined. Either RAN-DAF will be a Control functionality at RAN or it can be a management or for some implementations as a SON functionality. With the proposed SBA as envisioned for both control and management functionalities, for both implementations of RAN-DAF, the interface will be via the inter-domain message bus interface.

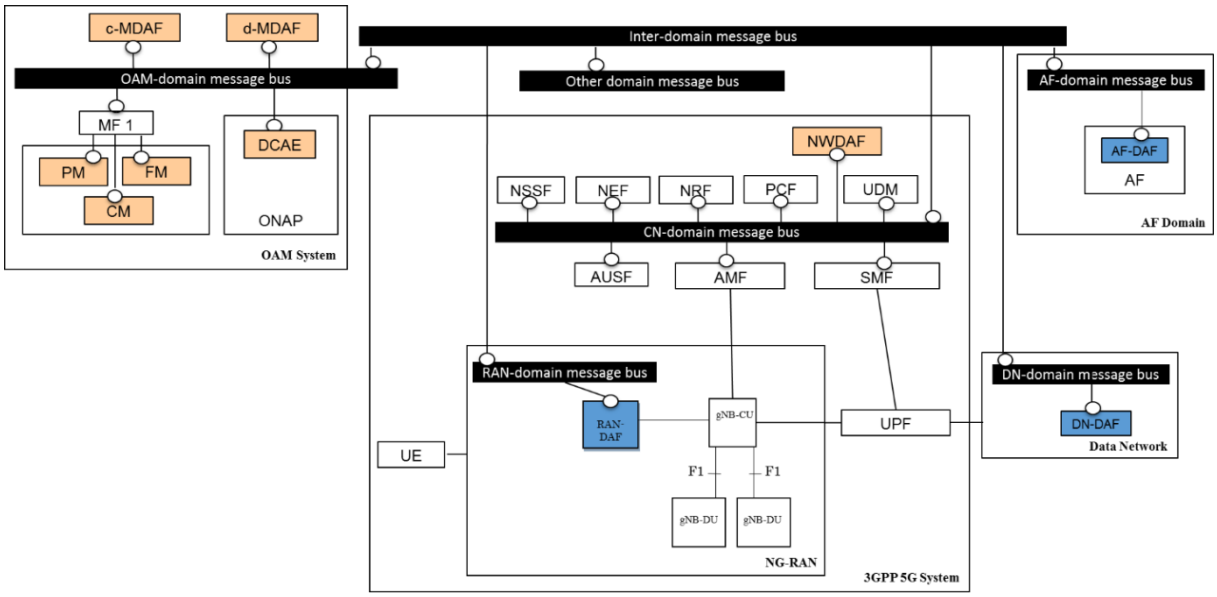


Figure 2-12: Integrated analytics architecture

Intra- and inter-domain message buses provide the functionality for registration, discovery, and consumption of services within a domain or across domains. Service registration and deregistration allows a service catalogue function to maintain an updated list of services available for consumption. Service discovery functionality allows to retrieve available services, refer requesting consumers to them and provide the means to access them. Service consumption functionality allows consumers to invoke services, e.g., by automatically routing requests and responses between service consumer and producers. This may include platform-like functionality, such as, load balancing, failover, security, message delivery rules, or protocol conversion / adaptation, and exposure of services to the inter-domain message bus and its service catalogue.

Given the types of analytics and the proposed architecture enhancements (which are further elaborated in [PMM+19]), Table 2-5 provides some exemplary functionalities which can be defined and configured based on different slice requirements and network conditions. For example, for some real-time analytics regarding parameters like UE mobility, wireless backhaul and access resource conditions/availability, as well as parameters for certain critical service types (e.g., URLLC and V2X), the deployment of analytics function in RAN domain could help providing more sophisticated and service-tailored RAN optimisation mechanisms in very short time-scale granularities and with minimum signalling overhead.

Table 2-5: Analytics functionality placement and classification [PMM+19]

	Parameter	**Type**	**Placement**	**Time-scale**
A. UE-related parameters	Mobility	Prescriptive Analytics	RAN-DAF	Real-time
		Descriptive Analytics , Predictive Analytics	AF-DAF	
	Interference level	Predictive Analytics, Prescriptive Analytics	RAN-DAF	Real-time

	UE QoS	Diagnostic Analytics, Predictive Analytics	NWDAF	Real-time / non-real time
		Prescriptive Analytics	AF-DAF	
B. Network-related parameters	Radio Resource Situation (conditions, usage, availability)	Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics	RAN-DAF	Real-time / non-real time
	Traffic / Load Situation	Diagnostic Analytics, Predictive Analytics	NWDAF	Non-real time
		Prescriptive Analytics	MDAF	
	Backhaul Conditions / Availability (e.g. BS neighbourhood change)	Descriptive Analytics , Predictive Analytics	RAN-DAF	Non-real time
		Diagnostic Analytics,	NWDAF	
		Prescriptive Analytics	MDAF	
	Cell Density	Diagnostic Analytics, Predictive Analytics	NWDAF	Non-real time
		Prescriptive Analytics	MDAF	
	Network QoS	Diagnostic Analytics, Predictive Analytics	RAN-DAF, NWDAF	Real-time / non-real time
		Prescriptive Analytics	AF-DAF	
C. Service-related Parameters	New / Modified Slice	Prescriptive Analytics	UE, AF-DAF	Real-time / non-real time
	NW assistance (for V2X)	Predictive Analytics	RAN-DAF, NWDAF	Real-time / non-real time
		Prescriptive Analytics	AF-DAF	
	UE Route / Trajectory (for V2X)	Prescriptive Analytics	AF-DAF	Real-time / non-real time
	Level of Automation Change (for V2X)	Prescriptive Analytics	UE, AF-DAF	Real-time / non-real time
D. Management-related parameters	PM	Descriptive Analytics, Diagnostic Analytics	MDAF	Non-real time
	FM	Descriptive Analytics, Diagnostic Analytics	MDAF	Non-real time

3 5G-MoNArch Enabling Innovations

Deliverable D2.1 [5GM-D2.1] and D2.2 [5GM-D2.2] have provided a gap analysis (cf., Appendix A) with respect to ongoing 5G system architecture design efforts in the industry and academia, as perceived from the 5G-MoNArch. As presented in Chapter 1, 5G-MoNArch contributions are based on innovation elements, each one composed of one or more enablers and grouped into three fundamental enabling innovations: *telco-cloud enabled protocol stack, inter-slice control and managemen*t, and *experiment-driven optimisation.*

The project's so-called innovation elements and enablers[5] as outlined in the following sections aim at enhancing, extending, and modifying the SotA 5G architecture concepts in order to close these gaps. Each section will therefore detail the innovation element, highlight the novelties of the concept, describe the involved NFs of the overall architecture (existing and novel NFs, cf. Chapter 2) and describe the work flow (using MSCs) to realise the innovations. Table 3-1 depicts an overview of the innovation elements and enablers, as well as their mapping to involved NFs and layer(s) of the overall architecture. On this basis, Table 3-1 together with Table 2-2in Chapter 2 provide an overview the interrelations of the 5G-MoNArch enablers and innovation elements within the 5G-MoNArch overall architecture. Table 3-1 presents also the physical layer interaction of 5G-MoNArch enablers/innovation elements for two main purposes such as collection of measurement reports and re-configuration for various parameters (e.g., for performance improvement and satisfying slice requirements).

Table 3-1: Mapping of enabling innovation, innovation elements, enablers and their Physical Layer Interaction to the 5G-MoNArch overall architecture layers

Enabling Innovation	**Innovation elements**	**Enablers**	**Involved/affected novel 5G-MoNArch components and interfaces from the 5G-MoNArch overall architecture**	**Physical Layer Interaction**
Cloud enabled protocol stack	Telco cloud-aware protocol design	Telco-cloud-enabled protocol design	Controller layer functions and selected UP VNFs in Network layer	Elastic VNFs requires adaptive Physical layer processing based on the available CPU resources
	Telco-cloud-aware interface design and requirements analysis	Telco-cloud-aware interface design and requirements analysis	RAN-level UP (and CP) VNFs, Xn and F1 interfaces, Network layer	No explicit interaction
	Terminal-aware protocol design	Terminal-aware protocol design	RAN-domain CP VNFs and interfaces incl. F1	No explicit interaction
Inter-slice control and management	Inter-slice Context-aware Optimisation	Inter-slice context sharing and optimisation	CN-level UP and CP NFS, M&O layer functions	No explicit interaction
		Inter-slice coordination	CN-domain CP NFs, service-based interfaces, Network layer	Requires flexible air interface where physical layer parameters eMBB and URLLC slice

[5] As illustrated in Chapter 1, a 5G-MoNArch enabling innovation consists of one or more innovation elements, where an innovation element can be constructed by one or more enablers depending on the needed level of granularity for designing the innovation element.

				can be tuned for resource efficiency
		Terminal analytics driven slice selection / control	CN-domain CP NFs (AMF, NSSF, NWDAF), interfaces to UE and Itf-X to M&O layer	No explicit interaction
	Slice-aware Functional Operation in RAN	Inter-slice RRM for Dynamic TDD Scenarios	RAN (Inter-slice RRM, IM, and Unified Scheduler)	Requires physical layer parameters and measurements
		Context-aware relaying mode selection	RAN (Dynamic RAN Control Unit at RRC), M&O Layer (Cross-slice M&O)	Requires frequent measurements from backhaul, access and direct links
	Inter-slice resource management	Slice-aware RAT selection	RAN-domain CP NFs, Network, Controller and M&O layer as well as associated interfaces	Requires Physical layer measurements related to the two RATs
		Inter-slice RRM using the SDN framework	RAN-domain NFs, XSC/ISC and applications, Network layer, Controller layer, interfaces: NBI, SoBI, MOLI	Requires measurements such as current radio resource status and channel feedbacks in the particular time interval decides by the Inter-Slice RRM algorithm
		Big data analytics for resource assignment	CN (NWDAF), M&O layer (Cross-slice M&O)	Requires measurements such as PRB usage, allocated MCS and channel quality reports
	Inter-slice Management & Orchestration framework	Framework for slice admission control	NSMF (Cross-slice M&O), NFVO, M&O layer and respective interfaces (e.g., Os-Ma-Nfvo)	No explicit interaction
		Framework for cross-slice congestion control	NSMF (Cross-slice M&O), NFVO, M&O layer and respective interfaces (e.g., Os-Ma-Nfvo)	No explicit interaction
		Slice admission control using genetic optimisers	NSMF (Cross-slice M&O), NFVO, M&O layer and respective interfaces (e.g., MOLI, Os-Ma-Nfvo)	No explicit interaction
Experiment-driven optimisation	ML-based optimisation using an extended	ML-based optimisation using an extended	RAN-domain VNFs (CP and UP), Network layer	Requires various parameters such as MCS allocation and

	FlexRAN implementation	FlexRAN implementation		channel quality reports
	Computational analysis of open source mobile network stack implementations	Computational analysis of open source mobile network stack implementations	RAN-domain VNFs (CP and UP), Network layer	Requires MCS allocation report for deciding computation resource requirements
	Measurement campaigns on the performance of higher layers of the protocol stack	Measurement campaigns on the performance of higher layers of the protocol stack	Higher-layer RAN VNFs (CP and UP), Network layer	No explicit interaction

The presentation of the innovation elements is structured using the three enabling innovations *telco-cloud-enabled protocol stack* (Section 3.1), *experiment-driven optimisation* (Section 3.5), and *inter-slice control & management*, where the latter is split into three sub-groups each representing an innovation element, namely inter-slice context-aware optimisation (Section 3.2), inter-slice resource management (Section 3.3), and inter-slice management & orchestration (Section 3.4).

The two enabling innovations *telco-cloud-enabled protocol stack* and *experiment-driven optimisation* have a special role since they do not follow the classical approach of designing a NF for a specific purpose, e.g., optimisation of resource utilisation. Rather, they propose a completely new approach to system architecture design. While the former focuses on minimising cross-functional dependencies (e.g., telco-cloud-ready function and interface designs in the RAN), the latter uses observations and results from operational networks to enhance the architecture and the behaviour individual NFs (e.g., resource orchestration algorithms used in the M&O layer). Therefore, these innovations do not always have an immediate representation in the functional architecture.

3.1 *Telco-cloud-enabled protocol stack*

The expected advantages brought by a cloud-enabled protocol stack design are backed by the relative maturity of current software initiatives (such as, Open Air Interface [OAI] or SRS LTE [SRSLTE]) and the recent increase in the pace of their updates. Moreover, this also provides the motivation and the means for researchers to investigate possible enhancements of technologies that have only been available in rather proprietary manner in the past, such as, cellular radio protocol implementations.

In future, fully softwarised and cloudified mobile networks will necessarily build on cloud-aware protocol stacks. Both network management and the resulting overall performance will benefit from making VNFs aware of being executed on shared resources by means of virtualisation environments such as virtual machines or containers. In this section, the main challenges to achieve this vision are discussed, while and describing possible implementations of functionality that builds on this cloud-awareness.

This approach entails two main challenges, namely (i) redefining the interactions between VNFs, relaxing as much as possible their temporal and logical connections, and (ii) support an elastic operation, to efficiently cope with changing input loads while running in an infrastructure of resources that is not over-provisioned. The functional requirements of these novel design strategies are detailed in what follows, before discussing why they will also require the formal definition of novel KPIs cf. [5GM-D6.1].

Given the high flexibility provided by the NFV approach, the deployment of such cloud-aware protocol stack does not have a direct negative implication on the provided telecommunication service per se. The re-definition of the interactions among VNFs allows for a more flexible service orchestration, while the re-design of VNF internals may be easily provided by a code refactoring in a much faster way than the current tightly coupled HW-SW PNF approach. While having a cloud-aware protocol stack will benefit any kind of telecommunication service, this may be particularly relevant for the extreme ones. For

example, a mission critical VNF can be optimised to reduce its memory footprint, while low latency services may exploit especially tailored orchestration patterns involving edge computing facilities.

In the following, we describe the telco-cloud enabled protocol stack in details. First, in Section 3.1.1, we explain the 5G-MoNArch approach to the telco-cloud-aware protocol design. Then, in Section 3.1.2 we discuss the architectural aspects of such new paradigm. In particular, we focus on the orchestration aspects and how the i) CPU consumption of the different modules shall be considered by the orchestration and ii) how the maximum allowed delay between them impose orchestration choices. Finally, we discuss the telco-cloud-enabled extension from the UE mobility perspective, and how group mobility can substantially improve the performance of the core network functions.

3.1.1 Telco-cloud-aware protocol design

Concept

The approach described in above provides several advantages, as it allows heterogeneous deployments for different services (i.e., mMTC and eMBB), which are tailored to their specific requirements. For example, depending on the latency, bandwidth, and/or computational requirements of the service, it may be better to locate certain VNF towards the edge of the cloud rather than in a central location. How to place VNF across the cloud is a network orchestration problem, which is constrained by the split into modules described above. However, this typical NF decomposition for the RAN protocol stack was not designed for its cloudification, and therefore the potential gains are limited. This issue is discussed in more detail in the following. Also, the deployment of VNF in computational resources constrained environments, such as edge clouds, takes advantage of this enabler.

One key assumption of network stack designs is that certain functions are implemented in the same physical space, e.g., within the same chip (maybe on a different chip, but surely on the same HW). So, non-ideal links with non-negligible delays are a problem for physical network elements that need to be decomposed into several NFs. Interfaces among them, thus, were designed considering communication links spanning some microns of silicon, and not several miles of fibre as in the case of, e.g., C-RAN.

In this way, the possible inter-dependencies between these functions are overlooked, as the delivery of information between them is practically immediate. However, as argued above, to fully benefit from a network-wide orchestration of a cloudified stack, VNF should support their execution on different nodes. But the design of traditional protocol stacks does not support such flexible placement of VNF, as those with heavy inter-dependencies may introduce very high coordination overheads or may not be even possible due to infeasible network requirements. These limitations severely constrain network orchestration, which compromises the overall gains obtained from the flexible function allocation. This is flagrant for e.g., the introduction of centralised RAN functions, where long delays in the information exchange between radio access points and the central cloud result in serious performance deterioration.

Position in 5G-MoNArch architecture & Protocol Implications

Because of the above, the full protocol stack (and, in particular, the RAN) has to be re-designed with the goal of leveraging the benefits of the flexible function decomposition and allocation, so as to cope with non-ideal communication (i.e., non-zero and varying delay, limited throughput) between the nodes in the cloud. Specifically, a cloud-aware protocol stack should relax as much as possible, or even completely remove, the logical and temporal dependencies between VNF, such as very tight timing constraints for the HARQ, to enable their parallel execution and provide a higher flexibility in their placement.

One of the most immediate and 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. However, it is also likely that in certain occasions the resource assignment across the cloud cannot cope with the existing traffic due to some peaks of resource demands. This is particularly true for C-RAN deployments, which have to deal with demand loads known to be highly variable. In this scenario, allocating resources based on peak requirements would be highly inefficient, as this design jeopardises multiplexing gains in particular when cloud resources may be scarce (e.g., a "flash crowd" at an edge cloud): here any temporal shortage might result in a heavy congestion or even a system failure. VNF, instead, shall efficiently use the resources they are assigned with. Thus, they have to become elastic, i.e., adapt their operation when temporal changes in the resources available

occur, in the same way they have a long-established manner of dealing with outages such e.g. channel errors. Therefore, to fully exploit the benefits of softwarising the network operation, the NF design has to take the potential scarcity into account and be prepared to react accordingly.

This enabler does not have a direct implication on the 5G-MoNArch architecture per-se, but it provides the fundamental building blocks (i.e., cVNFs and uVNFs) on the network layer, that are used by the controllers and M&O to achieve elasticity or resilience.

However, enabling Telco-cloud aware protocol stack means to design specific interfaces that gather the resource utilisation of a given function at any time. That is, through the controllers for RAN functions, the MANO shall be able to get precise information about the resource utilisation of each VNFs running in a specific container / VM. Information such as memory utilisation or CPU utilisation, possibly broken down into specific function utilisation (i.e., encoding or decoding functions). Also, the amount of available resources shall be communicated to the VNF itself, through system parameters that are configured by the controllers or the Element Manager of the VNFs.

Evaluation and analyses

In the context of wireless communications, the concept of elasticity usually refers to a graceful performance degradation when the spectrum becomes insufficient to serve all users. However, in the framework of a cloudified operation of mobile networks that has to deal with elasticity under resource shortages, also other kinds of resources need to be considered that are native to the cloud environment such as computational, memory, and storage assets available to the containers the VNF are bound to. This has hardly been a problem for traditional NFs, that were designed to run over a given HW substrate with exclusive access to the resources and requires the definition of novel interfaces that will provide the amount and type of available cloud resources at a given point in time, just like, e.g., the accessible spectrum is a parameter for a RAN function.

Elasticity has also been considered by non-VNF cloud operators, but the presented 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 more scattered in the presented scenario (e.g. they are distributed across the "edge clouds"), which reduces the possibility of damping peaks by aggregating resources.

To better illustrate the benefits of elasticity in the cloudified mobile network operation context, firstly the notion of “computational outage” is considered, i.e., the unavailability of the required resources to perform the expected operation. In a traditional, non-elastic operation, there is a 1-to-1 mapping between outages and performance loss, as Figure 3-1 illustrates: if the resources are not available 20% of the time, there is a 20% performance degradation, as the function is unable to operate under any shortage. In contrast, an elastic design supports what hereafter is referred to as graceful performance degradation, which causes that the VNF would still work under a resource shortage (with reduced performance, though), this resulting in the "gains" qualitatively illustrated in the Figure 3-1. Making a protocol stack cloud-aware through elastic VNF requires hence a paradigm shift in their design, moving away from the tight HW-SW co-design as discussed before, to a flexible operation in which the amount of available resources is an additional parameter.

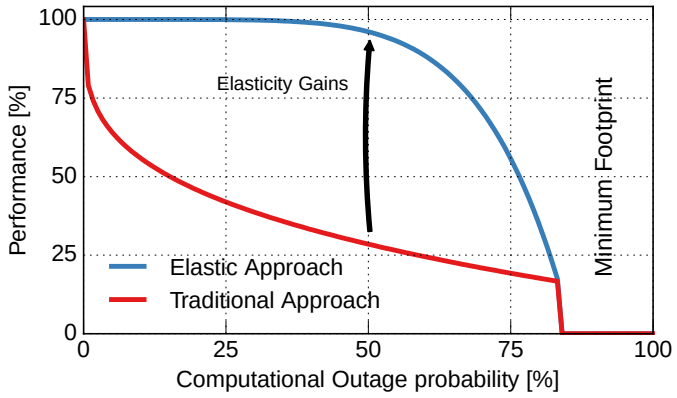


Figure 3-1: Telco-cloud-aware (elastic) VNF operation (illustrative)

To fully take advantage of elastic VNF, a detailed analysis of their operation is required: first, a thorough assessment of the resources consumed during execution, including statistics about temporal variations over time; second, a characterisation of the correlations between VNF operations, to serve as input for the orchestration algorithm, so it could e.g. dynamically assign resources to resilient VNFs and quickly "rescue" them when outages happen. Indeed, the quest for cloudification will end up with novel orchestration algorithms. Specific algorithms are defined in [5GM-D4.2], but the overall operation can be generalised as depicted in Figure 3-2.

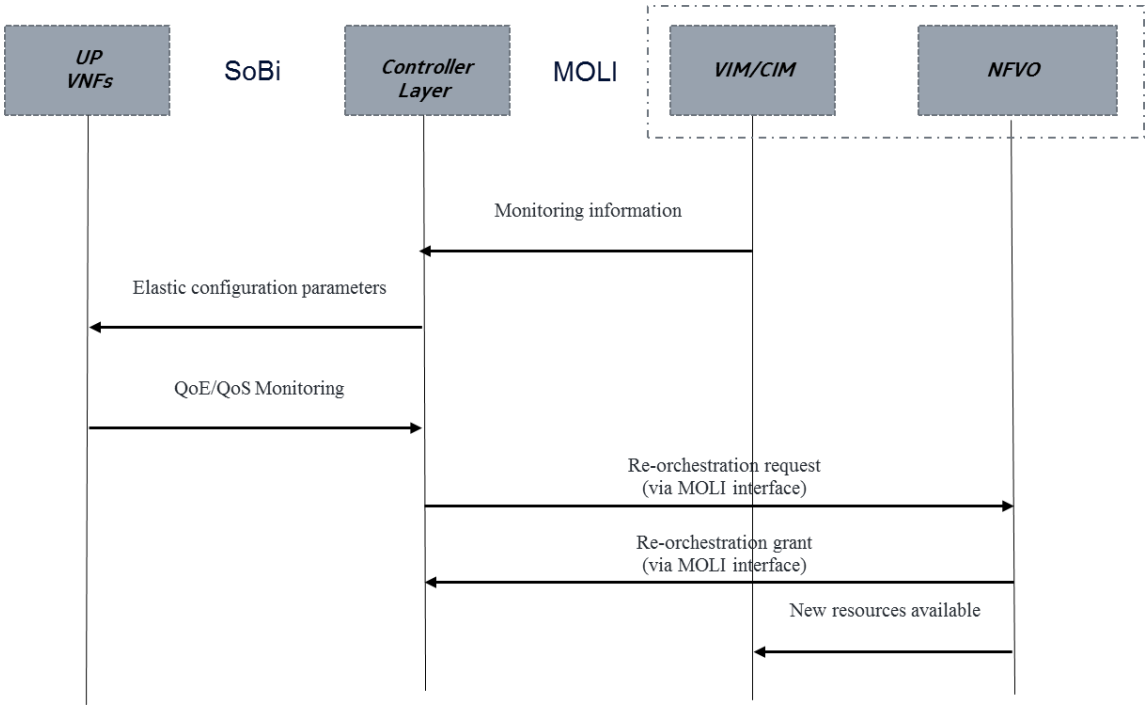


Figure 3-2: Telco-cloud-aware protocol stack operation for elasticity

3.1.2 Telco-cloud-aware interface design and requirements analysis

Concept

Besides the novel design of elastic VNFs, one of the most challenging tasks to introduce a cloud-enabled RAN protocol stack is to derive requirements regarding the interface among VNFs which include separated RAN functionalities such as the MAC scheduler, or PHY layer procedures. It is also for further study to what extent the RAN protocol stack can be cloudified. Especially in the MAC and PHY layer delay requirements and interdependencies among functions are critical.

A first study related to these targets and challenges is presented in [5GM-D4.1]. The basic architecture to introduce a cloud enabled protocol stack is illustrated in Figure 3-3. On the one hand, an extension to 3GPP's Xn and F1 interfaces to provide much higher flexibility is required as well as extensions regarding the transport protocol to interconnect multiple, e.g., containers among multiple physical machines is needed. The yellow boxes represent the virtualised environment of the either CP or UP functionality illustrated as blue boxes. One of the major upcoming tasks is to derive subgroups of functionalities dependent on the limitations defined by acceptable additional delay and interdependencies of functionalities which then will be virtualised.

Position in 5G-MoNArch architecture & Protocol Implications

Telco-cloud-aware interface design affects the 5G-MoNArch RAN functions and interfaces. It mainly analyses UP functions, but also potential interdependencies of CP sub-functions are covered. Moreover, the proposed concept affects the interfaces Xn and F1 as defined by 3GPP for Release 15. In the 5G-MoNArch context, both interfaces reside in the Network layer.

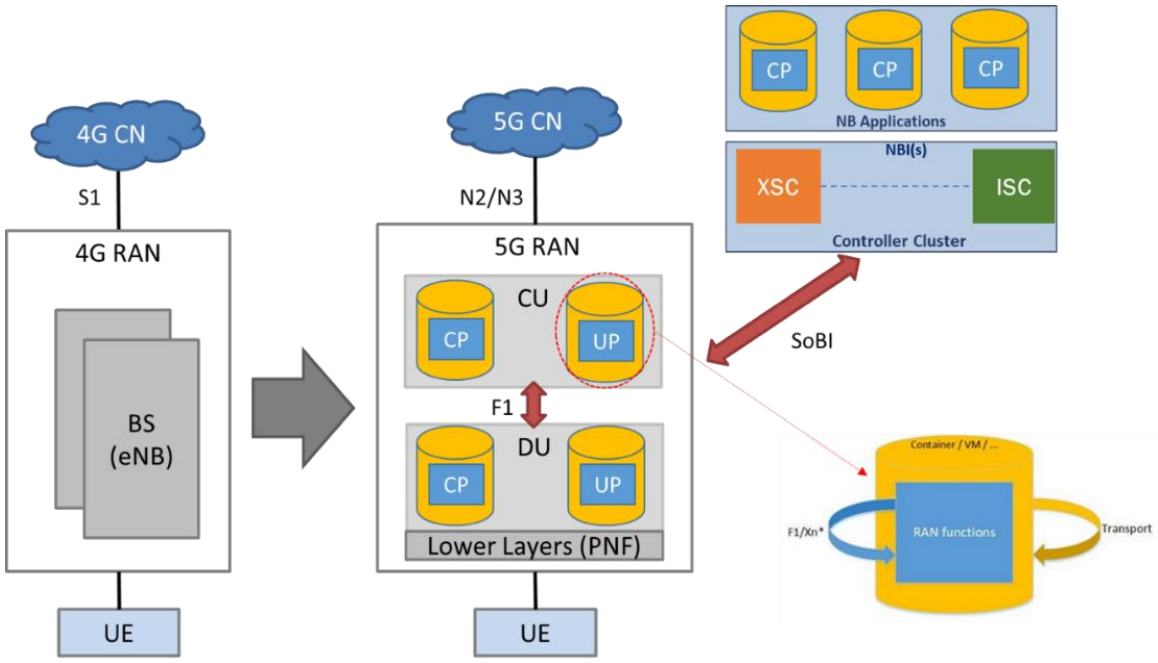


Figure 3-3: Cloud-enabled protocol stack architecture

Novel F1 information elements considering computational resources

The F1 interface interconnects CU and DUs within a gNB, as specified in [3GPP TS 38.401]. The corresponding F1 application protocol (F1AP) is specified in [3GPP TS 38.473].

As state-of-the-art techniques in 5G NR, centralised radio resource coordination or load balancing can be performed based on information exchange between gNBs, CUs and DUs. This can be applied by means of the X2 application protocol (X2AP), as specified in [3GPP TS 36.423] and the XnAP respectively, specified in [3GPP TS 38.420] and [3GPP TS 38.423]. This requires an RRM functionality/algorithm placed at a central entity which could be represented by either a master gNB or CU.

The first CU/DU split option that is specified by 3GPP introduce a split between PDCP and RLC layer, as described in Option 2 in Section 11.1.1 of [3GPP TR 38.801]. This means that PHY, MAC and RLC will be located in the DU while PDCP and SDAP plus RRC will be located in the CU, SDAP for the user plane protocol stack and RRC for the control plane protocol stack.

To support balancing or coordination under consideration of computational resource requirements between CUs and DUs for gNBs, e.g. in a telco environment comprising centralised and edge clouds, the F1 interface as well as the Xn-C interface need to carry additional information about computational resource usage, such as for example CPU, memory, and network interface utilisation.

An extension of the interfaces between DU and CU is essential for the consideration of computational resources in addition to traditionally considered radio resources as additional KPI. The utilisation of such status reporting regarding the computational resource utilisation of individual entities within the DU (e.g. PHY, MAC, etc.) can yield significant performance if combined with corresponding parameter adaptation algorithms within the CU.

An extension of additional information elements for both F1 and Xn-C interface regarding computational resource utilisation is described in the following.

Based on the specified structure in [3GPP TS 36.423] subclause 9.2.117, a possible solution to extend NG, F1 and Xn-C with additional information elements, regarding the consumption of computational resources might be defined, as Table 3-2 shows.

Table 3-2: Novel information elements for communication addressing computational resource utilisation

IE/Group Name	Presence	IE Type & Reference	Semantics Description
ID[6]	M		
CPU utilisation information	M	BIT STRING (4…)	Each bit string combination represents a threshold related to the CPU utilisation used by the gNB, CU/DU respectively. The length of the bit string is an integer defined based on the necessary number of thresholds defined. The CPU utilisation information is continuously/periodically reported.
Memory utilisation information	M	BIT STRING (4…)	Each bit string combination represents a threshold related to the memory utilisation used by the gNB, CU/DU respectively. The length of the bit string is an integer defined based on the necessary number of thresholds defined. The memory utilisation Information is continuously reported.
Network interface utilisation information	M	BIT STRING (4…)	Each bit string combination represents a threshold related to the network interface utilisation used by the gNB, CU/DU respectively. The length of the bit string is an integer defined based on the necessary number of thresholds defined. The network interface utilisation Information is continuously reported.
Storage utilisation information	M	BIT STRING (4…)	Each bit string combination represents a threshold related to the storage utilisation used by the gNB, CU/DU respectively. The length of the bit string is an integer defined based on the necessary number of thresholds defined. The storage utilisation Information is continuously reported.
Measurement Time Window	M	BIT STRING	The bit string encodes a time window that has been used for the measurement of the computational resource utilisation by the gNB, Cu/DU respectively. The values range from one TTI to N_max TTIs
Time Stamp	M	BIT STRING	The bit string encodes a time stamp that indicates the start of the time window that has been used for the measurement of the computational resource utilisation by the gNB, CU/DU respectively. The bit string could encode an absolute time value in a synchronised radio access network, or a TTI number according to the e TTI numbering scheme applied for 5G NR

[6] ID may be either the NR Cell Global Identifier (NCGI), gNB-DU ID or gNB Identifier (gNB ID) dependent on the unit to send/receive the novel information elements ([3GPP TS 38.300] subclause 8.1, 8.2, [3GPP TS 38.401] subclause 6.2.1, 6.2.2)

Estimated first moment of the processing time distribution	M	BIT STRING (4…)	Each bit string combination represents an estimate of the first statistical moment of the processing time used by the gNB, CU/DU respectively. The length of the bit string is an integer defined based on the necessary resolution. The moment estimation is continuously reported.
Estimated second moment of the processing time distribution	M	BIT STRING (4…)	Each bit string combination represents an estimate of the second statistical moment of the processing time used by the gNB, CU/DU respectively. The length of the bit string is an integer defined based on the necessary resolution. The moment estimation is continuously reported.

The novel information elements introduced in Table 3-2 could additionally be used for the communication with a data base as a virtual element for the computational resource utilisation of individual functions within a gNB or eNB. This computational elasticity data base (CEDB) and the corresponding interfaces to CU within a gNB are shown in Figure 3-4. The interface (CE-C) will facilitate the support of advanced control algorithms for computational elasticity by means of having a common data base for all controlled entities (functions within CU and DU) and entities control the computational resource utilisation of these controlled entities. Such a controller could for example be located in a CU that controls the computational resource utilisation of multiple DUs within the same gNB. It would furthermore be possible to control the computational resource utilisation of multiple CUs and/DUs in different gNB by a single computational resource master controller within the CU of a master gNB. This could also be relevant in case of an edge cloud deployment with multiple gNBs without having a master gNB. The CEDB can furthermore be extended with additional interfaces that provide access to it for processes running outside the gNB or eNB hosting the CEDB.

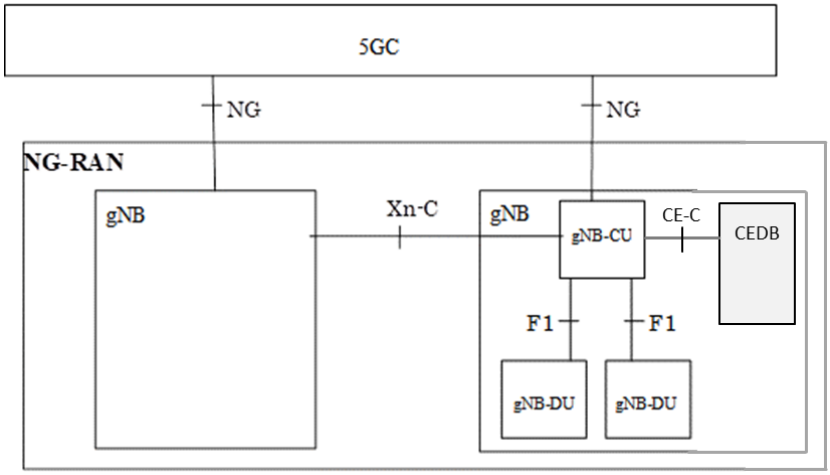


Figure 3-4: Overall NG-RAN architecture with CEDB extension

Entries in the CEDB could for example comprise:

- gNB identifier (gNB-ID)
- Location of that function within the NG-RAN (FN-LOC) (e.g. ,CU or DU)
- Function identifier (FN-ID) (e.g. downlink resource allocation)
- Time stamp
- Time window

- Current computational resource utilisation of that function (FN-CRU) (e.g., CPU and memory)

The first three entries represent a unique identification of a specific CU or DU within the radio access network. The time stamp indicates for each entry when it has been updated last time, and the time window indicates the duration over which the computational resource utilisation was measured. The third part indicates the computational resource utilisation for the specific function with the indicated time window.

Latency requirements analysis

An important factor regarding the CU/DU split is the latency requirement for the interface between these units. Specific requirements have already been discussed at 3GPP within the scope of a study item [3GPP TR 38.801]. Together with contributions [3GPP R3-161813] and [3GPP R3-161784], the report provides estimations for the allowable latency between different functional split options.

The analysis in this section provides a more detailed evaluation on how the latency between suggested split options affects the UE throughput and the stability of an exemplary virtualised eNB implementation. Although, a virtualised eNB instead of a gNB is used for the experimentation due to the availability of open-source protocol stacks, the findings are relevant for the gNB as well, thanks to the similar protocol stack structure. Specifically, interface latency values that may lead to a system crash or perceptible degradation in the throughput performance depending on different CU and DU split options. Figure 3-5 shows the investigated splits options for CU and DU functions based on the definition in [3GPP TR 38.801].

The processing time evaluation conducted during the performance study is based on the moment estimation related information elements described in Table 3-2.

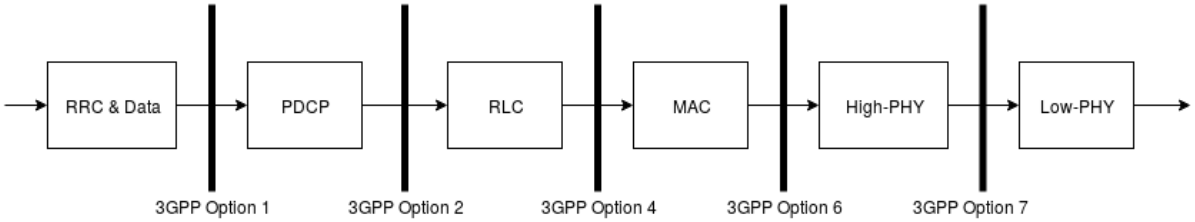


Figure 3-5: Functional split options within the downlink transmission chain

The corresponding functions have been identified in the testbed based on srsLTE [SRSLTE]. In order to emulate the different functional splits, additional latency has been introduced before or after every function call in order to model the corresponding interface latency. The information in [3GPP TR 38.801] provides the maximum allowed latency for different functional splits as summarised in Table 3-3.

Table 3-3: Mapping Split options to srsENB functions

3GPP split option	**srsENB function delay location**	**Maximum allowed latency according to 3GPP TR 38.801**
1	before downlink PDCP function	10 ms
2	after downlink PDCP function	1.5~10 ms
4	before downlink MAC scheduler	approx. 100 us
6	before downlink PHY encoding	250 us
7	after downlink PHY encoding	250 us

Table 3-4 describes the system parameters which were used for this evaluation. Here it has to be considered that these assumptions differ from the ones specified in [3GPP TR 38.801]. The latter consist of 256 QAM, 8 MIMO layers and 32 antenna ports, which are currently not supported by the testbed.

Table 3-4: System parameters for latency requirements evaluation

Parameter	Value
Central frequency	2660 MHz
System bandwidth	5 and 10 MHz
Modulation	64QAM
Number of antenna ports	1
Number of PHY threads	1
Traffic generator	1200 and 2500 bytes for 5 and 10 MHz bandwidth, respectively, with 1 ms mean time between two sent packets (exponential distribution)
Every measurement duration	10 seconds
eNB host machine CPU	Intel Core i7-6700K with 4 GHz x 4

The flowchart in Figure 3-6 shows the automated measurement procedure. The configuration for a test-run of the system comprises a specific latency for the modelled CU/DU split interface, a packet size for the traffic generator. Every test iteration runs for a predefined number of seconds.

The srsENB, srsUE and the srsEPC processes are running on different hosts connected via a 1 Gb/s Ethernet interface. After each system test run, the latency is increased until it either reaches a configured maximum latency or the eNB stops working. The latter is most critical for the lower layer split options when the eNB is unable to process subframes (transmission time intervals) in one millisecond.

The evaluation shows that for the higher layer split options (3GPP Options 1 and Option 2), the increased latency does not lead to a crash of the eNB, even for high values of the additional latency (up to 100 ms). However, the throughput is exponentially decreasing as can be observed in Figure 3-7.

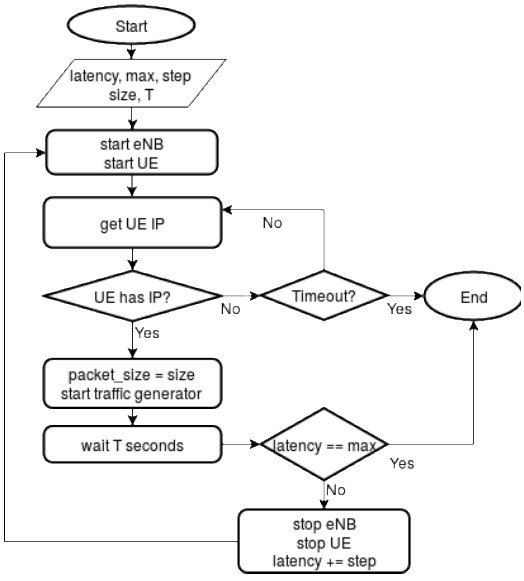


Figure 3-6: Measurement procedure for the latency evaluation for functional splits

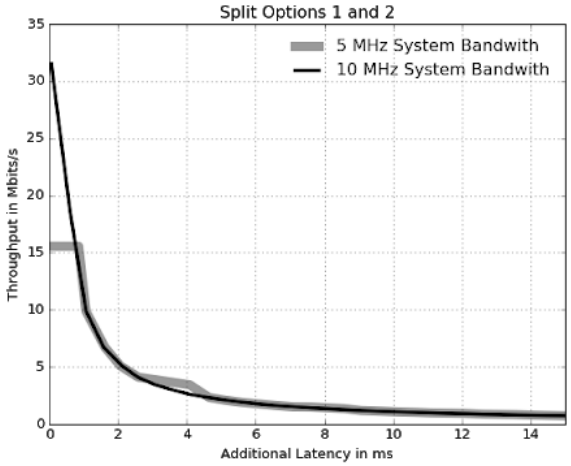


Figure 3-7: Throughput depending on the additional latency for 3GPP split Option 1 and Option 2

Additional latency for Option 4, Option 6 and Option 7 is not affecting the throughput. However, from certain points on it yields the eNB to crash due to late samples transmitted to the remote radio head (RRH). Table 3-5 shows the maximum allowed additional latency, for the corresponding split options for a system running with 5 MHz bandwidth and the maximum modulation and coding scheme, as observed during test runs of the testbed.

Table 3-5: Maximum latency values for Option 4, Option 6 and Option 7

3GPP split option	Maximum latency
4	806 us
6	820 us
7	830 us

The results show that for Option 4, Option 6 and Option 7 the critical latency is around 800 us. Figure 3-8 shows the estimated probability density function (Gaussian) based on the mean and the standard deviation estimations of the downlink encoding processing time for MCS 20 and 28. The figure shows furthermore the defined threshold at 800 us. The histogram shows the actual observed frequencies of processing time intervals based on samples collected every subframe (transmission time interval). The results show that the assumption of a Gaussian distribution provides sufficiently accurate estimations for the upper tails of the processing time distributions.

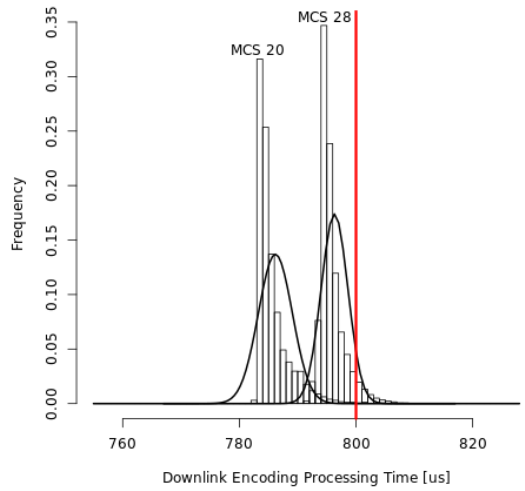


Figure 3-8: Probability density function for 5 MHz bandwidth with 755 us additional latency

Based on the complementary cumulative distribution function the probability of exceeding a configured threshold (e.g. 800 ns) is calculated. Figure 3-9 shows the probability for different values of additional interface latency introduced for the splits according to 3GPP Option 6 and Option 7.

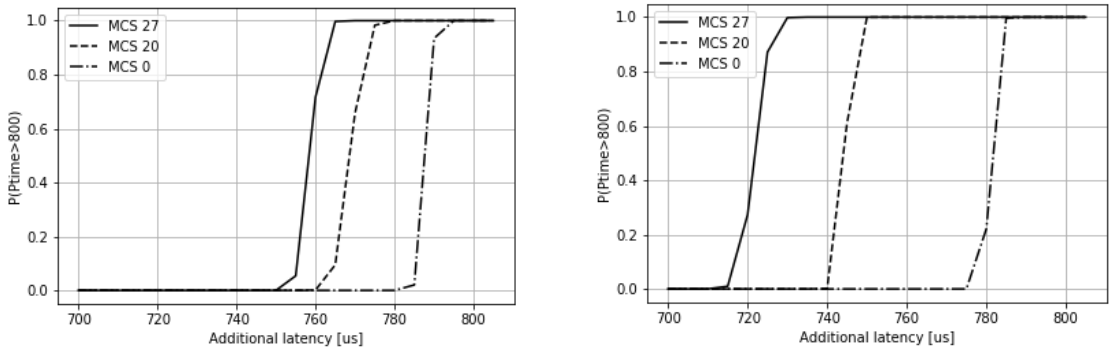


Figure 3-9: Probability of exceeding the 800 us threshold for different values of additional latency for 5 MHz (left), 10 MHz (right) system bandwidth

The processing time values depend on hardware configuration of the eNB and might differ for different CPUs and availability of computational resources.

The evaluation in this section shows that additional interface latency for lower layer splits based on 3GPP Option 4, Option 6 or Option 7 might have a more critical impact on the stability of the eNB when it is unable to fulfil the transmission time interval requirement of one millisecond. The interface latency for higher layer splits corresponding to 3GPP Option 1 and Option 2 yields graceful performance degradations.

3.1.3 Terminal-aware protocol design

Conventionally, a user terminal or User Equipment (UE), as defined in 3GPP, plays important role in measuring and reporting channel state information for access and mobility management. Gradually, a user terminal has transitioned into more prominent roles in line with new developments in standards, e.g. on Device to Device (D2D) Communication**.** D2D communications facilitates an enabling innovation for further support of service continuity and smooth mobility (beyond the network edge). This can be realised via Group Mobility, an area currently under study in standards (mainly for wearable devices) where a Relay UE acts as the surrogate of handover signalling messages for Remote UEs when they move along together.

In practice, the linkage between a group of Remote UEs and a Relay UE may not be exclusive or permanent due to, e.g. non-uniform mobility patterns followed by them. Therefore, different mobility scenarios can be envisioned beyond those followed in current standard discussions. For example, in case of stationary Internet of things (IoT) devices, Remote UEs do not necessarily move along a Relay UE through which they communicate. Hence, a new mechanism to efficiently handle group mobility in such scenarios is required.

Concept

Flexible Group Mobility via floating mobility anchors

A novel group mobility paradigm is proposed with floating mobility anchor as shown in Figure 3-10, not necessarily "pinned" to a single Relay UE. Instead, the anchor and corresponding mobility group can be dynamically changed based on the mobility patterns, relative channel quality fluctuations and the level of support needed.

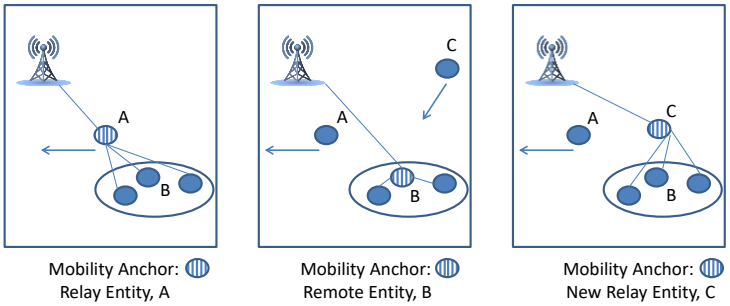


Figure 3-10: Concept of proposed floating mobility anchor

The above concept can enhance telco grade support in a group of Remote UEs (e.g. from scalability perspective) in line with Gap #4 to facilitate offloading some signalling at the RAN level (from direct signalling to the gNB to indirect signalling between anchor and Remote UEs) (cf. Appendix A).

Position in 5G-MoNArch architecture & Protocol Implications

Mobility and handover can be seen as a cross-slice NF (as part of AMF) within 5G-MoNArch architecture which should be commonly supported for all the sessions of a specific UE. The above solution can be confined to RAN domain where the mobility management is handled by a gNB. However, the gNB needs to be aware of anchor assignment and associated Remote UEs per group. Furthermore, coordination is needed if an anchor and / or part of the Remote UEs leave the group for example due to changes in neighbourhood / mobility pattern. Enhancements could be envisioned on carried information over F1 interface (in CU/DU functional split) to update information as above between gNB-CU and anchor (e.g., over an enhanced F1 termed as F1*) in case of group status change or mobility at RAN-level.

Evaluation and analyses

The detailed description of Flexible Group Mobility is as follows:

- Group anchor assignment / Group formation: a group is defined as a set of UEs or other entities (in either Remote or Relay modes) that can be bundled together based on proximity, good inter-entity channel quality, similarity and correlation in supporting cell, mobility pattern, service profile or slice-driven characteristics. A group anchor is defined per group considering multiple criteria (good intra-group connectivity to maximum number of group members, good link connectivity towards mobility management function/ entity, sufficient power limit or processing capability). In presence of a Relay entity within a group, the Relay can be the natural group anchor as it will satisfy the relevant criteria. A UE or any other entity may participate in multiple groups assuming multiple-service/ slice profiles. As a result, a Remote entity in one group can be a Relay entity in another group. The exclusivity or generality of the groups is subject to network operator's decision. A group anchor aggregates individual group member signalling messages related to mobility management (anchor change or handover) towards/ from a gNB or relevant gNB-CU based on the level of support needed.
- User Update: should a group member leave the group, the gNB-CU and the group anchor send an updated list of associated Remote UEs to each other to coordinate on the changes.
- Anchor change: If the current group anchor leaves the group (e.g. due to changes in mobility pattern, channel quality or service/slice profile), a reassignment procedure is triggered so a new group anchor is designated by the network (gNB-CU) and the old anchor is released. Therefore, the anchor role can dynamically float across candidate members.
- Group handover: If a handover procedure is triggered via group anchor (e.g. due to the changes of channel quality to the gNB), the gNB or relevant gNB-CU decides on remote UEs to be shifted to another anchor (via anchor change procedure) or alternatively Remote UEs to be handed over (along the old anchor) to another gNB. The anchor change or handover of remote UEs should precede any old anchor handover. Afterwards, the old anchor handover can be followed.

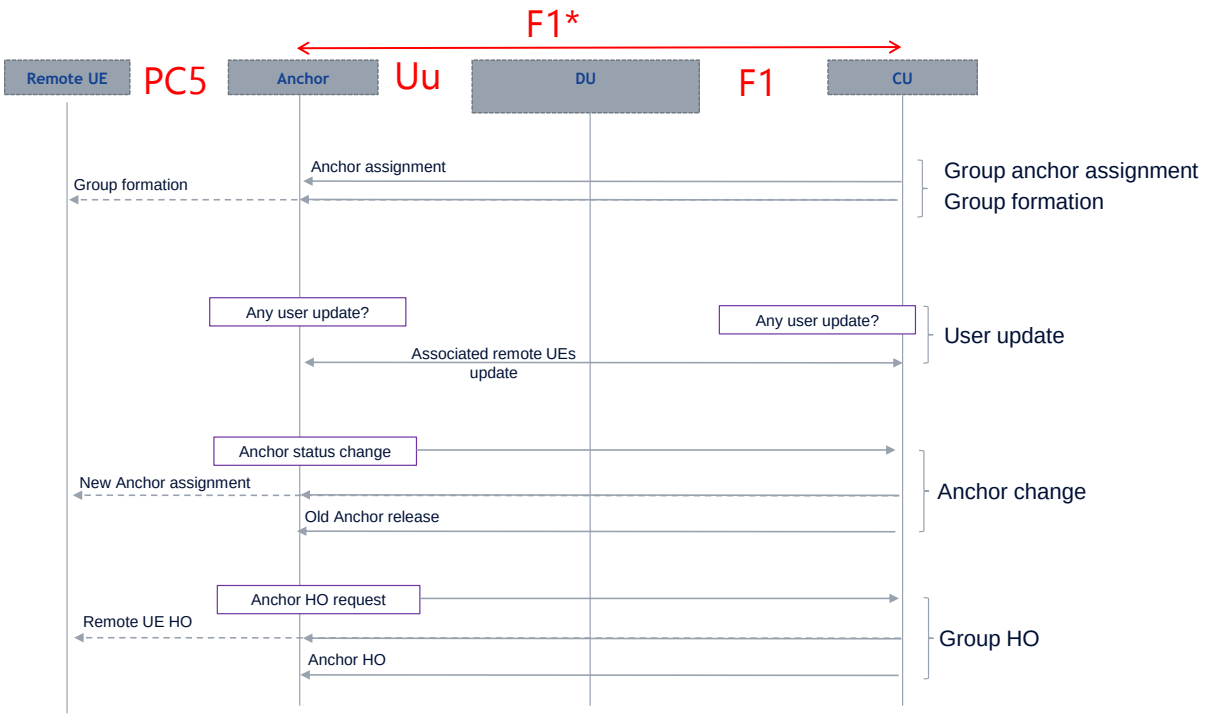


Figure 3-11: Message sequence chart of the anchor change / group handover concept

Figure 3-11 shows the Message Sequence Chart for the proposed concept for different stages, in particular for anchor change and group handover as described above.

3.2 *Inter-slice context-aware optimisation*

This section describes the three solutions for inter-slice context-aware optimisation. The first solution focuses on the inter-slice context sharing and optimisation. The key aspect of this solution is to enable the cross-slice optimisation based on different parts of the system such as Network layer and M&O layer to allow a coordination among distinct decision-making loops in the system. The second solution describes the inter-slice coordination focusing on inter-slice optimisation within the same part of 5GS, in this case in Control Plane of the network layer. This solution is associated with the components described in Section 2.2.2: **ISCF** (new NF proposed in this solution), and enhancements on PCF, AMF, and SMF. The third solution is related to slice selection for UEs based on analytics information. This solution is associated with the components described in Section 2.2.2: Enhancements on AMF, NSSF, and NWDAF.

3.2.1 Inter-slice context sharing and optimisation

Concept

5G Systems Phase 1 (i.e., Release 15) defines NWDA function in the network layer to perform per slice data analytics for service assurance. While cross slice context sharing and E2E cross-slice optimisation is not fully supported. This work focuses on defining entities in a mobile network that can operate with context awareness to close the loop of decision-making between entities of the Network layer and M&O layer in order to optimise intra and inter network slice operations. The work covers the following aspects:

- Enable information about status of entities of the system kept in the M&O layer to be considered in the decision making of control plane (CP) functions in the Network layer. This concept has been successfully delivered to 3GPP standard, as listed in Table 2-1. We introduced during the eNA Study Item and most recently in the normative phase for 5G Networks for Release 16 in 3GPP, the capability of NWDAF, a CP function, to collect data from M&O Layer as defined in Clause 6.2.3 in TS 23.288 (V0.3.0) [3GPP TS 23.288].
- Reduce the chances of unnecessary or multiple changes at M&O layer and CP for solving a situation involving related entities, or entities in the same geographical region. For instance, if PCF, SMF, O&M consume the same data analytics about prediction of probably reduction on throughput in a certain area of the network slice, how to prevent that SMF triggers UPF relocation, PCF triggers changes in policies to reduce traffic, and M&O layer triggers auto scaling of NFs, when these actions are affecting the same area of the network slice. This coordination problem has been acknowledged in 3GGP study item on Enhanced Network Automation (eNA SID Rel. 16) for Release 16 5G Specification as defined in TR 23.791 V16.0.0 [3GPP TR 23.791]. This is a direct result of our contribution associated with Key Issue #4 (Key Issue 4: Interactions with OAM for Data Collection and Data Analytics Exposure). This contribution is listed in Table 2-1.
- Potential to reduce the need for long term capacity planning and pre-provisioning of infrastructure resources in order to guarantee the expected performance of mobile network services. In 4G, the QoS Class Identifier (QCI) has a budget of delay that is expected to be provisioned at the infrastructure by the M&O layer. This means that today it is first necessary to understand the characteristics of the mobile traffic and then dimension and pre-provision the network up front for such demand. This can lead to over- or under-provisioning of the network and this will impact the service performance. The presented enhancements tackle this problem by using context awareness to assure E2E QoS and at the same time improve the usage of the mobile network from the point of view of the operators.

Position in 5G-MoNArch architecture & Protocol Implications

5G-MoNArch reference architecture is updated to enhance the Network layer and M&O layer, and provide enablers for inter-slicing optimisation (e.g., inter-slice context sharing and optimisation by enhanced NWDAF).

5G-MoNArch reference architecture is updated to enhance the coordination of slice and cross slice optimisations and M&O layer optimisations and provides enablers for coordination of Network layer

and M&O layer optimisation (e.g., cross-slice and M&O layer context sharing, and optimisation by enhanced NWDAF).

Evaluation and analyses

Main analysis required towards the definition of mechanism for 5G Phase 2 (i.e., further 3GPP release introducing significant changes beyond Release 15) for inter-slice context sharing and optimisation are:

- Analysis of required inter-slice interfaces and procedure enhancements (i.e. Phase 1 gap analysis), relating to CP, UP, and M&O layer, NWDAF function enhancements.
- Inter-slice interfaces and procedure solutions design.
- Analysis of required context information, information source, information format, to be collected and used for optimisation.
- Analysis of the potential applications and optimisation for NWDAF.
- Scenario and requirements analysis for inter-slice coordination of optimisations as well as coordination of optimisations across Network layer and M&O layer.
- Architecture analysis on the conflicts of parallel actions performed by network layer and M&O layer.
- Analysis on the related interface/procedure enhancement mapping to the current standards and 5G-MoNArch high level architecture.

Figure 3-12 illustrates examples of enhancements to be included into NWDAF functionalities to support the mechanisms listed above. The enhancements of NWDAF proposed to tackle the issues of coordination of slice/cross-slice optimisation and M&O layer optimisations is illustrated in Figure 3-13.

In this proposal, NWDAF can generate feedback and coordination notifications. These coordination notifications are messages sent by NWDAF in order to inform, NFs and OAM of situations in which a generated feedback consumed by a certain entity might generate effects on another entity. The entity suffering the effects becomes aware that it might not trigger changes, before just triggering actions, the entity suffering the effects might trigger some back off time to avoid unnecessary changes in the system. In addition to NWDAF, the NFs and M&O layer functions are enhanced in order to support such coordination triggered by the NWDAF.

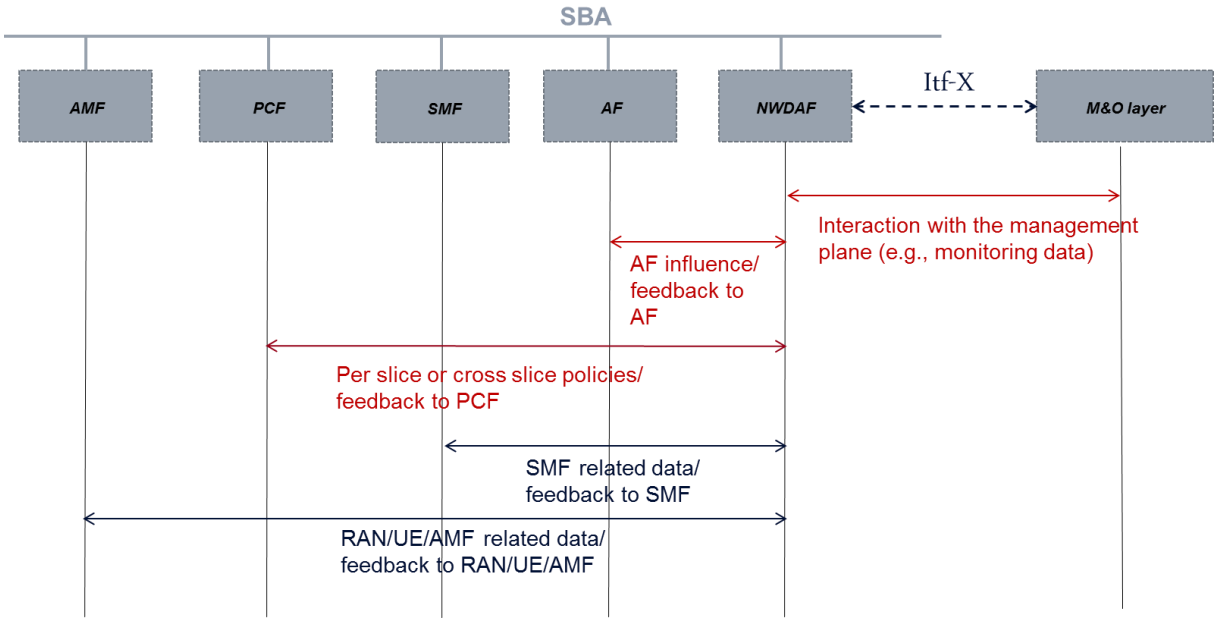


Figure 3-12: 5G-MoNArch enhancements of NWDAF (red-coloured parts)

The proposal defines a minimal set of types of contexts that can be generated by the NWDAF. Furthermore, defined is a format for describing a context type as a tuple (<Entity#1> - <Entity#2>), where the first entity indicates who enforces a certain change in the operation of the system, and the

second entity indicates which entity might be influenced by consequences of the enforced changes. For instance, if CP changes the gateway from the users, M&O layer will see a reduction of traffic in a part of the network and an increase in another part. This is represented as a CP-M&O layer type of context.

The minimal set of types of contexts are described as follows:

- **CP only:** related to changes to be enforced in CP with minimal or no effect in M&O layer; trigger no notification
- **CP-M&O layer:** related to changes in CP that probably will affect M&O layer; trigger notification message to M&O layer functions
- **M&O layer-CP:** related to changes in M&O layer that will probably affect CP; triggers notification message to CP functions

Figure 3-13 shows the interactions among NWDAF, NFs, and M&O layer for coordination of actions due to the generation of the proposed types of contexts.

- Step 0: NFs and OAM register to receive feedback from NWDAF
- Step 1: NWDAF collects data from NFs and/or M&O layer in order to generate feedback
- Step 3: NWDAF generates feedbacks and determine the type of context associated with each feedback

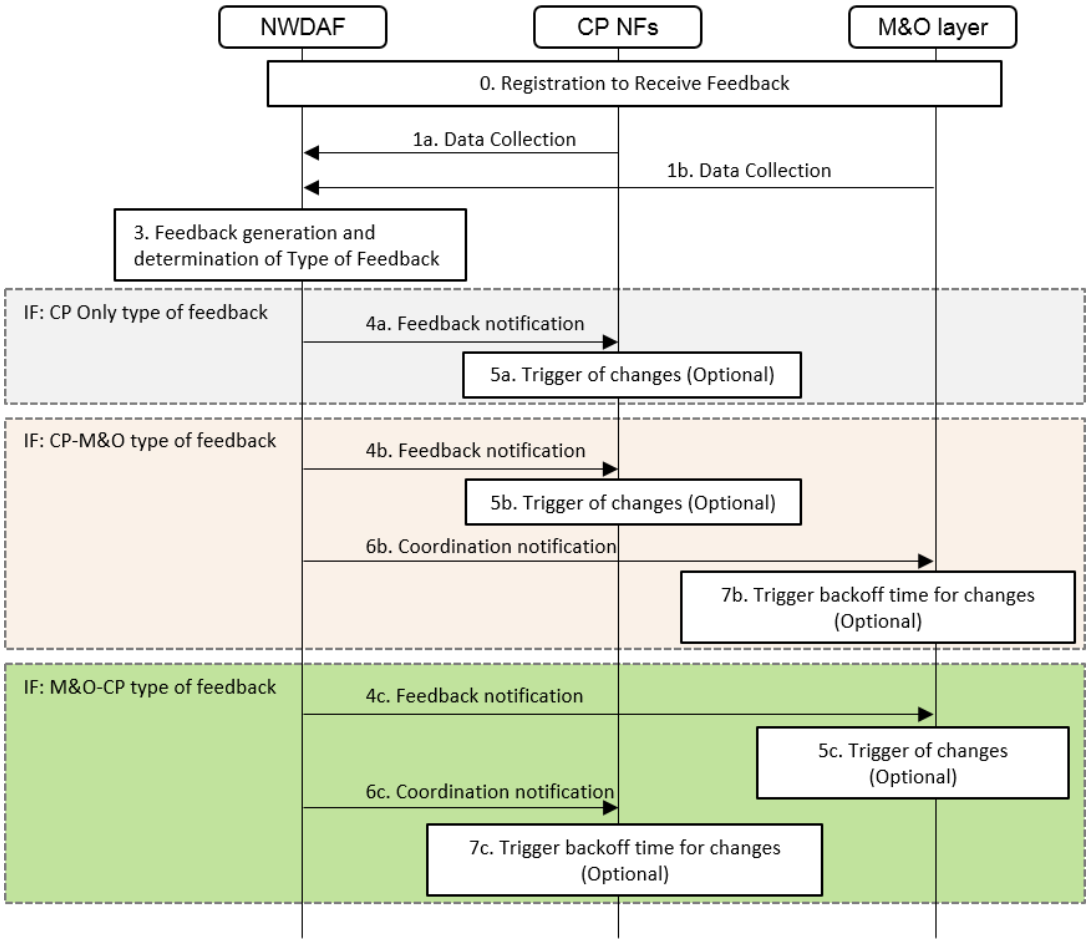


Figure 3-13: NWDAF enhancements for coordination of feedback usage between network layer and M&O layer

- If CP Only type of feedback:
  - Step 4a: NWDAF notifies NF with the feedback it registered to receive
  - Step 5a: NF that received a feedback will decide if changes need to be triggered based on the received feedback

- If CP-M&O layer type of feedback:
  - Step 4b: NWDAF notifies NF with the feedback it registered to receive
  - Step 5b: NF that received a feedback will decide if changes need to be triggered based on the received feedback
  - Step 6b: NWDAF will generate a coordination notification message to M&O layer about the generated feedback that might affect M&O layer operation
  - Step 7b: M&O layer upon receiving the coordination notification message decides whether a back-off timer for changes should be triggered (to avoid conflicting or unnecessary changes) or not.
- If M&O layer-CP type of feedback:
  - Step 4c: NWDAF notifies M&O layer with the feedback it registered to receive
  - Step 5c: M&O layer upon receiving a feedback will decide if changes need to be triggered based on the received feedback
  - Step 6c: NWDAF will generate a coordination notification message to NFs about the generated feedback that might affect their operation
  - Step 7c: NFs upon receiving the coordination notification message decide whether a back-off timer for changes should be triggered (to avoid conflicting or unnecessary changes) or not.

The specific interfaces between NWDAF, NFs, and M&O layer, are currently under discussion at the SA2 study item on enhanced Network Automation [3GPP TR 23.786]. The discussion includes the type of services offered and consumed by NWDAF, NFs, and OAM to enable both the data collection as well as the consumption of feedback generated by NWDAF. In addition, the actual parameters to be considered by these interfaces is included in this discussion.

Framework for Evaluation of Solutions

In order to evaluate the advantages of the enhancements proposed for NWDAF for inter-slice context sharing and optimisation, we propose an initial description of a framework for the comparison of 3GPP Release 16 capabilities with the capabilities of the proposed solutions.

The proposed framework is illustrated in Table 3-6 below.

Table 3-6: Initial Framework for evaluation of solutions for inter-slice context sharing and optimisation

	Gain			Cost		
	System	UE Level	Application	System	UE Level	Application
Parameters	Signalling Ratio Resource Consumption	QoE	QoE QoS	Signalling Ratio Resource Consumption	QoE	QoE QoS

The comparison between 3GPP Release 16 versus the proposed solution is based on the identification of the trade-off between the gains versus the costs for a given feature. For instance, a gain on the reduction of the signalling for avoiding simultaneous and unnecessary changes, needs to be put in relationship to the signalling cost to collect the data, and coordinate the entities of the system.

The direct comparison can be performed on use case basis. An example of use case is depicted in Figure 3-14.

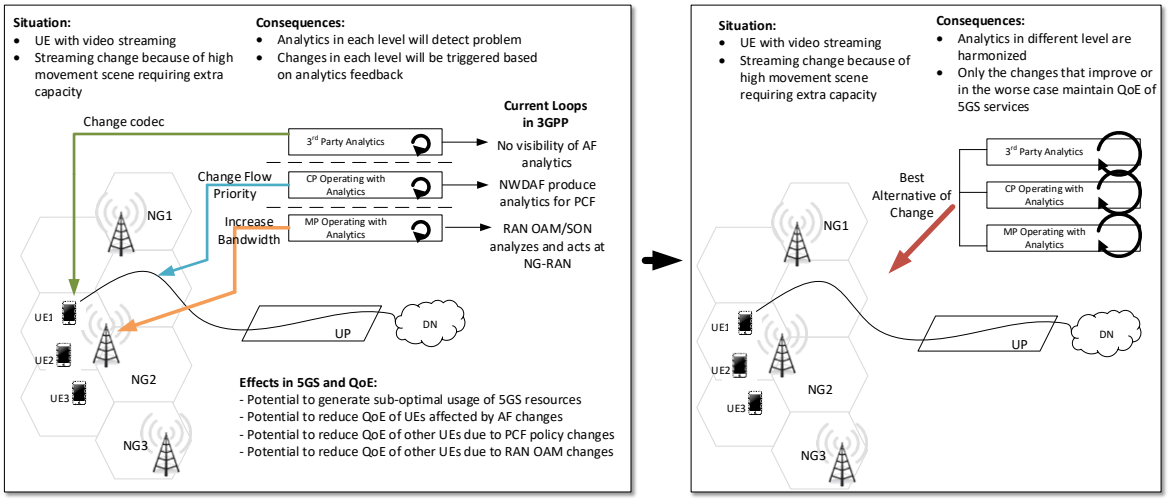


Figure 3-14: Use case for comparison of coordination of feedback usage between network layer and M&O layer

3.2.2 Inter-slice coordination

The previous section discusses mainly cross slice optimisation via the interaction between different layers or parts of the network. This section focuses on the inter-slice coordination at the same layer or part of the network.

Concept

3GPP defines some basic network slice types, e.g., eMBB, mMTC, and URLLC, where each network slice type is designed for a group of services sharing similar service requirements. However, some applications/services may require multiple service flows. Such multiple service flows can be implemented by different QoS flows, different PDU sessions or even different network slices. For instance, in remote driving case (cf. Figure 3-15), the High Definition (HD) video requires high throughput which is supported by eMBB slice. While, the on-vehicle sensor data and vehicle control signalling require low latency and high reliability which is supported by URLLC slice. Similarly, in Touristic City scenario the VR/AR application may require an eMBB slice to transfer HD video contents from the video server, meanwhile a URLLC slice maybe needed to exchange the haptic interaction between the tourist and the guide [5GM-D6.1].

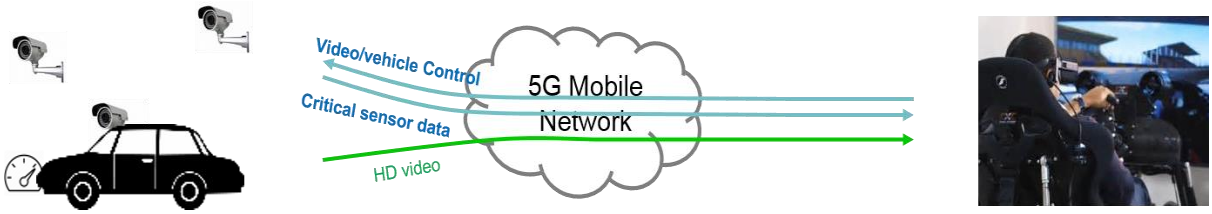


Figure 3-15: Remote driving use case

For services using multiple slices, different slices can be fully isolated, and their performance are independent to each other. However, the actual performance of individual independent slice will be affecting the same service. This results in the correlation between requirements of different network slices for the same service/applications, more specifically:

- Performance of one slice affects the required KPIs of another slice. For instance, in remote driving/AR/VR, long latency of video control (direction of view) signal makes the transmission of HD streamed video from the vehicle site useless.
- The KPI (e.g., Latency) budget is actually shared between different slices. The user experience is affected by the summary of the latency from multiple slices (i.e., the latency of the control signalling to the vehicle and the latency of the video/sensor report from the vehicle).

Obviously, exploring the service correlation of different service flows can help to increase network efficiency and improve user experience.

Position in 5G-MoNArch architecture & protocol implications

This work proposes a CP NF ISCF in 5GC to bind the services flows from the same application/service. There are two options to implement this function, see Figure 3-16: 1. At the AMF 2. At the PCF. In option 1, ISCF binds the service flow of the applications by intercepting the session establishment requests from the UE. In option 2, ISCF gets the service flow binding information from the AF, i.e., via interactions with verticals/applications.

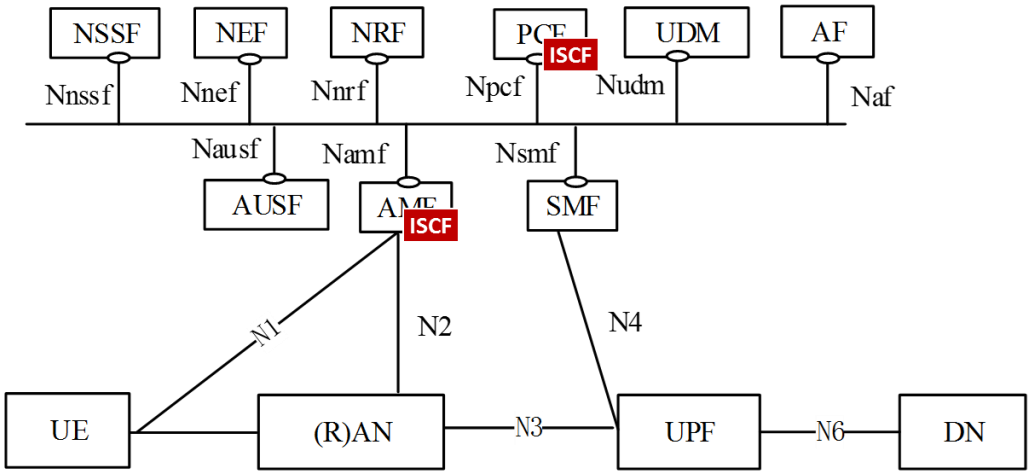


Figure 3-16: Two options of ISCF implementation in SBA

Evaluation and analyses

Except for the binding of services flows, such binding information should be distributed to the NF, e.g., in RAN, or M&O layer where such information is used for network optimisation. When ISCF gets the service flow binding information, it can provide it via SBI to other NFs for optimisation purpose (e.g., SMF and PCF), or perform data analytics (e.g., NWDAF), or RAN via N2 interface, or further to the management layer (e.g., MDAF).

Meanwhile, network optimisation needs to be based on the correlated KPIs of the bound services flows. This information can come from the CSMF in the management layer and stored at PCF as correlated QoS profile of different traffic flows. This information can also come from the AF via influencing the QoS profile at PCF.

The following Figure 3-17 shows an example procedure in case the ISCF is located at AMF:

(1) Network Slice Selection Policy (NSSP) at the UE maps one service/application into multiple network slices (i.e., with different S-NSSAIs). Triggered by an application, UE will send the session establishment request with multiple PDU sessions each indicate the PDU session ID and corresponding S-NSSAI to the AMF via NAS message.

(2) AMF marks down related PDU session IDs and S-NSSAIs for this request, and also the related SMFs. It sends the binding information of the PDU sessions to the related SMFs.

(3) The related SMFs decides on the QoS flows of these bound PDU sessions and further bind the QoS flows.

(4) SMFs indicate the PDU session/QoS flow binding to PCF (or other NFs where such information is needed, or to other NFs to RAN, UE, management layer, etc.)

(5) PCF decides on per QoS flow/per PDU session policy and sends this decision to the related NFs.

As indicated above, such correlated QoS information can be exploited by RAN to fulfil the target service requirements as well as to increase the resource utilisation efficiency. In the following, two analyses are provided where the correlation information is utilised by RAN.

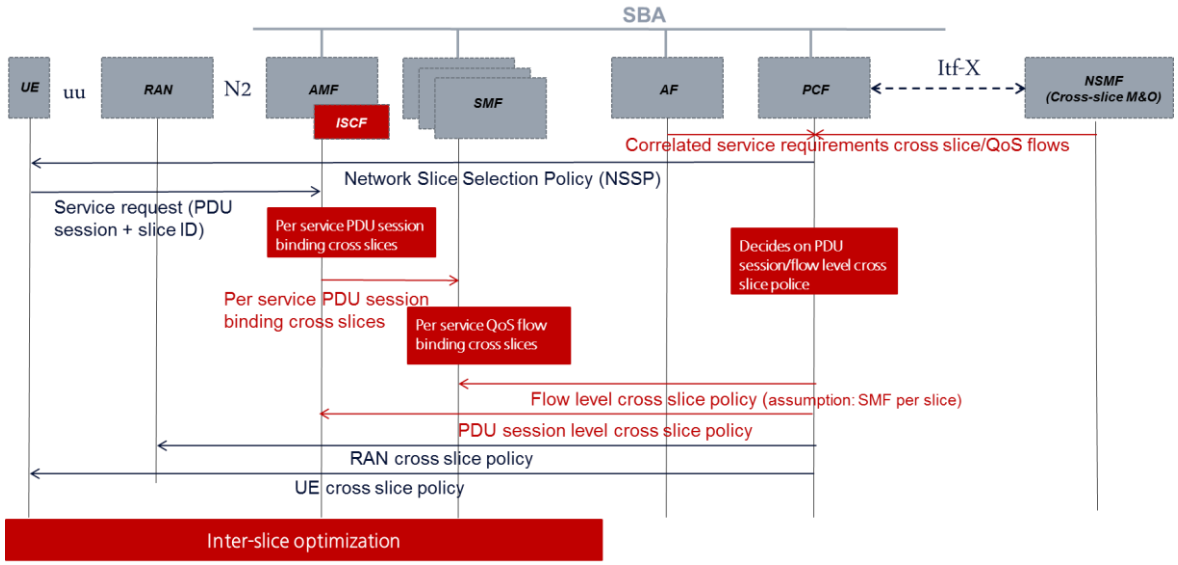


Figure 3-17: Example message sequence chart for inter-slice service correlation (5G-MoNArch enhancements are indicated in red colour)

Joint delay budget handling in case of same link (UL)

As shown in Figure 3-18(a), e.g., in remote driving aka tele-operator driving (ToD) application, video and haptic information from the vehicle need to be synchronised on UL. In particular, scheduling period the URLLC is mini slot (2, 4 or 7 OFDM symbols) while that of eMBB is 1ms TTI. In case of correlated QoS configuration awareness at the scheduler, the latency constraint of haptic information (which is an URLLC service) can be relaxed to be synchronised with eMBB scheduling intervals. That is, the haptic information needs to be jointly processed with the respective eMBB service information and the haptic information can be transmitted within the delay budget of the eMBB packet. With relaxed latency constraints known at the scheduler level, appropriate scheduling interval to meet the joint delay budget can be chosen. This results in higher resource utilisation efficiency. Namely, if the correlation information is not available, the scheduler tries to allocate resources to the URLLC data as soon as possible, although this is not needed due to the service characteristics.

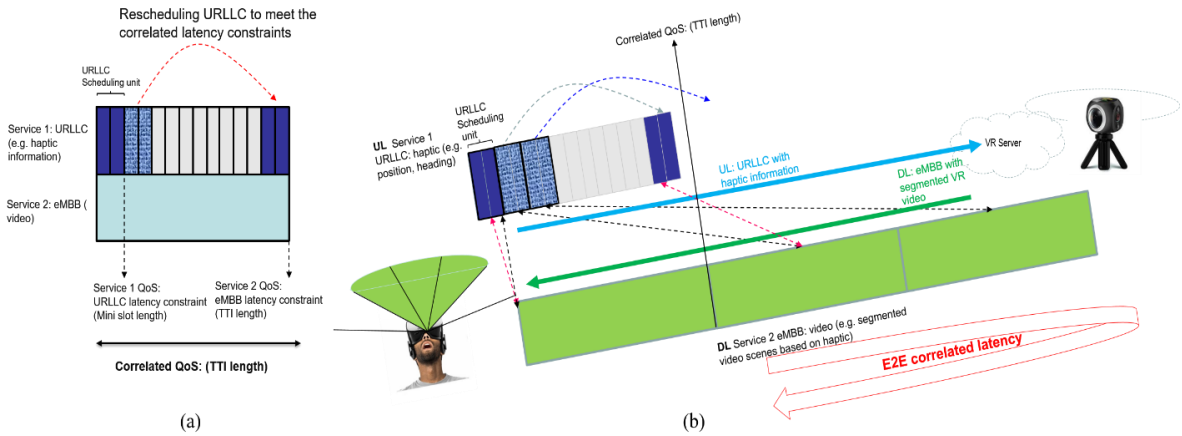


Figure 3-18: Utilisation of correlated QoS information in RAN; (a) example scheduling for correlated QoS treatment on UL for the case of remote driving aka ToD and (b) example scheduling for correlated QoS treatment on UL and DL jointly for the case of AR/VR application

E2E joint delay budget handling in case of different links (DL and UL)
As shown in Figure 3-18(a), e.g. in AR/VR application, QoS flows of DL and UL are correlated to meet the QoS. In particular, the scheduler with the knowledge of DL/UL correlation, allocates DL traffic (i.e., segmented video) according to the UL information (i.e., direction of view) such that it meets the E2E latency constraints.

3.2.3 Terminal analytics driven slice selection and control

There are already established services within next generation service-based CNs (as outlined in 3GPP SA) to access operator-specific analytics (NWDAF). The current defined services are mainly envisioned to share such information within CN between different NFs of the same slice.

UEs are natural data collection points to gather more localised analytics within the network. Examples of data that the UE can provide are positioning information (e.g. collected from inertial sensors of the UE, geo-referenced radio data from Wi-Fi) or user profiling information (e.g. when a UE changes environment from outdoor to indoor or from vehicular to pedestrian mode). Such information may help the 5GS to make more intelligent decisions on slice selection (e.g. to switch from a slice with more flexible resources to a resilient one or vice versa).

Concept

As UEs can simultaneously connect to or switch across different slices (e.g. in case of mobility), they can have more prominent role for data preparation for the network to provide relevant localised contextual information and to identify earlier any changes in the network compared to the past intra-slice and/or cross-slice information they have gathered. The outcome processed information can also be used for network slice selection for the UEs. This can be utilised to address Gap #6 (cf. D2.2 [5GM-D2.2]) to further optimise cross-slice operations.

As an example, 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 common NFs change subject to operator policy.

Position in 5G-MoNArch architecture & Protocol Implications

As captured above, there are intra-slice services (i.e., NWDAF) for sharing analytics between NFs within 5GC. PCF and NSSF at cross-slice level can be seen as the consumer of such services.

It is assumed that procedures for collection of terminal analytics are installed on the UE and, if there are multiple procedures, each can be identified. So, the NWDAF knows the identities of these procedures. Collection of terminal analytics data by these procedures can be turned on by default, or the UE Configuration Update procedure can be used to enable or disable these procedures. A new Terminal Analytics Data (TAD) Setting is introduced for this purpose.

In order to share TAD generated with NWDAF, we propose to get the information from the UE via established N1 signalling messages to the AMF as a new Information Element (IE) added to the Registration Request of the UE.

Two options can be considered on how the NWDAF gets the information from the AMF. The corresponding procedures have been captured in in Figure 3-19.

Option 1: The NWDAF asks the AMF to be informed of all Registration Requests. Each time a Registration Request is received by the AMF, it informs the NWDAF and the NWDAF uses a newly proposed Namf_TAD_GET Request message to retrieve the Terminal Analytics. The notification applies for all UEs so either the Namf_TAD_GET request would be for a single UE (uniquely identified by a 5G Subscription Permanent Identifier- SUPI) or it would be a batch request. The NWDAF could specify a particular SUPI when it subscribes for notifications.

Option 2: A new Terminal Analytics Data Update Event is defined for the AMF. The NWDAF subscribes to be informed of this event. This subscription can be for an individual SUPI, or set of SUPIs, or all. When a UE sends a Registration Request that includes Terminal Analytics Data, this triggers the notification of the event towards the NWDAF together with the new Terminal Analytics Data.

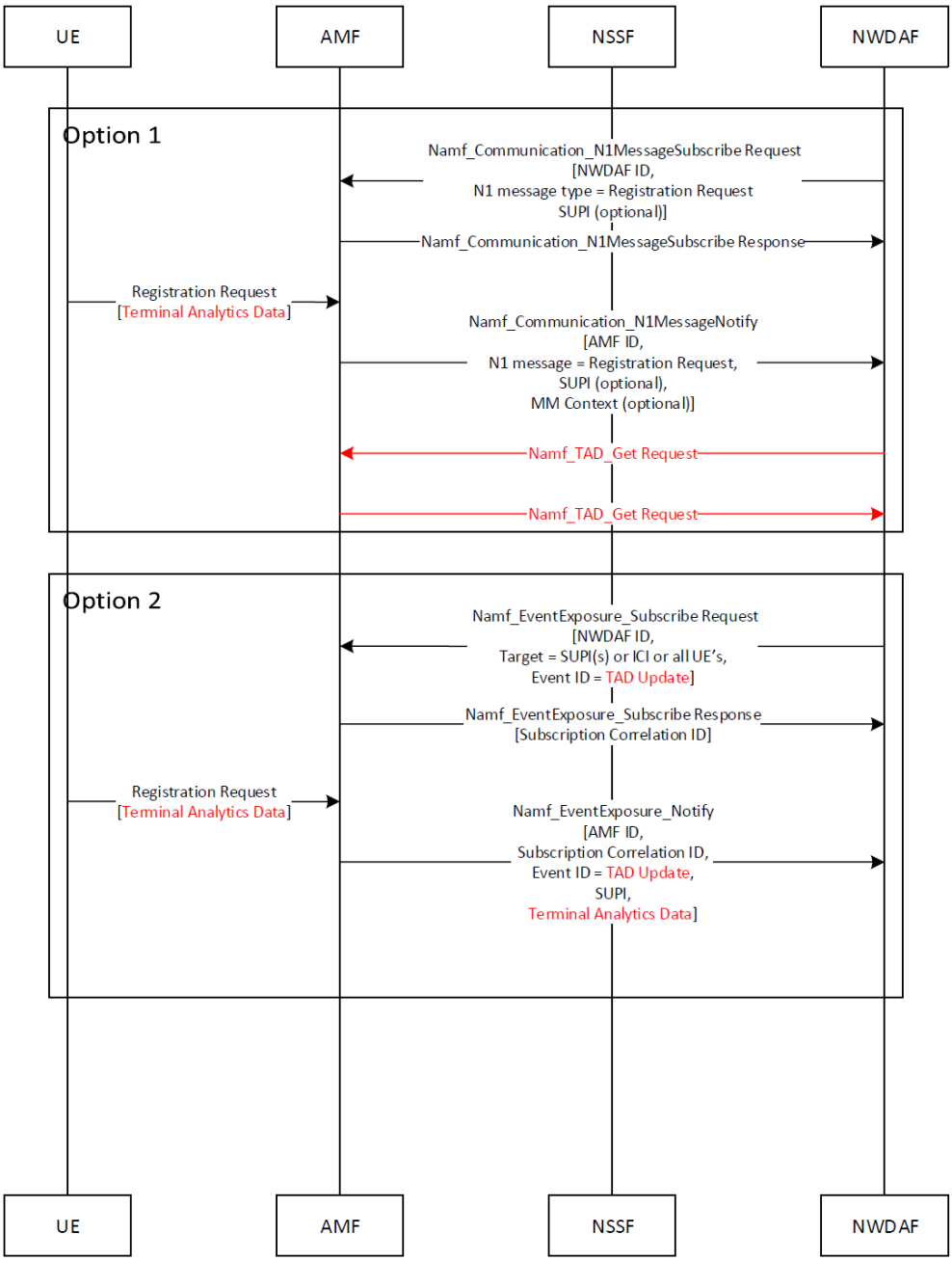


Figure 3-19: Options of sharing TAD between UE and NWDAF

Evaluation and analyses

As the above proposed procedures to share Terminal Analytics with NWDAF might be required more frequently than a single Registration Request or at specified times, we propose two types of triggers for the above procedures.

Event-based: Events can be defined (e.g. mobility of the UE into a new area, periodic registration expiry) which invoke a Registration Request by a UE. These may originate from changes in device QoE e.g. due changes in network or slice-level loading status that can be measured by certain KPIs, such as Average or median of packet latency, Packet error rate, Average or median of packet jitter.

Timer-based Triggers: Alternatively, time intervals can be defined as Triggers to invoke UE Registration Request. The intervals can be set for an individual SUPI or set of SUPIs or for all.

Figure 3-20 shows an example Message Sequence Chart for Terminal Analytics Sharing based on option 2 with Timer-based triggers.

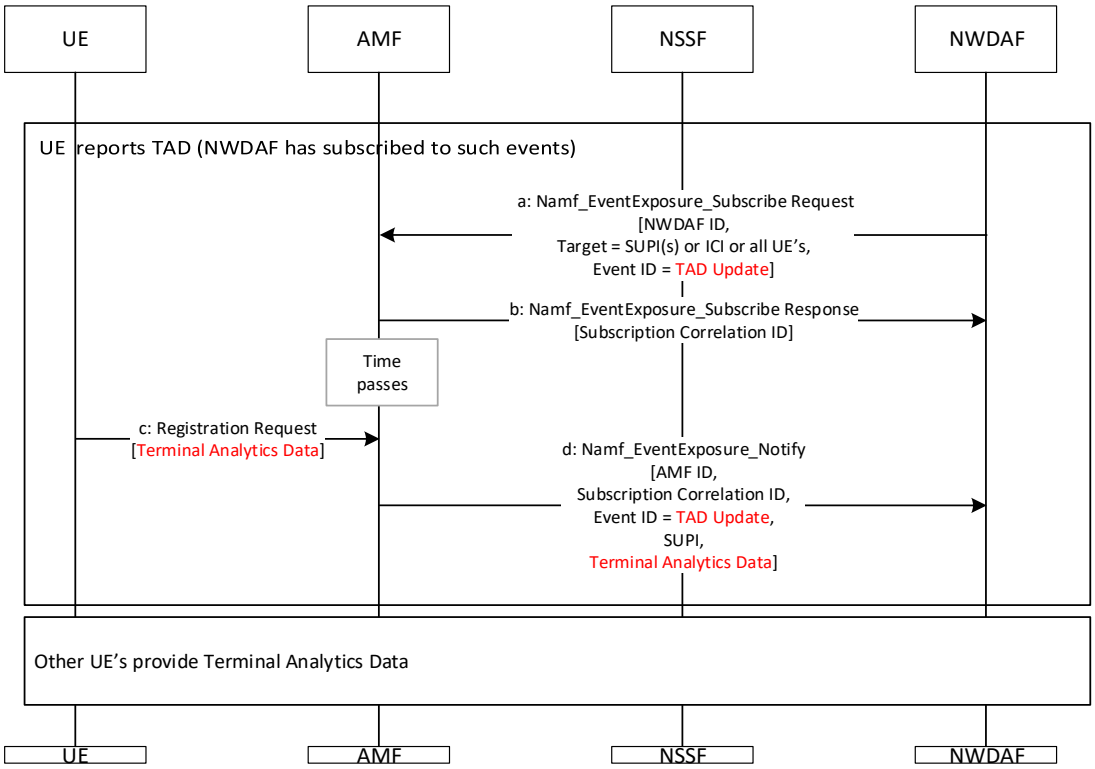


***Figure 3-20: Example message sequence chart for Option 2 with timer-based trigger*Inter-slice resource management**

Inter-slice resource management is a key innovation enabler in 5G-MoNArch architecture for optimising performance by allocating resources among slices which may share the same spectrum bands in access networks. This section presents enabling resource management solutions to accommodate various use cases and with different dynamicity requirements, as previously discussed in Section 2.2.1. On this basis, first, the radio resource allocation mechanism is exemplified, which is followed by the slice-aware functional operation that considers not only the radio resources but the extended notion of a resource including dynamic access nodes. Slice-aware mobility management is presented next. A controller layer implementation of the slow RRM mechanism is discussed. Thereafter, the use of big data analytics for slice resource allocation is detailed.

3.3.1 Inter-slice RRM for dynamic TDD scenarios

Concept

As bursty data services such as IoT, social media, etc., with asymmetric UL/DL demands are widely adopted, TDD systems gain momentum for 5G networks considering dense or hot spot deployments. This enabler introduces the notion of network slicing in 5G TDD networks, considering a multi-service environment with asymmetric traffic conditions. Network slices are formed on-demand with the allocated resources being dynamically adjusted with the objective to enhance the resource utilisation efficiency. Each network slice is customised to accommodate distinct service types by allowing each tenant to adopt a different TDD frame enabling a distinct UL/DL ratio, which can be re-configured independently reducing the loss of multiplexing gain. Although such TDD oriented network slicing framework is analysed in [SSS+16] considering an SDN-based architecture that enables multi-service and multi-tenancy support, the allocated slices have a fixed resource size for the entire duration of the service request, occupying only specific isolated sub-carriers.

This enabler builds on top of this slicing framework considering more dynamic slice allocations for dynamic radio topologies (addressing identified Gap #6 in D2.2 [5GM-D2.2] (E2E cross-slice optimisation not fully supported)), where slice resources can be adjusted during the time of a session request introducing the following planned contributions.

Initially, the generic optimisation problem for multi-slice multi-user and multi-cell UL and DL resource optimisation is formulated. The problem is translated to two sub-problems (P1 and P2) to allow for solving it with lower complexity and enhanced modularity. P1 involves the link activation selection per time instance, given the slice / traffic requirements and the TDD patterns given dynamic radio topologies. In addition, P2 takes as input the link selection and per time-slot aims to optimise performance by allocating resource blocks to the active links, in a way that the KPI is optimised.

A graph-based solutions framework is proposed for both problems to optimise slice performance while keeping the signalling overhead and complexity. For P1 a constraint-based greedy algorithm is provided, whereas for P2 the problem is solved by a novel bi-partite graph-colouring based solution, which aims to perform adaptive frequency partitioning per time slots in a way that interference due to resource conflicts is avoided and at the same time resource utilisation efficiency remains in high level. Initially, a bi-partite graph is translated to a line colouring graph, where each node is a combination of link and transmission time interval (TTI) (edge of the bi-partite graph). The edge between two nodes in the line colouring graph appears only if a conflict exists at the receiving end of the bi-partite graph, which is equivalent of having two or more links being assigned to the same TTI. The graph-colouring algorithm assigns a different colour to a node only in case of a conflict, which means that different sub-bands will be scheduled to avoid interference. Based on this algorithm, the output is a time-table where each link is assigned to different bands (e.g. F1, F2, …, F6), within distinct TTIs to ensure interference-free transmission/reception. In fact, this algorithm provides a flexible dynamic adaptation, where different parameters like number of users, slice KPIs and resource availability can be altered accordingly.

Solution to P1: Slice-aware TDD pattern Activation: For P1 a heuristic solution is provided as illustrated in Algorithm 1 (Figure 3-21) for activating the links in a time window based on the slice demand and aforementioned constraints. Initially in Step 0, a list of permitted timeslots for UL and DL is introduced per slice considering the TDD configuration pattern where a link can be activated only for a given Transmission Time Interval (TTI). A weight *f*(*e*,*s*) is also defined based on the slice traffic demand and a list of conflicting links, considering the half duplex constraint. In Step1, a random link is chosen to be included in a Candidate List (CL) for the first TTI and then the next link is identified with the minimum demand, provided that it does not violate the above rules. Once selecting a link, in Step 3 it is added to the CL and reduce its weight by 1. This is repeated in Step 4 and Step 5, till no more links exist for this TTI and then this process is repeated till all TTIs are considered (Step 6).

Step 0: ∀ slice (s):
- Set list of allowable timeslots per slice for DL: AM_DL (s, TTI) based on confDL(s) and for UL: AM_UL (s, TTI) based on confUL(s).
- Set vectors of Links (E) and Traffic Demand per link: fe,s
- Set List of conflicting links for each link e: Conf (e, s) and CL={}

Step 1: Start from random link e0 , add to CL={e0}

Step 2: Add the link e* with the lowest f(e*,s) to CL list $\forall e^*: e^* \notin AM\_DL, AM\_UL$ or $\forall e^*: e^* \notin Conf(\{CL\})$ → CL=CL+{e*}

Step 3: Reduce f(e*,s) by 1. If f(e*,s) is 0, remove e* from E list

Step 4: Go to Step 2 till E={} or no link can be added

Step 5: Store CL as FL(i) and reset CL={}. i=i+1 and repeat Steps 1-4

Step 6: Stop when i=T

Figure 3-21: Algorithm 1: slice-aware TDD pattern activation for inter-slice RRM

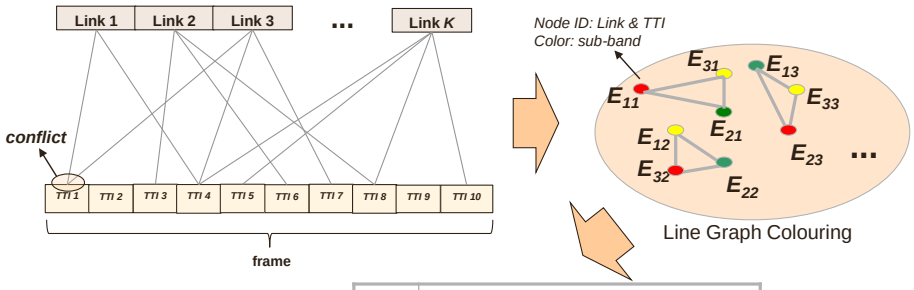
- Set FL as [#Links x #TTI] matrix from Algorithm 1
- Set a color set Color and maximum number of colors Cmax and Clist={}

for TTI=1:T
  if FL(1:links, TTI)==y≤ $Cmax$
    Set randomly $y \in Colors$ different colors for the links
    connecting to TTI
  end if
    Store color indices for all links for TTI in a matrix as:
    Coloring(Link, Color Index, TTI)
end for
for color_index=1:Cmax
  CList(color_index) =Coloring (1:links,color_index,1:T)
end for
for bands=1:RB and color_index=1:Cmax
  Map bands to CList(color_index) that maximizes weighted sum-rate
end for
end for

Figure 3-22: Algorithm 2: graph-based resource allocation

Solution to P2: Graph-based Resource Allocation: For P2 a graph theoretic approach is considered. The outcome of the solution to P1 gives an allocation of links to TTIs. However, it is still unknown how many and which resources can be assigned to these links in order to avoid inter-cell and cross-link interference assuring the desired slice performance. The proposed P2 solution is illustrated in Algorithm 2 (Figure 3-22).

Initially, a bi-partite graph is created including the set of links and the set of TTIs. Based on this bi-partite graph, the resource allocation problem is translated into a time-tabling problem, where a number of activate links are required to occupy a number of different TTIs. A small cell Access Point (aka s-AP) has to create a time-table according to its availability in a way that no collision occurs in each slot. A graph-colouring is adopted to assign different colours, so as to restrict the allocation of links to conflicting TTIs in distinct sub-channels. As shown in Figure 3-23, a bi-partite graph is translated to a line colouring graph, where each node is a combination of link and TTI (edge of the bi-partite graph).



Link No.	Time Periods					
		1st	2nd	3rd	...	10th
	1	*F1*	-	*F4*		*F3*
	2	*F2*	-	*F1*		*F2*
	3	*F3*	-	*F3*		-
	...					
	K	*F1*	-	*F2*		*F1*

Figure 3-23: Graph colouring algorithm overview

Position in 5G-MoNArch architecture & Protocol Implications

- A common RRC functionality is required which configures TDD patterns in a slice-aware manner and the link activations in long term.
- Inter-slice RRM functionality at MAC or Unified Scheduler (which can be interpreted as an overarching layer on top of MAC for inter-slice dynamic scheduling) is considered for dynamic resource allocation among slices based on the configured TDD patterns. The placement of this functionality depends on the dynamic radio topology and on the functional operation/capabilities/supported split of the DUs (which can be planned / unplanned small cells).

Figure 3-24 shows the message sequence chart for inter-slice RRM in a dynamic TDD scenario.

Evaluation and analyses

Monte Carlo system level simulations are provided for a 5G Ultra Dense Networks (UDNs) where resources can be shared by multiple slices with diverse KPIs (example for throughput, reliability). The evaluation study focuses on an outdoor small cell deployment of 4 s-APs covering a hotspot area, using the 3GPP as baseline for simulations (24 users uniformly distributed, 3GPP UMi channel, ideal backhaul). In each s-AP the corresponding users (6 users per cell) are randomly distributed. MATLAB Monte Carlo simulations and random user drops for 500 snapshots are run. Assumed are 4 slices, whereas each slice has different TDD pattern as slice requirement (Slice 1: 80/20, Slice 2: 70/30, Slice 3: 60 /40, Slice 4: 50/50). At each snapshot, randomly 6 users are selected out of 4 cells to be connected to each slice, and a random traffic demand (1-10Mbps per user for both UL and DL) is applied.

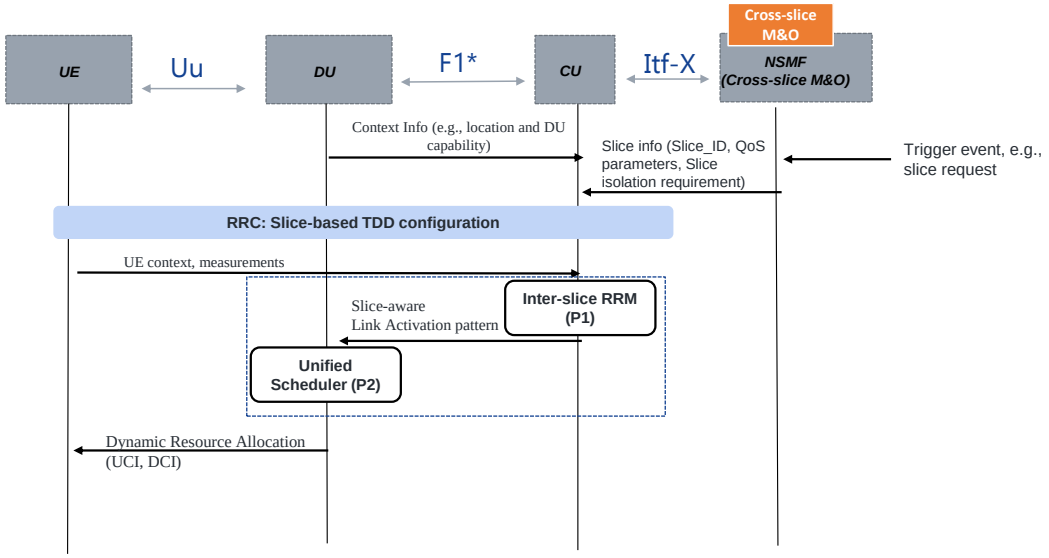


Figure 3-24: Message sequence chart for inter-slice RRM in dynamic TDD scenario

For the simulation comparison are considered:

- **Benchmark 1** is the cell specific dynamic frame re-configuration (CSDR) [SKE+12] without slicing where each s-AP can adopt a different TDD pattern, while using the same spectrum resources, with inter-cell and cross-link interference potentially deteriorating performance.
- **Benchmark 2** is the service-oriented TDD slicing [CSS+16], where slices are assigned a constant amount of resources (¼ of resource blocks in the simulations) and different TDD patterns are used independently for each slice. This solution provides a high spectral efficiency due to the interference isolation, but at the cost of lower resource utilisation, which can limit the peak throughput.
- **Algorithms 1 and 2** are used to select links and allocate resources over the entire range of resource blocks, while keeping interference at low levels.

Figure 3-25 shows the comparison of CDF curves of DL throughput per user as well as for UL throughput respectively (averaging it over the allocated TTIs) for all snapshots. In Figure 3-25 (upper part) it can be observed that the proposed solution outperforms Benchmarks 1 and 2 since it better addresses the trade-off between interference isolation vs resource utilisation. Benchmark 1 shows the worst-case interference scenario, whereas Benchmark 2, uses orthogonal resources for different slices. For Benchmark 2, the DL rate for all slices is aggregated collectively and it is shown that for the median and the 90th percentile of the CDF, the average throughput can be increased by more than 150%. The proposed solution shows a significant gain even over the second Benchmark, due to the fact that it achieves higher spectral efficiency with more resources being allocated to DL links based on the corresponding demand (in Benchmark 2, some resources may be wasted). In Figure 3-25-(lower part), a similar trend is observed for the CDF of the UL throughput. The proposed solution shows similar performance at the 10th percentile of the CDF (cell-edge performance), whereas at median and 90th percentile (cell-centre) it outperforms both Benchmarks 1 and 2 respectively. This gain is mainly due to the fact that a better UL spectral efficiency can be achieved, and at the same time allocate more resources to links based on the actual demand, so as to maximise the total performance.

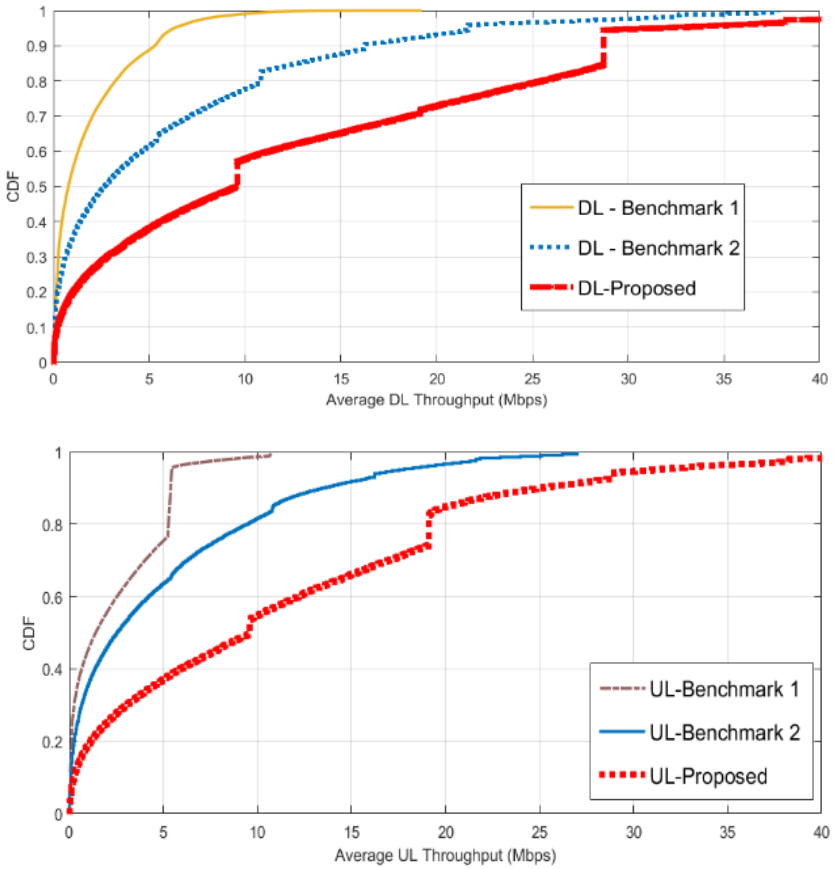


Figure 3-25: CDF of average user throughput illustration for DL (upper part) and UL (lower part)

3.3.2 Context-aware relaying mode selection

Concept

As highlighted in [5GM-D2.1][5GM-D2.2] and in Appendix B, for the fulfilment of network slice SLAs, an extended notion of a resource shall be taken into account, where the availability of wireless access nodes and the network topology shall be jointly considered along with the network slice requirements. This becomes particularly important when the network topology is changing as in case of self-backhauled DSCs, e.g., VNNs. The dynamic network topology can be exploited to better adapt to changing traffic conditions over time and space in cost-efficient way.

The wireless backhaul link of the DSCs can be reached by employing a relaying functionality. A fixed relay can be typically deployed as fixed radio frequency (RF) amplify-and-forward /repeater or layer 3 (L3) decode-and-forward node [3GPP TS 36.300]. As opposed to fixed functional operation in the SotA, slice-awareness and 5G tight KPIs can necessitate on-demand flexible SC operation. Slice-based target KPIs can comprise throughput / spectral efficiency for eMBB communications, high reliability and low latency for URLLC, and connection density for mMTC. Network slices may have different requirements in terms of throughput and latency, which necessitate enabling different operations for different types of traffic to meet certain KPIs. To this end, additional context can be utilised, such as, the position of the DSCs at different parts of the cells and the associated channel link qualities. Furthermore, different functional operations of DSCs can have different E2E latencies (e.g., amplify-and-forward relaying imposes less latency compared to decode-and-forward relaying thanks to fewer processing steps of the signals). On this basis, as illustrated in Figure 3-26, the rational of this enabler is to analyse and determine the appropriate relaying mode (i.e., functional operation) of DSCs, based on, e.g.,

- Slice requirements, such as latency and required data rate;
- Resultant performance of selected mode (e.g., throughput and latency);
- Location of DSCs in the target service region (e.g., cell edge, cell middle, and cell centre).

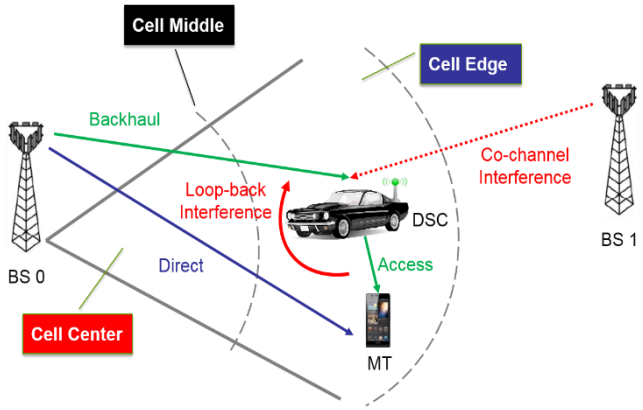


Figure 3-26: Example factors that can influence the functional operation of the DSCs

Different example functional splits among donor macro-site and DSCs are depicted in Figure 3-27. As mentioned above, different possible modes can be identified given the per-slice requirements, the backhaul channel (between macro and DSC) and the RAN conditions. In this context, the first option is the L3 DSC with full protocol stack, i.e., the L3 DSC can control the cell under its coverage, e.g., with a physical cell ID (PCI). In case of L2 DSC, a PDCP-level split can be employed.

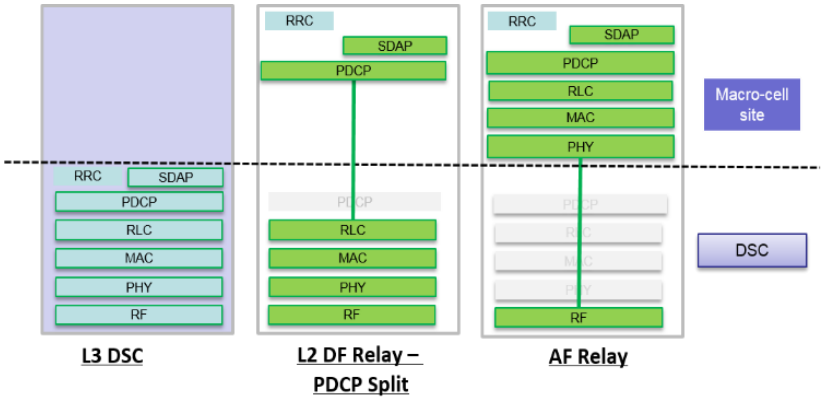


Figure 3-27: Illustration of analysed functional operations/modes at DSC, i.e., L2 decode-and-forward relay and amplify-and-forward relay, and illustration of L3 DSC as reference; amplify-and-forward and decode-and-forward are marked as AF and DF, respectively

The PDCP split could be more applicable in cases of frequent fast handovers (e.g. high mobility users) between the macro and DSCs, since PDCP re-transmissions would be required more often, and PDCP should be centralised for fast traffic forwarding. Decode-and-forward relaying option typically applies half-duplex operation to isolate backhaul and access links, and thanks to signal regeneration, there is no noise or interference amplification. Another option is the DSC to act as radio frequency (RF) amplify-and-forward which functions as half duplex; however, amplify-and-forward mode may especially suffer from interference amplification, e.g., loop-back interference between backhaul and access links. These modes may not be confined to protocol stack layers, i.e., some of the functionalities at each protocol stack layer may also be split. For example, MAC functionality of HARQ may be at the DSC, while another MAC functionality multiplexing/de-multiplexing may reside at the macro-cell BS.

This enabler is part of the 5G-MoNArch enabling innovation Inter-slice control and management. It targets the identified Gap #3 (The functional operation of small cell networks is fixed) and Gap #6 (E2E cross-slice optimisation not fully supported), cf. Appendix A.

Position in 5G-MoNArch architecture and Protocol implications

The mode selection can be based on a **dynamic RAN control unit** which can be located at the donor BS (e.g., CU) to which the wireless backhaul link connection is established. It is worth noting that the

dynamic RAN control unit can take into account the information and/or commands provided by the **slow Inter-slice RRM App in the controller layer** (see Figure 2-4 and Figure 2-5). Such a control functionality can be considered as an extension to RRC protocol layer, as highlighted in Section 2.2.1. In addition, network slicing management functions (e.g., **Cross-slice M&O function**) can also be considered, which are responsible from RAN configuration. An example operation is depicted by an MSC in Figure 3-28. Therein, slice-aware mode selection is performed by the dynamic RAN control unit, where the needed slice information and QoS parameters can be obtained from the 5GC, where DSC-related context information can be additionally utilised to decide on the appropriate DSC mode and the associated QoS parameters. The DSC mode thus can comprise radio bearer configurations. The necessary information elements are transmitted from the CU to the DSC in a configuration signalling message. The DSC can thereafter acknowledge the reception and application of the configuration.

In terms of other protocol implications, the selection of the appropriate serving DSC to be activated from the available set of candidates, more frequent link quality measurements would be required from the PHY. The link quality can pertain to the wireless backhaul link, where the backhaul link can be impaired by both fast fading and shadowing. Therefore, with frequent measurements on the backhaul link event fast fading impacts can be elevated. This has the trade-off increased overhead on the backhaul link signalling. In order to ensure slice requirements E2E, in addition to the backhaul link quality measurements, measurements from access link (i.e., between UE and DSC) and direct link (i.e., between UE and macro BS) can also be considered.

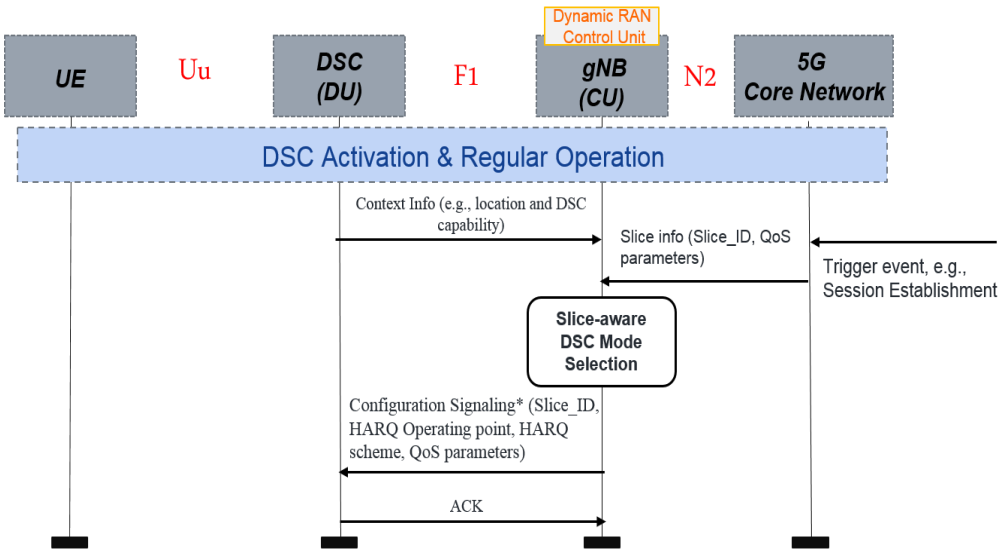


Figure 3-28: Message sequence chart for the operation of the context-aware relay mode selection

Evaluation and analyses

The evaluation of this enabler comprises a joint optimisation of the achievable data rate and the induced protocol processing delays considering different relaying modes and the location of the DSC in a macro cell. The simulation set-up and further analyses are provided in Appendix B. In order to decide on the final DSC mode, the slice requirements shall also be considered in addition to the performance of the different modes. On this basis, in Figure 3-29, two network slices with different requirements on the spectral efficiency and latency are depicted along with the performance comparison of amplify-and-forward and decode-and-forward modes at the cell middle. It is to be noted that amplify-and-forward mode induces lower E2E latency compared to decode-and-forward mode, because amplify-and-forward mode includes fewer amount of processing functions (i.e., only RF) and does not include a decoding of the signal (i.e., from RF up to RLC/PDCP). Additionally, the amplify-and-forward mode is typically full duplex, which implies no delay is introduced due to, e.g., half-duplex time-division multiplexing (TDM). When the UE access link SNR is 18 dB, as marked in Figure 3-29, the slice 2 requirement on

the spectral efficiency can already be fulfilled by the amplify-and-forward mode. As the amplify-and-forward mode induces shorter latency, and slice 2 has strict latency requirement, for slice 2, the amplify-and-forward mode shall be configured. On the other hand, slice 1 spectral efficiency requirement can only be fulfilled by the decode-and-forward mode and as the slice 1 has relaxed latency requirement, for slice 1, the decode-and-forward mode shall be configured. Under the light of these analyses, it can be concluded that the performance of different modes, e.g., in terms of throughput performance, E2E latency, and reliability, shall be considered and based on the slice requirements, the DSC mode can accordingly be determined.

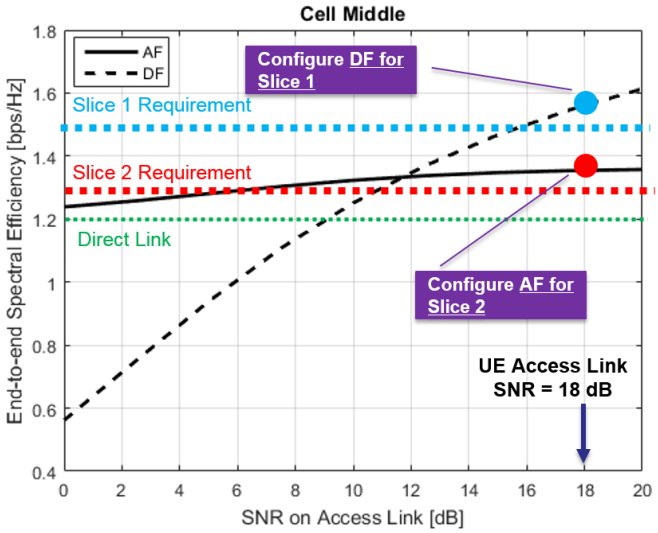


Figure 3-29: Slice-aware DSC mode selection which considers link qualities and slice requirements; Slice 1 requires 1.5 bps/Hz with relaxed latency requirement while Slice 2 requires 1.3 bps/Hz with strict latency requirement; amplify-and-forward and decode-and-forward are marked as AF and DF, respectively

3.3.3 Slice-aware RAT selection

Concept

Network slicing enables to tailor a network instance to the specific requirements of a future 5G service. In this context, RRM will be a complex task, because the 5G network will integrate different RATs, each one with its specific characteristics in terms of e.g. coverage and capacity (see Figure 3-30). The appropriate configuration of the RAT and the management of the associated resources is a challenging task, when considering the heterogeneous requirements of the diverse 5G services.

In LTE system, cell range expansion has been used as a way to offload traffic from macro cells to small cells and boost the network capacity [ONY+11]. The concept foresees a similar scheme to balance traffic across multiple RATs, where biased received powers related to different RATs are compared at the UEs in order to select the most appropriated RAT to use.

In LTE, the same bias is used at different UEs to compare the received powers and associate to a nearby small cell or the macro cell, accordingly. However, although some UEs may benefit of the improved capacity offered by small cells, other UEs may rather require reliable coverage, offered by the macro cell signal. In addition, when considering millimetre-wave (mmW) small cells in future 5G networks it is of paramount importance to consider their propagation characteristics, high path loss and sensibility to blockages, which can be detrimental for the user performance. This is particularly true for URLLC and Vehicle to Everything (V2X) communications. Therefore, this study focuses on Gap #5 of D2.2 [5GM-D2.2].

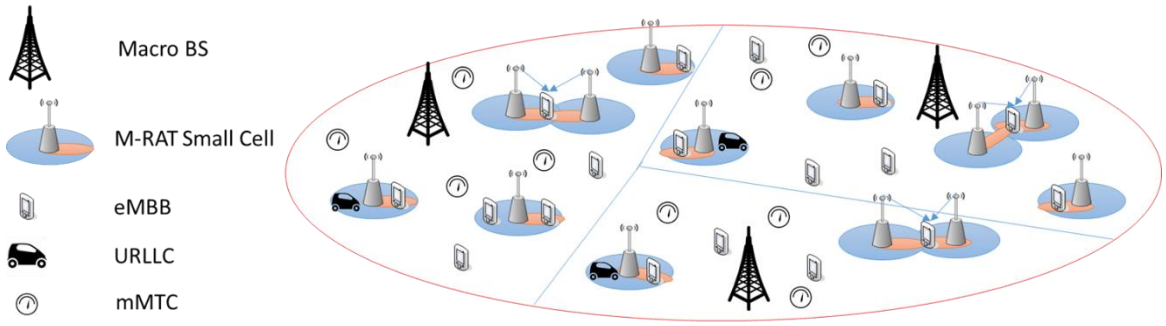


Figure 3-30: 5G multi-RAT deployment for heterogeneous service provisioning

An exemplary pseudo code for the implementation of the proposed approach is shown in Figure 3-31. In this example, with consider a multi RAT network deploying, eMBB, URLLC, and mMTC slides, each one characterised by different constraints in terms of SINR ($\boldsymbol{P_C}$) and data rate ($\boldsymbol{P_R}$) distribution, as well as for blocking probability. Based on context-aware related information (network deployment density, user density, and vehicle traffic in the area), the CU computes, by using stochastic geometry [GDC18] tools, the value of the RAT selection biases that satisfy these constraints, and then selects the appropriate slice-related bias accordingly. The bias is the transferred to the end users attached to each slice, and finally used to connect to the optimal RAT.

Algorithm 1 Network-Side Pseudo-code

1: Obtain the data about expected vehicular density in the service area.
2: **for** each slice of QoS triplet $(\mathcal{B}, \mathcal{P}_C, \mathcal{P}_R) \in \mathcal{T}$ **do**
3: Identify the set of biases $(0, Q_B)$ that satisfy $\mathcal{B}$
4: Identify the set of biases (Q_{C1}, Q_{C2}) that satisfy $\mathcal{P}_C$.
5: Identify the set of biases (Q_{R1}, Q_{R2}) that satisfy $\mathcal{P}_R$.
6: Obtain $Q_R^* \in (1, Q_B) \cap (Q_{C1}, Q_{C2}) \cap (Q_{R1}, Q_{R2})$ for maximizing $\mathcal{P}_C$ if URLLC/mMTC slice or for maximizing $\mathcal{P}_R$ if eMBB slice, using random restart hill climbing.
7: Broadcast Q_R^* within the slice.
8: **end for**

Algorithm 2 User-Side Pseudo-code

1: Measure downlink sub-6GHz received powers, $P_{tv\mu}$, from all BSs.
2: **if** $P_{Mv\mu 1} \geq P_{Sv\mu 1}$ **then**
3: Associate to the strongest MBS.
4: **else**
5: Associate to the strongest SBS and measure the mm-wave power from it (P_{Svm1}).
6: Obtain the RAT bias Q_R^* for the associated slice.
7: **if** $P_{Sv\mu 1} \geq Q_R^* P_{Svm1}$ **then**
8: Start service from SBS in sub-6GHz band.
9: **else**
10: Start service from SBS in mm-wave band.
11: **end if**
12: **end if**

Figure 3-31: Slice-aware RAT selection pseudo-codes

Position in 5G-MoNArch architecture and Protocol implications

A shared NF located at CU is defined inside the mobility management (MM) module, which controls the load balancing and the user association in the RAN, such that the slice requirements are considered. Such a control functionality can be considered as an extension to RRC protocol layer, as highlighted in Section 2.2.1. The slice related physical layer network requirements are signalled to the CU from the M&O layer, through an interface that is currently under definition in 5G-MoNArch and denoted as Itf-X. Accordingly, the MM module computes the RAT selection, which are transferred to the distributed RAN units and the signalled to the end users. Finally, the UEs use this information jointly with the RAT related measurements (e.g. the strength of the received signal in a specific band) to select the appropriated RAT. The MSC of the proposed solution is shown in Figure 3-32.

Evaluation and analyses

We summarise here the main contributions of this study, and then discuss some salient results. More detail can be found in [GDC18]. The major contributions of this work are enumerated below:

- We use the Poisson line process to model the roads of an urban scenario on which multi-RAT small cells, operating in both sub-6GHz and mm-wave bands are deployed to serve pedestrian UEs.
- We propose an mm-wave interference model for small cell deployment along roads.

- We consider the effect of the vehicles that cause a temporary blockage in the line of sight link between an outdoor UE and the small cells. This enables the operators to properly dimension the network so as to cater to the needs of reliability constrained applications.

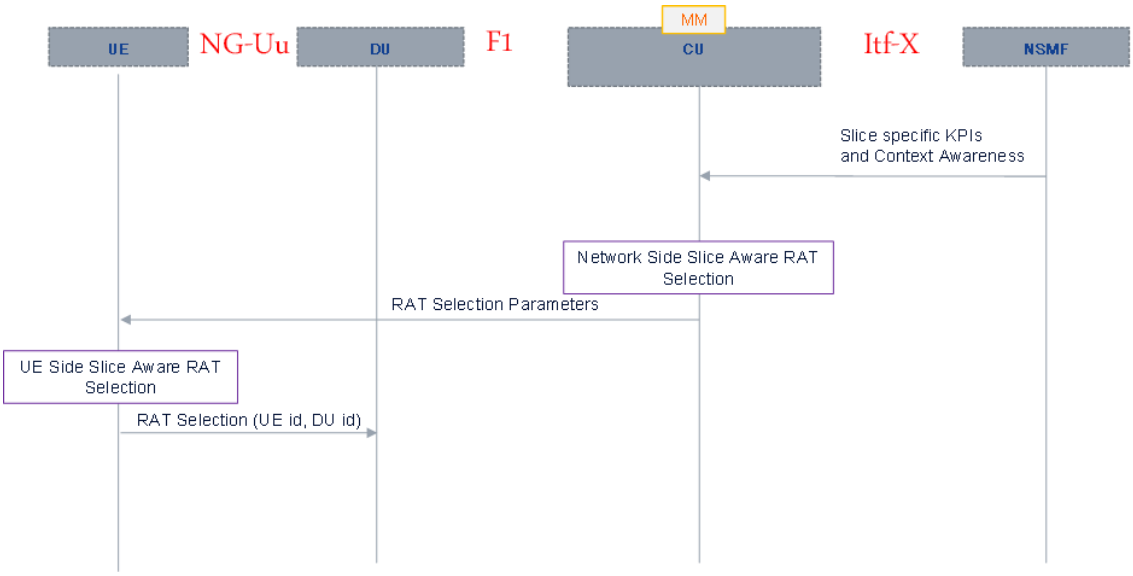


Figure 3-32: Proposed slice-aware RAT selection mechanism

Our results show that for a given density of the small cells and vehicles, the optimal RAT selection bias should vary for addressing different service requirements, e.g., vehicular blockage, coverage, and data rate.

In Figure 3-33, we show the optimal RAT selection in the sub-6GHz and mm-wave bands as a function of the vehicular density λ_V for three exemplary services with different requirements: 1) a slice for a URLLC service with requirements set equal to ($\boldsymbol{B}$ =0.001, $\boldsymbol{P_C}$ =0.85, $\boldsymbol{P_R}$=0), 2) a slice for an mMTC service with requirements equal to ($\boldsymbol{B}$ =0.1, $\boldsymbol{P_C}$ =0.9, $\boldsymbol{P_R}$=0), and 3) A slice for an eMBB service with requirements equal to ($\boldsymbol{B}$ =0, $\boldsymbol{P_C}$ =0.85, $\boldsymbol{P_R}$=0.7).

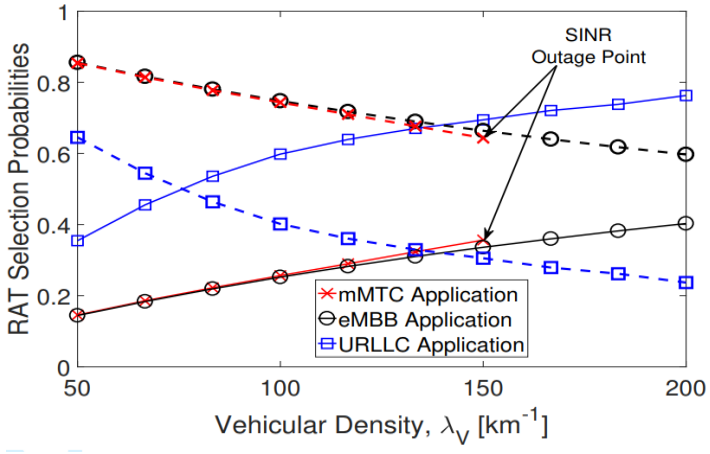


Figure 3-33: Slice-aware RAT selection probabilities for different use-cases, solid lines denote the sub-6GHz association probabilities and the dotted lines denote the mm-wave association probabilities respectively

We assume a network with small cell density of λ_S = 10 km−1, deployed in an urban area with street density of λ_R = 15 km−1. For URLCC, $\boldsymbol{B}$ is 0.1%, which results in an optimal RAT selection bias Q_R^*= 19.7 dB. This leads in a lower mm-wave selection probability as compared to the other applications. As

the vehicular density increases, the maximum allowable Q_R^* to satisfy the vehicular blockage constraint gradually decreases, thereby further decreasing the mm-wave selection probability. For mMTC, the $\boldsymbol{B}$ is less stringent (10%), whereas $\boldsymbol{P_C}$ is tighter (the outage constraint is equal to 10%). In case of low λ_V (e.g., 50 km−1), the optimal bias is Q_R^* = 25 dB, which achieves an SINR coverage of more than 91%, with an mm-wave selection of over 80%. This is considerably higher than the URLLC applications. However, for $\lambda_V \geq 150$ km−1, no feasible bias exists to satisfy the outage constraint, and the application cannot be supported with current network dimensioning. The vehicular density value after which the network is not able to sustain outage below 10% is shown in Figure 3-33. Finally, the eMBB service does not have any vehicular blockage constraints. Thus, the bias for eMBB applications aims to maximise the rate coverage probability, while satisfying the outage constraint (here 15%). For λ_V = 50 km−1, the optimised bias (Q_R^* = 26.21 dB) results in a slightly higher mm-wave selection probability than the mMTC application. As the vehicular traffic increases, the optimal bias value decreases. However, as the outage constraint is not as stringent as the mMTC application, the UE can be served even under very high vehicular densities (e.g., λ_V = 200 km −1) in the mm-wave band.

3.3.4 Inter-slice RRM using the SDN framework

Concept

In order to realise the concept of network slicing, it is important to design and validate an appropriate RRM strategy to share stringent radio resources between slices that have different SLAs. There is a lack of propositions in the current literature for enabling such strategies in the SDN/NFV driven 5G architecture (Gap #12, cf. D2.2 [5GM-D2.2]). The inter-slice RRM strategy in the 5G-MoNArch architecture needs to consider the two layers of control, i.e., Controller layer and M&O layer and needs to identify and separate the strategical decisions to be deployed in those layers. As an example, the RRM decisions between slices deployed in the same domain can be from the Controller layer and those across different domains need to be from the M&O layer. The advantage of SDN such as E2E network abstraction, programmable user plane and centralised control plane benefits mobile network architecture by designing and deploying applications/algorithms that can control/manage stringent resources from the centralised vantage point. 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).

The proposed inter-slice RRM approach as depicted in Figure 3-34 is a cross layer optimisation technique to improve the overall utilisation of radio resource between slices by considering SLAs of slices, current back-haul network latency, current radio resource usage and RLC buffer status information. In this framework, the proposed approach is the “slow inter-slice RRM” approach in addition to the fast-inter-slice RRM approach typically in the Network layer. This is due to the fact that it is impossible for the SDN controller to interact with the RAN scheduler for every scheduling period (~1ms) with new optimised parameters (latency in communication and processing). In summary, the proposed approach is the cross layer as well as two level inter-slice RRM technique.

Position in 5G-MoNArch architecture & Protocol Implications

As shown in Figure 3-34, the Inter-Slice RRM can be deployed as NB application on top of the controller framework in the 5G-MoNArch architecture. The controller collects matrices such as RLC buffer status, network latency and radio resource status information via SB interface and update the dynamic network topology in the controller. Inter-Slice RRM application uses that information available in the controller data-store along with SLAs of those slices under consideration by interacting with M&O layer via a dedicated interface. The MSC in Figure 3-35 explains the interaction between various functions during the operation of inter-slice RRM application in the SDN framework.

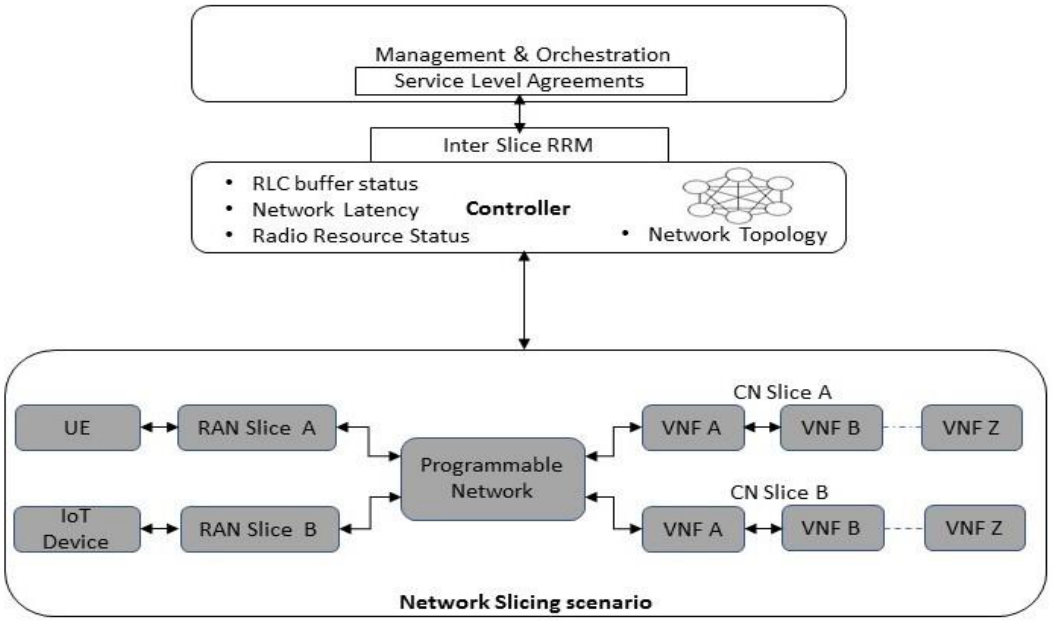


Figure 3-34: Inter-slice RRM using SDN framework

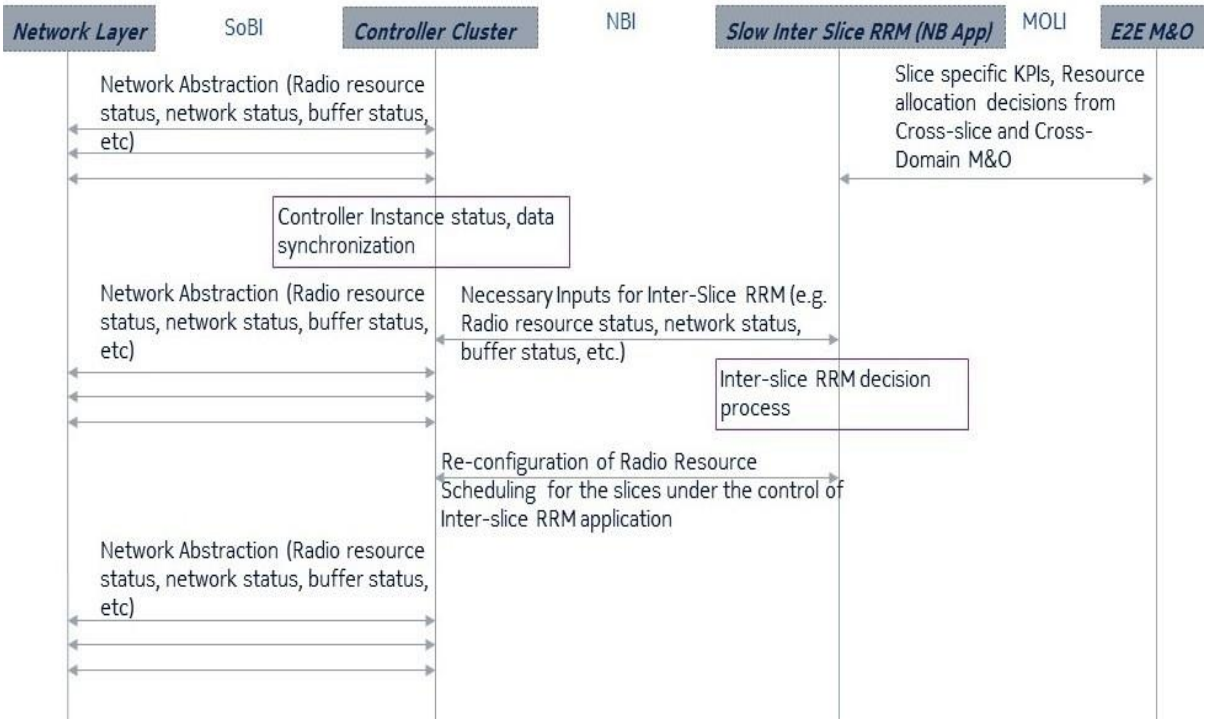


Figure 3-35: Message sequence chart of the proposed inter-slice RRM using SDN framework

Evaluation and analysis

The experimentation platform for evaluation of inter-slice RRM using SDN framework is shown in Figure 3-36. The platform is built using srsLTE RAN and EPC stack [GGS+16], which is an open source end-to-end mobile network stack supporting 3GPP release 9. The used physical machine for the experimentation has an Intel® Core™ i7-4790 CPU @ 3.60GHz and 16 GB RAM. The machine is operated using ubuntu 18.04 LTS with Kernel version 4.15.0-29- low latency. Both RAN and Core stack of srsLTE is virtualised and deployed in docker container with version 18.06.1-ce. The USRP B210 is used as a transceiver. The scalable controller framework developed in WP3 of 5GMonArch is utilised

here with SB protocol extensions to support high performance and scalability. By exploiting the concept of virtual queues and flow-based scheduling, the inter slice scheduler algorithm on top of the controller can schedule flows corresponding to different services based on their level of priority. This approach is L4 pre-scheduling of services before the L2 MAC scheduling. The OpenVswitch (OVS) used in this platform is an extended version with deep packet learning feature [OVS]. The user experience along with performance of the system with and without RRM strategy is measured and compared to validate such approach.

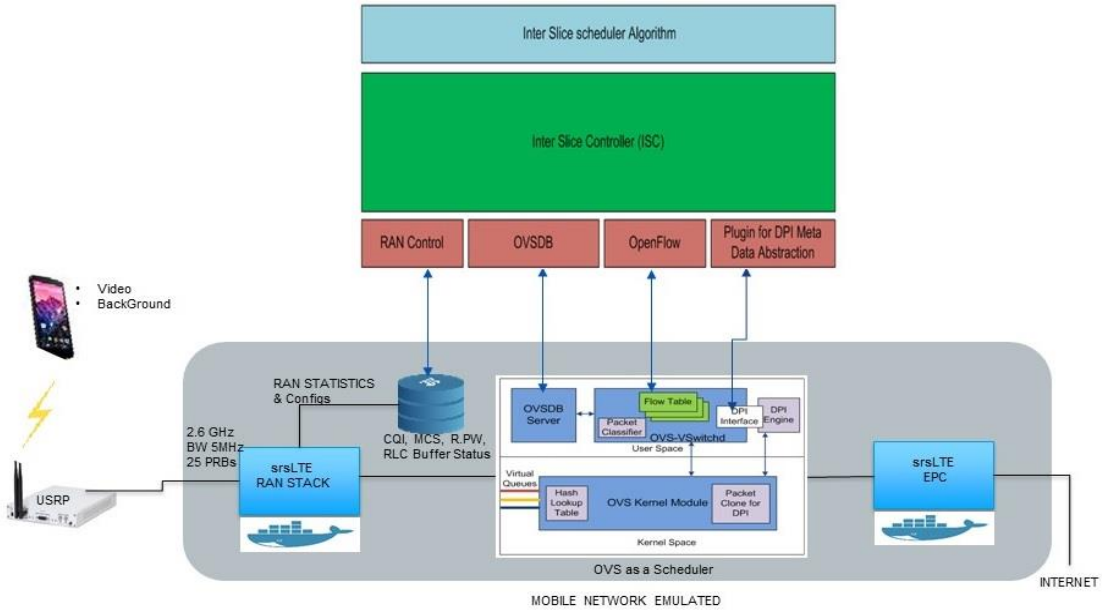


Figure 3-36: Experimentation platform

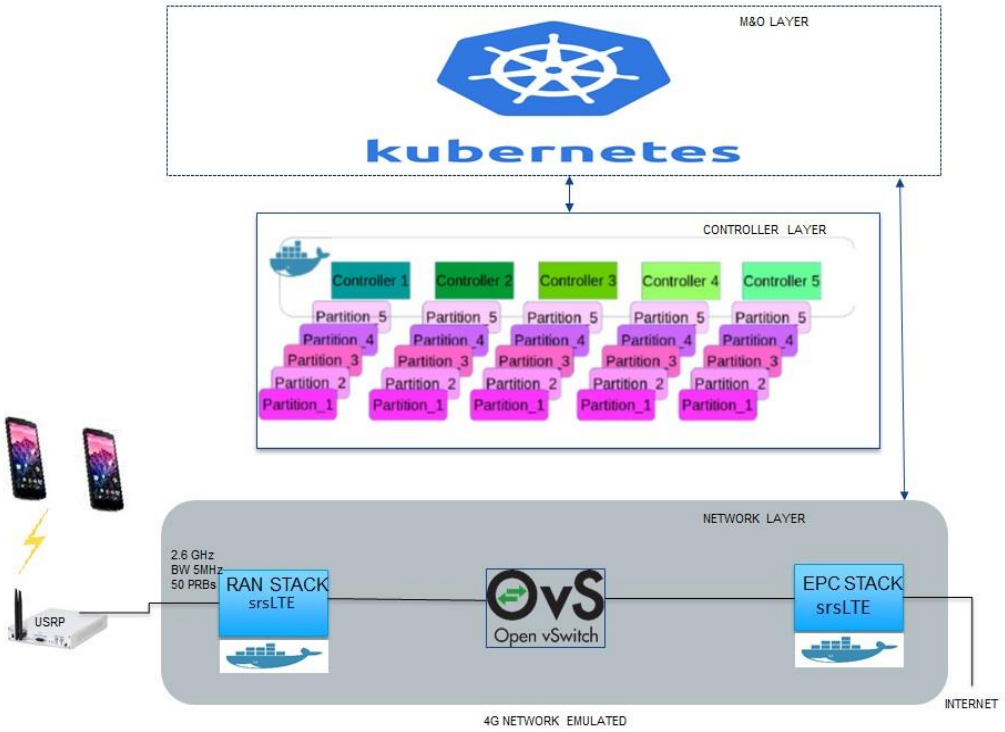


Figure 3-37: Experimentation platform integration with scalable controller & orchestration framework

The integrated view of the experimentation platform with scalable controller developed in WP3 of 5G-MoNArch and one of the open source orchestration frameworks such as Kubernetes is shown in Figure 3-37. In the final experimentation platform all components such as srsLTE RAN, srsLTE EPC and Controller framework is containerised with docker complaint and stored in the local repository. The experimentation platform consists of 2 PCs (1 for Kubernetes Master and 1 for Worker node).

The average instantiation time of end-to-end network (network to be on live with EPC, RAN and USRP) via Kubernetes deployment is measured to be around 7 seconds in a single PC with physical configuration Intel® Core™ i7-4790 CPU @ 3.60GHz and 16 GB RAM. The inter-slice scheduler application on top of the SDN controller creates two virtual queues on the OpenvSwitch one for video traffic and another for non-video traffic. The queues are assigned to operate on different priority on scheduling packets on out-port of the switch (e.g. 75% for video traffic and 25% of non-video traffic) The application setup rules on the switching table by learning header information of flows to map traffic on different in-bound queues according to their priority. Such approach improves the performance of high priority flows.

3.3.5 Big data analytics for resource assignment

Concept

Chapter 2 described an overall architecture for instantiating multiple network slices, along some possible optimisations of the interactions among the functions in a VNF chain. However, when setting up a slice without stringent service requirements, one of the key desired features will be that of elasticity; this is needed in all cases where resource overprovisioning is not a valid option either due to the actual resource availability (e.g., in the edge of the network) or due to the dynamic nature of network load, which makes an efficient network slice dimensioning difficult. In those cases, temporal and spatial traffic fluctuations may require that the network dimensions resources such that, in case of peak demands, the network adapts its operation and re-distributes available resources as needed.

These load fluctuations usually characterise each slice. In this context, statistical multiplexing gains can be improved by applying elasticity to simultaneously serve multiple slices using the same set of physical resources (in conjunction with the cloud-aware protocol stack described in Section 3.1.1). This has a direct impact on the number of network slices that can be hosted within the same infrastructure and, in turn, allows to exploit complementarities across slices, yielding larger resource utilisation efficiency and high gains in network deployment investments (as long as cross-slice resource orchestration is optimally realised).

Position in 5G-MoNArch architecture & Protocol Implications

This behaviour is implemented by orchestration algorithms implemented in the Cross-slice M&O module. NFV-Os orchestrate VNFs on the available resources according to the available resources and their elasticity. For example, resources may be equally shared initially, then, in case of peak demands, the Cross-slice M&O can re-assign resources taking advantage of different distributions of loads. In this case, resources are borrowed from slices not in peak load. The behaviour of the various elastic slices when the resources needed to accommodate their peak demands exceed the originally assigned ones is driven by the elastic operation.

Big Data engines can be used to perform the operation described above in an automated fashion. By studying the past load of different network slices, this engine can identify the most usual time interval and locations in which a network slice experience higher peak demands or, on the other hand, lower activities. Summarising, the foundation of this work lies in the network slice characterisation.

An accurate characterisation of traffic demands over a given area supports a more efficient planning of network resources. For example, in case of capacity-limited deployments, accurate characterisation supports a very efficient deployment of resources over time. Therefore, this kind of analysis will be beneficial also for the economic feasibility of multi service deployments. As depicted in Figure 3-38, the resource assignment procedure considers inputs coming from data monitoring modules deployed in the core network (such as the NWDAF).

As discussed already for the cloud enabled protocol stack, the Big Data analytics for resource assignment needs a thorough monitoring of all the resources used by the VNFs, both on the networking side (i.e., latency, throughput, number of connected devices) and the infrastructure side (i.e., CPU, Memory,

location of VNFs). All this information shall then be stored at the NFVO and passed to the Cross-Slice M&O module. Here, the decision about the resource allocation is finally taken, possibly using input from the ENI module, as discussed in Appendix B.

Evaluation and analyses

The evaluation performed for this activity is in two steps. Firstly, using a large-scale dataset, the activity patterns of different network slices are evaluated [5GM-D42], identifying possible complementarities in the load they impose on the network. Besides the network metrics, also other metrics such as cloud resources consumption and the related costs are evaluated. Secondly, based on these finding, it is assessed what would be the needed interfaces towards the orchestration and the network control layers that a Big Data driven resource assignment algorithm needs.

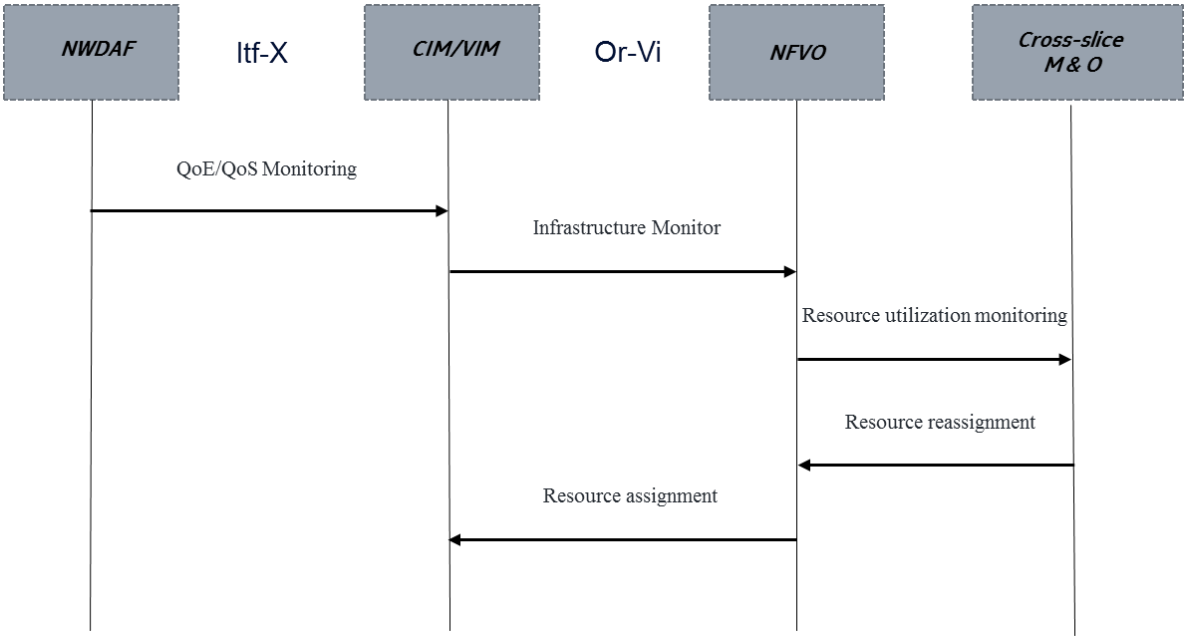


Figure 3-38: Big data resource assignment operation

3.4 *Inter-slice Management & Orchestration*

In the previous section, we have defined inter-slice mechanisms for radio resource management that take advantage of RAN and Controller layer architecture components introduced by 5G-MoNArch. In this chapter, we focus on 5G-MoNArch M&O layer. In particular, in Section 3.4.1 we describe two general frameworks for slice admission control and cross-slice congestion control. Both solutions cover the phase of setting up and commissioning a new network slice instance and therefore are closely related. Further, in Section 3.4.3, a concrete implementation for slice admission control using genetic optimisers is presented. Moreover, Section 3.4.2 describes how the slice congestion control is executed within the 5G-MoNArch architecture for deploying multi-slice networks.

3.4.1 Framework for slice admission control

Concept

This section presents a *Framework for slice admission control* that will render the slice admission procedure easier, by analysing the available infrastructure resources and their remaining capacity for the accommodation of a new slice. The proposed solution uses an existing enabler proposed in WP4, namely the *Multi-objective Resource Orchestration* enabler that is based on multi-objective optimisation using evolutionary algorithms. The enabler aims to efficiently answer the question "*Can a new slice be served efficiently using the current resources?".*

The implementation of slice admission control must ensure that after the admission of a new NSI, the resource allocation methods can optimise the network utilisation while also meeting the SLAs of each NSI. Towards this end, multiple factors must be considered, such as: slice SLA constraints, service

requirements per slice, demand of the slices, computational resources, and requested demand for the new slice. A high-level architectural diagram of the proposed approach is shown in Figure 3-39. Given the aforementioned factors as input, as well as a resource orchestration module developed in the context of WP4, the framework decides if the new NSI can be deployed or not.

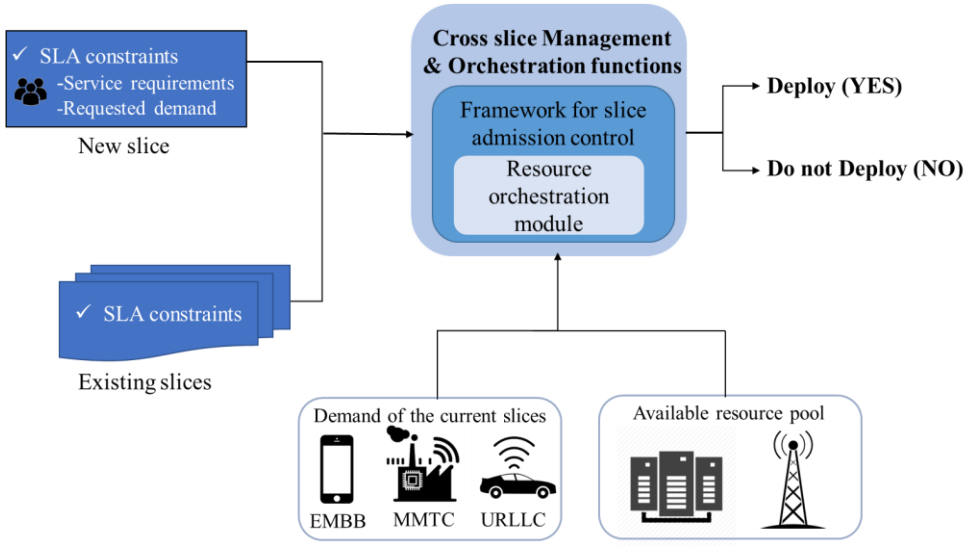


Figure 3-39: The architectural diagram of the proposed framework for slice admission control

Evolutionary algorithms for Multi-objective optimisation problems

In our approach we used an algorithm proposed by Zhang and Li [ZL07] the multi-objective evolutionary algorithm by decomposition (MOEA/D). Evolutionary algorithms such as MOEA/D are a subset of evolutionary computation. Along with other techniques and methods such as artificial neural networks, fuzzy logic or swarm intelligence it belongs to Computational Intelligence field of study, also known as soft computation, a sub-discipline of the Artificial Intelligence field [FJ08], [BHS97]. A common characteristic of these approaches is that they are can effectively represent numerical knowledge, easily adapt to changes of the input data while efficiently producing solutions in computationally hard problems.

In general, evolutionary algorithms evolve a population of candidate solutions. The fitness of each individual belonging to the population is computed through use of problem specific objective functions, and the individuals with the highest function scores are used as the basis of a new generation of the population. In MO problems, optimising the value of a single objective function can result to undesirable results with respect to the other objective functions. Instead, the solutions outputted satisfy the objectives without being dominated by other solutions, i.e. none of the values of the objective functions examined can be improved in value without the value of some other objective deteriorating. These solutions are called non-dominated or Pareto optimal.

In MOEA/D each objective function is decomposed to a single problem and then evolutionary operators are exploited to achieve combinations of the best solutions of each sub-problem while maintaining a record for all non-dominated solutions found. While weighted decomposition can be used with MOEA/D, in the proposed approach Tchebycheff decomposition is chosen since it does not require any input of arbitrary weights by the user and performs well both in convex and non-convex problems.

Real world problems often have constrains imposed by the inputs of the problem, e.g. the limited amount of resources available to the network. There are various methods employed in order to handle the problem constrains. A frequent one is using some penalty function to penalise all population solutions that violate the chosen constrains; such a penalty function with fast convergence properties was proposed by Kuri-Morales and Quezada in [KMQ98].

Position in 5G-MoNArch architecture and Protocol implications

As already mentioned, the *Framework for slice admission control* partly uses a resource orchestration module proposed in WP4 [5GM-D4.2]. Specifically, this resource orchestration module is utilised in order to identify the quantity of the resources utilised by the existing slices and how many resources are free on average, so as to take the decision of if the new slice can be served efficiently using the current resources. This means that the proposed enabler is implemented at the orchestrator level, and in particular in the Cross-slice Elasticity Mgmt. module inside the NSMF (see Sections 2.2.3 and 4.2.2) in order to manage virtual resources across different slices and make sure that the requirements of accepted slices are satisfied. The call flow of the enabler is shown in Figure 3-40. Each slice requires the allocation of both computational (i.e., cloud) and radio resources in the core and RAN. Details on the implementation of slices and the slice admission process are available in Sections 4.2.3 and 4.2.2 respectively.

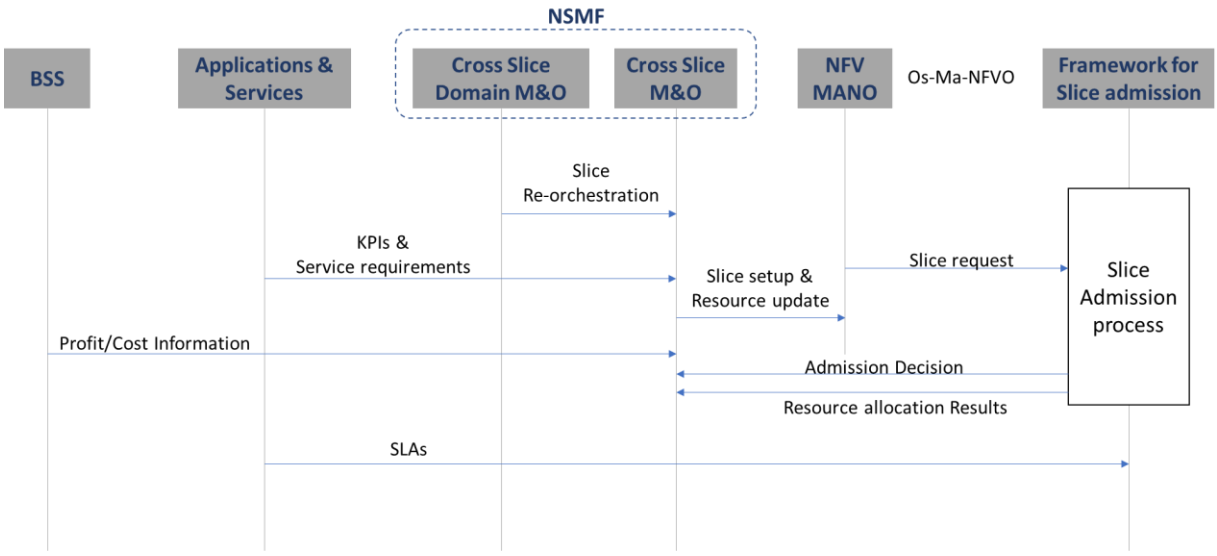


Figure 3-40: MSC for framework for slice admission control

Evaluation and Analyses

The slice admission simulator was developed in Python and uses the *PYGMO2* framework [BIY10] to solve the multi-objective optimisation problems. Since data from a real-world use case is not available, a synthetic data set was created to evaluate the proposed mechanism. For a network offering a number of different services in a number of slices, in this case web browsing, video streaming, internet of things (IOT) sensor communications, an augmented reality (AR) app and a virtual reality (VR) app, for each service request we generate a time-stamp and the resource consumption required to serve each request along with the 'aggregated' cost for these resources, see Table 3-7. Details for the dataset creation can be found in Appendix B.2.

Additionally, we use three different types of SLAs, as defined in D4.1 [5GM-D4.1]:

- **Guaranteed slice (GS):** where the offered service quality is guaranteed to be kept in between the minimum and maximum level.
- **Best-effort with a minimum guaranteed slice (BG):** the SLAs of these type of slices guarantees them receiving the minimum QoS. However, the higher QoS will be provided in a best-effort manner.
- **Best-effort (BES):** the slice will be served in best effort manner and the guaranteed on the QoS is offered beside as soon as resources become available the slice will be served

In the scenario examined we use eight slice ‘templates’ as shown in Table 3-8. The user requests describe the ideal number of users that the slice owner would serve if all resources requested are allocated (max number of users) and the minimum acceptable number of users to be served.

Table 3-7: Request resource requirements for each slice type

Service	CPU	Bandwidth	Memory	Datacentre Power	Transmission power	Cost
Web Browsing	170 -200	225 – 270	136 -160	0.85-1	1.5-1.8	0.51 – 0.6
Video Stream app	250 - 300	300 – 375	200 -240	1.25 – 1.5	2-2.5	0.75 – 0.9
IOT sensors	20 - 30	15– 30	16 – 24	0.1- 0.15	0.1-0.2	0.06 –0.09
VR app	900-1000	750-900	720-800	4.5 - 5	5-6	2.7 -3
AR app	550-950	375-450	440-760	2.75 – 4.75	2.5 – 3	1.65 - 2.85
Maximum Available	95000	91500	60000	375	610	Not Applicable

To verify our approach, we used two more methods of slice admission:

(1) Slices are admitted if there is enough space to fully accommodate them, i.e. the network is non-elastic

(2) Slices are partially admitted if there is enough space to accommodate them, provided the SLAs are maintained, i.e. the network is elastic

In both cases the sequence of slice arrival is random, and a first-come, first-serve approach is used while the slices are governed by the same SLA agreements as in the data used in the optimised example. These slice admission schemes where run for n = 100 and the KPI presented in the next slides are the averaged results. For the optimisation solution the values of the objective functions for the Pareto set where decomposed into a single value using weighted decomposition with equal weights (w = 0.2) and the solution with the lowest objective function value was chosen.

Two KPI where used in order to evaluate the results: Let $R = \{r_1, r_2, \ldots, r_n\}$ the available network resources and $C = \{c_1, c_2, \ldots, c_n\}$ the total consumed resources at some time T. Then we define the resource utilisation efficiency KPI as:

$$\sum_{i,j=1}^{n} \left(\frac{c_1}{r_1}\right) + \left(\frac{c_2}{r_2}\right) + \cdots + \left(\frac{c_n}{r_n}\right)$$

Additionally, let D be the set of requests not served per network slice $D = \{d_1, .., d_m\}$ and $P = \{p_1, p_2, \ldots, p_m\}$ a set of penalties that correspond to the SLAs that apply to each slice. Then we define the SLA violation penalty KPI as:

$$\sum_{i,j=1}^{n} d_1 * p_1 + d_2 * p_2 + \cdots + d_m * p_m$$

A flow chart of the process examined is presented in the Appendix B.2.

Table 3-8: Slice templates used for evaluation

Slice	Type	Minimum number of users	Maximum number of users	SLA	Penalty for SLA violation
Slice 1	VR app	8	8	GS	50
Slice 2	IOT sensors	300	300	GS	50
Slice 3	IOT sensors	450	450	GS	50
Slice 4	AR app	15	35	BG	20
Slice 5	Web Browsing	120	150	BG	20
Slice 6	Video Stream app	35	60	BG	20
Slice 7	Web Browsing	-	140	BE	10
Slice 8	Video Stream app	-	140	BE	10

Concerning the SLA violation penalty KPI, the proposed method produces the lowest values throughout our simulation while it competes with the greedy method for the resource utilisation efficiency KPI, as shown in Table 3-9. We can distinguish three phases: The first is between the 11:00 to 12:00 where the two BG slices are checked for optimisation to free resources for the admission of a new GS slice where the elastic non-optimised scheme produces values close but still higher than the proposed method for both KPIs. Between the 13:00 and 18:00 timeslots, the system tries to accommodate three competing BG slices and the MO optimisation method outperforms the other two methods. In the third phase between the 18:00 to 22:00 slots, where two BE slices are checked for admission and optimisation, the elastic non-optimised scheme produces results in better resource utilisation values but worse values in the SLA violation KPI.

In all cases, the methods that use elasticity outperform the slice admission without elasticity enabled. Additionally, the optimised method, on average performs better than both the greedy and unelastic admission schemes (cf. Table 3-10).

Table 3-9: Resource utilisation efficiency & SLA violation KPI values

	Resource utilisation efficiency				**SLA violation penalty KPI**		
Time	**Elastic**	**Unelastic**	**Optimised**		**Elastic**	**Unelastic**	**Optimised**
08:00 – 09:00	4.319	4.319	4.319		0.000	0.000	0.000
09:00 – 10:00	4.319	4.319	4.319		0.000	0.000	0.000
10:00 – 11:00	4.319	4.319	4.319		0.000	0.000	0.000
11:00 – 12:00	4.846	3.494	4.851		211.485	2073.267	180.000
12:00 – 13:00	4.319	4.319	4.319		0.000	0.000	0.000
13:00 – 14:00	4.756	4.170	4.823		2170.693	3282.178	460.000
14:00 – 15:00	4.731	4.156	4.823		2450.693	3344.554	460.000
15:00 – 16:00	4.742	4.190	4.823		2349.703	3223.762	460.000
16:00 – 17:00	4.774	4.153	4.823		2315.248	3334.653	460.000
17:00 – 18:00	4.695	4.000	4.823		3119.208	4478.218	460.000
18:00 – 19:00	4.974	4.319	4.949		1815.248	2600.000	1760.000
19:00 -20:00	4.975	4.319	4.949		1774.653	2600.000	1760.000
20:00 -21:00	4.981	4.319	4.949		1798.416	2600.000	1760.000
21:00 -22:00	4.978	4.319	4.949		1805.149	2600.000	1760.000

Table 3-10: Average KPI results percentage difference between slice admission schemes

	Resource Utilisation Efficiency	**SLA violation penalty**
Optimised - Unelastic	11.708	-77.782
Optimised - Elastic	0.47	-50.6
Elastic - Unelastic	11.241	-34.545

Apart from the values of the KPIs measured, the results for the actual admission process show that the proposed method accepts more slice instances that the two schemes used for comparison. As expected, the unelastic method has the most instances where a slice is not served, while in the greedy method there are instances where all slices could be served but they are not e.g., Web_2 slice in 17:00 to 22:00 timeslot, as seen in Figure 3-41.

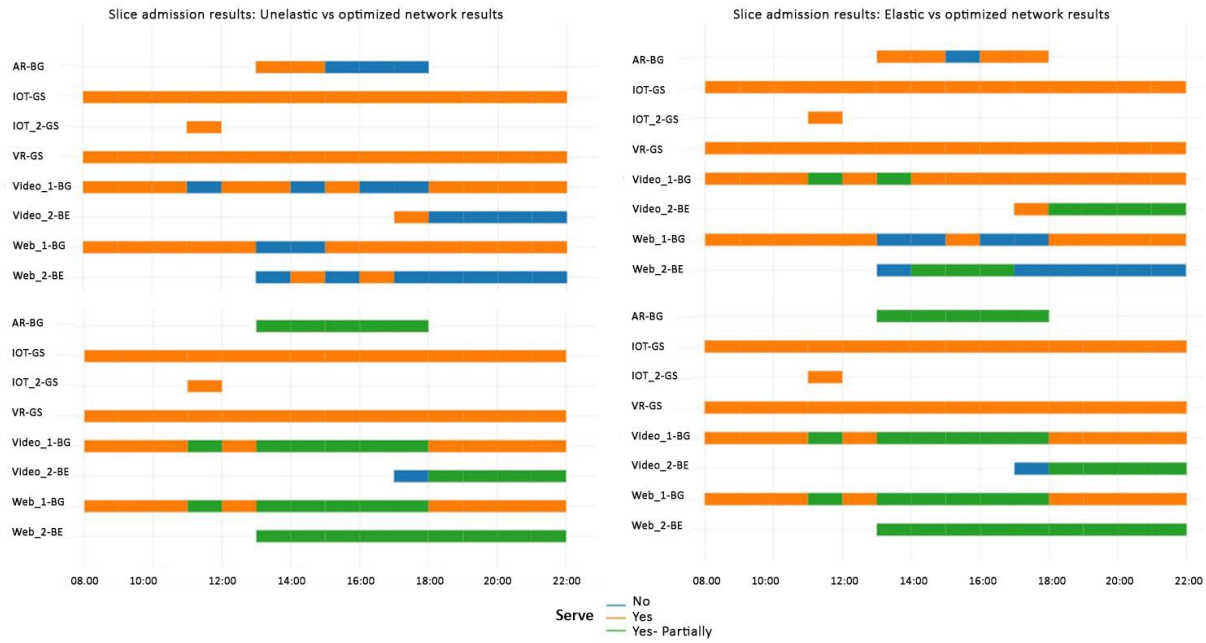


Figure 3-41: Slice admission results comparison of optimised, unelastic and elastic network schemes

3.4.2 Framework for cross-slice congestion control

Concept

Network slicing is realised by deploying a set of VNFs requiring resources such as radio, computing, and storage resources. The Cross-slice Congestion Control (CSCC) function shown in Figure 3-42 is responsible of accepting or dropping a new slice request by controlling resource availability, slice priorities, and queue state.

The CSCC may decide, based on the service level requirements of a class, to scale down the allocated resources to one or multiples slices in order to accept a larger number of requests, which have high priority. The proposed CSCC has to be able to foresee the impact of a decision on the overall system performance [PJD+15]. This intelligence is ensured by using reinforcement learning (RL) techniques that allow to make the optimal decisions maximising resources utilisation [GBL+12]. Therefore, this study focuses on Gaps #5, #6, and #12, cf. D2.2 [5GM-D2.2].

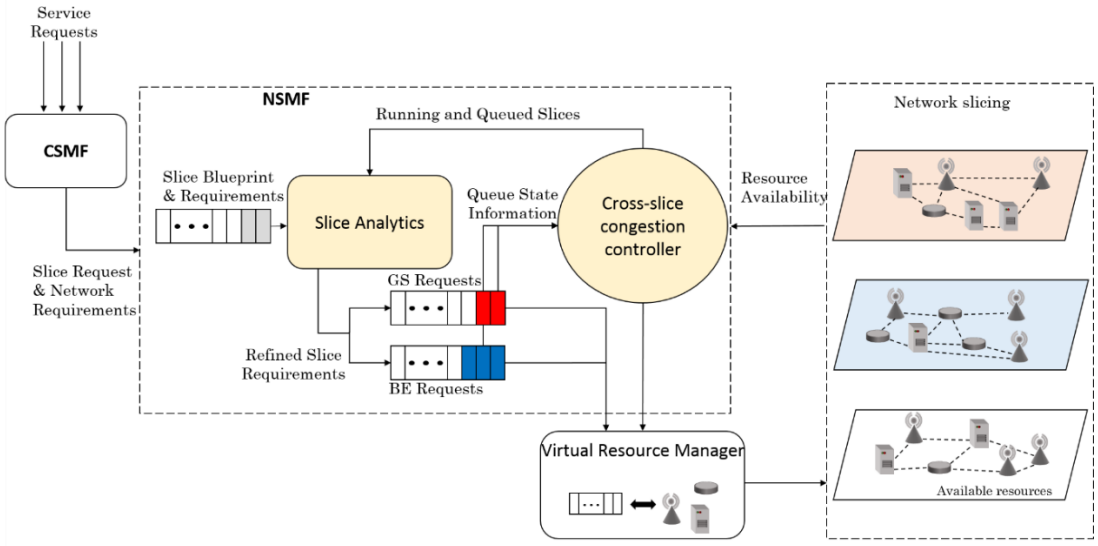


Figure 3-42: Proposed cross-slice admission and congestion control framework

Position in 5G-MoNArch architecture and protocol implications

The CSCC is implemented at the orchestrator level, and in particular in the Cross-slice Elasticity Mgmt. module inside the NSMF (see the MSC shown in Figure 3-43) as it enables to manage virtual resources across different slices such that the overall resource utilisation efficiency is maximised, dropped slices are minimised, and the requirements of accepted slices are satisfied. It is important to highlight that decision taken at the NSMF will likely to have an impact at the domain level and potentially at the level of the controller/NF. The requirements for each slice requests are related not only to the computational and storage resources to be allocated in the cloud architecture but also to the communication resources to be allotted in the 5G core and RAN. For more details on its implementation in the 5G-MoNArch architecture, see Section 4.2.3.

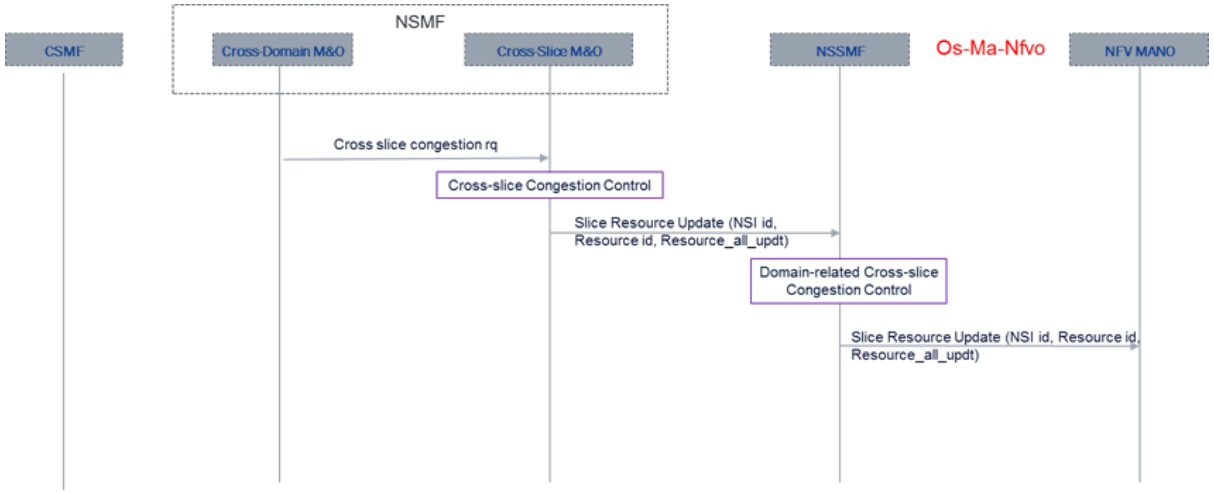


Figure 3-43: MSC for the Cross-slice congestion control

Evaluation and analyses

At this stage, two slice classes are defined: best effort (BE) and guaranteed service (GS). In order to prioritise the deployment of GS requests, a higher reward is assigned for accepting their requests. It is important to note also that negative rewards will be considered when dropping a GS request so that the policy is pushed toward deploying more GS requests rather than BE slices. In contrast to the Q-Learning scheme implemented during the first months of the project, we have now elaborated a more advanced solution based on SARSA, which integrates linear function approximator [MMR+08] in order to find a reliable solution in problems with large state space dimension. The learning procedure in SARSA with linear function approximator is described in Figure 3-44.

Input: Set of features $\boldsymbol{\phi}^{\intercal}(s,a)$, and their random weights $\boldsymbol{\theta}^{\intercal}(s,a)$.

1: Initialize all $Q_{\theta_0}(s,a) = 0$

2: **for** $t := 1$ to T **do**

3: From current state-action pair (s,a), execute action a and get reward R and new state s'.

4: Select an action a' from the action space using a policy derived from Q-learning policy e.g., ϵ-greedy.

5: Let $\boldsymbol{e}_t = \boldsymbol{\phi}(s,a)$, and

6: $k_t = R_t(s,a) + \gamma Q_{\boldsymbol{\theta}_t}(s',a') - \gamma Q_{\boldsymbol{\theta}_{t-1}}(s,a)$

7: Update weights as follows:

8: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t k_t \boldsymbol{e}_t$

9: Replace $s \leftarrow s'$ and $a \leftarrow a'$

10: **end for**

Figure 3-44: Learning phase in SARSA with linear function approximation

In the results shown in Figure 3-45, we show the improvement led by the congestion controller in terms of average reward and dropped slice requests. The results show that the proposed solution is able to improve the resource utilisation enabling to increase the percentage of BE accepted slice request without negatively affecting the performance at the GS slices.

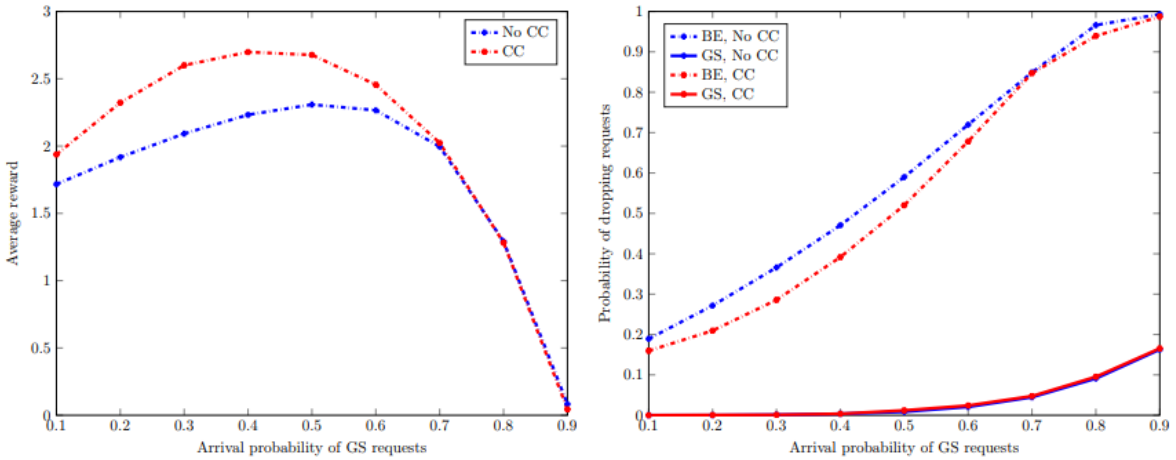


Figure 3-45: Average reward (left) and probability of dropping a slice request (right) with and without congestion control as a function of GS arrival probability

3.4.3 Slice admission control using genetic optimisers

Concept

Inter-slice control based on tenant request and binary decision:

As a specific form of public cloud service in context of sliced 5G telecommunication networks, multi-tenancy network-slicing improves the sharing efficiency and the resource utilisation rate. Generally, network resources (both physical and logical) are bundled by the MNO into slices of different predefined types. Depending on the slice type, different slices have various utility efficiencies and periodical payments. Tenants can propose requests to create new slices upon their specific demands for network resources. The MNO, according to its current idle resource pool and the network's overall utility statistics, makes an individual binary decision to every arriving request, i.e. if to accept or to decline the request. Once a request is accepted, a new slice will be created to serve the requesting tenant. The corresponding portion of network resources remain occupied to maintain the created slice, until the service level agreement (SLA) is terminated and the network slice is released. The mechanism is briefly summarised in Figure 3-46.

Concept and optimisation of admission decision strategy:

A consistent decision strategy is defined as a binary decision function $d = (\mathbf{s}, n)$, where $\mathbf{s}$ is the set of current reserved resource bundles for active slices under maintenance, and n denotes the type of requested resource bundle. d=0 means the MNO will decline the request, and d = 1 stands for acceptance. By adjusting the decision strategy, the MNO is able to statistically optimise the overall utility rate of the entire network in long term. Every consistent decision strategy can be encoded into a binary sequence, where every bit represents the decision that the MNO can freely make, given a certain combination of current network resource pool status and incoming request. An example design of such encoders is illustrated in Figure 3-47.

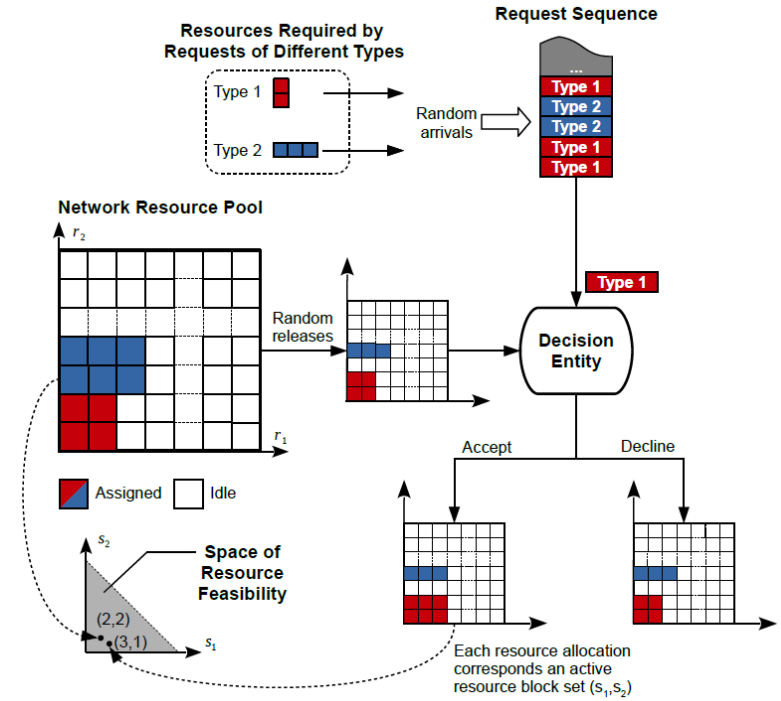


Figure 3-46: Inter-slice control based on requests and binary decisions

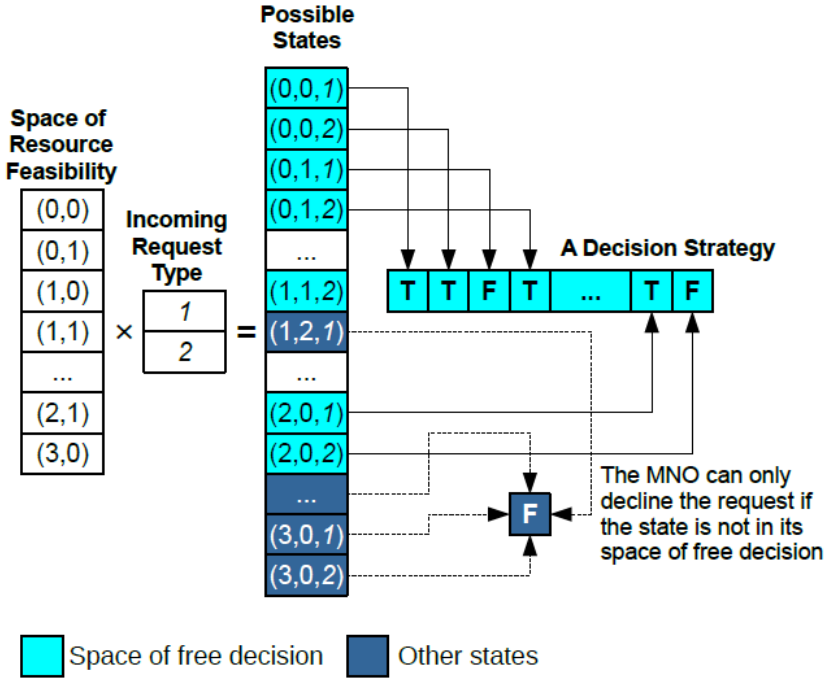


Figure 3-47: A codec design for slice admission decision strategies

Mechanisms to handle declined requests:
Declines to requests can be caused by two different kind of reasons: 1) hard constraint of the MNO's resource pool; and 2) low estimated utility rate (especially revenue rate) of the requested slice with respect to the opportunity cost (the utility of slices that may potentially be created in the future with the network resources required by the current request).

In the first case, it is impossible to immediately satisfy the tenant's demand without upgrading the resource pool. To mitigate rejecting the tenant and therefore losing the client, the MNO can offer a delayed service. Possible approaches to implement this include:

- A random delay protocol where the tenant resends its declined request after a random delay (similar to the random-access procedure in RAN);
- A queuing mechanism where the declined requests are buffered in a queue (or a pool) to wait for released resources.

The first option is easy to implement via a simple protocol, and able to deliver resource efficiency and fairness among tenants. However, it lacks control of the service priority, e.g. it is unable to realise a first-come first-served (FCFS) admission policy. The latter one with queuing mechanism is therefore generally preferred.

In the second case, besides the delayed service, a bidding mechanism can be integrated where a tenant can keep increasing the payment it offers for the requested slice until it exceeds an upper bound or eventually gets accepted by the MNO.

Position in 5G-MoNArch architecture & protocol implications

The Genetic Slice Admission Control particularly affects the 5G-MoNArch M&O layer. The overall procedure involves the NFV MANO functions, Cross-slice M&O function within NSMF, XSC in the Controller layer as well as BSS functions, applications, and services of the Service layer. The call flow for slice admission control using genetic optimisation is depicted in Figure 3-48.

Evaluation and analyses

A genetic method is proposed, where each feasible slicing strategy is encoded to a binary sequence. A population of randomly generated strategies are initialised and parallel evaluated in real-time with respect to their long-term network utility rate. A genetic algorithm (GA), which includes the three steps of reproduction, crossover and mutation, then applies to the current population, so that a new generation of candidate strategies will be created. This process runs in iterations so that the entire population evolves to a good set of strategies with high utility rates, and the best strategy in the population approaches to the global optimum through a winding process. The overall procedure is illustrated in Figure 3-48. The proposed method is model-free, can be flexibly applied to different (and even heterogeneous) constructions of utility function. It was verified to be effective, fast-converging and robust against inconsistent environment.

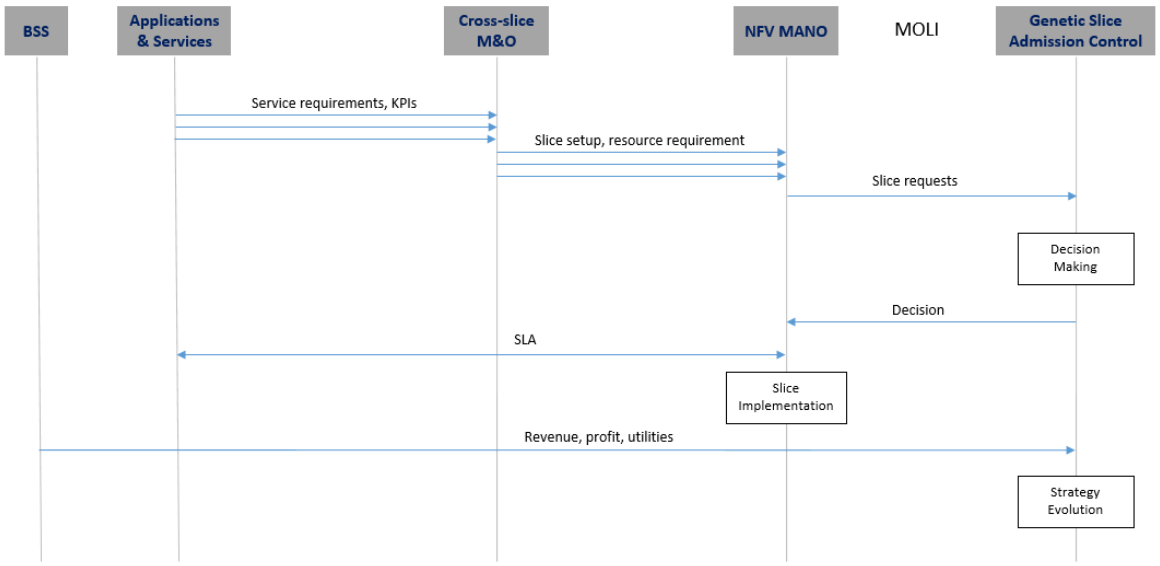


Figure 3-48: Message flow chart of deploying genetic slice admission control, where the cross-slice M&O is assisted by the CSMF for translation from service requirements to slice requirements

Figure 3-49 shows the performance evolution of the entire population of strategies. Figure 3-50 shows the performance of the best (deployed) candidate in non-consistent network traffic scenario, three "naïve" strategies and a static optimum are also tested as benchmark. More details about the proposed method, the simulation design, more evaluation results, and further analysis can be found in [HJS18].

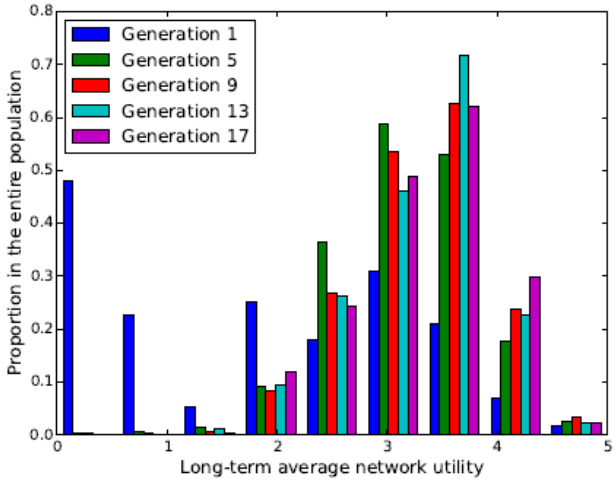


Figure 3-49: A population of 50 randomly selected slicing strategies evolves over 17 generations

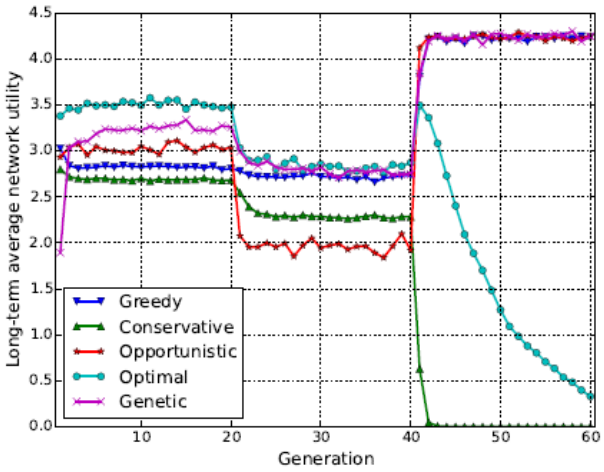


Figure 3-50: Proposed genetic optimiser remains on an almost-optimal performance level in inconsistent service environment, outperforming all "naïve" benchmarks and the static reference optimum

3.5 *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 being virtualised the cloud infrastructure providers have to develop an experimental procedure to be able to meet the QoS requirements of each VNF optimally. Scaling and elasticity decisions (either vertical, by adding more resources to the same machine or horizontal, by adding more machines) 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 this context, the

experiment-driven modelling and optimisation is a key innovation enabler for the 5G-MoNArch project filling the current gap on experiment-based E2E resource management for VNFs. However, 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 gaps (listed in D2.2 [5GM-D2.2]). From a more general perspective, the innovation element mentioned above brings a new paradigm in network management and orchestration. The two other enablers of the project (telco-cloud-enabled protocol stack and inter-slice control & management) as well as the functional innovations of the project (resource elasticity, resilience) are fed and optimised with experiment-based inputs. That is, the orchestration can be tailored with very accurate models of the real VNF consumption in terms of CPU. In the following we describe three possible aspects of the experiment driven optimisation related to the profiling of lower layers (Sections 3.5.1 and 3.5.2) and higher layer (Section 3.5.3) of the RAN protocol stack.

3.5.1 ML-based optimisation using an extended FlexRAN implementation

The aim of this experimental optimisation is to develop a Machine Learning (ML) based approach to manage the NFs, which are virtualised and implemented on the Commercial Off-The-Shelf (COTS) computers or data-centres. The two main steps are (i) profiling the NF computational complexity in bare-metal (i.e., without virtualisation) and container-based environment (since containers have relatively lower computational overhead and are more suitable for VNF in RAN with tight processing delay budget), (ii) developing ML agent(s) optimising the network based on the real-time reports and measurement.

Step 1 - Profiling of bare-metal container-based implementation of RAN:

Studying the complexity of RAN network complexity in terms of processing time is the focus of this step. After initial evaluations with Open Air Interface [OAI], the demonstrator development of RAN functions which forms the basis for the experimental evaluation has been moved to srsLTE [SRSLTE] which relies on object-oriented code design and is therefore more suitable for CU/DU splitting implementations. The EPC is still the one provided by OAI. The latest version of the demonstrator works with both srsLTE and OAI in combination with Docker.

In addition to previously conducted processing time evaluations depending on PRB and MCS selection, the evaluation is now done depending on the transport block size (TBS) of resource allocations within a transmission time interval (TTI) which corresponds to an LTE subframe. An exemplary evaluation of the downlink encoding processing time per LTE frame is given in Figure 3-51.

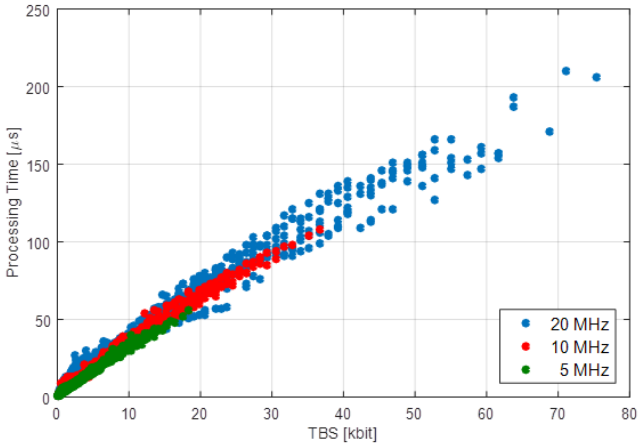
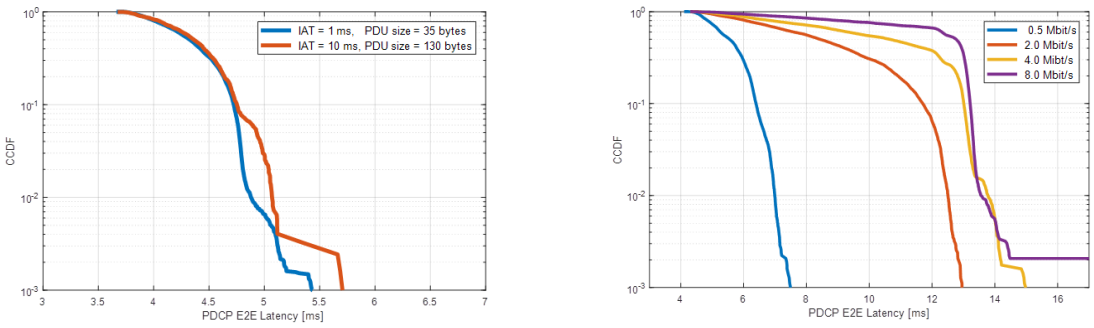


Figure 3-51: Downlink processing time for transport block encoding with srsLTE

These performance studies provide the essential input for computational elasticity algorithms based on machine learning strategies, they will for example be used for processing time requirement prediction in case of massive multiuser MIMO transmissions that involve a significantly increased number of transport blocks per transmission time interval. The latter will be evaluated by means of system level simulations.

Further development activity was devoted to the implementation of end-to-end latency evaluations between the PDCP (Protocol Data Convergence Protocol) instances of the LTE eNB and the UE. This forms a crucial basis for comprehensive studies concerning the impact of CU/DU split implementations which might involve a trade-off between processing time enhancement and increased end-to-end latency due to additional communication interfaces between protocol stack entities running on different hosts. Figure 3-52 (A) shows an exemplary downlink latency evaluation in case of traffic load with an exponential distribution of packet inter-arrival times, and Figure 3-52 (B) shows the corresponding measurement results with traffic load generated with iPerf3 [IPERF].



(A) ***(B)***

Figure 3-52: PDCP latency evaluation

All measurements within the protocol stack NFs are directly written into a database connected to the RAN which allows for live monitoring direct adaptation of individual NF parameters such as modulation and coding scheme or assigned number of resources within a transmission time interval. The communication between the data base and the virtualised NFs makes use of the messages described in Section 3.1.2.

Step 2: Developing Machine Learning Algorithm

In the next step, the reports and measurements from the experimental setup are fed to the machine learning agent. The aim is to first update the models with the measurement and report on the run-time. The processing time of a specific function is highly dependent on the implementation techniques (e.g., number of memory access). Hence, the ML-based approaches are needed to adopt the complexity models of NFs to the reports and measurements. These models can be used later for optimisations and improvement of networks or used by other ML agents.

In addition, ML-approaches are one of the candidate solutions for the elastic allocation of resources (radio or computational) to different network slices with different requirements. Both radio and computational resource management are dependent on traffic demands. Shifting resource management from the passive mode, i.e. observing a change in traffic demand and react to the change, to the active mode, where the changes are predicted, can improve the resource utilisation.

Estimating the processing time using a Deep Neural Network (DNN):

As it is presented in the former section, the processing time of the RAN NF may vary as the channel condition varies and different MCS (Modulation and Coding Scheme) indexes are selected. Processing the subframes with higher MCS index are relatively more complex. Figure 3-53 presents the measured processing time for different MCS index when the maximum MCS index is limited to 5, 10, 15, and 20.

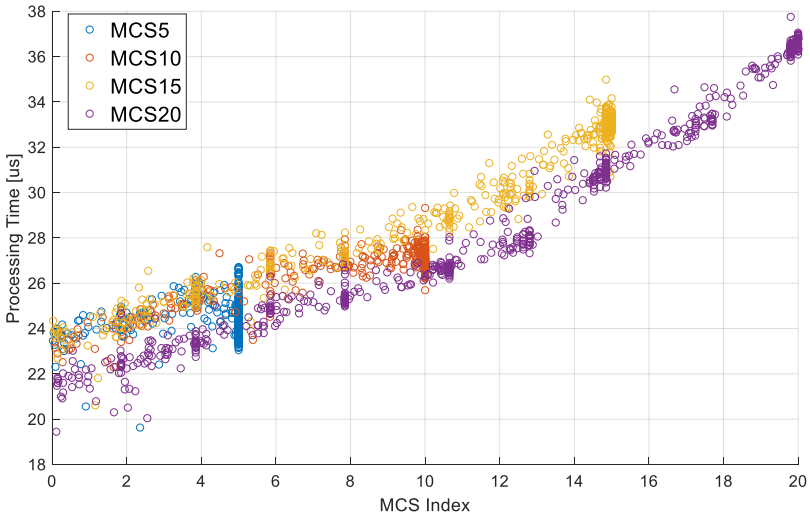


Figure 3-53: Measured processing time for different MSC index

Based on the provided measurement, a deep neural network is trained to estimate the processing time per PRBs based on the selected MCS index and applied MCS limitation. The DNN has seven dense layers with Leaky ReLU activation functions, where the number of neurons in the layers are relatively 300, 600, 1500, 1500, 500, 300, 1. Figure 3-54 illustrates the estimated processing time versus the measured data used as the input for two cases where the MCS index is limited to 10 and 15. It is apparent from the plots that the DNN have a fair estimation on processing time given the variance of input data.

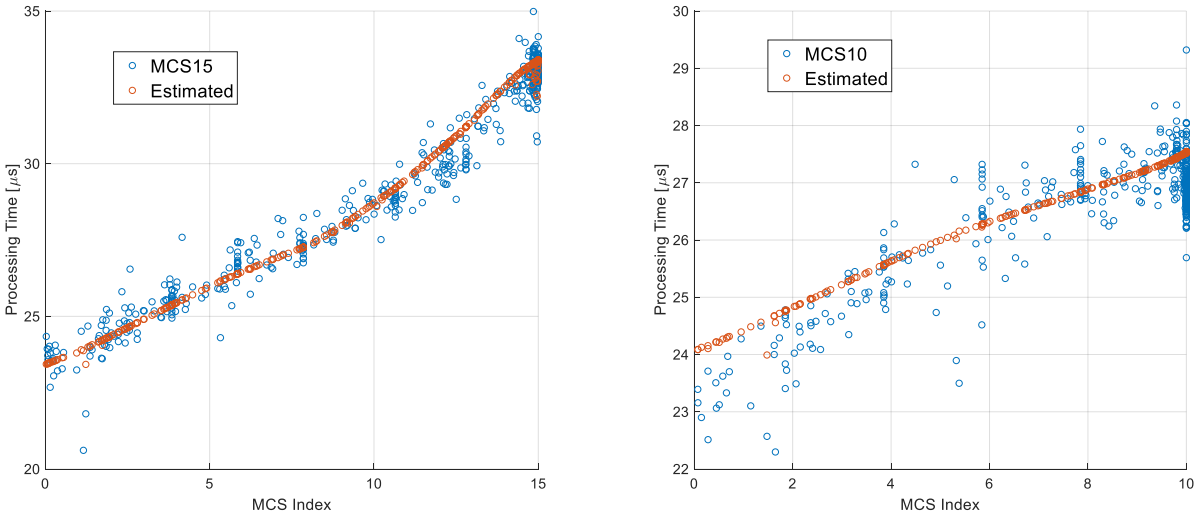


Figure 3-54: Estimated processing time versus the actual measured data

Detecting the Computational Congestion with Fuzzy Logic DNN

In the next step, the estimated processing time has to be used to determine if the measurement report coming from the network indicate any computational congestion. The goal is not only to know if there are any congestions or not (i.e. binary logic) but to have more detailed decision. Hence, four classes of congestions are defined, which are “No Congestion”, “Low congestion”, “Medium Congestion”, and “High Congestion”. To decide on the network situation based on the processing time readings, another DNN is designed that get the MCS index, MCS index limitation, and the estimated processing time and marked the input with one of the four possible labels. This DNN have five dense layers with Leaky ReLU activation functions. The output layer uses a softmax activation function.

To train the DNN, three set of random processing delay with Gaussian distribution are generated. While they have identical variance, their mean values are different. These delays are used to generate four set of training data sets.

Figure 3-55 presents the output of the DNN deciding the computational congestion for in the four aforementioned situations. The plot shows that in general the DNN is successful in determining the state of the computational pool. However, there are more inaccuracy in decision of DNN when there are low load and the MCS index in less than 5.

Finally, the output of this step enables enforcement of different policies in to manage the computational resource for cloud-based protocol stack. One may consider scaling out when there is high load and scaling down when there is low load.

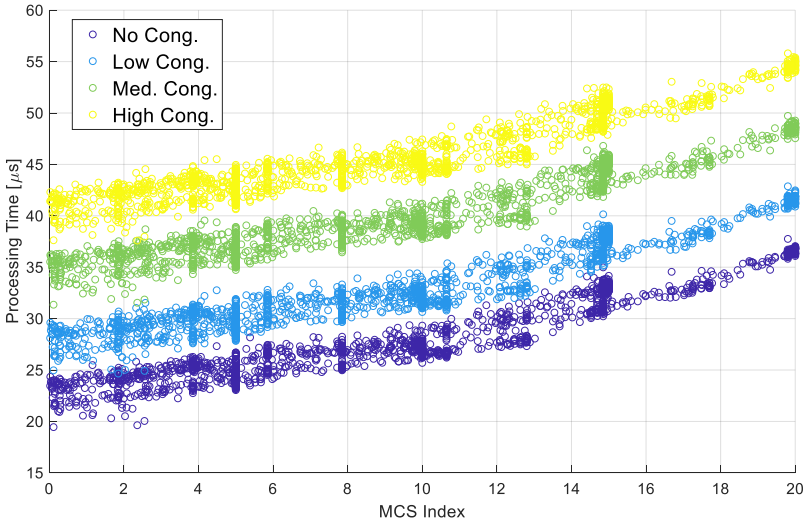


Figure 3-55: DNN output labelling the measured processing time

3.5.2 Computational analysis of open source mobile network stack implementations

Experiment-driven optimisation necessarily builds on top of thorough measurements of software modules. 5G-MoNArch uses, especially for the testbeds, a mixture of open source and ad-hoc developed solutions. Therefore, some well-known implementation of the RAN: OAI [NMM+14] and srsLTE (SRS) [GGS+16] are measured.

While in the previous version of the deliverable we observed the CPU consumption of different open source mobile network stack implementation, in this section we focus on the internals of the VNF implementation, to understand the CPU dynamics of each implementation.

For these experiments we use an LTE USB dongle as user equipment (UE) and a software RAN VNF that includes srsLTE and a software-defined radio (SDR) USRP B210 as RF interface. While we set up UDP traffic at full buffer with good SNR channel conditions. On the other hand, we limit the available CPU through the cgroup API available in Docker, from two CPU cores downwards. We also fix the used Modulation and Coding Scheme to different ones and study the decoding time. Results are depicted in Figure 3-56.

The results motivate the need from a cloud-enabled protocol stack, empowered with algorithms such as the ones proposed in [5GM-D4.2], namely the “Elastic RAN Scheduling”. Decoding times are constantly growing when the available CPU starts to be scarce (maybe due to orchestration policies), being higher when MCS are higher as well. This has a translation into the perceived throughput. For instance, when the CPU decreases below 0.5 cores, it is actually better to use a lower MCS, having then a cloud-enabled protocol stack.

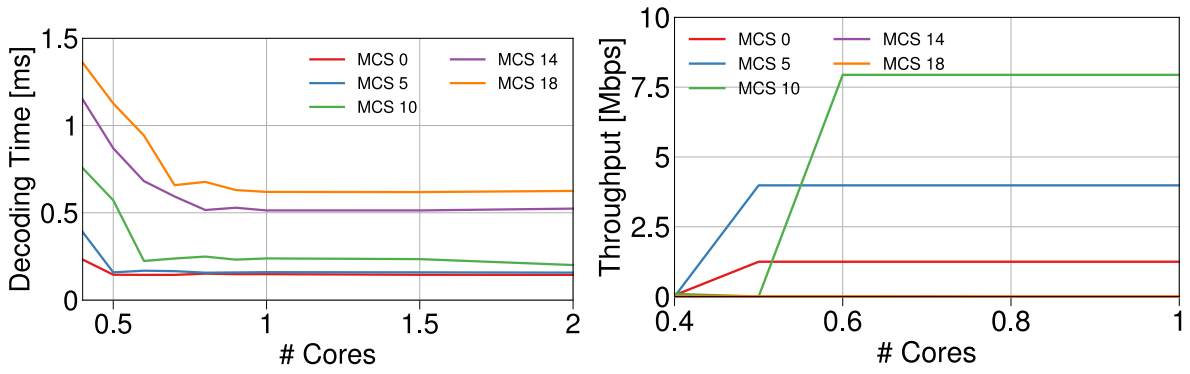


Figure 3-56: Decoding time (left) and Throughput (right) vs CPU Share

3.5.3 Measurement campaigns on the performance of higher layers of the protocol stack

While the previous two sections have dealt with the profiling of lower layer functions (i.e., that relate to decoding of subframes), in this one, we focus on the profiling of the higher layers of the RAN protocol stack, exploiting the implementation made available for the Touristic City testbed.

As has been described in [5GM-D2.1], to take advantage of the experiment-driven modelling and optimisation in a cloud enabled network, new challenges arise. A key requirement is 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 (a key choice for the cloud-enabled networking). As one of the main options for functional split at the RAN protocol stack assigns protocols above the MAC layer to a CU, in this work, more emphasis has been placed on the higher layer protocols in RAN, i.e., PDCP and the RLC in a virtualised environment.

The evaluation of such an approach can be based on the actual testbed implementations. Key target is the quantification of the computational and memory resources (CPU/RAM load) that are consumed by the higher layer protocols in the RAN protocol stack as well as to investigate the impact that a function split at the RLC level can provide in terms of delay to a provided service.

In this context, the PDCP/RLC functionality was implemented in a stateless way using python on top of the following SW/HW:

- Processor type: CPU(s) 4 x Intel(R) Core (TM) i7-6700 CPU @ 3.40GHz / 1 core was allocated to the VM that run the protocols
- Cache memory size: 8 MB SmartCache
- Memory assigned: 4 GB
- Hypervisor/OS PROXMOX Virtual Env 5.1-41 / Ubuntu 16.04.4 kernel:4.4.0-31 generic

Measurements were extracted in two different scenarios. First, with 4K video streaming, to assess the CPU and RAM consumption while a demanding application is running, and second with increasing traffic using Iperf to depict the relation of load and CPU consumption. The results are depicted in Figure 3-57 and Figure 3-58, respectively.

Takeaways, from initial campaign are listed below:

- The implementation affects the performance, meaning that optimisations of the code can provide elasticity (as defined in WP4 of the project) of the PDCP/RLC functions. For instance, in the current implementation the allocated CPU supports up to 65Mbps (after that the CPU is fully loaded and unstable).
- The split of higher layers from the MAC/PHY adds CPU load due to the interfacing between the two network nodes. It is noted that, in the current set up the load for interfacing takes approx. 90-95% of the CPU load and similar percentage of the total latency.

- The CPU load can be used as a trigger for applying resource elasticity. However, it is not as an indication for the service performance. Practically, the impact of overloading the CPU on the service performance, as revealed from the tests, is not visible to the app layer till the point that the packets cannot be served. This is due to the notion of the targeted functions, which actually perform a kind of 'forwarding', compared to more sophisticated NFs like the MCS selection or the decoding.

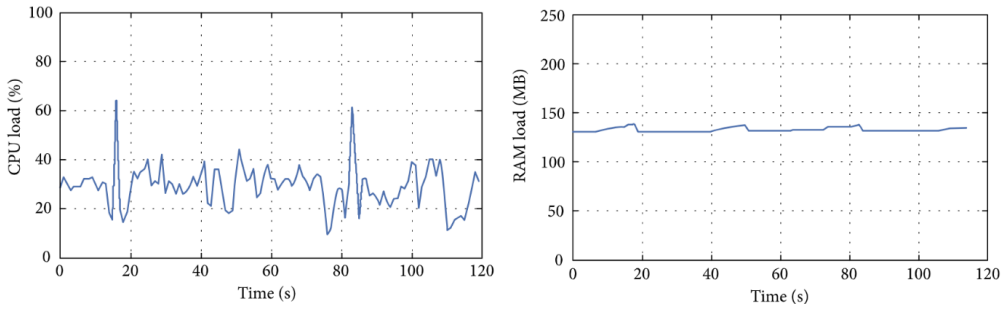


Figure 3-57: CPU and RAM consumption from PDCP/RLC functions including the required interfacing to forward the packets to lower layers that reside in a separated network node

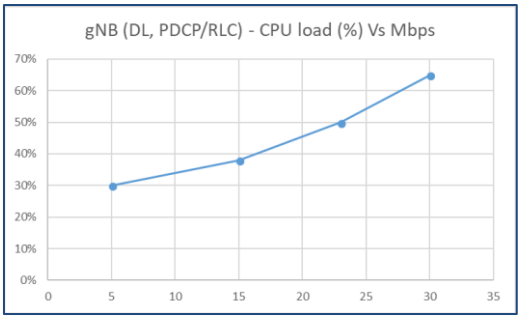


Figure 3-58: CPU consumption from PDCP/RLC functions for increasing input traffic (the measurements include the required interfacing to forward the packets to lower layers that reside in a separated network node)

4 Architectural Extensibility and Customisation

In Chapter 2 and 3, we have respectively presented the 5G-MoNArch overall architecture and its enabling innovations. These chapters detail the elements, functions, and interfaces to provide baseline slice deployment, control, and orchestration. In contrast, this chapter details the identified innovation elements and network functionalities towards flexible extensibility and customisation of both network slice functionality and network infrastructure. In a first step, the chapter describes the general framework, where the 5G-MoNArch Network Slice Blueprint concept is used to design network slices incorporating customised NFs. Subsequently, based on the 5G-MoNArch overall architecture, and the M&O functionalities defined in Section 3.4, we present the procedure and the signalling for the 5G-MoNArch Network Slice Allocation and Network Slice Congestion Control concepts, which demonstrate how a single common infrastructure can efficiently host multiple network slice instances. Finally, the presented concepts are applied to the two 5G-MoNArch testbed use cases. For the Smart Sea Port scenario, specific NFs for resilience and security extend 'standard' URLLC and mMTC slices to be deployed on a common infrastructure in the city of Hamburg. For the Touristic City scenario, specialised NFs for resource elasticity customise typical eMBB functionality towards the requirements of interactive consumer applications.

4.1 *General means for extensibility and customisation*

In 5G-MoNArch, the Network Slice Blueprint concept is the major means to generate customised network slices that are capable of realising the performance and functional requirements of the addressed service.

GSMA has started a work to standardise a GST, where every Network Slice can be fully described by allocating values (or ranges of values) to each relevant attribute in the GST. GSMA would also standardise some GST for specific vertical use cases. 5G-MoNArch has collaborated with GSMA Network Slicing Task Force, where 5G-MoNArch vision on the slice blueprint design starting from such a GST has been highlighted and inputs to the GST design, e.g., in terms of 5G-MoNArch testbed use case requirements have been provided. Accordingly, 5G-MoNArch work on the slice blueprint design shows how a GST can be adopted for the actual E2E slicing realisation and network slice deployments.

As illustrated in Figure 4-1, 5G-MoNArch foresees that starting from the GST, an Operator could make some more customisation to build up some more specific basic templates that consider its own specific network deployment. In this way, the Operator could have a catalogue of slice templates to start from when the Operator has to deploy a specific NSI to meet the customer specific requirements for a network slice. Figure 4-1 provides the illustration, where 5G-MoNArch slice blueprint is placed in the GSMA flow (a), and also details the process that creates the slice blueprint in 5G-MoNArch environment (b).

For each vertical (e.g., related to eMBB and URLLC services), the Operator starts from the specific GST for that vertical, customises it according to the peculiarity of the network infrastructure and creates a catalogue. When a customer asks for a communication service, the Operator picks the most adequate template from the catalogue and extends and customises it according to the customer specific requirements.

This process, ideally automated as far as possible, leads to the concrete definition of an NSI in terms of specific configurations that are used by the M&O layer to deploy the NSI. This concrete definition of the specific NSI requested by the customers, with specific extensions and customisation, is the 5G-MoNArch Network Slice Blueprint.

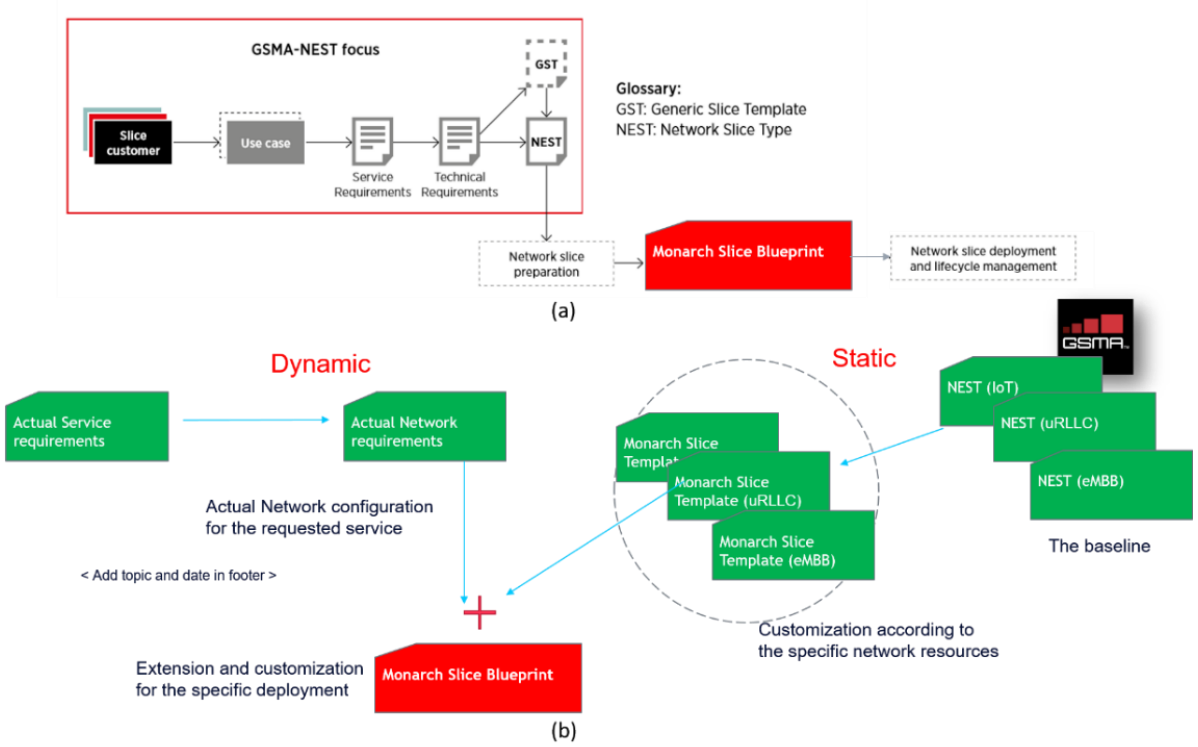


Figure 4-1: 5G-MoNArch network slice blueprint for slice extensibility and customisation; GSMA focus and 5G-MoNArch slice blueprint interaction on high-level (a), and detailed development of the 5G-MoNArch slice blueprint comprising GSMA NEST

4.1.1 Network slice description

Network Slicing aims to provide the tenant with a network solution that meets the demands of the tenant and the tenant's applications to the greatest possible extent. Before a new network slice can be designed, tenant and service provider / network operator have to align on what kind of service is expected by the tenant and how the new network slice shall look like. In this alignment, various aspects need to be discussed and some parameters have to be agreed. While GSMA is still working on the definition of its GST aka Network Slice Template (NEST), the following groups of parameters can be identified to be relevant.

Purpose of the network slice for the intended application:

The purpose that a new network slice is intended for, does not go directly into the slice blueprint that is to be designed. However, a sound understanding of the application intended to run in the network slice helps the service provider / network operator to imagine how ICT technology contributes to the overall application and which communication services the network slice shall provide. In this way, the description of the intended application is an important contribution to the design of the network slice and for the best possible integration of ICT technology and application.

Slice Type:

In 3GPP [3GPP TS 23.501], currently three Slice / Service Types (SST) have been predefined: MBB, URLLC, and MIoT. Aside of that, further tenant- / purpose-specific types can be defined in addition. GSMA intends to specify a NEST for each of these SSTs. The selection of an SST respectively NEST according to the purpose of the application supports the discussion between service provider / network operator and tenant, with a high-level characterisation of the network slice and predefined parameter values / value ranges.

Functionality of the network slice:

In many cases, tenants may expect data transmission with a certain performance, but are not interested to know or influence the detailed network functionality. Therefore, we propose to divide network functionality into two categories: Tenant-specific functionalities and generic network functionalities.

Regarding tenant-specific functionality, e.g., details on the expected kind of data transmission and terminal mobility, should be specified, as well as details on requested data processing and storage capabilities. Furthermore, the tenant should state if the tenant expects the service provider / network operator to take specific security measures, like encryption, tunnels, firewalls, etc.

Generic network functionality comprises NFs covered by standards like 3GPP. In many cases, it may be sufficient to state that the tenant expects, e.g., data transmission "as usual" according to 3GPP Release 16. Less common deviations, e.g., the need for unidirectional transmission or multi-cast, should be mentioned explicitly in this context.

The description of the expected network functionality is the basis to identify the required NFs as well as their interconnection within function chains (NF Forwarding Graphs). In the slice design process, this functional description has to be transformed into a description that can be evaluated and implemented by the orchestrator.

Geography:

Network Slices may cover the whole access network of a network operator, selected geographical parts of that (e.g., the port area of Hamburg in case of the Hamburg testbed) or even integrate access networks of other network operators (e.g., in case of a global network slice). The description of the terminals' locations is needed to configure which radio cells shall be accessible in the context of a certain network slice. This allows to exclude terminals from service when they are in the wrong location. Furthermore, the geographic distribution of terminals is important to estimate traffic load and resource consumption of the network slice on a per-cell basis.

In the transport network and in cloud data centres, it may be necessary that a network slice is forced to use certain resources (affinity rules) or to abstain from incorporating certain resources (anti-affinity rules), e.g., to comply with specific security, data privacy or resilience demands.

Traffic Profile

The data traffic that is generated by terminals and servers is an important influence factor on the resource consumption of a network slice. For the design of a new network slice, it is therefore essential to assess the expected traffic volume as precisely as possible. Relevant data are

- the average as well as the peak data rate per terminal,
- traffic in UL as well as in DL direction,
- size and frequency of the generated data packets, and
- dependencies between multiple terminals in the generation process of the data packets.

Aside the data rate, also the sensitivity of applications with respect to latency, packet loss and transmission faults affect the resource consumption of a network slice.

Terminal devices

The access technologies that are supported by the terminals obviously limits the choice of access technologies that have to be supported by the network slice.

The expected number of terminal devices in combination to the traffic profile influences the total traffic load and thus the resource demand of the network slice.

Security / privacy requirements:

With respect to security and privacy requirements, two aspects are of interest: On one hand, the expected threats and vulnerabilities, and on the other, the protection measures that the tenant expects the service provider / network operator to take.

Regarding threats and vulnerabilities, in particular, those should be mentioned that result from the specific purpose of the network slice and that go beyond the "usual" threats. A protection of the network slice against the "usual" threats shall be done anyway, of course.

Topics to be addressed with respect to the protection mechanisms are the extent and mechanisms of slice isolation (physical separation, logical separation by tunnels, etc.) against other slices as well as mechanisms for authentication and access control. Further topics are encryption and integrity protection if they are not provided by the tenant on application level anyway.

Operational aspects:

In the context of operational aspects, it has to be specified to which extent the tenant wants / has to manage the network slice on his own. Depending on the business model of tenant and service provider / network operator, different offer types are possible [5GN-D3.3]. For example, a tenant can restrict himself to managing only functionality on application layer, while the network management is done completely by the service provider or network operator. Alternatively, a tenant could also manage the function chains in his slice and control slice-internal resource allocation.

Frequency and duration of slice initiation affect design and performance of the M&O system. A high frequency or a short initiation time imply that orchestration has to be highly automated.

Pre-emption affects what happens in case of resource shortage when the slice is initiated. Important slices, e.g., for public safety communications, may be assigned the "capability" to "steal" resources from other, less important and thus "vulnerable" slices. The priority of a network slice is relevant for resource sharing between network slices in case of resource shortage during their runtime (i.e., independent from slice initiation). Furthermore, it has to be defined which performance indicators / key performance indicators can be monitored by the tenant or are reported to the tenant, and how the tenant is informed about actual alarms and fault situations.

In order to investigate whether the above groups of parameters are suitable to describe a tenant's expectations on a new network slice, this description methodology has been applied on the network slices setup in the 5G-MoNArch testbed in Hamburg. During the project, the set of parameters has been improved. For example, asking for details on expected RAN or CN functions will likely overburden most tenants; they simply expect wireless connectivity. This led to the differentiation between generic vs. tenant-specific network functionality as explained above. Later on, the terminology has been adapted to the wording defined by GSMA [GST], as far as possible. In doing so, many similarities have been identified, but also enhancements to the GSMA GST. Details are to be provided in 5G-MoNArch deliverable D5.2 [5GM-D5.2].

4.1.2 Representation of slice geography

The 5G-MoNArch project has exemplarily developed a method for the description of the area that shall be covered by a network slice[7]. Slices can differ with respect to their functionality, the performance and service quality, and their coverage area. As said in the previous section, a proper description of the geographical area of coverage is beneficial for multiple reasons: For the resource planning of a new network slice, for offering regionally restricted network services, and it may be helpful to hinder terminals to access a network slice when they are in the wrong location. Thus, a good method for describing a slice's geographical area is valuable. According to [3GPP TS28.541], a network slice comprises one or more slice subnetworks, and each of these again may consist of one or more coverage areas. Figure 4-2 shows this object hierarchy.

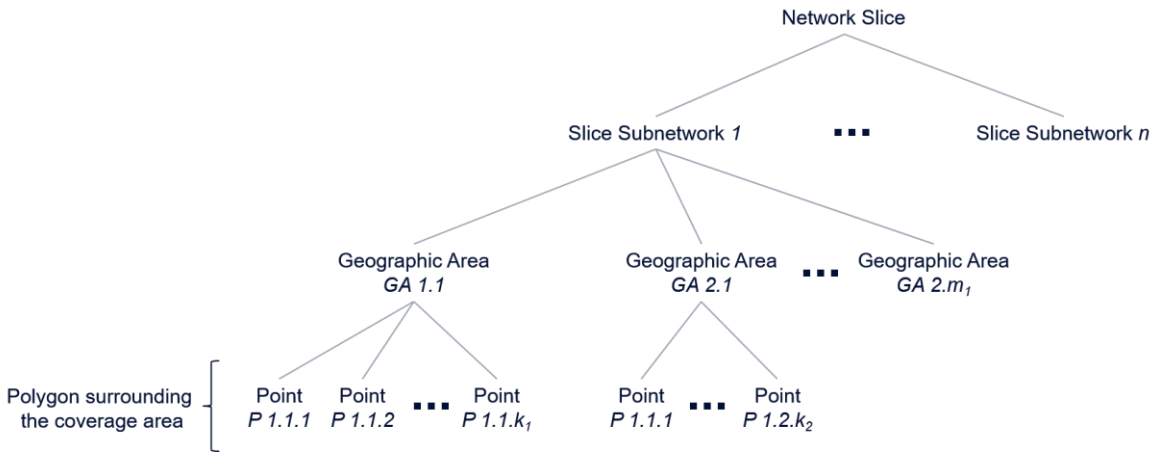


Figure 4-2: Hierarchy of geography-related slice objects

[7] Development of a full specification of all parameters for a slice description is a task for standardisation bodies like 3GPP or organisations like GSMA.

Each geographical area can be assigned multiple parameters (non-exhaustive list):

- Radio coverage inside the coverage area may be required, forbidden or permitted. Typically, the tenant defines where he requires radio coverage. Around this area, there may be an area where no radio coverage is required by the tenant, but which is partially covered by base stations needed to cover the required area. Areas with forbidden coverage may be needed to protect against unwanted access via the radio interface. Radio coverage can be described by a binary variable (available or not available), discrete signal levels for achieving pre-defined data rates (e.g., low, medium, good, very good), or by means of more fine-grained numerical parameters, e.g., SINR.
- Permitted access technologies, frequency bands and similar radio parameters should be specified per geographical area.
- Density and traffic profile of the terminals inside the geographical area are necessary for planning air interface capacity and amount of processing resources consumed by a network slice.
- The geographical boundaries of each geographical area have to be described.

Listing parameters for network slice description as in 5G-MoNArch deliverable D5.2 [5GM-D5.2] or as in GSMA's GST [GST] suggests a simple data structure, in which all parameters are practically independent. However, this impression is misleading: Since there may be multiple geographical areas with different access technologies, radio frequencies etc., a data structure is needed that takes the underlying relationship between parameters into account.

The boundary of a geographical area can be described in several ways. The easiest way is the use of existing pre-defined boundaries, e.g. national boundaries. For industrial applications, however, it may be more desirable to restrict coverage to e.g. a tenant's production plant. Then a description by means of a polygon is needed.

Specification 3GPP TS 23.032 [23.032] provides a suitable data format. It is based on the World Geodetic System 1984 (WGS84) and has multiple data types, among others point, polygon and points with altitude value. Coding rules are defined in the specification for the mapping of WGS84 coordinates into integers. The main drawback of this data format is its limitation to maximum 15 points per polygon. It is easy to imagine situations where polygons with 15 points are too coarse to describe a geographical area with sufficient precision.

An alternative not suffering from this limitation is the shapefile format defined by ESRI [ESRI]. It has become a de-facto standard, and files in shapefile format can also be read by network planning tools. Shapefile knows multiple data types, including polygons. Polygons are sets of rings, where each ring consists of at least 4 points. A ring is well-suited to describe the boundary of a geographical area. Points and polygons can also comprise altitude values. As in specification 3GPP TS 23.032, coordinates are measured according to the WGS84 coordinate system. Coordinate values are represented in IEEE double precision format (8 bytes / 64 bits).

4.1.3 Slice design process

The objective of the blueprint design is to identify and pre-allocate suitable infrastructure resources in a way that during runtime, the slices will satisfy the above requirements of its tenant. Aside the requirements of the tenant, there are further inputs to this design process: The available hardware and software (i.e., VNFs incl. details on their implementation and configuration), information on other slices running on the same infrastructure, as well as business policies of the service provider / network operator. Figure 4-3 shows schematically and on a high level how such a design process can be structured.

In the upper part of the process, it is determined which and how many resources are needed for the implementation of the slice. In the lower part, the available resources for the slice setup are identified. For the blueprint, some of the available resources have to be selected and integrated such that later during runtime the slice will meet its tenant's expectations. Possibly several iterations are needed for this resource pre-allocation.

For the generic network functionality that is supported in the service provider's / network operator's infrastructure, VNF descriptions should be available as template catalogue. These VNFDs and the

corresponding VNFFGDs depend on their implementation by the software vendor as well as on their configuration by the service provider / network operator. Hence, they will typically differ for each service provider / network operator. The tenant-specific network functionalities are combined and integrated with the generic functionalities and yield a complete functional setup of the slice. Taking the traffic profile into account, the necessary resources for the execution of this functional setup can be determined.

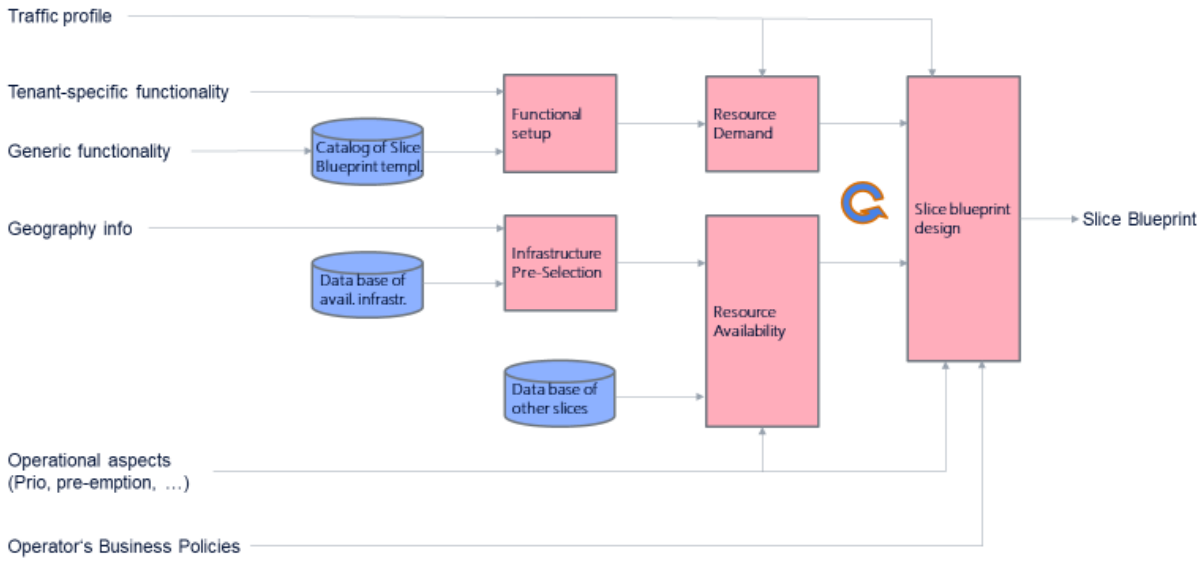


Figure 4-3: Transformation of slice description into slice blueprint

In parallel, the Infrastructure Pre-Selection determines which radio cells are needed to cover the geographic area requested by the tenant. This requires a data base of the available antenna sites (including their radio frequency layers) respectively a map of the geographic areas covered by the radio cells. Furthermore, if the tenant's application requires a low latency, the distances in the transport network must not become too long. Data centres that are too far away from the base stations and incur a too large delay in the transmission network must be excluded from the selection of suitable infrastructure resources. Furthermore, affinity rules and anti-affinity rules must be considered in selecting the suitable data centres.

As soon as all suitable infrastructure components have been identified, it has to be checked which transmission, processing and storage resources they can provide and to which extent these resources are already occupied by other slices. This yields the resource availability.

For the slice blueprint, available resources have to be selected such that the resource demand is met. If this is not possible at all or the selection is not desirable from the service provider's / network operator's viewpoint, additional measures like changing the resource allocations of already existing slices can be taken into consideration. In this case, Operational Aspects like pre-emption and priority rules for the new slice as well as for the already existing slices need to be observed. Furthermore, if the slice design process results in several possibilities or allows to vary parameter settings in a certain range, business policies of the service provider / network operator can be used to select the preferred design and optimise the parameter setting.

4.1.4 5G-MoNArch network slice blueprint concept

The 5G-MoNArch Network Slice Blueprint defines the Network Slice Instance in terms of NFs, their interconnection and configuration according to a specific service request. The chosen approach to define 5G-MoNArch Slice Blueprint is to refer and enhance what is already defined by SDO. Based on the SotA, a few considerations led to the definition of the 5G-MoNArch Network Slice Blueprint:

- 3GPP Network Management is well defined for release 14 but release 15 slice modelling is not yet complete and there is no clear indication on Slice modelling.

- ETSI MANO, instead, provides a consolidated documentation for the Network Service Descriptor and for the elements that compose the network service

The analysis highlighted several benefits in adopting a MANO based Slice Blueprint. First and foremost, the fact that the resulting blueprint would directly be MANO compatible. Since MANO represents a fixed point in the 5G-MoNArch architecture, this would allow to minimise the number of intermediate functions that would need to translate the 5G-MoNArch Slice Blueprint into a MANO Network Service Descriptor. The MANO model for NS description is based on a set of tables that represent the descriptor for an entity, e.g. MANO defines the Network Service Descriptor and the Virtual Network Function Descriptor. 5G-MoNArch uses the same "Descriptor" based approach for network slicing. 5G-MoNArch network slice blueprint is implemented as a collection of Descriptors (MANO style) that are tables containing all the needed information to deploy an NSI.

Because of the composition of an NSI, 5G-MoNArch Network Slice Blueprint is composed by the following descriptors, see Figure 4-4:

- The NSI: it is described by the 5G-MoNArch Network Slice Descriptor (MNSD)
- The NSSI: it is described by the 5G-MoNArch Network Slice Subnet Descriptor (MNSSD)
- The NF: it is described by the 5G-MoNArch Virtual Network Function Descriptor (MVNFD)
- The connectivity: it is described using the standard defined by MANO for connectivity using Virtual Link Descriptor (VLD) and VNF Forwarding Graph Descriptor (VNFGD).

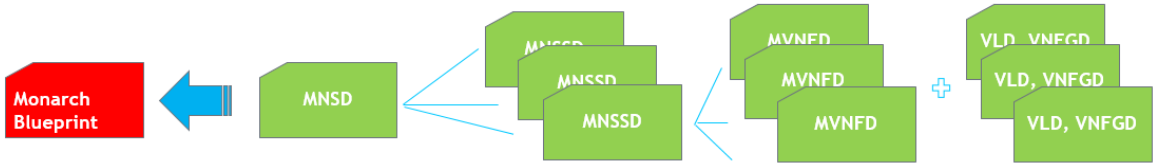


Figure 4-4: 5G-MoNArch network slice blueprint composition with descriptors

From a management point-of-view, an NSI is a collection of NSSI that are defined by the information to setup the virtualised part of the contained NFs, the configuration on the application-level of the NFs (both physical and visualised) and by the information about the connectivity among the NFs. Those are all the information needed to provide a Network Service plus the information about the configuration of the application part of the NFs.

Since the first information needed maps directly to the ETIS NFV MANO Network Service Descriptor, in 5G-MoNArch it was agreed to create the blueprint as an extension of the NSD with application-level information needed to provide the requested service. Figure 4-5 describes the idea behind the extension of the NSD with the inclusion of Application information and configuration in order to obtain a full Blueprint of the Network Slice Subnet.

The connectivity between network slices subnets (NSS), as shown in Figure 4-6, can be described with VNFFGD and VLD. The connectivity among NSSI is implemented through the connectivity among the NFs of one NSSI and the NFSs of another NSSI. Currently ETSI NFV MANO defines also the support for physical links for the VLD. The 5G-MoNArch Network Slice Blueprint represents the collection of NSSIs and their links, so it is a collection of MNSSD and VNFD/VLD, cf. Figure 4-7.

To extend the ETSI MANO model toward the Blueprint it is useful to look at what Open Source Mano (OSM) is working on. To implement the 5G-MoNArch Network Slice Blueprint, the most important steps is to implement the MVNFD. As described above, the MVNFD is an extension of the MANO VNFD including the Application configuration for the VNF. OSM too is working in that direction, starting from MANO specifications and extending them to include Application configuration. This extension is made by OSM using charms. A charm is a collection of scripts and metadata that encapsulate a specific configuration for a particular product. Charms are included in OSM VNFD with a similar approach 5G-MoNArch includes the configuration details for the Application part of the VNF into the MVNFD. OSM doesn't give specification about the parameters, they are up to implementation. The new field defined by OSM for the application configuration is named vnf-configuration, the main

idea, for the implementation of MNVFD (see next section), is to start from this new field ad adapt it according to 5G-MoNArch view.

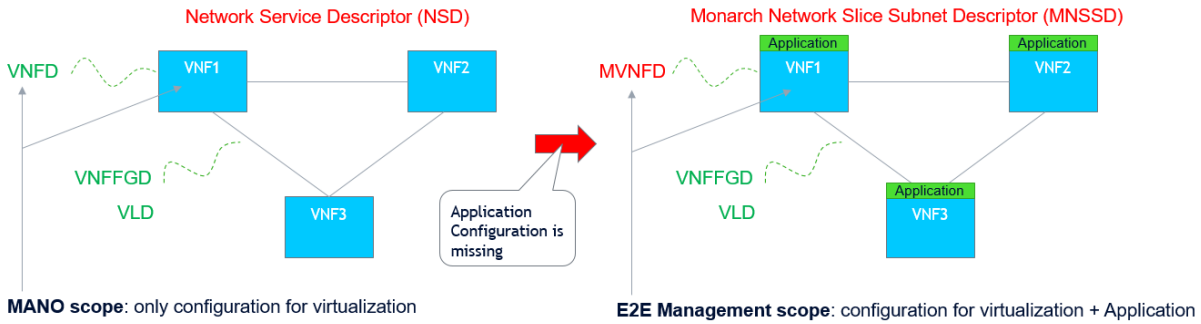


Figure 4-5: Generating a mobile network slice subnet descriptor from a network service descriptor

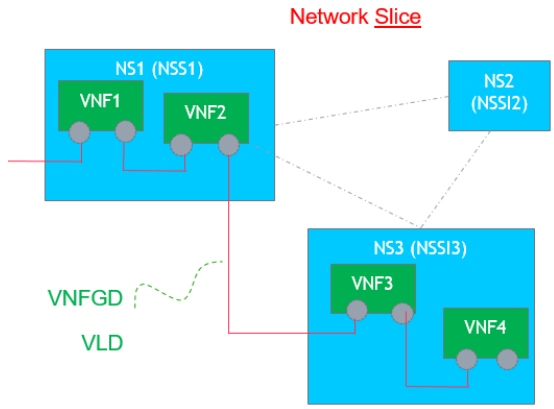


Figure 4-6: Links among network slice subnet instances

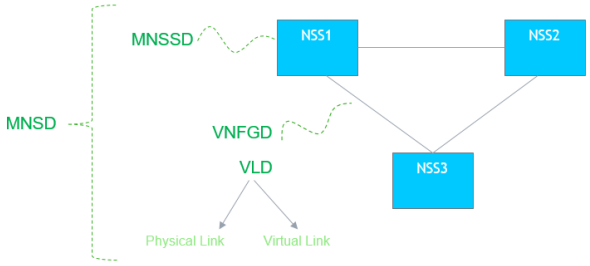


Figure 4-7: 5G-MoNArch network slice blueprint

4.1.5 5G-MoNArch network slice blueprint implementation

This section presents the implementation proposed for 5G-MoNArch Blueprint as a collection of enhanced ETSI NFV MANO NS descriptors as described in Section 4.1.4. The reference structure used for the definition of each descriptor and its components is the ETSI NFV MANO framework. The schema in Table 4-1 depicts the 5G-MoNArch VNF Descriptor. The introduced novelties over the ETSI NFV MANO VNFD are highlighted.

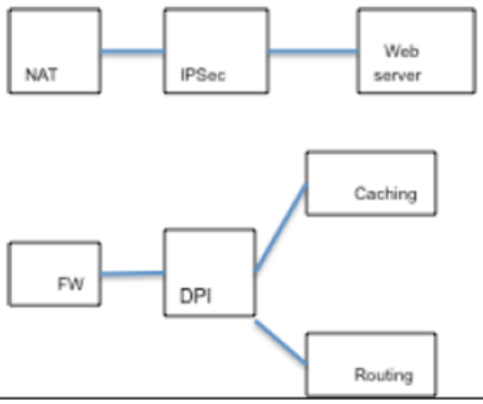
MVNFD is the enhancement of ETSI NFV MANO VNFD with the addition of application level configuration data. 5G-MoNArch introduces a new raw vnf_configuration in order to accommodate such information coherently with the approach used by OSM for the same purposes. The novelties, with respect to OSM, is that the configuration parameters for the application part of the VNF will be compliant with the configuration parameters and their modelling as specified by 3GPP. The configuration parameters are the one specified in the Network Resource Model (NRM) for each 3GPP NF [3GPP TS 28.541], [3GPP TS 28.540]; optionally some vendor specific configuration parameter could also be introduced.

Table 4-1: 5G-MoNArch VNF descriptor

Identifier	Type	Cardinality	Description
Id	Leaf	1	ID (e.g. name) of this VNFD.
vendor	Leaf	1	The vendor generating this VNFD.
descriptor_version	Leaf	1	Version of the VNF Descriptor.
version	Leaf	1	Version of VNF software, described by the descriptor under consideration.
vdu	Element	1...N	This describes a set of elements related to a particular VDU, see clause 6.3.1.2.
virtual_link	Element	0...N	Represents the type of network connectivity mandated by the VNF vendor between two or more Connection Points, see clause 6.3.1.3.
connection_point	Element	1...N	This element describes an external interface exposed by this VNF enabling connection with a VL, see clause 6.3.1.4 (see note).
lifecycle_event	Leaf	0...N	Defines VNF functional scripts/workflows for specific lifecycle events (e.g. initialization, termination, graceful shutdown, scaling out/in, update/upgrade, VNF state management related actions to support service continuity).
dependency	Leaf	0...N	Describe dependencies between VDUs. Defined in terms of source and target VDU, i.e. target VDU "depends on" source VDU. In other words sources VDU shall exists before target VDU can be initiated/deployed.
monitoring_parameter	Leaf	0...N	Monitoring parameters, which can be tracked for this VNF. Can be used for specifying different deployment flavours for the VNF in a VNFD, and/or to indicate different levels of VNF service availability. These parameters can be an aggregation of the parameters at VDU level e.g. memory-consumption, CPU-utilisation, bandwidth-consumption, etc. They can be VNF specific as well such as calls-per-second (cps), number-of-subscribers, no-of-rules, flows-per-second, VNF downtime, etc. One or more of these parameters could be influential in determining the need to scale.
deployment_flavour	Element	1...N	Represents the assurance parameter(s) and its requirement for each deployment flavour of the VNF being described, see clause 6.3.1.5.
auto_scale_policy	Leaf	0...N	Represents the policy meta data, which may include the criteria parameter and action-type. The criteria parameter should be a supported assurance parameter (vnf:monitoring_parameter). Example of such a descriptor could be: • Criteria parameter → calls-per-second. • Action-type → scale-out to a different flavour ID, if exists.
vnf_configuration	Leaf	0…N	Configuration parameters for the Application part of the VNF

Since a new MVNFD structure has been defined, also the 5G-MoNArch Network Slice Subnet Descriptor (MNSSD) needs to be defined following the same approach as per the VNFD, the ETSI NFV MANO Network Service Descriptor was analysed and used as a reference. In this case, since all the information need to set up a network service are already present, the only parameter that needs to be changed to setup an NSSI is the reference to the VNFD, which, in the case of the model used here, will point to the MVNFD, cf. Table 4-2.

Table 4-2: 5G-MoNArch Network Slice Subnet Descriptor (MNSSD)

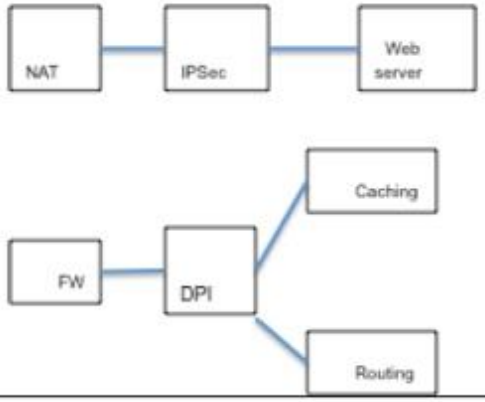
Identifier	Type	Cardinality	Description
Id	Leaf	1	ID of this Network Service Descriptor.
vendor	Leaf	1	Provider or vendor of the Network Service.
version	Leaf	1	Version of the Network Service Descriptor.
vnfd → mvnfd	Reference	1...N	VNF which is part of the Network Service, see clause 6.3.1. This element is required, for example, when the Network Service is being built top-down or instantiating the member VNFs as well.
vnffgd	Reference	0...N	VNFFG which is part of the Network Service, see clause 6.5.1. A Network Service might have multiple graphs, for example, for: 1. Control plane traffic. 2. Management-plane traffic. 3. User plane traffic itself could have multiple NFPs based on the QOS etc. The traffic is steered amongst 1 of these NFPs based on the policy decisions.
vld	Reference	0...N	Virtual Link which is part of the Network Service, see clause 6.4.1.
lifecycle_event	Leaf	0...N	Defines NS functional scripts/workflows for specific lifecycle events (e.g. initialization, termination, scaling).
vnf_dependency	Leaf	0...N	Describe dependencies between VNF. Defined in terms of source and target VNF i.e. target VNF "depends on" source VNF. In other words a source VNF shall exist and connect to the service before target VNF can be initiated/deployed and connected. This element would be used, for example, to define the sequence in which various numbered network nodes and links within a VNF FG should be instantiated by the NFV Orchestrator. 

The same approach is used to define 5G-MoNArch Network Slice Blueprint (MNSB), see Table 4-3, that is a collection of MNSSD and links among them. Summing things up, MNSD will result in a schema which, again, will be derived from MANO's network service descriptor. Nevertheless, also at this level, a few elements need to be introduced:

- ID will be replaced by the S-NSSAI plus the NSI ID, defined by 3GPP in [3GPP TS 23.501].
- 5G-MoNArch Network Slice Blueprint will replace the reference to the VNFD with a reference to the MNSSD.

Table 4-3: 5G-MoNArch network slice blueprint

S-NSSAI + Network Slice ID (NSI ID)

Identifier	Type	Cardinality	Description
Id	Leaf	1	ID of this Network Service Descriptor.
vendor	Leaf	1	Provider or vendor of the Network Service.
version	Leaf	1	Version of the Network Service Descriptor.
vnfd → mnssd	Reference	1...N	VNF which is part of the Network Service, see clause 6.3.1. This element is required, for example, when the Network Service is being built top-down or instantiating the member VNFs as well.
vnffgd	Reference	0...N	VNFFG which is part of the Network Service, see clause 6.5.1. A Network Service might have multiple graphs, for example, for: 1. Control plane traffic. 2. Management-plane traffic. 3. User plane traffic itself could have multiple NFPs based on the QOS etc. The traffic is steered amongst 1 of these NFPs based on the policy decisions.
vld	Reference	0...N	Virtual Link which is part of the Network Service, see clause 6.4.1.
lifecycle_event	Leaf	0...N	Defines NS functional scripts/workflows for specific lifecycle events (e.g. initialization, termination, scaling).
vnf_dependency	Leaf	0...N	Describe dependencies between VNF. Defined in terms of source and target VNF i.e. target VNF "depends on" source VNF. In other words a source VNF shall exist and connect to the service before target VNF can be initiated/deployed and connected. This element would be used, for example, to define the sequence in which various numbered network nodes and links within a VNF FG should be instantiated by the NFV Orchestrator. 

In the following, an example of MVND extension for the new raw on application configuration is provided. The modelling used in the example is YANG as specified by 3GPP in [3GPP TS 28.541] and also used by OSM. A possible implementation of the modelling could be done with JSON.

The YANG model for the MVNFD and for 5G-MoNArch Network Slice Blueprint, is the combination of the YANG definition for a VNFD from OSM and the YANG model for the NRM of a VNF from 3GPP.

A YANG example for the VNFD modelling, from OSM can be found in [OSM]. In [3GPP TS 28.541], in the normative Annex H, it is specified the use of YANG to define the NRM for the application configuration of a 5G VNF. OSM YANG model has a specific raw for VNF application configuration using charms: vnf_configuration. For MNSBP that raw is filled with the YANG specification of SA5 NRMs.

4.2 *Network slice lifecycle management*

This section describes the overall network slice lifecycle management process, captured in Figure 4-8. The network slice lifecycle management process is composed of two phases, namely the network slicing pre-operation phase and the network slicing operation phase. In the former phase, the NSMF produces

the network slice blueprint based on the network requirements of the slice and with the support of predefined templates related to standardised slices. Upon this decision the NSMF should proceed with slice resource pre-selection considering the available resources, needed computational power, network topology, currently operating NSIs along with their demands, etc.

In the network slicing operation phase, the NSMF should proceed, using this as input, with the actual slice deployment. In this phase, the actual functions' configuration is setup in each domain. Such configuration includes functions' parameterisation and placement according to the available resources. The operation phase also includes the slice performance monitoring sub-phase. That is, the NSI/NSSI performance monitoring modules (cf. Figure 2-10) continuously control whether the slice is able to meet the SLA requirements and, if not, to inform the corresponding functions; eventually, an alarm can be triggered. If the changes relate to updates in the functions' configuration and proper placement then the slice configuration functionality should be triggered. Similar procedures can be implemented when the slice requirements are updated at the user side, i.e., at the CSMF. Section 4.2.1 provides the most relevant use case on slice allocation focusing on NSI and NSSI sharing. Section 4.2.3 provides a detailed example of this procedure, related to the network slice allocation, focusing on the involved M&O layer functions. Section 4.2.4 describes the procedure for managing the performance for a slice subnet by adapting its configuration due to traffic or resource availability changes.

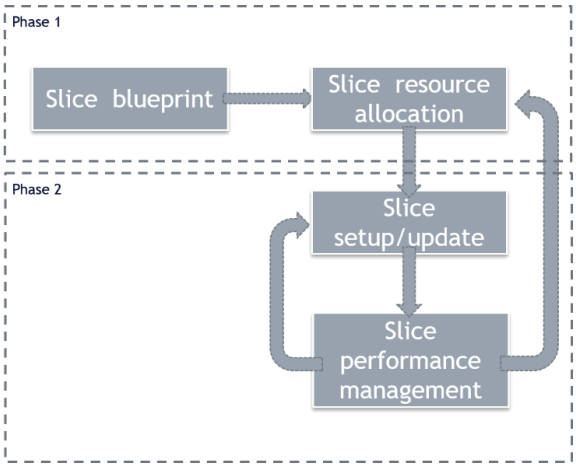


Figure 4-8: Network slice lifecycle management process in 5G-MoNArch

4.2.1 Cross-slice orchestration with shared NF

Figure 4-9, from [3GPP TS 28.530], represents the two main scenarios related to resource sharing in network slicing:

- More communication services share the same NSI
- More NSI share some NSSIs

According to those two scenarios 5G-MoNArch M&O layer has to support the following use cases:

- The allocation of an NSI to serve a new communication service. The allocation process is intended to create a new NSI or reuse an existing NSI.
- The creation of a new NSI reusing existing NSSIs sharing them among NSIs
- Updating the communication service requirements in the case of sharing the NSI or the NSSIs

According to the functional split of 5G-MoNArch M&O layer, as in Section 2.2.3, some management functions are specifically defined for the cross-slice orchestration, especially when shared NFs are involved. In the Cross-Slice M&O block are present the following management functions: Cross Slice Requirement verification and Cross Subnet Requirement Verification. These management function support the use cases proposed in this section. The network slice and subnetwork slice sharing scenarios proposed by 3GPP have been evaluated and analysed with the conclusion that the following user stories has to be supported by 5G-MoNArch M&O layer trough the new management function introduced in the Cross-Slice M&O function.

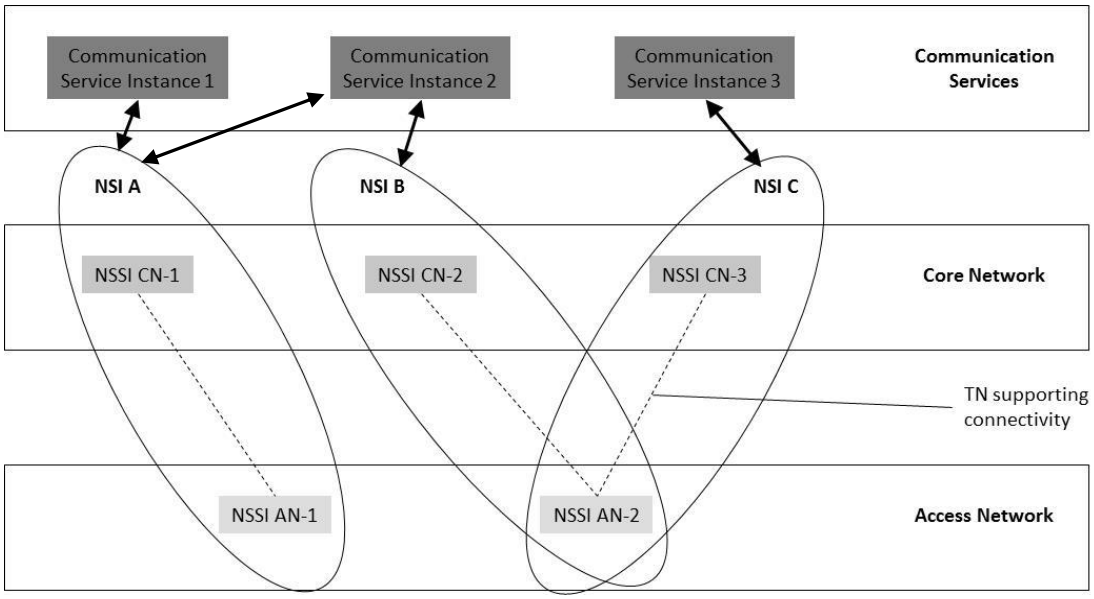


Figure 4-9: Slice subnetwork sharing across two or more network slices

Network slice allocation using an existing NSI

The M&O layer has to provide a network slice instance that fits the requested network requirements. The allocation process foresees to use an existing network slice instance (NSI) to optimise the network resource usage or to create a new NSI.

The M&O layer verifies if other existing NSIs support the requested communication service. An identified existing NSI has to be compatible with the network requirements, the network management policies let it to be shared and it still supports the new overall performance, capacity and lifecycle management requirements for all the communication services it has to provide. If no existing NSI can be used the M&O layer has to create a new one, if possible.

Detailed description

- The 5G Operator receives the request to provide a new communication service. The 5G Operator uses the M&O layer to provide an NSI to satisfy the request. The M&O layer performs the following steps:
- Verify if there is an NSI that is compatible with the network requirements.
- If no compatible NSI exists, create a new NSI and associate the requested communication service to it.
- If any compatible NSI exists, verify if the network management policies (e.g. related to sharing) allow using it.
- If the policies don't allow using any of the identified NSIs, create a new NSI and associate the requested communication service to it.
- If the M&O layer finds an existing NSI that can be used, verify if the identified NSI supports the overall performance, capacity and lifecycle management requirements for the communication services it has to serve.
- If yes, use it to satisfy the current communication service request.
- If none of the identified NSIs support the overall performance, capacity and lifecycle management requirements for the communication services, verify the network management policies to decide to reconfigure one of the identified NSIs or to create a new NSI.
- In case the network management policies don't allow to reconfigure any of the identified NSIs, create a new NSI and associate the requested communication service to it.
- In case the network management policies allow reconfiguring one of the identified NSIs, proceed defining the new requirements for the NSI according to the overall performance, capacity requirements and lifecycle management for all the communication services.

- Verify if the new overall requirement for the NSI are still compatible with the network management policies.
- If the verification is positive, associate the requested communication service to it otherwise create a new NSI.
- In the case of creating a new NSI, verify if the original updated network requirements are compatible with the network management policies and the resource availability.
- If yes, the new NSI can be created otherwise the provisioning request is denied.

Network slice creation using existing NSSIs

The M&O layer has to create a new network slice instance (NSI) that meets the requested network requirements. The M&O layer tries to use existing network slice subnets instances (NSSIs), sharing them, to optimise the network resource usage.

The M&O layer has to provide the constituent network slice subnets that will be used for the network slice. The allocation process verifies, for each requested network slice subnet, if there are sharable NSSIs available that support the requirements, otherwise I have to create a new one.

Detailed description

The M&O layer has already identified that the requested communication service cannot rely on an existing NSI, so it is proceeding to create a new NSI.

As part of a new NSI creation process, The M&O layer decomposes the network slice requirements into network slice subnet requirements.

The M&O layer allocates the network slice subnets using existing NSSIs or creating new ones, if possible. The M&O layer verifies if there are already deployed NSSIs that can be shared in terms of network management policies and that are compatible in terms of requirements.

To provide each requested NSSI, the M&O layer performs the following steps:

- Verify if there is an NSSI that is compatible with the network subnet requirements.
- If there is no compatible NSSI, create a new NSSI.
- If there is a compatible NSSIs, verify if the network management policies (e.g. related to sharing) let me use it.
- If the network management policies don’t allow me to use any of the identified NSSIs, create a new NSSI.
- If any compatible NSSI exists, verify if it supports the overall performance, capacity and lifecycle management requirements for all the NSIs it has to serve.
- If yes, use it to satisfy the current request for network slice subnet allocation.
- If none of the identified NSSIs supports the overall performance, capacity and lifecycle management requirements requested by all the NSIs, verify the network management policies (e.g. related to an NSSI maximum capacity) to decide if to reconfigure one of the identified NSSIs or to create a new NSSI.
- In case the network management policies allow to reconfigure one of the identified NSSIs, proceed defining the new requirements for the NSSI according to the overall performance, lifecycle management and capacity requirements for all the NSIs it has to serve.
- Verify if these new overall requirements for the NSSI are still compatible with the network management policies.
- If the verification is positive, reconfigure the NSSI, otherwise create a new NSSI.
- In the case of creating a new NSSI, verify if the original updated network requirements are compatible with the network management policies and the resource availability.
- If yes, the new NSSI can be created otherwise the NSI allocation request is denied and, consequently, the NSI creation request is denied.

Requirements update when the NSI is shared among services

The M&O layer has to modify a network slice instance (NSI) according to a request of network requirements update.

The M&O layer verifies if the current NSI already supports the new requirements and if it still supports the new overall performance, capacity and lifecycle management requirements for all the communication services it has to provide. If needed and if it is possible accordingly to the network management policies (e.g. related to an NSI maximum capacity), The M&O layer reconfigures the NSI, otherwise the operator creates a new network slice instance to support the communication service.

Detailed description

The M&O layer receives the request to update the network requirements of a communication service provided by an NSI.

The M&O layer performs the following steps:

- Verify if the current NSI is still compatible with the new network requirements
- If yes, verify if the new overall performance, capacity and lifecycle requirements for all the communication services are still compatible with the NSI.
- If yes, use it to satisfy the current communication service request.
- If no (no compatibility with the network requirements or with the overall performance, capacity and lifecycle requirements), evaluate, according to the network management policies and to the requirements of the other services, if updating the current NSI or if allocating a new one.
- To update the current NSI, define the new requirements according to the overall performance, capacity and lifecycle requirements for all the communication services.
- Verify if these new overall requirements are compatible with network management policies.
- If yes, reconfigure the NSI.
- Otherwise, proceed allocating a new NSI to support the updated communication service.
- If the M&O layer has provided a new NSI to fulfil the new requirements, evaluate the network requirement and the performance, capacity and lifecycle requirements for the remaining communication services that are still using the old NSI (if any) to decide if that NSI has to be reconfigured.

Requirements update when some NSSI is shared among NSIs

The M&O layer has to modify an existing NSI according to a request of network requirements update. Alternatively, if there is another NSI which could support the new network requirements, the M&O layer may decide to use the alternative NSI.

The M&O layer verifies if the current NSI already supports the new requirements. If the NSI doesn't fit the new requirements, The M&O layer evaluates if reconfiguring the current NSI or using some other existing NSI that fits the new requirement.

Detailed description

The M&O layer receives the request to update the requirements of an NSI. This NSI is not shared with other communication services but some NSSIs are shared with other NSIs.

The M&O layer performs the following steps:

Verify if the current NSI is still compatible with the new network requirements and if the shared NSSIs are compatible with the overall performance, capacity and lifecycle requirements for all the NSIs they are supporting.

If the current NSI and shared NSSIs are still compatible, continue using them.

If not, verify, according to the new requirements and to the network management policies (e.g. related to sharing or NSI capacity), if reconfiguring the current NSI or using some existing NSI already compatible with the new requirements to provide the update communication service.

If the M&O layer decides to use an existing NSI, the chosen NSI has to be:

- Sharable according to the network management policies.
- Compatible with the new requirements.
- Compatible with the overall performance, capacity and lifecycle requirements for all the communication it has to serve.
- If the M&O layer decides to reconfigure the current NSI it has to update the subnets requirement and to verify if it is possible to update the current NSSIs and/or create new NSSIs.

- For each requested subnet, the M&O layer has to verify if the current NSSI is still compatible with the new network requirements
- If yes, and if the NSSI is shared, verify if it is compatible with the new overall performance, capacity and lifecycle requirements for all the NSIs it has to serve.
- If yes, use it to satisfy the current update request.
- If no (no compatibility with the network requirements or with the overall performance, capacity and lifecycle requirements), evaluate, according to the network management policies and to the requirements of the other NSIs using it, to update the current NSSI or to provide a new one.
- To update the current NSSI, define the new requirements according to the overall performance, ì capacity and lifecycle requirements for all the NSSIs that are using it.
- Verify if these new overall requirements are compatible with network management policies.
- If yes, reconfigure the NSSI.
- Otherwise, proceed allocating a new NSSI to support the updated communication service.
- If the M&O layer has provided a new NSSI to fulfil the new requirements, evaluate the network requirement and the performance, capacity and lifecycle requirements for the remaining services that are using the old NSSI (if any) to decide if the NSSI has to be reconfigured.
- If the M&O layer decides to use an existing NSI, the old NSI has to be dissociated from the communication service.

This procedure may be partially automatised by implementing the slice analytics enabler investigated in WP4 (see IR4.2 [5GM-IR4.2]).

4.2.2 5G-MoNArch network slice allocation

With reference to the M&O functional split described in Section 2.2.3, this section describes a high-level call flows for slice allocation, in order to give an example on the interaction among the management function. The call flow (see Figure 4-10) shows how the management system proceeds allocating a Network Slice Instance (NSI) to support a communication service seeking for an existing NSI to share or creating a new NSI. This call flow is simplified to give a first overview of the interaction among the management functions.

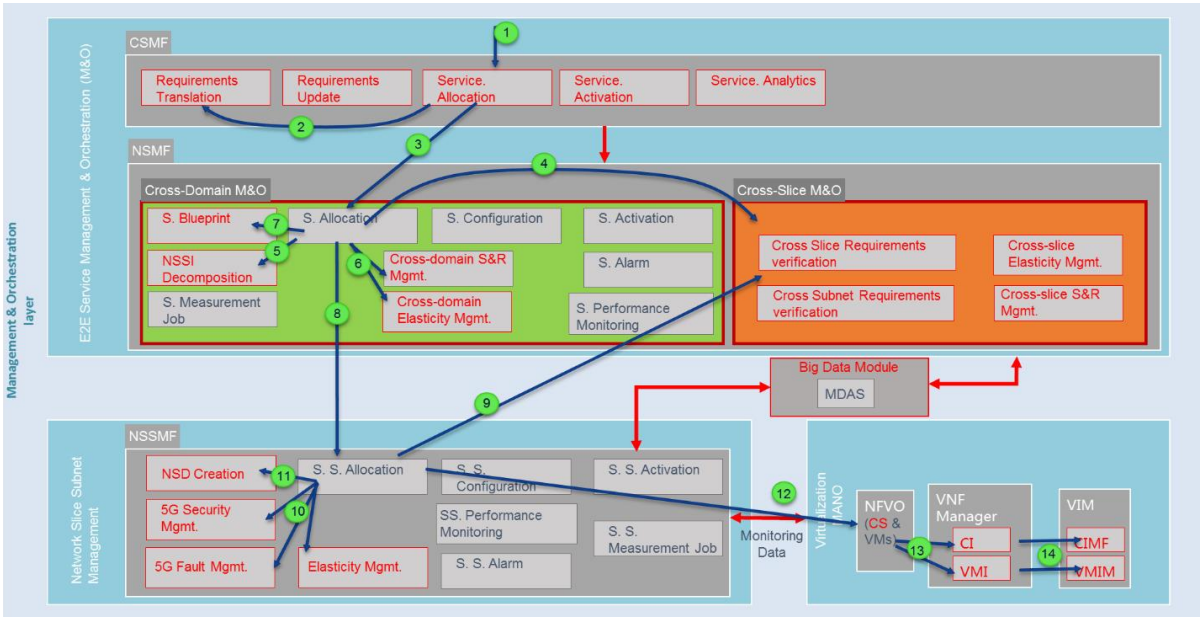


Figure 4-10: Network slice allocation flow

(1) The Communication Service Allocation function receives the request for the allocation of a new communication service with the related service requirements.

(2) The Communication Service Allocation function triggers the Requirements Translation function to translate the service requirements into network requirements.

(3) The Communication Service Allocation function triggers the Slice Allocation function inside the Cross-Domain M&O of the Network Slice Management Function requesting the allocation of an NSI

(4) The Slice Allocation function triggers the Cross-Slice Requirement Verification function to verify if an existing slice that fits the purpose.

(5) If none existing slice is available, the Slice Allocation function triggers the NSS Decomposition function to define the network slice constituents in terms of network slice subnets.

(6) To optimise the network slice, depending on the slice requirements and network condition the Slice Allocation function triggers the Cross-slice S&R and Elasticity Mgmt.

(7) The Slice Allocation function triggers the Slice Blueprint function to completely define the slice in terms of its constituents (e.g. NFs, connectivity and topology) and their configuration.

(8) To deliver the network slice a set of network slice subnets has to be allocated, this means recusing existing NSSIs that fits the requirements or creating new NSSIs. To do this the Slice Allocation function triggers the Slice Subnet Allocation function inside the NSSMF. 5G-MoNArch management system foresees the possibility to have more NSSMFs, maybe for different domains, so the Slice Allocation function triggers all the NSSMF responsible for the subnet allocation that are requested for this network slice. The following steps are repeated for each requested slice subnet.

(9) The Slice Subnet Allocation function triggers the Cross-Subnet Slice Requirement Verification function to verify if an existing slice subnet fits the purpose. If an existing NSSI is available, it is used maybe after updating it.

(10) If none existing slice subnet is available, the Slice Subnet Allocation function has to create a new one, to optimise it triggers the 5G Security Mgmt., 5G Fault Mgmt., and Elasticity Mgmt., depending on the slice requirements and network condition.

(11) If none existing slice subnet is available, the Slice Subnet Allocation function has to create a new Network Service to provide the slice subnet, to do this the Slice Subnet Allocation triggers the NSD Creation function. This function produces a Network Service Descriptor that will be the input for the request toward MANO for the network service creation

(12) The Slice Subnet Allocation function triggers MANO, with the appropriate NSD, to create the network service to support the network slice subnet. In the NSD are the requirement for the connectivity inside the network service and for the connection among the other subnets.

4.2.3 5G-MoNArch network slice congestion control

This section describes a high-level flow for slice congestion control. The flow shows how the management system reacts to the increasing resource requirements of a given slice. In this case, the need for additional network resources is related to perceived performance reduction at the slice level, due, for instance, to an increased slice load. This description gives an example of the interactions among the management functions, related to possible implementation of the cross-slice congestion algorithm detailed in Section 3.4.1.

In Figure 4-11, a first flow is represented. In this case, the S. Alarm module in the NSMF, informs the Cross-domain Elasticity Mgmt. that a given slice performance is decreasing. Accordingly, the Cross-domain Elasticity Mgmt. verifies whether there is a need for additional network resources according to the slice requirements defined at the S. Blueprint. The Cross-domain Elasticity Mgmt. monitors the network resource availability and allocates additional resources to the slice accordingly. After this step, it exchanges related information at the Elasticity Mgmt. modules related to each slice domains. Finally, the Elasticity Mgmt. verifies whether the new resource allocation is compatible with the resource availability at the domain level, if necessary it update the resource allocation decision, and accordingly demands its implementation at the MANO level.

A slightly different flow is described in Figure 4-12. In this case, the Cross-domain Elasticity Mgmt. cannot allocate additional resources to the slice according to the feedback received by the S.

Measurement Job, as the system may be overloaded. Accordingly, the Cross-domain Elasticity Mgmt. requests the Cross-slice Elasticity Mgmt. to initiate a cross slice congestion procedure. Therefore, the latter 1) identifies the slices with looser requirements for which the amount of allocated resources can be reduced and 2) updates their resource allocation. After this step, it exchanges related information at the Elasticity Mgmt. modules for each of the domains of the involved slices. Finally, the Elasticity Mgmt. verifies and potentially adjusts the new resource allocation plan and demands its realisation at the MANO level.

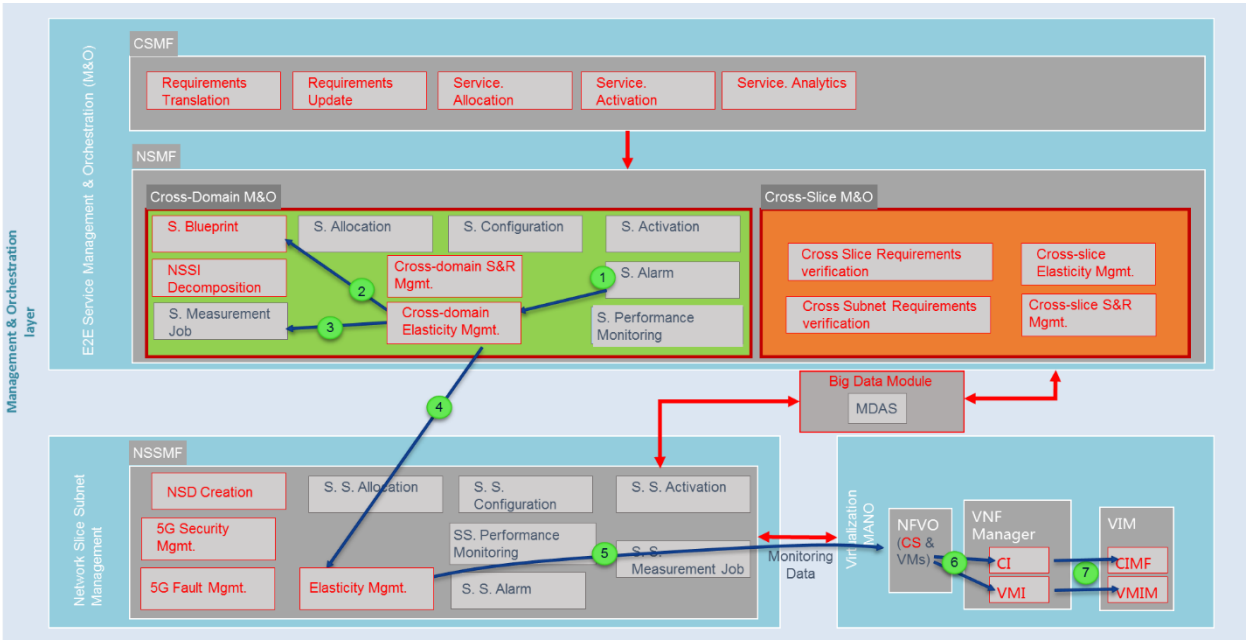


Figure 4-11: Network slice congestion control flow

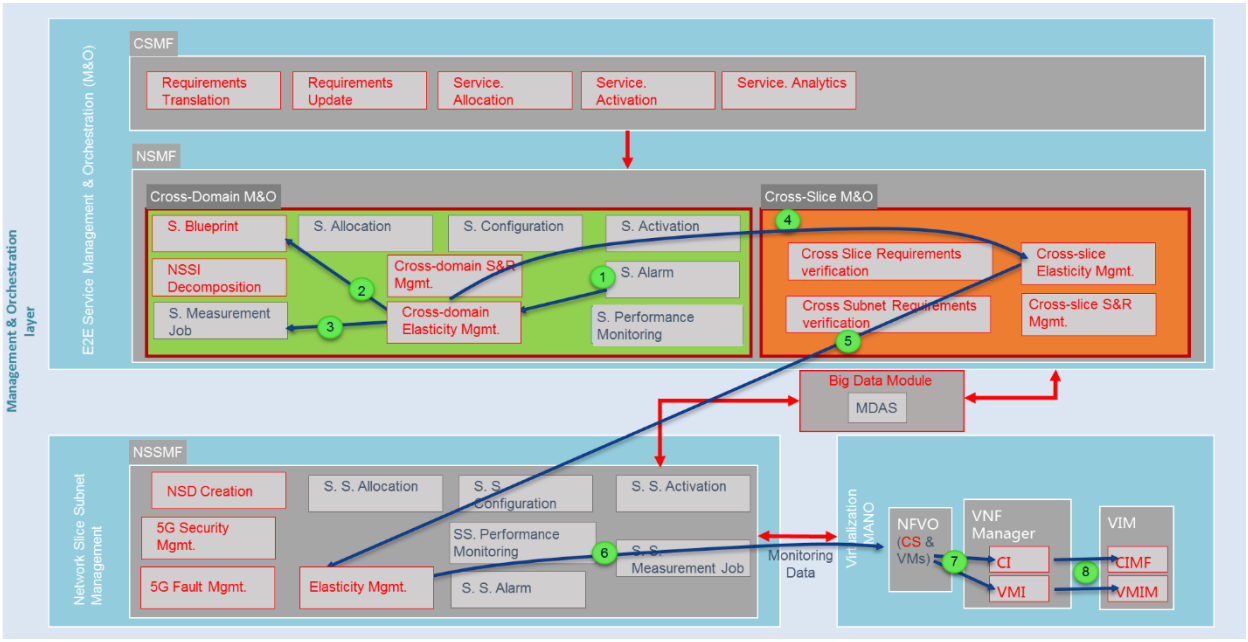


Figure 4-12: Cross-slice congestion control flow

4.2.4 5G-MoNArch network slice subnet performance management

This section describes the procedure for managing performance for a slice subnet by adapting the configuration due to traffic or resource availability change. One key example is the Radio Access

Network (RAN) slice-subnet which may require adaptation due to the fact that the radio resources and traffic demand may strongly affect the configuration of the AN-NSSI.

RAN Configuration Adaptation at the RAN Management layer (e.g. AN-NSSMF) may be needed to cope with changes in different time scales. This may include:

- Slow changes, e.g., a change of the traffic situation at the RAN, group UE mobility, a change of backhauling capacity, etc. One example for adaptive configuration requirement due to slow changes, is the adaptive placement of RRM functions/ RRM Splits in a slice-aware manner by NSSMF due to wireless backhaul situation change (e.g. in IAB scenarios or dynamic radio topologies) [PP17].
- Fast / On-demand changes, e.g., communication protocol of individual traffic flow, dynamic RAN topologies, individual UE mobility, critical slice traffic, new application requirements etc. One particular example of requiring some fast adaptation is the V2X slicing scenario where the RAN-NSSI needs to be adapted either due to dynamic network condition changes or traffic changes at RAN (e.g. group handovers) or new application requests (e.g. change of level of automation for a session). This requires the tight interaction between the M&O and Controller layers to adapt the RAN part of the slice.

Below, the functional flow is described for the slice subnet performance management, with particular focus on AN-NSSI as example:

Pre-condition: The NSI Performance Management function (at NSMF) decomposes and sends the NSSI related KPI to the NSSI Performance Management function (at NSSMF) in pre-operation phase.

(1) The NSSI Performance Monitoring function (at NSSMF) receives monitoring of the resource / traffic situation by the RAN subnet either periodically or on-demand or by an event triggered by RAN that may indicate a change of context with respect to a change of information on a process, a protocol function, and/or a resource and/or traffic at the RAN.

(2) The NSSI Performance Monitoring function may trigger a request to the NSSI Performance Management function, to adapt certain parameters for the NSSI. For example, it may identify that the capacity of one or more supporting network slice subnets need to be modified with X1 amount, X2 amount, X3 amount or the RRM policies between slices in RAN may need to change (different level of isolation or split of resources).

(3) The NSSI Performance Management function adapts the RAN-NSSI based on the up-to-date RAN KPIs and the received monitoring report / trigger.

(4) The NSSI Performance Management function informs NSI Performance Monitoring function and NSI Allocation function in case of further impact to other subnets

(5) The NSSI Performance Management function applies the adaptation of the RAN-NSSI (functions and/or resources at RAN) and sends this information to the RAN

(6) RAN may enforce the adapted configuration it received based on the real time resource situation

4.3 *Integration of functional innovations for 5G-MoNArch use cases*

This section motivates and illustrates the use-case-specific functional extensions to the 5G-MoNArch architecture in light of the service requirements of the Hamburg Smart Sea Port (resilience, security) and the Turin Touristic City (resource elasticity) scenarios.

4.3.1 Resilience and security

4.3.1.1 Basic concepts and required architecture components

A network slice intended to support URLLC services needs to fulfil high resilience, reliability and security requirements. Such specialised service requirements coming from the customer need to be carefully translated into the resource-facing service description, e.g. by including in the slice template the NFs and their corresponding configurations that can support the specified requirements. Furthermore, the slice template needs to contain the instructions for actual deployment, management,

orchestration and control of specialised NFs that will be executed by different functions of 5G-MoNArch architecture, e.g. instructions on LCM of specialised VNFs that will be executed by VNFM.

The following paragraphs elaborate on the main specialised NFs needed for enabling high reliability, resilience and security, along with their placement in 5G-MoNArch architecture as depicted in Figure 4-13 and discussed in Section 2.1.

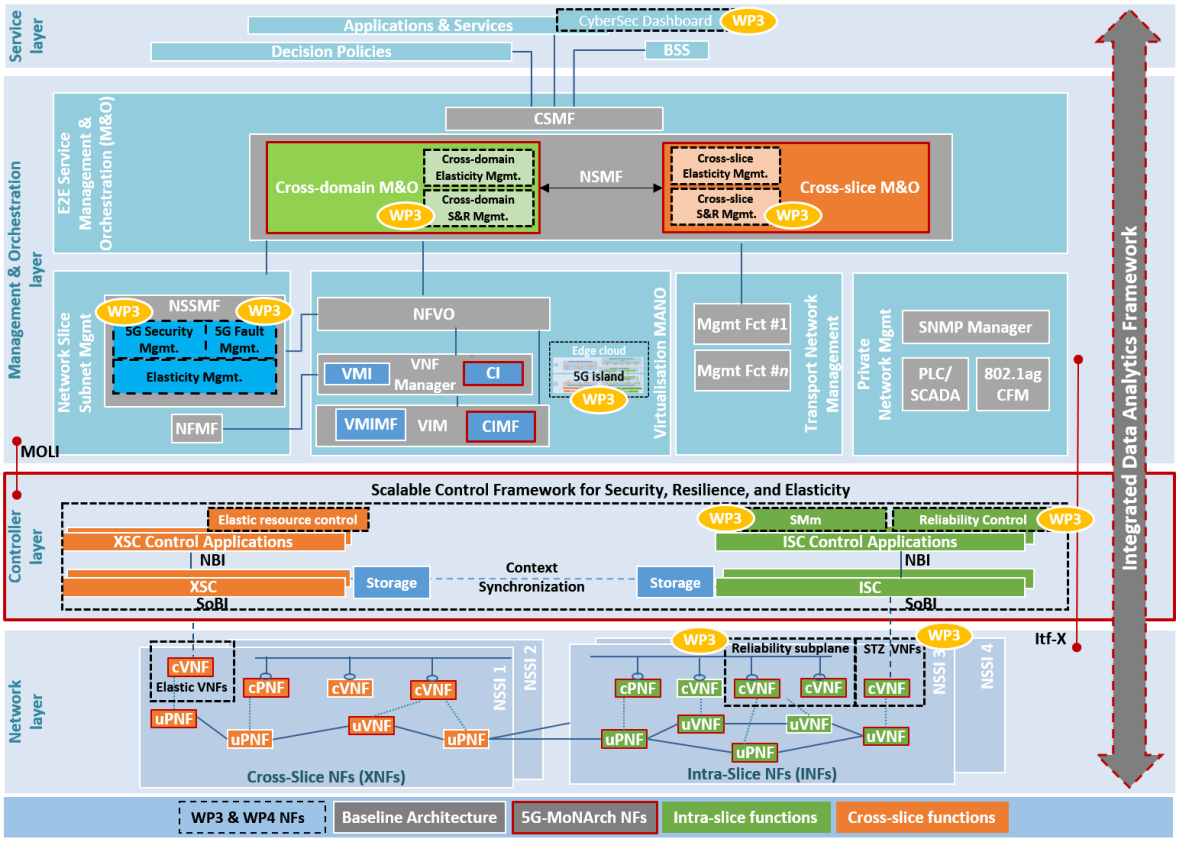


Figure 4-13: 5G-MoNArch overall architecture with the marked WP3 modules to enable resilience and security functional innovation [5GM-D3.2]

The details of the highlighted specialised NFs are available in [5GM-D3.1], [5GM-D3.1], and are summarised as follows.

4.3.1.2 RAN reliability functional innovations

The *RAN reliability sub-plane* comprises the functions for multi-connectivity (data duplication) and network coding for improved RAN resilience. It appears as the pool of resilience-enabling NFs which can be on-demand, dynamically instantiated and configured based on the actual network context, resilience requirements as well as agreed SLAs with the slice tenant. Such sub-plane resides in the Network layer, which provides the required UP and CP functionality. Optionally, a corresponding control application ('Reliability Control') in the Controller layer can be instantiated as well, cf. Section 2.2.1.3. This function is responsible for instantiation, activation and control of the data duplication and network coding functions in the Network layer.

4.3.1.3 Telco cloud resilience functional innovations

Functions for improving the telco cloud resilience include *FM* functions specialised for resolving the network issues in 5G virtualised networks by jointly handling the faults coming from virtualised and physical infrastructure and considering slice requirements. Therefore, the 5G network FM needs to leverage on the information available at E2E Service M&O and the NFVO entities, c.f. Figure 4-13. The 5G network FM can be seen as a part of the 3GPP Network Management module including considerable

extensions compared to the legacy FM in order to incorporate slicing and virtualisation awareness, as well as automation and cognition aspects in order to improve the flexibility and adaptability of the FM process.

Furthermore, the 5G FM developed in 5G-MoNArch needs to realise *enhanced event correlation* in order to perform the adequate troubleshooting of NFs, slice subnets and network slices that can have interdependencies (e.g. due to resource sharing) and can be deployed under different management and administrative domains. Additionally, the "5G island" is the robust solution which relies on the edge cloud for "standalone" network operation, i.e., without permanent connectivity to the central cloud. This concept aims at estimating the need to migrate certain NFs from central cloud in order to have it available at the edge cloud once the connectivity towards the central cloud is lost. Such approach can mitigate the effects of backhaul connection outages, as well as cyber-attacks, planned outages and unintentional errors at central cloud.

Finally, a crucial building block for improving the telco cloud resilience is the load-aware '*Scalable and Resilient Control Framework*', as depicted in Figure 4-13. This framework allows automatic scaling of controller nodes with respect to the underlying traffic in the network. It improves the distributed states management and synchronisation of newly added controller nodes. Such seamless multiplication of controllers in the network enables the scalability but also ensures the high availability of controllers.

4.3.1.4 Functional innovations on security

As security is one of the fundamental requirements of ultra-reliable services it needs to be supported within all components of the E2E service. Such *holistic security approach* comprises the security of end devices, network security and network slice security. In other words, the service-tailored security assets need to be deployed i) at the *end-device*, ii) within the *5G NFs* and iii) to all *software and hardware components that mechanise the required network slices* in the considered use-case. Furthermore, a set of specialised NFs for improving the level of security needs to be deployed in order to achieve the required level of service reliability.

Security Trust Zones (STZs) define a logical area of infrastructure and services where a certain level of security and trust is guaranteed. Security level corresponds to the quality of being protected against threats, whereas trust assures that certain expectations will be met throughout a defined period of time. Details on the use of STZs are provided in the corresponding deliverables of WP3 of 5G-MoNArch [5GM-D3.1], [5GM-D3.2], as well as in [STM+19], [MGG+18]. In short, the main properties of STZ are detection, prevention and reaction which describe the capabilities of the STZ to achieve the promised security and trust levels. Various methods for achieving the STZ capabilities are analysed within 5G-MoNArch; for instance, a graph-based method is utilised for anomaly detection, along with deep machine learning techniques [5GM-D3.2].

Different *STZs with specific security level* and means of achieving required security level can be deployed all over the network. Within different STZs, the NFs responsible for Security Threat detection, protection or reaction, as well as the Threat Intelligence Exchange (ThIntEx) NFs (collectively referred to as STZ VNFs) can be deployed based on the actual security level that needs to be implemented. Furthermore, the function that coordinates the activity of Security Trust Zones (SMm – Security Monitoring Manager) deployed across network slices needs to be present.

SMm component is located at the *Controller layer* (c.f. Figure 4-13) in order to reach for different network slices and coordinate threat intelligence exchange by means of the corresponding ThIntEx NF, e.g. to share the information about security incidents detected in different network slices. This element is in charge of improving the reaction against threats and avoiding propagation of threats across-slices. Furthermore, the SMm receives the data provided by the different detection, prevention and reaction components deployed through the STZs of the same network slice. Finally, the Cyber Security Dashboard, located in the Service layer (c.f. Figure 4-13), provides tenants with visualisation and awareness of the security status of the deployed network slices at any point in time.

4.3.1.5 Interactions of the resilience and security functional innovations

Due to many interrelations between security and resilience considerations as well as common tools for detection of security and network issues, the *Cross-slice M&O function* responsible for inter-slice management will incorporate Cross-slice Security and Resilience Management function ('cross-slice

S&R Mgmt', c.f. Figure 4-13) specialised for addressing jointly the security and resilience considerations. Cross-domain M&O function is taking care of the coordination/negotiation between different management domains (e.g., RAN, CN) within a single slice and it can incorporate the functionality for joint dealing with security and resilience issues, i.e. Cross-domain Security & Resilience Management ('cross-domain S&R Mgmt', c.f. Figure 4-13). Such joint dealing with security and resilience issues may include, for instance, resource management aspects such as joint provisioning and handling of virtualised resources used for recovery from security and resilience related threats.

An example of such interaction between the resilience and security functional innovations relates with the joint study of fault management and security management, as documented in [5GM-D3.2]. This joint study stems from the virtualisation aspects of 5G, leaving thus room for flexible as well as cost-effective implementation of virtualised network elements, which potentially share common virtualised resources. As a result, joint actions can be taken as regards common mitigation approaches for tackling, for example, common root causes with detrimental effects in both the fault management and the security domain.

This joint consideration of fault management and security management is particularly relevant in network-slicing deployment approaches, where resilience (by means of network fault management) and security are typically deployed in a common slice due to the common application requirements they are associated with. Consequently, in the 5G-MoNArch approach the resources allocated for resilience and security purposes are dealt with in a common framework, such that the *cross-domain S&R mgmt.* and *cross-slice S&R mgmt.* architecture modules (c.f. Figure 4-13) determine the resources allocated within the resilience and security slice and across the resilience and security slice and potentially other slices in the network, respectively. Such common framework allows for identifying synergies between the concepts of resilience and security, towards a more efficient usage of the available virtualised resources [5GM-D3.2].

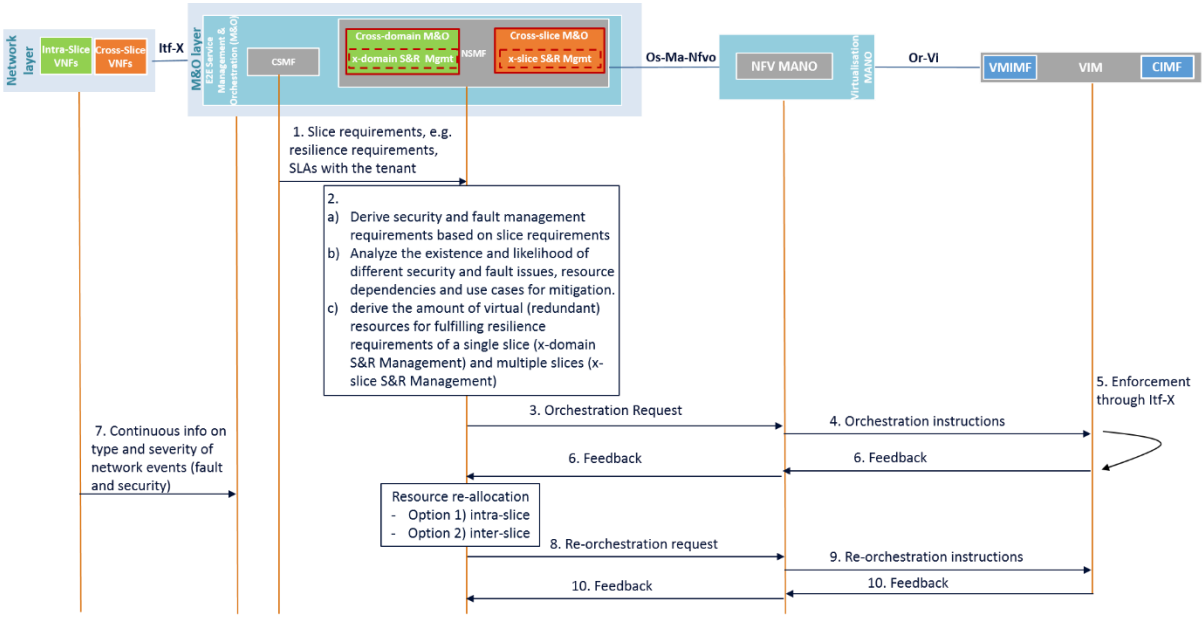


Figure 4-14: Message sequence chart for joint fault management (resilience) and security management operations

Figure 4-14 sheds light onto the process associated with signalling and information exchange between the involved network entities, by providing the message sequence chart of the joint fault management and security study. In particular, Figure 4-14 extracts those architecture enablers illustrated in Figure 4-13 that are involved in the joint process and highlights the corresponding signalling exchange between such entities. In a detailed view, Figure 4-14 underlines the following process steps:

- Derivation of the slice requirements at the CSMF: This step comprises obtaining the resilience and security requirements from the tenant of the slice. These are typically specified in the respective service level agreements (SLA)s between the tenant and the infrastructure provider.
- The analysis of the existence and likelihood of network fault and security issues at the NSMF, and the derivation of the virtual resources for satisfying the given requirements of the slice.
- The orchestration requests and corresponding feedback are exchanged between NSMF, NFV MANO and VIM entities. Such orchestration requests are enforced by means of the Itf-X.
- The potential resource re-allocation both at an intra-slice and inter-slice level.
- The iteration of the above process leading to corresponding re-orchestration requests and respective feedback.

Another example of the deployment of NFs based on the required resilience and security levels of different network slices is illustrated in the ensuing section. There, the target functional architecture to be utilised for the Hamburg *Smart Sea Port* use case is described. The ensuing section further explains how selected security and resilience functions are integrated into the Smart Sea Port use case.

4.3.1.6 Network architecture for the Smart Sea Port use case

For the *Smart Sea Port* use case deployed in the Smart Sea Port testbed of 5G-MoNArch, a customised architecture instance of the general overall architecture is utilised. It employs the service-based representation for the Service and M&O layers as well as for the control plane. Further, a subset of the 5G-MoNArch enabling common network functionality as well as selected functions of the use-case-specific functions as developed in WP3 (cf. Section 4.3.1.1) are utilised. The latter include cross-domain and cross slice security and resilience management, 5G Fault Management functions, as well as multi-connectivity-enabled RAN for increased reliability (RAN reliability sub-plane.)

The *Smart Sea Port* target architecture instance is depicted in Figure 4-15. It shows the NFs in each layer, Network layer, M&O layer, and Service layer.

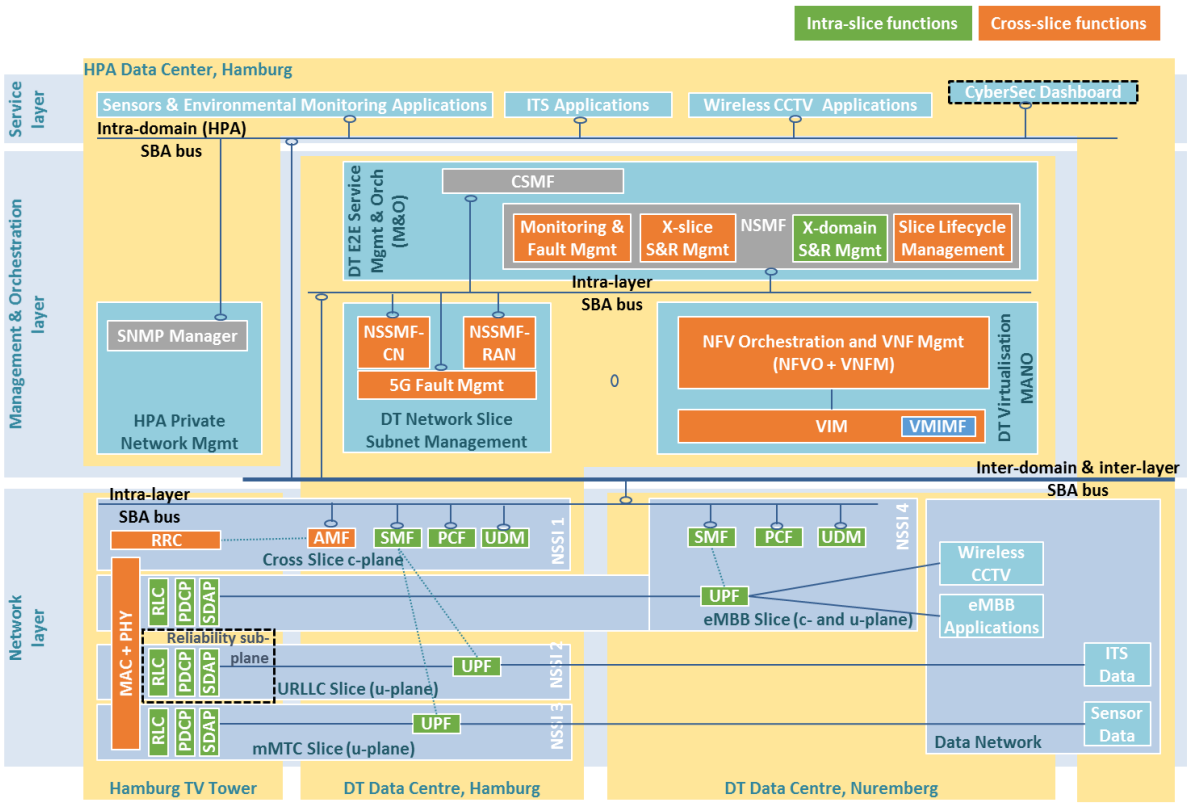


Figure 4-15: Targeted functional architecture for the Smart Sea Port use case

While the 5G-MoNArch controller layer is not utilised in the *Smart Sea Port* testbed, it realises two intra-layer SBA buses, one intra-administrative-domain SBA bus, and one inter-layer/inter-domain SBA

bus. The NFs are distributed across four locations in Hamburg and Nuremberg. In Hamburg, there are the Deutsche Telekom (DT) Data Centre, the Hamburg Port Authority Data Centre (and private networks), and the TV Tower hosting the base station. In Nuremberg, one of the central office Data Centres ('central cloud') of DT is hosted. The four locations are depicted in yellow in Figure 4-15.

In the Network layer, the testbed implements three network slices, enhanced Mobile Broadband (eMBB) communication, Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine-Type Communication (mMTC) delivering the Augmented Reality (AR), Intelligent Transport Systems (ITS), and Environmental Sensing use cases, respectively. They have the following deployment characteristics:

eMBB network slice: The eMBB network slice is utilised to carry the AR traffic (e.g., augmented maintenance for HPA service staff) as well as providing eMBB services like Internet access or video streaming to cruise ship tourists. In the RAN, the slice uses the common PHY and MAC layers of the testbed radio infrastructure. SDAP, PDCP, and RLC layers are slice-specific due to customisations reflecting service requirements. Further, RRC is common for all deployed slices. In the control plane (CP), the AMF is shared with other network slices, while PCF, UDM, and SMF are dedicated to the eMBB slice. Core network user plane (UP) function(s) are dedicated and therefore service-specific. Besides AMF, all core NFs of the eMBB slice (from CP and UP) run in DT's central cloud data centre in Nuremberg, one of DT's central office sites. Further, the AR applications (and other eMBB-like applications) in the Data Network that process the incoming user data are also hosted in Nuremberg.

URLLC network slice: The URLLC network slice is utilised for ITS applications in the sea port area, in particular for traffic light control. Similar to the eMBB slice, the URLLC slice uses the common RRC and lower radio layers (MAC and PHY) and service-specific upper radio layers (RLC, PDCP, SDAP) in the RAN. One such service-specific customisation comprises the WP3 reliability sub-plane for multi-connectivity, thus increasing reliability in the radio network. In the core network, AMF is shared with all three deployed slices, while SMF, PCF, and UDM are shared among the slices deployed in the local edge cloud (DT Data Centre, Hamburg), i.e., URLLC slice and mMTC slice. The core network UP uses a dedicated, customised UPF instance. Due to latency requirements for traffic light control, all network functionality in CP and UP is deployed locally. Therefore, also the ITS Data applications in the Data Network are operated in the local HPA Data Centre in Hamburg.

mMTC network slice: The mMTC slice is used to carry traffic from environment sensors deployed in the Hamburg sea port, particularly from the barges patrolling through the sea port. The slice has the same setup as the URLLC slice in terms of deployment of network (CP and UP, RAN, and CN) and application functions. Nevertheless, upper layer radio functions (RLC, PDCP, and SDAP) and core network UPF are realised as dedicated instances with customised behaviour.

The M&O layer comprises DT's functionalities for managing public land mobile networks (PLMN), particularly Network Slice Subnet Management Functions and novel advanced 5G Fault Management functions as developed in WP3. For the virtualisation MANO, the Hamburg sea port testbed utilises a VM-based virtualisation approach. The deployment uses a streamlined ETSI NFV MANO architecture, i.e., VIM and an NFV lifecycle management component integrating NFV Orchestrator and VNF Manager. For e2e M&O, NSMF and CSMF incorporate according monitoring as well as fault and slice lifecycle management functions. From WP3, cross-domain and cross-slice security and resilience management functions are incorporated into NSMF. A lightweight CSMF implementation provides mediation capabilities between NSMF and the Service layer. Beyond these M&O layer functions operated by DT in their Hamburg Data Centre, the deployment comprises the management functions for HPA's private networks, namely SNMP Managers running in HPA Data Centre in Hamburg. The latter functions manage the largely wireline network infrastructure of HPA which is also used to connect the UPs of the local network slices with the HPA Data Centre. More specifically, the *ITS* and *Environmental Monitoring/Sensor Data applications* run in the HPA Data Centre in Hamburg where the UP data coming from the URLLC and mMTC slice, respectively, are forwarded to. Form the mobile network perspective, these application functions belong to the Data Network outside the operator domain. Only in case of the eMBB slice, the application (*AR Data*) is hosted in the DT Data Centre in Nuremberg. Finally, each of the three applications also has a management and control component residing in the service layer, executed in the local HPA Data Centre. They interact with the CSMF to provide the specific service requirements used to customise the slice instances and to receive latest performance and configuration details about the network slice hosting the respective service.

4.3.2 Resource elasticity

4.3.2.1 Basic concepts and required architecture component

As discussed in Section 3.3.5, resource assignment in the network should avoid overprovisioning and assign resources just where and when they are needed. This flexibility is referred to as resource elasticity, which includes the ability (i) to scale resources according to the demand, and (ii) to gracefully scale the network operation when insufficient resources are available.

Elasticity has been traditionally implemented in the context of communication resources, where the network gracefully downgrades the quality for all users if its communication resources (e.g., spectrum, radio link capacity) are insufficient. In the framework of a softwarised network, a new paradigm for resource elasticity that comprises processing power, memory, and storage resources is needed. While cloud frameworks typically aim at guaranteeing that the computational resources required by a function are always there, in the orchestration environment considered here this may not be possible, since (i) the timescales involved in RAN functions are much tighter than those considered in cloud solutions, which cannot prevent outages at such short timescales, and (ii) cloud resources are typically limited at the edge, preventing cloud solutions to exploit multiplexing gains. Adding more elasticity to the resource consumption of NFs requires an understanding of the nature of the different resources and performance trade-offs, leading to different dimensions of elasticity which are described next.

This section provides a set of ideas on how to provision resource elasticity, in particular the technical challenges in the virtualised architecture of 5G systems that resource elasticity is meant to address, as well as design hints on the type of solutions or mechanisms that could address those challenges. Table 4-4 provides a summary of the content of this section.

Table 4-4: Innovation areas, challenges, and potential solutions towards an elastic architecture

Innovation Areas	Challenges	Potential Solutions
Computational elasticity	Graceful scaling of computational resources based on load	Elastic NF design and scaling mechanisms
Orchestration-driven elasticity	NF interdependencies	Elastic cloud-aware protocol stack
Slice-aware Elasticity	E2E cross-slice optimisation	Elastic resource provisioning mechanisms exploiting multiplexing across slices

4.3.2.2 Challenges that had to be solved

A first challenge in virtualised networks is the need to perform graceful scaling of the computational resources required to execute the VNFs according to the load. In that respect, the computational elasticity innovation refers to the ability to scale NFs and their complexity based on the available resources. In case of resource outage, NFs would adjust their operation to reduce their consumption of computational resource while minimising the impact on network performance.

The second challenge can be illustrated with the current LTE design of the protocol stack, where the NFs co-located in the same node are inter-dependent, i.e., interact and depend on each other. One example of logical dependencies within the stack is the recursive interaction between MCS, Segmentation, Scheduling, and RRC. In addition to logical dependencies, traditional protocol stacks also impose stringent temporal dependencies, e.g., the HARQ requires a receiver to send feedback informing of the decoding result of a packet within 4 milliseconds after the packet reception. 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 this challenge, a new protocol stack, adapted to the cloud environment, needs to be designed. This new protocol stack relaxes and potentially removes the logical and temporal dependencies between NFs, with the goal of providing a higher flexibility in their placement. This elimination of interdependencies

among VNFs allows the orchestrator to increase its flexibility when deciding where to place each VNF, hence the name orchestration-driven elasticity.

A third challenge of the envisioned 5G architecture appears at the intersection of virtualisation and network slicing, i.e., the need for E2E cross-slice optimisation such that multiple network slices deployed on a common infrastructure can be jointly orchestrated and controlled in an efficient way while guaranteeing slice isolation. To address this challenge, it is important to devise functions that optimise the network and resource consumption 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.

4.3.2.3 The use of AI

As mentioned in Section 2.2.3, 5G-MoNArch foresees that all the above elasticity-related functionalities could be greatly enhanced through AI. In the context of elasticity, we propose two different taxonomies for learning, based on i) the data used for learning, and ii) the network slice lifecycle phase [5GM-D4.2]. Firstly, with respect to the data, learning techniques for the elastic network slice management can be categorised along two main directions, independently of the actual algorithm in place:

- *Inputs*: learning techniques shall learn features from the end user demand to the network, the infrastructure utilisation, and the slice policies. These inputs shall be conveniently measurable (and labelled in case of supervised techniques) in order to be applied in one of the outputs.
- *Outputs*: following the 3GPP definition [3GPP TR 28.801], lifecycle management is composed of four stages: preparation, instantiation, run-time and decommissioning. Hence, depending on the kind of algorithms, its target and the input features, the learning algorithm shall be employed in one of these phases.

The input direction can be further split along three dimensions, depending on the characteristics of the learned input feature. In Figure 4-16, we show this three-dimensional classification, highlighting its three main axes: the *demand*, the *infrastructure* and the *requirements*. Triangles in Figure 4-16 represent the needed monitoring granularity on each of the axes: the darker the colour the finer the granularity. Rectangles indicate the different operational point of the learning processes.

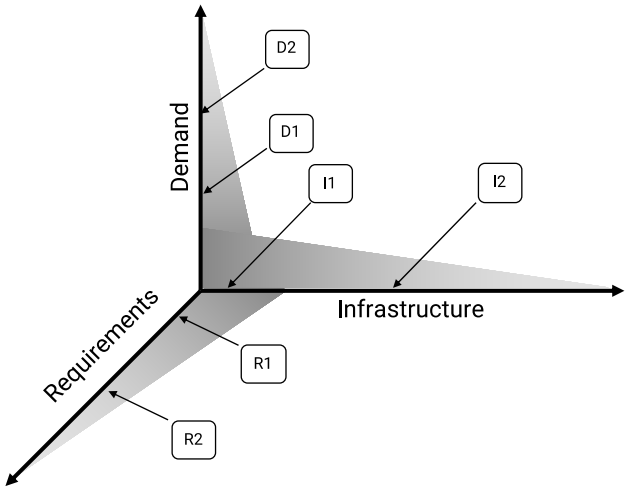


Figure 4-16: Learning taxonomy axes for slice lifecycle management

- **Demand.** This refers to the estimation of the temporal and spatial demands of services.
- **Infrastructure.** This is related to the understanding of how the underlying infrastructure reacts or limits elastic management/orchestration decisions.
- **Requirements.** Here Machine Learning is used to automatically translating consumer-facing service descriptions into resource-facing service descriptions.

On the output dimension, the proposed taxonomy refers to the network slice lifecycle phases, as various approaches can be adopted and applied in all the phases of the lifecycle of a slice instance [3GPP TR 28.801]. For example, slice behaviour analysis can be a critical asset for elasticity provisioning in the slice preparation phase, since statistics can be exploited to efficiently decide the basic configurations and set the network environment.

As examples, we provide a few insights and use cases on AI-based elasticity mechanisms that are applied in the instantiation and run-time phases:

- **Instantiation phase.** Here the AI mechanism ca decide the admission of new slices and potentially the re-configuration of the running slices in the network.
- **Run-time phase**. Here fast time scale adaptation is implemented, including reconfigurations at VNF or slice level.

4.3.2.4 Architectural considerations

Motivated by the prominent role that AI can play in modern telecommunication systems, 5G-MoNArch also decided to participate in the activities of the Experiential Network Intelligence (ENI) workgroup [WF18], created by ETSI. 5G-MoNArch research activities perfectly fit with the goals set by ENI and the ENI's system can integrate and be beneficial to 5G-MoNArch's architecture [5GM-D4.2]. Focused on optimising the operator experience, this engine would be equipped with big data analytics and AI capabilities that could enable a much more informed elastic management and orchestration of the network, often allowing proactive resource allocation decisions based on the history rather than utilising reactive approaches due to changes in load. Figure 4-17 depicts a 3GPP-compliant management and orchestration architecture, integrated with an ENI engine. The interaction and interoperability of ENI with an assisted system is determined by the latter's support of the ENI Reference Points. As depicted in Figure 4-17 below, the current NFVI Information allows ENI to be aware of the computational resources' capabilities (e.g., type of CPU, memory, data plane and accelerators) and availability (utilisation level), while in turn this enables ENI to influence and optimise placement decisions made by the VIM, while ensuring that 3GPP policies, resources allocation and SLA are adhered too. Moreover, by using this information, ENI can further optimise resource utilisation by i) enabling higher density for a given set of workloads under associated SLA, ii) anticipating and reacting to changing loads in different slices and assisting the VIM in avoiding resource conflicts, and/or iii) timely triggering of up/down scaling or in/out scaling of associated resources.

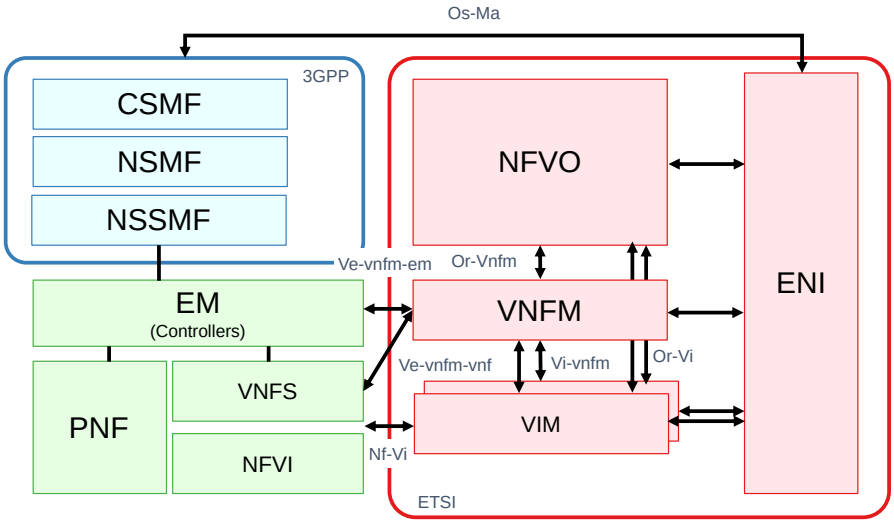


Figure 4-17: Joint ETSI – 3GPP management and orchestration architecture

Elasticity mainly addresses two domains: network M&O, through Cross-domain Elasticity Mgmt., Cross-slice Elasticity Mgmt., and the big data module, and network controllers, through the intra-slice controller (ISC) and cross-slice controller (XSC), which interact each other through the MOLI interface.

The former shall incorporate the elements needed to (i) flexibly assign resources to different slices and (ii) find the best location of a VNF belonging to a certain network slice within the infrastructure. The latter, instead, shall provide an inner loop control of NFs, to enforce elasticity at faster time scales, such as the ones needed in the RAN.

The overall view of the elasticity modules in the 5G-MoNArch architecture is provided in Figure 4-18 (WP4 specific modules are highlighted in yellow).

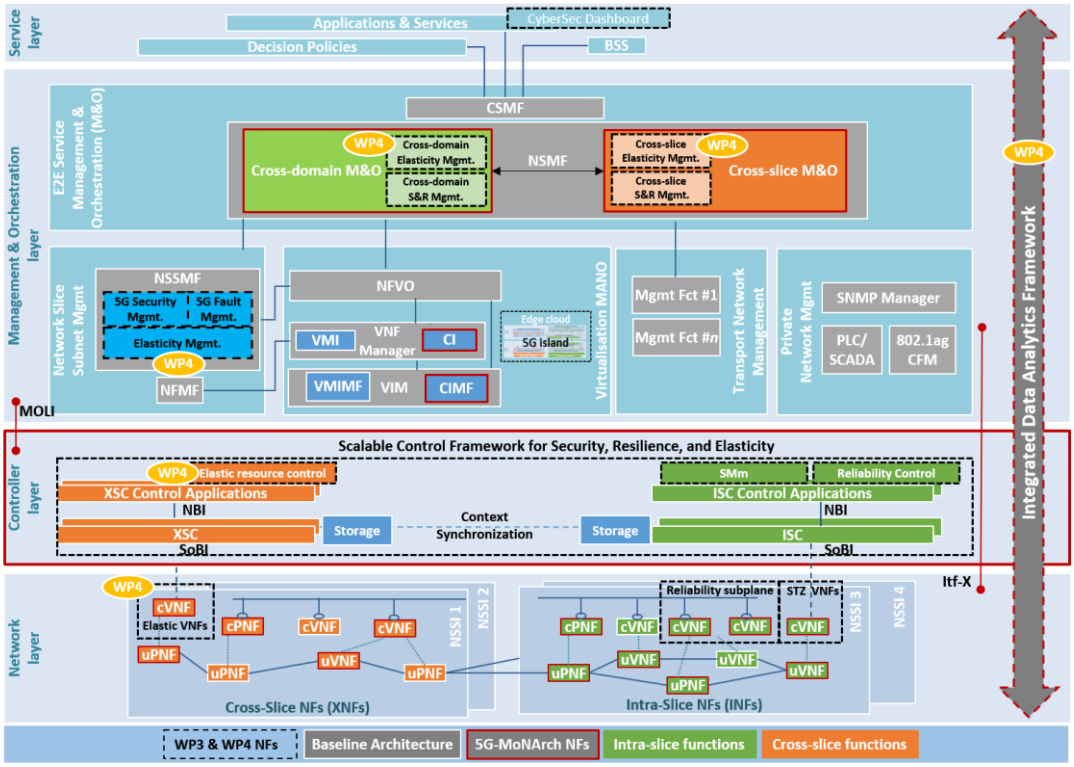


Figure 4-18: 5G-MoNArch overall architecture with the marked WP4 modules to enable resource elasticity functional innovation [5GM-D4.2]

The interactions between the M&O layer and the controller layer are described in Figure 4-19. The details on the depicted interfaces are provided in D4.1 [5GM-D4.1]. The following paragraphs provide some more detail on each of the identified innovation areas.

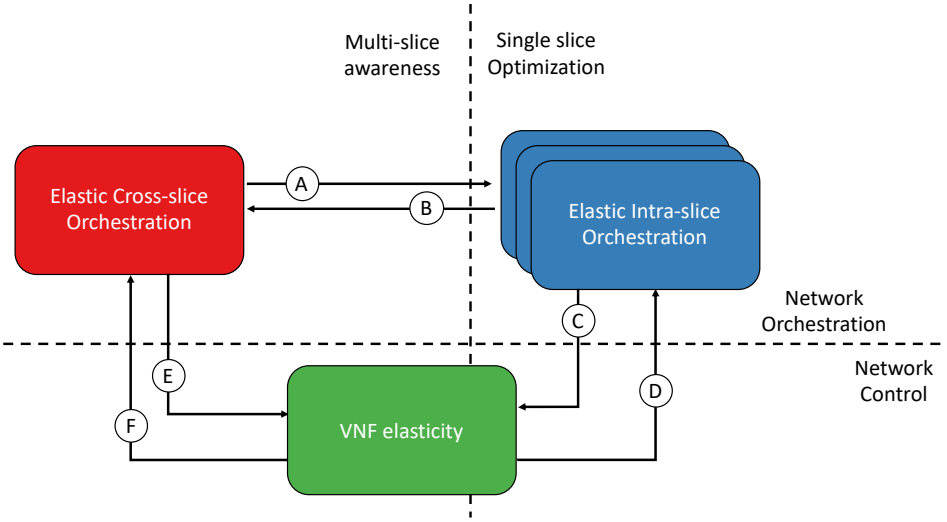


Figure 4-19: High-level interactions across elastic modules in the M&O and Controller layers

4.3.2.5 Computational elasticity

The goal of exploiting computational elasticity is to improve the utilisation efficiency of computational resources by adapting the NF behaviour to the available resources without impacting performance significantly. Furthermore, this dimension of elasticity addresses the notion of computational outage, which implies that NFs may not have sufficient resources to perform their tasks within a given time. In order to overcome computational outages, one potential solution is to design NFs that can gracefully adjust the amount of computational resources consumed while keeping the highest possible level of performance. RAN functions in particular have been typically designed to be robust only against shortages on communication resources; hence, the target should be directed at making RAN functions also robust to computational shortages by adapting their operation to the available computational resources. An example could be a function that chooses to execute a less resource-demanding decoding algorithm in case of resource outages, admitting a certain performance loss.

In addition, the scaling mechanisms, i.e., the modification of the amount of computational resources allocated to such computationally elastic NFs may help in exploiting the elasticity of the system if they are properly designed. There are two significant ways to scale a NF: (i) horizontal scaling, where the system is scaled up or down by adding or removing new identical nodes (or virtual instances) to execute a NF, and (ii) vertical scaling, where the system is scaled out or in by increasing or decreasing the allocated resources to the existing node (or virtual environment) [Wil12]. As an example, in the RAN domain, supporting higher system throughput by adding additional access points is referred as horizontal scaling, whereas an increase in operating bandwidth is referred as vertical scaling.

Figure 4-20 depicts a possible way to activate computational elasticity, including communication between the VIM and the I/XS Controller via the MOLI interface, which triggers a re-orchestration by the VIM. Alternative solutions that do not require an additional Controller layer can be envisioned too [5GM-D4.2].

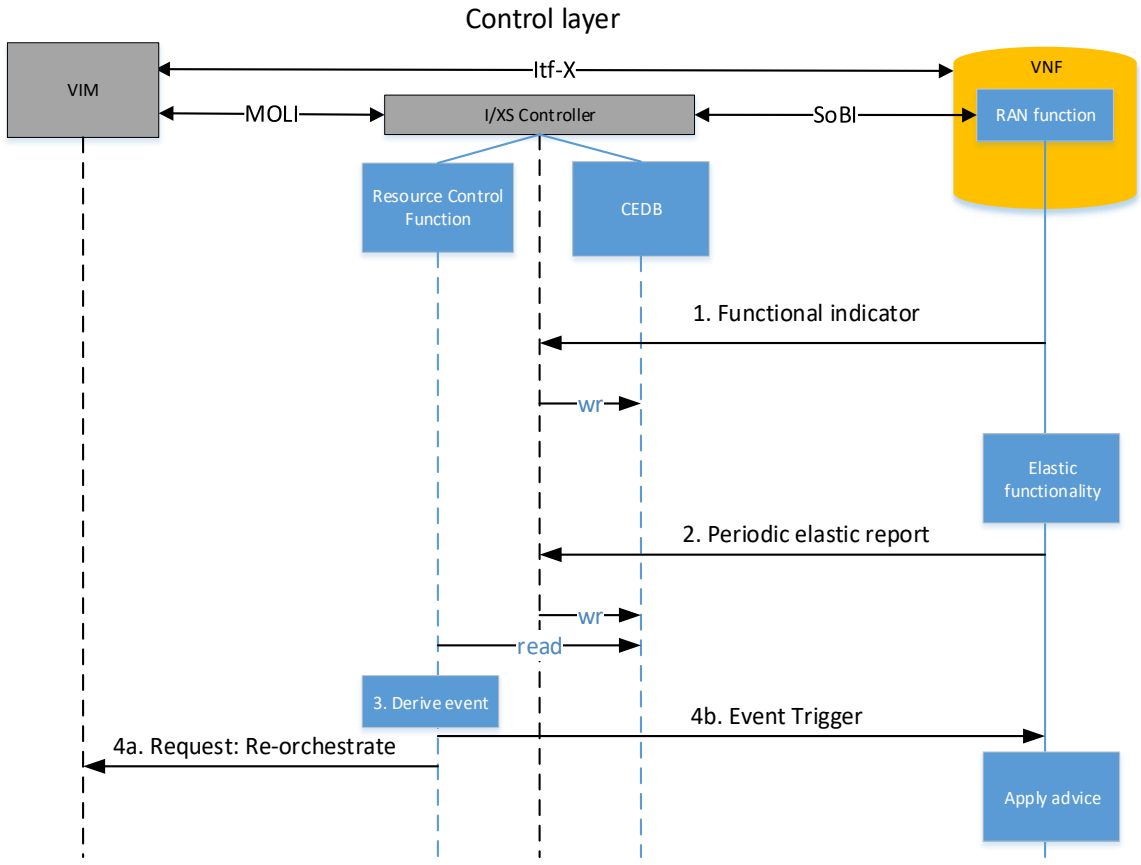


Figure 4-20 Message sequence chart for computational elastic operations [5GM-D4.2]

4.3.2.6 Orchestration-driven elasticity

This innovation focuses on the ability to re-allocate NFs within the heterogeneous cloud resources located both at the central and edge clouds, considering service requirements, the current network state, and implementing preventive measures to avoid bottlenecks. The algorithms that implement orchestration-driven elasticity need to cope with the local shortage of computational resources by moving some of the NFs to other cloud servers which are momentarily lightly loaded. This is particularly relevant for the edge cloud, where computational resources are typically more limited than in the central cloud. Similarly, NFs with tight latency requirements should be moved towards the edge by offloading other elastic NFs without such tight timescale constraints to the central cloud servers.

To efficiently implement such functionalities, special attention needs to be paid to (i) the trade-off between central and edge clouds and the impact of choosing one location for a given function, and (ii) the coexistence of Mobile Edge Computing (MEC) and RAN functions in the edge cloud. This may imply scaling the edge cloud based on the available resources, clustering and joining resources from different locations, shifting the operating point of the network depending on the requirements, and/or adding or removing edge nodes [OSB16].

Orchestration-driven elasticity mechanisms may be activated when the fulfilment of the requirements of provided services are jeopardised by an evolution of the network conditions under which the initial resource assignment was performed. This may happen for various reasons: an increase in the overall network work load, the worsening of radio communication conditions that impact the VNFs' operations, the modification of the requirements of other services instantiated in the same slice and the consequent re-adjustment of the allotted resources, the implementation of elasticity mechanisms at a VNF's or an inter-slice level, etc. Figure 4-21 provides a high-level description of how VFN re-orchestration is triggered and implemented in this case.

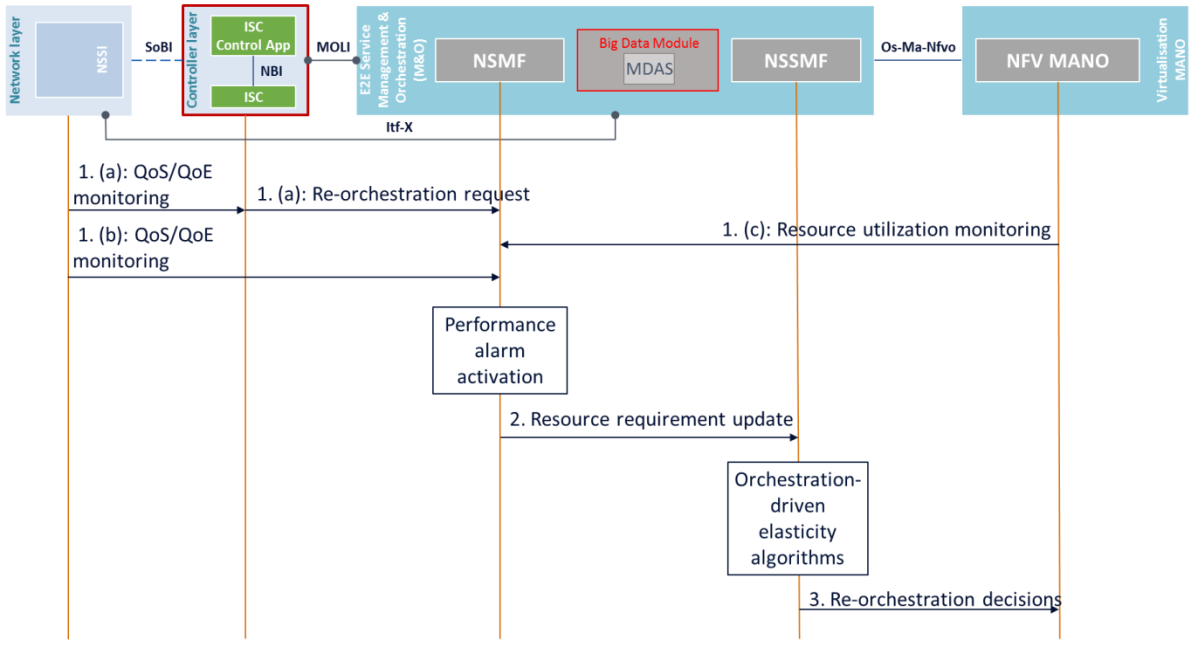


Figure 4-21: Flow of information for VNF re-orchestration in case of performance degradation [5GM-D4.2]

Slice-aware Elasticity

Finally, this dimension of elasticity addresses the ability to serve multiple slices over the same physical resources while optimising the allocation of computational resources to each slice based on its requirements and demands, a challenge earlier referred to as E2E cross-slice optimisation. Offering slice-aware elastic resource management facilitates the reduction of Capital Expenditure (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, as Figure 4-22 illustrates.

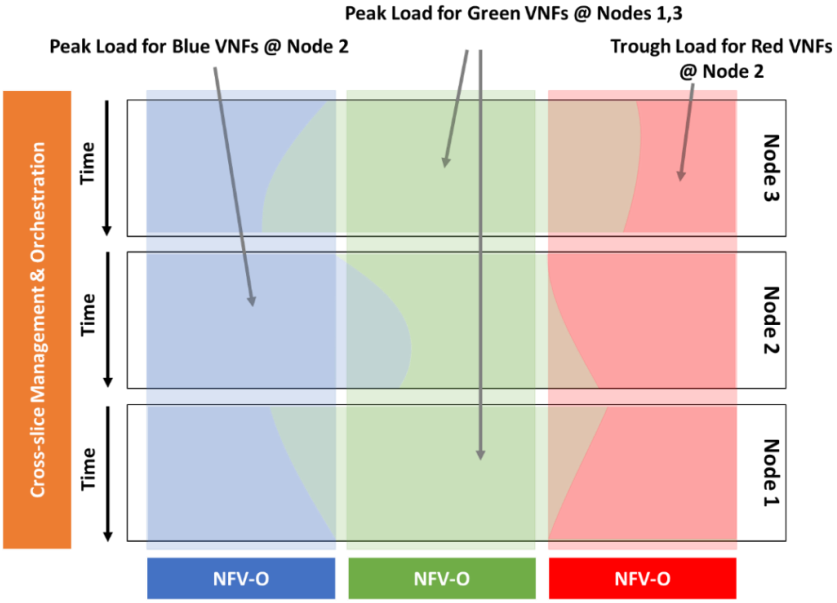


Figure 4-22: Illustration of slice-aware elasticity

Adaptive mechanisms that exploit multiplexing across different slices (when resource isolation is not needed) must be designed, aiming at satisfying the slice resource demands while reducing the amount of resources required. Hence, the solutions must necessarily dynamically share computational and communications resources among slices whenever needed. An elastic admission control system would be also required, as elastic slices need not have the same amount of available resources as e.g., a highly resilient slice where all resource demands must be fully satisfied at each point in time. Furthermore, in this context a monitoring module should be deployed to retrieve the information required to take optimal sharing decisions, considering trust relationships issues for slices managed by different tenants. Figure 4-23 shows an example of how slice-aware elasticity measures can be implemented, triggered by intra- and inter-slice processes or consequences of the implementation of other elasticity measures (computational elasticity or slice-aware elasticity). In such a case, a performance alarm comes from the infrastructure and is notified to the NFV-O. The latter notifies the Management system about the imminent network re-orchestration, which in turn grants the need for a re-orchestration back to the NFV MANO, charged of enforcing it.

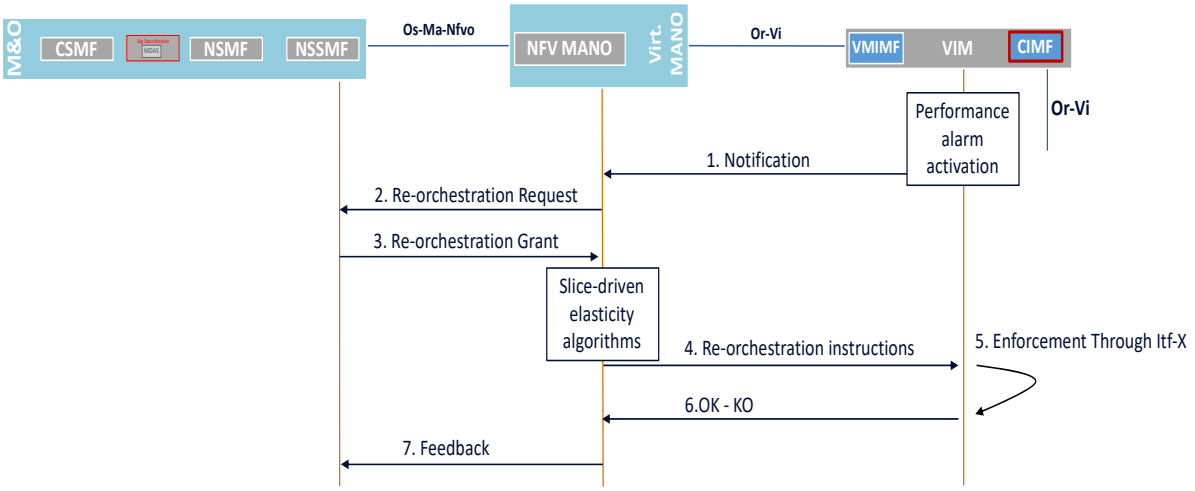


Figure 4-23: Flow of information for slice-aware re-orchestration in case of performance degradation [5GM-D4.2]

4.3.2.7 Network architecture for the Touristic City use case

As for the *Smart Sea Port* use case deployed in the Hamburg testbed of 5G-MoNArch, a customised architecture instance of the general overall architecture as described in Section 2.1 is deployed for the *Touristic City* use case. Also, for this scenario, a subset of the 5G-MoNArch architecture is deployed, with some specific modules developed in WP4 (cf. Section 4.3.1.1), among them the Cross-Slice elasticity module, the big data analytics module, and the slice subnet elasticity module.

The *Touristic City* target architecture instance is depicted in Figure 4-24 below that shows the NFs in each layer, Network layer, M&O layer, and Service layer. Currently, the optional 5G-MoNArch Controller layer is not part of the *Touristic City* testbed, but possibly some of the core network functionality may be implemented in such way. Due to the smaller extent, the NFs of this testbed are deployed in three main location: the antenna site (where the radio PNF are executed), the edge cloud, and a central cloud (that has a higher latency to the UE). They are depicted in yellow in Figure 4-24. In practice, the edge and the central cloud are two dedicated cloud infrastructure domains, connected via fibre to the antenna site. The central cloud emulates a farther processing site, with an increased latency but a lower operational cost. Both sites are deployed in the premises of the demo but are owned by the operator.

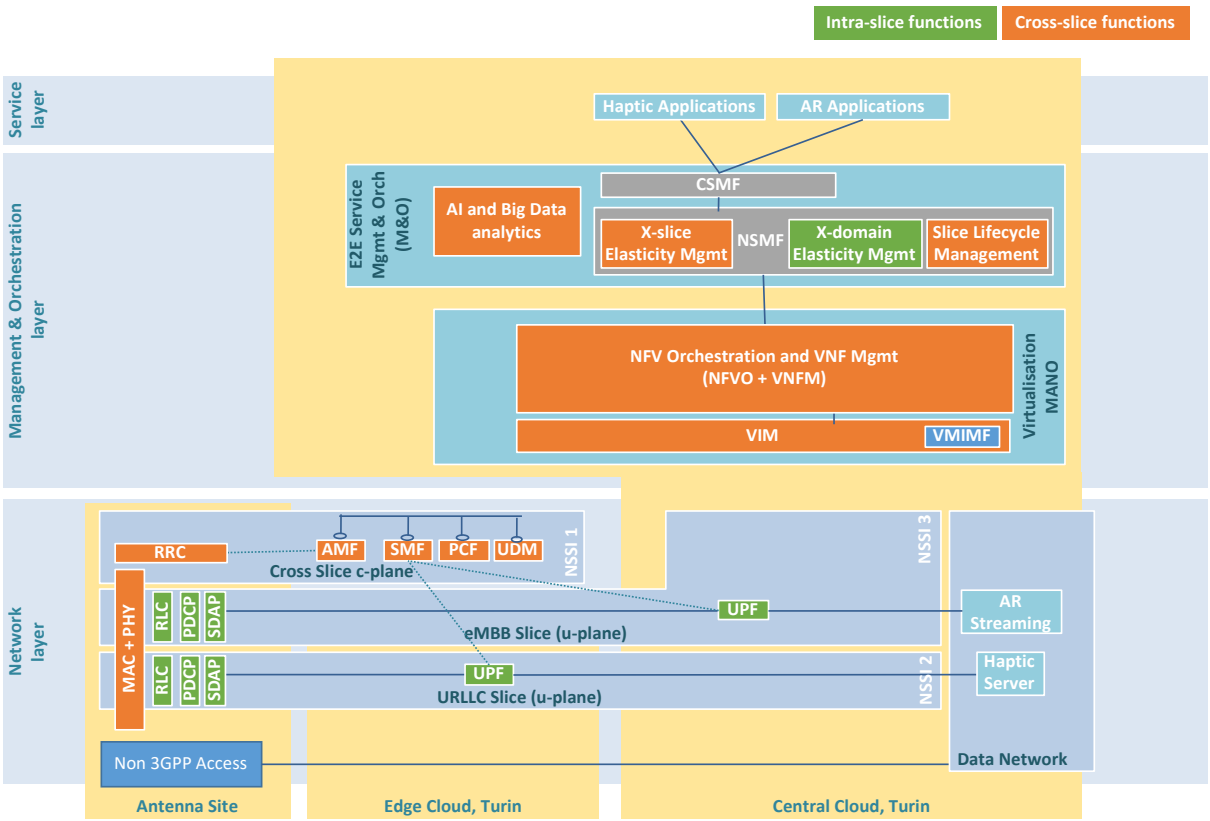


Figure 4-24: Targeted functional architecture for the Touristic city use case

In the Network layer, the testbed implements two network slices: an enhanced Mobile Broadband (eMBB) communication and an Ultra-Reliable Low-Latency Communication (URLLC), one. They are used to provide two different services: the high-res video streaming for the Augmented Reality applications and the haptic server that connects the avatars for their interactions. They have the following deployment characteristics:

- eMBB network slice: The eMBB network slice deliver the high resolution 360 video to the mobile user. In the Radio Access Network (RAN), the slice uses the common PHY and MAC layers of the testbed radio infrastructure, while the higher layers are slice-specific due to customisations reflecting specific service requirements. The RRC instead, is common to both slices. The CP functionality is shared across slices, while the UP function (UPF) is dedicated to

each slice. In terms of deployment, the core functions are deployed in the central cloud, as well as the UPF. Also, the application server run in the central cloud,

- URLLC network slice: The URLLC network slice is utilised for delivering the low latency haptic interactions among the avatars (one fixed and one mobile). The radio deployment is equivalent to the eMBB network slice. Also, the core function setup is similar in terms of sharing and deployment. However, the UPF may be moved from one cloud to the other according to the specific load of the network, according to the inputs coming from the elasticity modules deployed in the NFV-O.

The management of the network comprises an implementation of the 3GPP elements CSMF, NSMF and NSSMF that, in turn, include specific elasticity modules such as the network slice admission control. The testbed includes both PNFs and VNFs that are managed by a VM-based virtualisation approach and the related MANO modules. The MANO stack is a simplified one, which relies on a VIM and an ad-hoc implementation of the VNFM and NFV-O. Nevertheless, container-based virtualisation may be included, particularly for the radio functions. The loop is closed by monitoring modules that report the current load to the management modules that use this information (e.g., network load, CPU load) to trigger both cross-slice and intra slice elasticity algorithms.

Finally, there is another network element in the architecture that is related to the non 3GPP access network (mmWave links in this case). Although this part is not directly orchestrated as VNF (the mmWave access points are PNF), it is still linked to the rest of the network and managed by the implemented management system that takes care of its lifecycle management.

4.4 *5G-MoNArch enablement of the 5G ecosystem evolution*

In Section 4.3, we described how the 5G-MoNArch final architecture can be flexibly extended to meet the specific requirements of the Hamburg Smart Sea Port (resilience, security) and the Turin Touristic City (resource elasticity) use cases. To conclude Chapter 4, in this section, we present a business-oriented analysis that highlights the benefits provided by such a flexible architecture in two use cases inspired by the two 5G-MoNArch testbeds. In particular, we examine the new types of ecosystem that might emerge in these scenarios which are enabled by the flexible architecture and features proposed under 5G-MoNArch.

4.4.1 Recap of mobile business case and ecosystem evolution anticipated for 5G

As was highlighted in D2.2 [5GM-D2.2], network slicing in 5G is seen as an opportunity to:

- Increase revenues by introducing new mobile services. These new mobile services not only increase revenue by introducing new subscribers. As many of these new services can be tailored to the individual requirements of customers they are also more likely to be higher value (in terms of revenue per GB) business to business (B2B) services for verticals.
- Reduce the cost per GB compared with MBB and eMBB only networks by providing a wide range of services from a single multi-service network. This means that network providers can not only extend the benefits from economies of scale already seen by delivering higher volumes of traffic (as already seen for MBB) but can also benefit from economies of scope delivered by multi-service platforms.

The above two trends have implications of:

- Introducing new types of end users to the mobile ecosystem (some of whom may, unlike consumers, have access to existing infrastructure).
- Increasing shared use of infrastructure from a range of sources to drive costs down.

To achieve the above objectives any 5G network architecture must therefore not only be able to rapidly and flexibly offer new tailored services or slice instances but must also be able to adapt to new infrastructure or service provision partnerships that might be formed.

To illustrate the above points, in the next two sections we examine the following use cases, inspired by the two 5G-MoNArch testbeds, of an existing mobile broadband service provider engaging with:

- A port authority to provide mobile services related to improved operations and passenger experience in their sea port.

- A city council and city attractions and venues to provide improved visitor experience and mobile services particularly in tourism hotspot locations in the city

4.4.2 Example deployment model and infrastructure partners for a 5G network in a sea port use case

As also highlighted in D2.2 [5GM-D2.2], WP6, which examines verification and validation of 5G-MoNArch, have developed a series of evaluation cases based in Hamburg city and including the sea port area (see Figure 4-25). To illustrate the architectural flexibility that might be required to accommodate the evolving ecosystem mentioned in the previous section, Figure 4-26 illustrates the range of infrastructure owners and players that might be involved in providing mobile services across Hamburg city and also more bespoke mobile services to the sea port area considering the three example service areas indicated on Figure 4-25 which consist of:

- The densely populated city centre area north of the river (shown in red) with a population of approximately 160,000 residents.
- The industrial area south of the river (shown in blue) which is managed by the Hamburg Port Authority (HPA). This contains infrastructure specific to the Hamburg sea port such as the container terminals, cruise ship terminal etc. but also many public roadways and buildings. There are approximately 3,000 residents in this area.
- The Steinwerder cruise ship terminal within the HPA industrial area. This can have ships carrying up to 5,000 passengers and 1,500 staff arriving at any one time creating a temporary demand hotspot in this lightly populated industrial part of the city.

Figure 4-25 considers the communications infrastructure that might be available in these three example service areas.

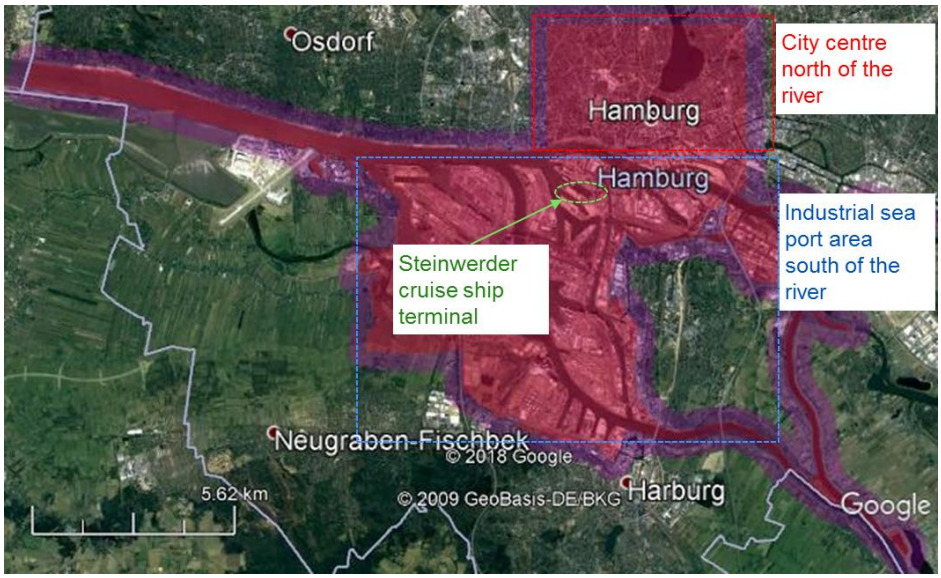


Figure 4-25: Hamburg area with the example three service areas shown

Firstly, eMBB services to consumer portable devices will be provided via the existing wide area network made up of MNO procured and managed antenna sites, edge cloud sites and fibre (illustrated by the red "MNO procured" blocks on the left of the diagram). Where appropriate these may be shared with other MNOs and/or be supported by a small cell layer particularly in the busy city centre areas if extra capacity is needed. This is the "business as usual" case of delivering today's services via a virtualised network and potentially using slicing to differentiate between different eMBB products on the basis of quality of service. In this case virtualising the network will likely deliver cost efficiencies by centralising some of the network processing away from the antenna sites and availing of diversity in traffic across antenna sites being served by the same edge cloud site. These savings will be increased further if elasticity can

be used to move NFs between edge cloud sites and make use of spare processing at edge cloud sites that may be quiet at particular times of day while other areas are experiencing peak demand levels.

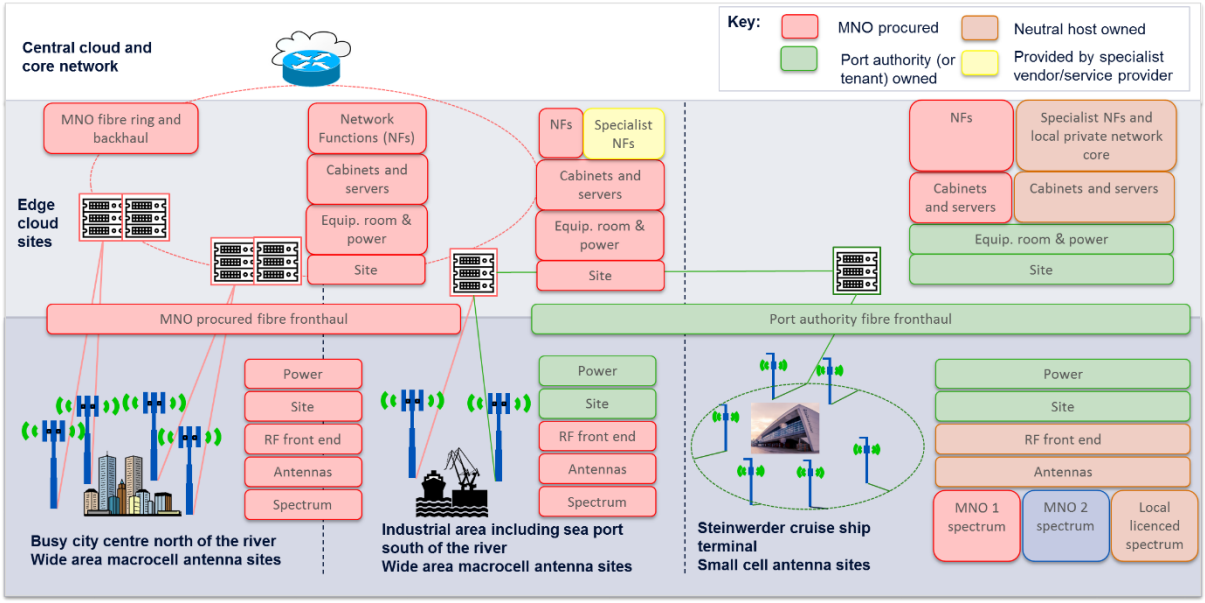


Figure 4-26: Potential range of infrastructure owners involved in the delivery of mobile services to Hamburg city and the sea port area

Considering the more industrial areas of the city south of the river, the "business as usual" eMBB services to consumer portable devices will continue to be provided via the existing MNO sites in the area in a similar way to the city centre. However, if the MNO engages with the port authority to provide bespoke wireless services to support operations in the port area this begins a relationship between the MNO and an end user who has significant existing infrastructure including land, buildings, street furniture, a fibre ring and data centre. As shown by the green "Port authority (or tenant) owned" boxes in the central part of the diagram, this means that should the MNO need to extend their network in the sea port area to acquire higher capacity or availability for the industrial grade sea port services that the MNO might be able to reduce the cost of adding new sites in this area by partnering with the port authority and making use of their street furniture for new site locations and existing fibre connectivity. This of course may be beneficial to the quality of the "business as usual" eMBB services being delivered in this area also.

Finally, considering demand hotspot services such as around the cruise ship terminal, as highlighted in D4.2 [5GM-D4.2], an MNO is unlikely to gain extra revenue and hence invest in infrastructure around the cruise ship terminal to address this temporary eMBB demand hotspot generating by passengers and staff arriving in that area. If MNOs are not willing to make the upfront CAPEX investment in a small cell deployment in hotspot scenarios such as this, another option is for the building owner themselves to commission a neutral host system or to contract the installation and running of such a system to a third party. There is some precedence for this already at the Steinwerder cruise ship terminal where the existing indoor Wi-Fi access points have been paid for and installed by MobyKlick. From this infrastructure set MobyKlick provide a Wi-Fi service free of charge in the terminal building. In the case of a 5G neutral host system, the neutral host or terminal operator (depending on the commercial arrangement) would take the risk of making the upfront CAPEX investment in the small cell antenna sites needed to boost capacity around the terminal building. This would leave the CAPEX infrastructure investment on the MNOs part to be limited to ensuring enough processing is available across their edge cloud sites to serve the hotspot at the time that it occurs in return for an OPEX fee for accessing the neutral host system around the cruise ship terminal. Additionally, as cruise ship liners and terminal operators continue to investigate ways of improving the customer experience and efficiency of transferring such large volumes of passengers from the cruise ship to the terminal and to onward transport, any provider of such a neutral host system around the terminal might also have the opportunity

to use localised spectrum to provide dedicated operational services to the cruise liners and cruise ship terminal or indeed more specialised applications to the passengers themselves.

4.4.3 Example deployment model and infrastructure partners for a 5G network in improving visitor experience in city centres and venues

In line with the second 5G-MoNArch testbed in Turin, the partnerships and deployment types needed to use 5G networks to improve visitor experience in a city centre or venue are next considered and illustrated in Figure 4-27.

In this scenario it is assumed that an existing MNO engages with the city council and/or city transport authorities to provide improved mobile services in city hotspot locations. The primary aim is to increase capacity and quality of experience of mobile services in busy city locations so that the experience of visitors to the city can be improved and new mobile tourism applications and experiences supported. Additionally, improved smart city services such as support for intelligent transport systems could be provided to the city council or transport authority under such partnership arrangements. Similar to the sea port scenario, a city council or transport authority will already have an existing set of street furniture, such as lampposts, CCTV poles and traffic lights, with many of these already having at least an existing power connection if not also a fixed network connection. Making this portfolio of assets available for small cell antenna sites under a partnership arrangement with MNOs or a third-party neutral host would significantly reduce the initial CAPEX investment by saving on site acquisition costs and dig costs for installing power and connectivity.

Enhanced wireless services to support improved visitor experiences are also expected to be consumed on a much more localised basis such as in a museum or venue. This scenario is shown on the right of Figure 4-27. Depending on the venue size, the associated revenues may not be enough for a single MNO to support a dedicated small cell deployment. As such, similar to the cruise ship terminal discussed earlier, a neutral host solution that including a private element for museum operation specific services, such as a push to talk (PTT) network for security and public safety staff, and a public element for allowing visitors to make use of applications on their own devices could be envisaged.

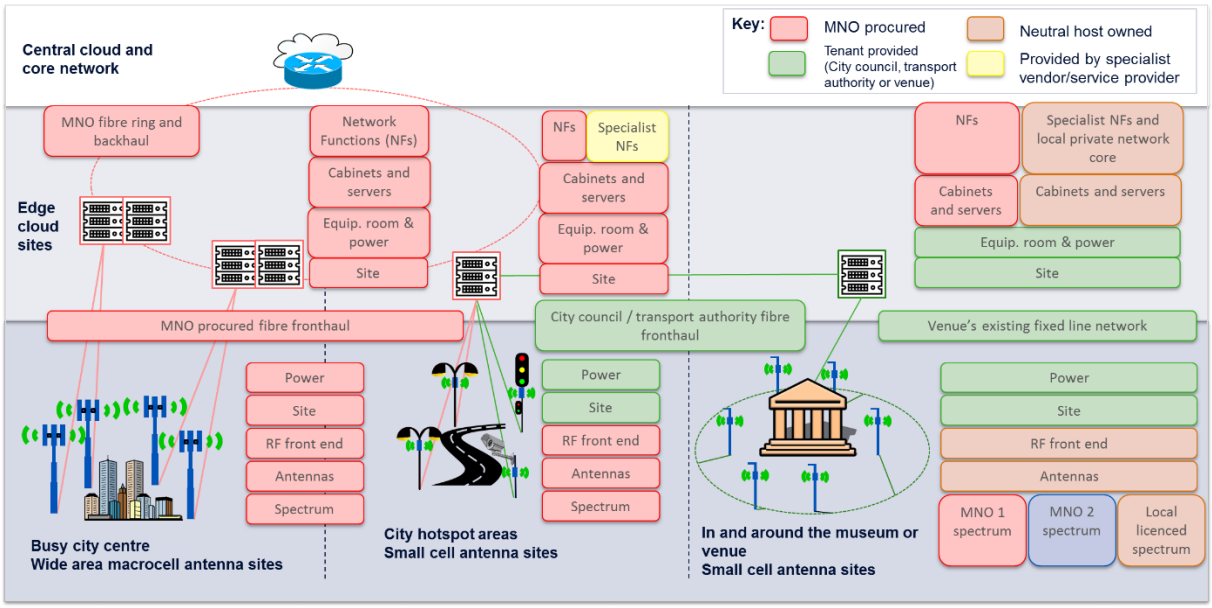


Figure 4-27: Potential range of infrastructure owners involved in the delivery of mobile services to improve visitor experience in a city and venues

4.4.4 Architectural implications of example deployment model examined

For the example deployments shown in the previous sections we can conclude that any 5G architecture will ideally need to:

- Be flexible and adaptive to new radio, transport and processing resources becoming available as new partnerships are formed with new tenants (e.g., the Inter-slice resource management enablers presented in Section 3.3).
- Be able to manage dimensioning of the available network resources optimally to meet the expected demand patterns of the services being provided and readily accommodate new services (e.g., the Inter-slice Management and Orchestration enablers presented in Section 3.4).
- Provide a protocol stack that can be partitioned and run by different players and partners in potentially different locations to accommodate different depths of slicing and infrastructure and service provision partnerships. This also requires open interfaces between NFs also (we study these aspects in the Telco-cloud-enabled protocol stack presented in Section 3.1).

These are all key elements of the 5G-MoNArch architecture described throughout this document.

5 Conclusions and Outlook

This deliverable has refined the initial architecture of 5G-MoNArch (from [5GM-D2.2]) towards the "Final Overall Architecture". Particularly, the design of the 5G-MoNArch overall architecture has followed the baseline requirements and related KPIs for 5G-MoNArch [5GM-D6.1] and has provided novel components and extensions to the baseline architecture to address the identified 5GS gaps presented in D2.1 [5GM-D2.1] (see Appendix A). In this regard, a summary of the enhancements that address these gaps is provided in Table 5-1. Therein, it is also highlighted how enabling and functional innovations have responded to the 5GS gaps. In particular, the design of the 5G-MoNArch overall architecture (i) takes into account the current SotA on 5G architecture, from previous 5GPPP Phase I collaborative projects as well as ongoing standardisation efforts, (ii) addresses the 5GS gaps identified via enabling and functional innovations, (iii) provides a complete architecture design for E2E network slicing realisation, comprising SotA and novel components as well as the descriptions of interfaces between them and (iv) provides optional integration of radio access network (RAN) control applications via RAN exposure functionality.

The proposed overall architecture consists of four different layers identified as Service layer, Management & Orchestration (M&O) layer, Controller layer, and Network layer. One of the key contributions is the detailed definition of the role of each layer, the relationship between layers, and the identification of the required internal modules within each of the layers. In the proposed overall architecture, multiple management domains for E2E network slice deployment and operation have been explored from 3GPP, ETSI NFV, and ETSI ENI perspectives. In particular, the proposed final overall architecture extends the reference architectures proposed by 3GPP and ETSI by addressing 5GS gaps identified within the corresponding baseline models. The overall architecture is presented both from a reference-point perspective and SBA perspective. The latter representation along with the extensions on unified SBA design is expected to be of particular relevance for future 5G releases beyond 3GPP Release 16. In this context, we have presented in detail the 5G-MoNArch Itf-X interface as well as its SBI representation, which enables interactions between functions on the M&O layer and network layer, in order to achieve enhanced flexibility and orchestration capabilities. In addition to the long-term expected impact, novel 5G-MoNArch solutions have already been agreed by various SDOs as presented in Section 2.1.4 with around 80 accepted/approved technical contributions grouped under 5G-MoNArch KTAs by the time of writing this deliverable. Capitalising on the SBA and SBI, an integrated data analytics framework has been proposed, where the roles of data analytics functions in different layers are sketched and the interactions among them are provided. As part of the integrated data analytics framework, we have proposed novel big data analytics module, which is accepted as the MDAF in 3GPP. Moreover, we have investigated how to extend the NWDAF introduced in 3GPP Release 15 to provide the 5G-MoNArch architecture with the ability to collect and analyse per-slice aggregated data, and to support E2E network optimisation. The data analytics capability has been also introduced to RAN, where optional exposure to the controller layer is presented.

The 5G-MoNArch novel components originate from the enabling and functional innovations, where the enabling innovations have been within the focus of this deliverable. Under three enabling innovations, 19 enablers have been developed where two of these enablers originating from the functional innovation resource elasticity are tightly related and thus are captured in detail herein. The analyses and evaluation results have been presented as means of highlighting their capabilities in meeting E2E slice requirements as well as slicing principles. The positioning of these enablers within the overall architecture has been described via protocols in terms of MSCs among the 5G-MoNArch novel components.

The ultimate goal of the proposed overall architecture is to allow for the instantiation of slices that can satisfy specific SLAs/requirements. Therefore, the proposed architecture accommodates potential NFs and solutions to achieve slice resiliency, security, and elasticity. These functions can be thus instantiated by the 5G-MoNArch architecture when deploying slices that need to provide the corresponding tailored services. Accordingly, we focus on the specific extension of the 5G-MoNArch architecture with respect to the testbed use cases investigated in WP3 [5GM-D3.2] and WP4 [5GM-D4.2], and the associated testbeds, to demonstrate network slicing elasticity, resilience and security. The 5G-MoNArch Network Slice Blueprint concept is thus the universal means for such service-specific design and operations of network slices. In this deliverable, we have described how to design and implement the network slice

blueprint starting from GSMA GST and how the slice blueprint can be used as input for the network slice M&O process.

The 5G-MoNArch enhancements addressing the 5GS gaps have provided novel components that substantially improve the baseline capitalising on the standardisation efforts in 3GPP, ETSI ENI, and ETSI ZSM along with collaborations with key industry fora GSMA and NGMN. Future standardisation efforts as of 3GPP Release 17 will be looking into the real deployment issues considering the cloud environment, vertical use cases, as well as enhanced data analytics. In this direction, ETSI ISGs have been focusing on the enhancements of the network management via data analytics as well as on the simplification of the network management via automation. On this basis, some of the aspects that can be or will be further studied are provided briefly in the following.

- The work herein has shown that the realisation of the E2E network slicing and the associated flexible and adaptive network architecture design require a cross-SDO/cross-industry collaboration. For example, the 5G M&O layer enhancements need collaboration among 3GPP working groups and ETSI ISGs. Further, the slice realisation starts with a slice template which is provided by the GSMA based on the thorough analyses of the vertical requirements. Therefore, cross-SDO/cross-industry collaboration will remain essential for the 5GS realisation.
- Although one of the main motivations of the work herein has been impacting the specifications prepared by the target SDOs, various enhancements have long-term implications. These include unified SBA, integrated data analytics, and network exposure. For instance, 3GPP SA2 currently finalises the scope of a new study item for eNA (Phase 2), commencing in Q4 2019, that several studies from 5G-MoNArch (KTA 1-KTA 5) may impact the outcome specifications[8].
- The proposed optional controller layer enables the inclusion of the control Apps that can support RAN control NFs. Joint optimisations of such control Apps shall be studied so that conflicting policies can be avoided.

Table 5-1: 5G-MoNArch enhancements to address the 5GS gaps

5GS Gap	5G-MoNArch Enhancement
GAP #1	5G-MoNArch has provided *Telco-cloud-aware protocol stack design* (see D2.3 Section 3.1) reducing the inter-dependencies between NFs and enabling flexible deployment of RAN NFs at centralised locations
GAP #2	5G-MoNArch has designed both the architectural elements (such as the Big Data Module and the needed interfaces), the algorithms (described in [5GM-D4.2]) and implemented a selection of them in the Touristic City Testbed (see [5GM-D5.2] for details)
GAP #3	5G-MoNArch has introduced a paradigm change from fixed functional operation of small cells toward slice-adaptive operation via *inter-slice resource management* (see D2.3 Section 3.3)
GAP #4	5G-MoNArch has proposed novel MM paradigm via floating mobility anchors to facilitate offloading some signalling at the RAN level (from direct signalling to the gNB to indirect signalling between anchor and Remote UEs).
GAP #5	5G-MoNArch has provided *Inter-slice RRM using SDN framework* (see D2.3 Section 3.3) to improve the QoS of high-priority slices. Moreover, the SDN framework is re-designed to provide telco-grade performance such as, high availability and load balancing.
GAP #6	5G-MoNArch has designed *E2E cross-slice optimisation mechanisms* for joint resource allocation, when multiple slices need to share the same network resources (e.g., spectrum pool), and also inter-slice coordination mechanisms for enabling the 5GS to support UE applications using multiple QoS flows in multiple network slices (see D2.3 Section 3.2, Section 3.3, and Section 3.4).

[8] These study item proposals are to be approved by 3GPP SA plenary meeting, i.e., SA#84 in June 2019.

GAP #7	5G-MoNArch has conducted extensive experimentation on open-source VNF implementations, whose results have been described in [5GM-D2.1] and Section 3.5 of this deliverable.
GAP #8	5G-MoNArch has conducted a 5G-specific security analysis that involves a classification of potential threats with respect to their impact on the 5G network performance. Particular emphasis is put on Hamburg Smart Sea Port testbed where particular elements of such 5G deployment are considered [5GM-D3.2].
GAP #9	5G-MoNArch has developed the concept of Security Trust Zones (STZ), which specify given parts of the network with common security requirements. This concept is applied to the network slicing environments where the requirements of a given slice determine the STZ characteristics [5GM-D3.2].
GAP #10	5G-MoNArch has developed concepts for data duplication and network coding which have been shown to improve the RAN reliability. In addition, the hybrid data duplication/ network coding scheme has been developed, which combines the benefits of both approaches [5GM-D3.2].
GAP #11	5G-MoNArch-has developed the concepts of enhanced fault management considering the virtualisation and slicing aspects. Furthermore, the mechanisms for improving the network controller scalability have been developed. Finally, the concept of 5G islands which applies NF migration between central and edge cloud for improved telco cloud resilience has been elaborated [5GM-D3.2].
GAP #12	5G-MoNArch-has developed concepts for *Inter-slice resource management* (see this deliverable, Section 3.3) enabling efficient (radio) resource sharing strategy for network slices considering slicing principles. Moreover, in [5GM-D4.2] a thorough explanation on how to achieve elastic RAN operation (both user scheduling and RAN control).

6 References

[3GPP R3-161784]	3GPP TSG RAN3#93 R3-161784, "The evaluation for different split options," Aug 2016.
[3GPP R3-161813]	3GPP TSG RAN3#93 R3-161813, "Transport requirement for CU&DU functional splits options," Aug 2016.
[3GPP R3-186014]	3GPP TSG RAN3#101bis R3-186014, "Slice support of IAB nodes," Oct 2018.
[3GPP R3-186544]	3GPP TSG RAN3#102 R3-186544, "Slice Requirement Assurance for IAB E2E Link," Nov 2018.
[3GPP RP180554]	3GPP TSG RAN #79 RP-180554, "Plan for finalising all NR architecture options," March 2018.
[3GPP RP182850]	3GPP TSG RAN #82 RP-182850, "Proposals to address RAN overload," Dec 2018.
[3GPP TS 23.032]	3GPP TS23.032, "Universal Geographical Area Description (GAD)," Release 15.
[3GPP TS 23.501]	3GPP TS23.501, "System Architecture for the 5G System; Stage 2," Release 15.
[3GPP TS 23.502]	3GPP TS23.502, "Procedures for the 5G System; Stage 2," Release 15.
[3GPP TR 23.786]	3GPP TR 23.786, "Study on architecture enhancements for EPS and 5G System to support advanced V2X services," Release 16.
[3GPP TS 23.288]	3GPP TS23.288, "Architecture enhancements for 5G System (5GS) to support network data analytics services"
[3GPP TR 23.791]	3GPP TR 23.791, "Study of enablers for Network Automation for 5G," Release 16
[3GPP TR 36.814]	3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects," v9.2.0, March 2017.
[3GPP TR 38.874]	3GPP TR 38.874, "Study on Integrated Access and Backhaul (Release 15)," v 0.5.0, October 2018.
[3GPP TS 28.530]	3GPP TS 28.530, "Management and orchestration of networks and network slicing; Concepts, use cases and requirements," Release 15.
[3GPP TS 28.533]	3GPP TS 28.533, "Management and orchestration; Architecture framework," Release 15.
[3GPP TS 28.540]	3GPP TS 28.540, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3," Release 15.
[3GPP TS 28.541]	3GPP TS 28.541, "Management and orchestration of networks and network slicing; NR and NG-RAN Network Resource Model (NRM); Stage 2 and stage 3," Release 15.
[3GPP TR 28.801]	3GPP TR 28.801, "Study on management and orchestration of network slicing for next generation network," Release 15.
[3GPP TS 36.300]	3GPP TS 36.300, "E-UTRA and E-UTRAN; Overall description; Stage 2," v 15.3.0, September 2018.
[3GPP TS 36.423]	3GPP TS 36.423, "E-UTRAN; X2 Application Protocol (X2AP)," v 15.3.0, September 2018.
[3GPP TS 38.211]	3GPP TS 38.211, "NR; Physical channels and modulation," v15.3.0, September 2018.
[3GPP TS 38.300]	3GPP TS 38.300, "NR; Overall description; Stage-2," v15.3.0, September 2018.
[3GPP TS 38.401]	3GPP TS 38.401, "NG-RAN; Architecture description," v 15.3.0, September 2018.
[3GPP TS 38.420]	3GPP TS 38.420, "NG-RAN; Xn general aspects and principles," v 15.1.0, September 2018.

[3GPP TS 38.423]	3GPP TS 38.423, "NG-RAN; Xn Application Protocol (XnAP)," v 15.1.0, September 2018.
[3GPP TS 38.470]	3GPP TS 38.470, "NG-RAN; F1 general aspects and principles," v 15.1.0, April 2018.
[3GPP TS 38.473]	3GPP TS 38.473, "NG-RAN; F1 Application Protocol (F1AP)," v 15.3.0, September 2018.
[3GPP TR 38.801]	3GPP TR 38.801, "Study on new radio access technology: Radio access architecture and interfaces," v 14.0.0, April 2017.
[3GPP TR 22.804]	3GPP TR 22.804, "Study on Communication for Automation in Vertical domains (CAV)", v16.1.0, September 2018.
[3GPP TR 22.886]	3GPP TR 22.886, "Study on enhancement of 3GPP support for 5G V2X services", v16.1.1, September 2018.
[3GPP S5-186486]	S5-186486, "Discussion about how SA5/ SA2 / RAN3 could work together to guarantee network slice SLA", October 2018.
[3GPP TS 28.533]	3GPP TS 28.533, "Management and orchestration; Architecture framework", v15.0.0, September 2018.
[3GPP TR 28.900]	3GPP TR 28.900, "Study on integration of Open Network Automation Platform (ONAP) at a Collection, Analytics and Events (DCAE) and 3GPP reference management architecture", v1.0.0, December 2018.
[3GPP TS 32.101]	3GPP TS 32.101, "Telecommunication management; Principles and high-level requirements", v15.0.0, Sep. 2017.
[5GARCH17-WPv2]	5G PPP WG Architecture, Architecture White Paper v2.0, "View on 5G Architecture," Dec. 2017.
[5GM-D2.1]	5G-MoNArch Deliverable D2.1, "Baseline architecture based on 5G-PPP Phase 1 results and gap analysis," Oct. 2017.
[5GM-D2.2]	5G-MoNArch Deliverable D2.2, "Initial overall architecture and concepts for enabling innovations," June 2018.
[5GM-D3.1]	5G-MoNArch, Deliverable D3.1, "Initial resilience and security analysis," June 2018.
[5GM-D3.2]	5G-MoNArch, Deliverable D3.2, "Final resilience and security report," March 2019.
[5GM-D4.1]	5G-MoNArch, Deliverable D4.1, "Architecture and mechanisms for resource elasticity provisioning," June 2018.
[5GM-D4.2]	5G-MoNArch, Deliverable D4.2, "Final design and evaluation of resource elastic functions," March 2019.
[5GM-D5.2]	5G-MoNArch, Deliverable D5.2, "Final report on testbed activities and experimental evaluation," June 2019.
[5GM-D6.1]	5G-MoNArch, Deliverable D6.1, "Documentation of Requirements and KPIs and Definition of Suitable Evaluation Criteria," September 2017.
[5GM-D6.2]	5G-MoNArch, Deliverable D6.2, "Methodology for verification and validation of 5G-MoNArch architectural innovations," July 2018.
[5GN-D2.3]	5G-NORMA, Deliverable D2.3, "Evaluation architecture design and socio-economic analysis - final report," December 2017.
[ATM18]	I.A. Alimi, A.L. Teixeira, P.P. Monteiro, "Toward an Efficient C-RAN Optical Fronthaul for the Future Networks: A Tutorial on Technologies, Requirements, Challenges, and Solutions," IEEE Comm. Survey & Tutorials, vol. 20, no. 1, 2018.
[BIY10]	Biscani, F., Izzo, D., and Yam, C. H., "A Global Optimisation Toolbox for Massively Parallel Engineering Optimisation," 4th International Conference on Astrodynamics Tools and Techniques, 2010.

[BHS97]	T. Back, U. Hammel, and H.-P. Schwefel, "Evolutionary computation: comments on the history and current state," IEEE Trans. Evol. Comput., vol. 1, no. 1, pp. 3–17, Apr. 1997.
[BRH+10]	A. Bou Saleh, S. Redana, J. Hämäläinen, and B. Raaf, "On the Coverage Extension and Capacity Enhancement of Inband Relay Deployments in LTE-Advanced Networks," Journal of Electrical and Computer Engineering, vol. 2010, Article ID 894846, 12 pages, 2010.
[BRR+09]	A. Bou Saleh, S. Redana, B. Raaf, T. Riihonen, J. Hamalainen, R. Wichman, "Performance of amplify-and-forward and decode-and-forward relays in LTE-Advanced," IEEE VTC 2009-Fall.
[BRZ+15]	Ö. Bulakci, Z. Ren, C. Zhou, et al, "Towards Flexible Network Deployment in 5G: Nomadic Node Enhancement to Heterogeneous Networks," ICC 2015, June 2015.
[BP19]	Ö. Bulakci, E. Pateromichelakis, "Slice-aware 5G Dynamic Small Cells," WCNC 2019, April 2019.
[CSS+16]	S. Costanzo, et al., "Service-Oriented Resource Virtualisation for Evolving TDD Networks Towards 5G," Wireless Communications and Networking Conference (WCNC), 2016.
[Doc]	Docker, https://www.docker.com/
[ESRI]	"ESRI Shapefile Technical Description," ESRI White Paper, July 1998; Online available https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
[ETSI ENI]	ETSI Experiential Networked Intelligence, Online available at http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp22_ENI_FINAL.pdf
[ETSI ENI17]	ETSI White Paper, "Improved operator experience through Experiential Networked Intelligence (ENI)," 1st Edition – October 2017.
[ETSI MEC16]	ETSI GS MEC 003, "Mobile Edge Computing (MEC); Framework and Reference Architecture, V1.1.1," March 2016.
[ETSI NFV13]	ETSI GS NFV 002, "Network Function Virtualisation (NFV) Architectural Framework, V1.1.1," October 2013.
[ETSI NFV16]	ETSI GS NFV-IFA 014, "Network Functions Virtualisation (NFV); Management and Orchestration; Network Service Templates Specification, V2.1.1," October 2016.
[ETSI NFV17]	ETSI GR NFV-EVE 012, "Network Functions Virtualisation (NFV) Release 3; Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework, V3.1.1," December 2017.
[FJ08]	J. Fulcher and L. C. Jain, Eds., "Computational Intelligence: A Compendium," vol. 115. Springer Berlin Heidelberg, 2008.
[FNK+16]	X. Foukas, N. Nikalein, M. Kassem, M. Marina, and K. Kontovasilis, "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks," 12th International Conference on emerging Networking EXperiments and Technologies, 2016.
[G12]	Gartner, "Magic Quadrant for BI platforms. Analytics Value Escalator," 2012
[GDC18]	G. Ghatak, A. De Domenico, M. Coupechoux, "Small Cell Deployment Along Roads: Coverage Analysis and Slice-Aware RAT Selection," submitted to IEEE Transaction on Communications, 2018.
[GDD18]	D. M. Gutierrez-Estevez, N. di Pietro, A. De Domenico, M. Gramaglia, U. Elzur and Y. Wang, "5G-MoNArch Use Case for ETSI ENI: Elastic Resource Management and Orchestration," 2018 IEEE Conference on Standards for Communications and Networking (CSCN), Paris, 2018, pp. 1-5.
[GGS+16]	I. Gomez-Miguelez et al., "srsLTE: an open-source platform for LTE evolution and experimentation," Proceedings of the Tenth ACM International

	Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterisation. ACM, 2016.
[GSA WP17]	GSA white paper on "5G Network Slicing for Vertical Industries," September 2017.
[GST]	GSMA: "Generic Network Slice Template," Version 0.02.
[HJS18]	B. Han, L. Ji, H. D. Schotten, "Slice as an Evolutionary Service: Genetic Optimisation for Inter-Slice Resource Management in 5G Networks," to appear in IEEE Access, 2018, https://arxiv.org/pdf/1802.04491.pdf
[iJOIN D3.3]	iJOIN (INFSO-ICT-317941), Deliverable D3.3, "Final definition and evaluation of MAC and RRM approaches for RANaas," October 2012.
[KMQ98]	A. F. Kuri-Morales and C. C. Quezada, "A universal eclectic genetic algorithm for constrained optimisation," Proc. 6th Eur. Congr. Intell. Tech. Soft Comput. EUFIT'98, pp. 2––6, 1998.
[Kub18]	The Kubernetes Authors, "Kubernetes – Production-Grade Container Orchestration," 2014-2018, online available at https://kubernetes.io/
[LPL+17]	Y. Li, E. Pateromichelakis, J. Luo, N. Vucic, W. Xu, "Resource Management Considerations for 5G millimeter-Wave Backhaul / Access Networks," IEEE Communications Magazine Special Issue on Agile Resource Management in 5G, July 2017.
[MGG+18]	D. S. Michalopoulos, B. Gajic, B. Gallego-Nicasio Crespo, A. Gopalasingham, and J. Belschner, "Network Resilience in Virtualized Architectures," Advances in Intelligent Systems and Computing book series, Interactive Mobile Communication Technologies and Learning, Springer, Feb 2018.
[MMR+08]	F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in Proceedings of the 25th International Conference on Machine Learning, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 664–671.
[NGMN15]	Next Generation Mobile Networks (NGMN) Alliance, "5G White Paper," February 2015, Online available at https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf
[NGMN18]	Next Generation Mobile Networks (NGMN) Alliance, "Service-Based Architecture in 5G," January 2018, Online available at https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2018/180119_NGMN_Service_Based_Architecture_in_5G_v1.0.pdf
[NI]	National Instruments, "USRP-2944, USRP Software Defined Radio Reconfigurable Device," Online available at http://www.ni.com/en-us/support/model.usrp-2944.html
[NNM+14]	N. Nikaein, et al., "OpenAirInterface: A flexible platform for 5G research," ACM SIGCOMM Computer Communication Review 44.5, pp. 33-38, 2014.
[OAI]	Open Air Interface, www.openairinterface.org
[ODK16]	T. O. Olwal; K. Djouani; A. M. Kurien, "A Survey of Resource Management towards 5G Radio Access Networks," IEEE Communications Surveys & Tutorials, no.99.
[ONF14]	Open Networking Foundation, "SDN architecture," Issue 1, ONF TR-502, June 2014.
[ONY+11]	K. Okino, et al., "Pico Cell Range Expansion with Interference Mitigation toward LTE-Advanced Heterogeneous Networks," IEEE ICC Workshops, 2011, pp. 1–5.
[OSB16]	J. Oueis, E.C. Strinati, and S. Barbarossa, "Distributed mobile Cloud Computing: A multi-user Clustering Solution," IEEE Int. Conf. on Communications (ICC 2016), 23-27 May 2016.

[OSM]	OSM Information Modes, Online available at https://osm.etsi.org/wikipub/index.php/OSM_Information_Mode
[PJD+15]	S. Parsaeefard, V. Jumba, M. Derakhshani et T. Le-Ngoc, "Joint resource provisioning and admission control in wireless virtualised networks," Wireless Communications and Networking Conference (WCNC), pp. 2020-2025, 2015.
[PMM+19]	E. Pateromichelakis, et al., "End-to-End Data Analytics Framework for 5G Architecture," in IEEE Access, 2019.
[PP17]	E. Pateromichelakis and C. Peng, "Selection and Dimensioning of slice-based RAN Controller for adaptive Radio Resource Management," Wireless Communications and Networking Conference (WCNC), 2017.
[PSW+17]	E. Pateromichelakis, K. Samdanis, Q. Wei, P. Spapis, "Slice-tailored Joint Path Selection & Scheduling in mm-Wave Small Cell Dense Networks," IEEE Globecom, 2017.
[Red18]	Red Hat, Inc., "OpenShift – The Kubernetes platform for big ideas," 2018, online available at https://www.openshift.com/
[SKE+12]	Z. Shen, et.al, "Dynamic Uplink-Downlink Configuration and Interference Management in TD-LTE," IEEE Communication Magazine, Vol.50, No.11, Nov. 2012.
[SRSLTE]	Open-source srsLTE, https://github.com/srsLTE
[SSC+17]	V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, A. Banchs, "Mobile Traffic Forecasting for Maximising 5G Network Slicing Resource Utilisation," IEEE INFOCOM, 2017.
[SSS+16]	V. Sciancalepore, K. Samdanis, R. Shrivastava, A. Ksentini, X. Costa-Perez, "A service-tailored TDD cell-less architecture," IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Valencia, 2016, pp. 1-6.
[STM+19]	D. Schinianakis, R. Trapero, D. S. Michalopoulos, and B. Gallego-Nicasio Crespo, "Security Considerations in 5G Networks: A Slice-Aware Trust Zone Approach," IEEE Wireless Communications and Networking Conference (WCNC) 2019.
[Wil12]	B. Wilder, "Cloud Architecture Patterns," O'Reilly Publications, 2012.
[XRAN]	xRAN: Next Generation RAN Architecture, http://www.xran.org/
[YT16]	F. Zarrar Yousaf, T. Taleb, "Fine-grained resource-aware virtual network function management for 5G carrier cloud," IEEE Network 30.2, pp. 110-115, 2016.
[WF18]	Y. Wang, R. Forbes, C. Cavigioli, H. Wang, A. Gamelas, A. Wade, J. Strassner, S. Cai, S. Liu, "Network Management and Orchestration using Artificial Intelligence: Overview of ETSI ENI," to appear in IEEE Communications Standards Magazine.
[ZL07]	Q. Zhang and H. Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition," IEEE Trans. Evol. Comput., vol. 11, no. 6, pp. 712–731, 2007.
[5GPPPW2C]	5GPPP, Software Network WG, "From Web Scale to Telco,", Whitepaper, Online available at https://5g-ppp.eu/wpcontent/uploads/2018/07/5GPPP-Software-Network-WG-White-Paper-23052018-V5.pdf
[ORAN]	O-RAN Alliance, https://www.o-ran.org/resources/
[OVS]	https://www.openvswitch.org/

Appendix A Summary of 5G System Gaps Identified by 5G-MoNArch

In the first deliverable from WP2 of 5G-MoNArch (D2.1) [5GM-D2.1], a baseline architecture has been delineated. This baseline architecture is 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 working groups (WGs). Following that delineation, a 5G system gap analysis was performed, identifying the additional modules/mechanisms that are required in addition to the baseline architecture to meet the 5G objectives. Furthermore, an overview of 5G-MoNArch innovations along with their mapping onto the identified gaps has been provided. A summary of the gap analysis is outlined as follows.

(1) **Inter-dependencies between NFs co-located in the same node**: Traditional protocol stacks have been designed under the assumption that certain NFs reside in the same node, i.e., fixed location and NF placement; 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. The logical and temporal dependencies between NFs should be relaxed and (as much as possible) removed to provide a higher flexibility in their placement.

(2) **Orchestration-driven elasticity not supported**: It is necessary for the architecture to flexibly shift NFs to nodes that better fit the specific requirements of each covered service; when doing so, it is necessary to take elasticity considerations into account.

(3) **Fixed functional operation of small cells**: In the current networks, the functional operation of small cells does not change relative to service requirements or the location of the small cell, which can be, e.g., unplanned and dynamic. 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. However, slice-awareness and 5G tight KPIs can necessitate on-demand flexible small cell operation.

(4) **Need for support for computational offloading**: Current architectures do not fully support delegating costly NFs beyond the network edge towards RAN (e.g., for cases like group mobility in D2D context). Addressing this gap can result in saving on energy consumption, signalling overhead or to offload resource demanding tasks when needed.

(5) **Need for support for telco-grade performance (e.g., low latency, high performance, and scalability)**: Most of management and orchestration technologies are inherited from IT world. Adopting such technologies in the telco domain without key performance degradation is a great challenge as the added functionalities in the control and M&O layer, as well as the more modular NFs, should still offer the same telco grade performance, without degradation.

(6) **E2E cross-slice optimisation not fully supported**: Architecture should allow for the simultaneous operation of multiple network slices with tailored core / access functions and functional placements to meet their target KPIs.

(7) **Lack of experiment-based E2E resource management for VNFs**: Current 5G systems are missing E2E resource management of VNFs that takes advantage of E2E software implementations on commodity hardware in a dynamic manner. Indeed, most of the proposals so far rely on simplifying assumptions that yield simple but possibly unrealistic models. To design algorithms that perform well in reality it is necessary to rely on more elaborate, experiment-based, models.

(8) **Lack of a refined 5G security architecture design**: There are various critical gaps in the literature and architectural deployments related to orchestration & management, accountability, compliance & liability, as well as performance and resilience.

(9) **Lack of a self-adaptive and slice-aware model for security**: E2E network slicing demands a revaluation and research on various aspects of traditional security (e.g., privacy, integrity, zoning, monitoring, and risk mitigation).

(10) **Need for enhanced and inherent support for RAN reliability**: RAN reliability should be a built-in solution/element of the architecture, through the application of mechanisms such as multi-connectivity and network coding.

(11) **Indirect and rudimentary support of telco cloud resilience mainly through management and control mechanisms**: The architecture should address resilience in a structured way considering different aspects (e.g., individual network elements (NEs)/NFs, telco cloud components, fault management, and failsafe mechanisms).

(12) **Need for (radio) resource sharing strategy for network slices**: While basic mechanisms for multi-slice resource management have been studied in 5G-PPP Phase 1 projects, elastic mechanisms need to be devised that improve the utilisation efficiency of the computational and radio resources by taking advantage of statistical multiplexing gains across different network slices.

Appendix B Relationship with standards and standardisation roadmap

In this section, we highlight the main relations of 5G-MoNArch with SDOs and pre-standardisation groups, considering RAN, core network, and management and orchestration planes. We recall that a summary of the 5G-MoNArch impact on the relevant SDOs has been provided in Section 2.1.

B.1 Radio access network

A baseline architecture including the RAN protocol stack and the essential functional elements has been provided in D2.1 [5GM-D2.1]. Therein, it is shown that a fundamental support for network slicing is provided in the RAN. Further standardisation progress has been captured in D2.2 [5GM-D2.2]. From the specification perspective, 3GPP Release 15 for next generation-RAN (NG-RAN) is frozen in June 2018 and a so-called late drop of Release 15, which includes further architecture options, was initially planned to be frozen by the end of 2018 [3GPP-RP180554] with an adjusted plan till the end of March 2019 [3GPP-RP182850]. This specification comprises slicing awareness in RAN via NSSAI including one or more S-NSSAIs, which allow to uniquely identify a network slice [3GPP TS 38.300]. While the fundamental slicing support is achieved by Release 15, e.g., granularity of slice awareness and network slice selection, various enhancements and optimisation can be considered for future releases. Such enhancements may imply, for example, specification-relevant signalling changes and implementation-dependent algorithms, e.g., related to resource management (RM) between slices. **Considering the latest specification progress, the 5G-MoNArch approach has focussed on both types of enhancements, where novel RAN components and interfaces are highlighted in Section 2.2.1.**

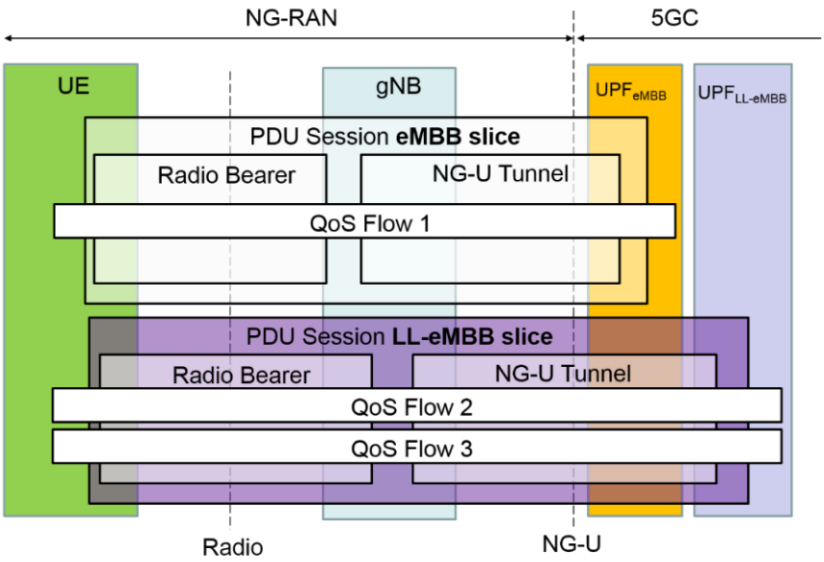


Figure B-1: Slice support in the 5GS

In principle, network slicing offers additional degree of flexibility, where NFs can be tailored according to the requirements of slice tenants. To this end, it can be expected that different tenants may have diverse network requirements. For instance, some slice tenants may only require a performance differentiation, e.g., in terms of Quality of Service (QoS) requirements, such as latency and data rate, which can be extended by further Service Level Agreement (SLA) requirements, such as number of connections for a given time and location.

Therefore, slice tenant requirements can be supported by different network slicing implementation variants, as elaborated in detail in D2.2 [5GM-D2.2]. In some of these variants, the whole RAN protocol stack can be shared by network slices where SLA differentiation can be performed with QoS enforcement. In particular, in line with the latest 5G Release 15 specification and as shown in Figure B-1, for a network slice instance one or more Protocol Data Unit (PDU) sessions can be established, where a PDU session belongs to one and only one specific network slice instance [3GPP TS 23.501].

Further, RAN maps packets belonging to different PDU sessions to different data radio bearers (DRBs), where within a PDU session there can be one or more QoS flows [3GPP TS 38.300]. On this basis, the RAN treatment of different network slices can be in terms of radio RM (RRM) schemes performed based on the QoS profiles of QoS flows mapped onto the respective DRBs, where QoS profiles can include performance characteristics, e.g., packet delay budget (PDB) and packet error rate (PER), and allocation and retention priority (ARP).

From RAN perspective, the slicing requirements can be mapped onto:

- **Spectrum Requirements**: Slices may require different chunks of radio resources (e.g., below 6 GHz and mm-Wave radio) to meet the slice performance needs. In order to meet the spectrum requirements from multiple slices, slice-tailored RM is required at RAN.
- **Functional Requirements**: Each slice may require different control plane/user plane functions, protocol parametrisations, and/or functional placements in order to provide optimised and agile performance at RAN. As stated in literature [5GARCH17-WPv2], network slices will allow for flexible functional placements and tailored NFs to meet the per-slice SLAs.
- **Isolation Requirements**: E2E slices shall be logically isolated. However, at RAN, the potential utilisation of common infrastructure might provide a bottleneck towards complete isolation. In addition, given the slice criticality, each slice might have different requirement for isolation at RAN domain. Hence, the different isolation requirement per slice at the RAN domain might require sophisticated RM to meet per-slice performance needs.

To meet the above requirements, inter-slice RM, aka multi-slice RM, is thus very important for improving the system efficiency, especially on shared infrastructure resources, which is a means for cross-slice optimisation. According to 3GPP, the inter-slice RRM can be supported by means of providing RRM split policies by the slice management system, as abstracted resource partitioning requirements for RAN [3GPP TS 38.300]. RAN is responsible of enforcing these policies dynamically based on the actual channel conditions, bearer load and users' demand. The inter-slice RM thus factors in the slice SLAs, e.g., to adapt the instantaneous radio resource allocation. **On this basis, 5G-MoNArch slice-aware RRM schemes presented in Section 2.2.1 and detailed in Section 3.3 are built upon these main slicing principles.**

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. This is referred to as the extended notion of a resource. Accordingly, **the slice-adaptive resource control as highlighted in Section 2.2.1 considers the changing radio topology including different access node types, e.g., micro-cells, pico-cells, relays, and Vehicular Nomadic Nodes (VNNs)**. One particular 5G deployment where the extended notion of resource can be applicable, is the central unit (CU)-distributed unit (DU) split of gNB (i.e., access node in 5G) for allowing flexible centralisation of RAN functions using adaptive topologies and functional placements [3GPP TS 38.300][3GPP TS 38.401]. In this context, slicing information can be exchanged between CU and DU, e.g., for bearer management in a slice-aware manner. Another possible deployment scenario is the Integrated Access Backhaul (IAB), which makes use of the CU-DU split architecture and has been recently studied in 3GPP [3GPP TR 38.874] and illustrated in Figure B-2. In IAB scenarios, IAB nodes can function as relays that can be connected to an IAB donor through a wireless backhaul link. The wireless backhaul link can employ below 6 GHz and above 6 GHz spectrum bands and can also support more than two hops (i.e., multi-hop operation). On this basis, an IAB node can include UE and DU functionalities, while the IAB donor can implement CU and DU functionalities. Therefore, the aforementioned small cell deployments (e.g., VNN) can employ the CU-DU architecture, where the small cells operate as DUs. Furthermore, as the E2E IAB link can comprise multiple-hops and UEs associated with different network slices, e.g., eMBB and low-latency (LL) eMBB, can be served, the slice requirements shall be ensured on the E2E IAB link. **In a further dimension, especially slow-timescale RAN control functions can be implemented as applications running in the Controller Layer shown in Section 2.1 and further described in Section 2.2.1.** Such applications

consider already standardised protocols and can provide enhancements within a cell or for neighbouring cells, see, e.g., [XRAN][9].

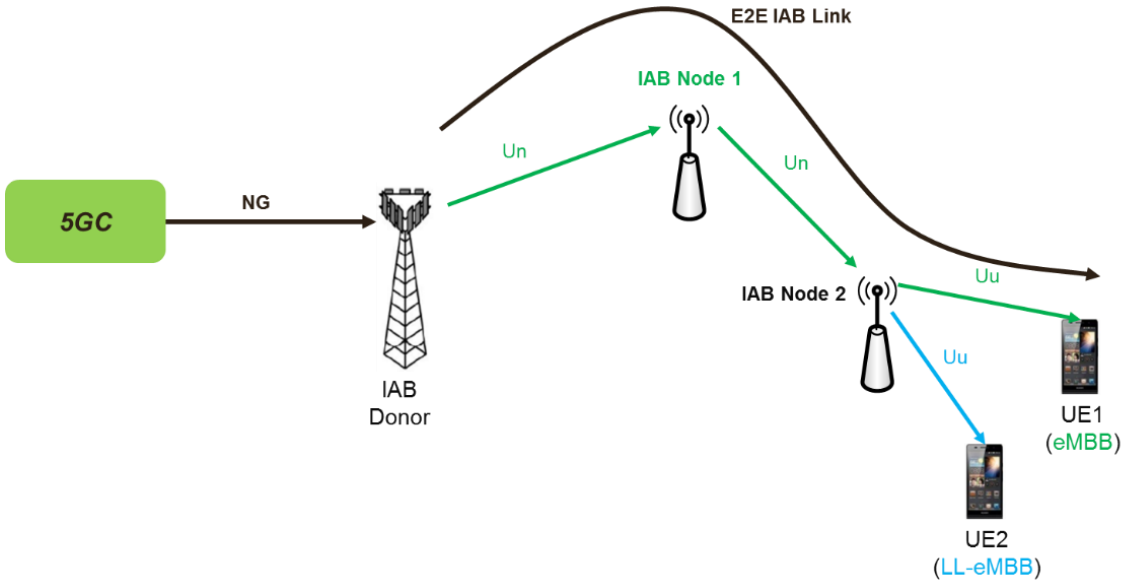


Figure B-2: The slice requirement shall be ensured on the E2E IAB link [3GPP R3-186014] [3GPP R3-186544].

The standardisation impact of the activities conducted within the WP3 framework of 5G-MoNArch is concentrated on data duplication approaches in RAN. Specifically, WP3 has proposed the concept of data duplication for both the data plane and the control plane, while such duplication can take place either in an inter-frequency or an intra-frequency setup. 3GPP has studied the case where data duplication is applied in a CU –DU architectural setup. **It is noted that this is the architecture adopted in 5G-MoNArch, as well.** As regards inter-frequency data duplication, there is a consensus in 3GPP standardisation activities that such duplication should be handled by the PDCP layer of the protocol stack, located at the CU. This implies that duplicate packets, i.e., PDCP PDUs initiate from the CU and sent to the corresponding DU entities. The most relevant standardisation input in this respect is [R2-1817582], where the details of the coordination of duplicate packets at the PDCP layer are specified. The main conclusion is that acknowledgment packets should be provided to the PDCP layer at the CU, confirming the correct reception of at least one packet from the UE receiver.

3GPP has also considered the case of intra-frequency data duplication, which mainly applies to mobility scenarios and concerns duplicating control plane messages. In such case, data duplication is used for ensuring minimal interruption during handovers, thereby providing mobility robustness. In this respect, data duplication is used as a redundancy method that minimises the chance of "too early" or "too late" handovers that may lead to potential radio link failures. The relevant standardisation input in this regard is [R2-1708588]. The principal idea in such document is the initiation of adding the target cell well before the usual time that this happens in conventional handovers and maintain the connection to two cells as long as both cells are seen with sufficient signal quality from the UE. Then, the roles of master and secondary cells are swapped, until the former master cell is removed, and the system reverts to the single connectivity mode with the new cell.

The work in the WP3 framework is in line with the aforementioned 3GPP activities, since both cases of user plane and control plane data duplication are developed in WP3. In addition to such considerations, WP3 has studied the threshold which applied in data duplication deployments, referred to as "link imbalance threshold". The link imbalance threshold is studied in WP3 as a means of controlling which access points are included in the duplication process. As such, a large link imbalance threshold implies a relaxed policy in adding access points to the duplication process, while

[9] Meanwhile, xRAN Forum has merged with C-RAN Alliance to form the ORAN Alliance.

low values of the link imbalance threshold limit the set of participating access points. By means of the link imbalance threshold, a prudent use of resources is achieved, such that resource overprovisioning scenarios are avoided.

B.2 Core network

The key technological components of the CN of 5G systems (5GS), i.e., 5G core (5GC) are architecture modularisation, CP and UP separation and Service-Based Interface (SBI). These are reflected in the SBA (crystallised in 3GPP Release 15 specifications [3GPP TS23.501]) where the CP NFs are interconnected via the SBI. Each NF, if authorised, can access the services provided by other NFs via the exposed SBI. As a set of examples, the Network Exposure Function (NEF) is a NF included in the 5GC which allows each NF (either internal or external) to expose its capability to other NFs; The Network Repository Function (NRF) is an NF included in the 5GC allowing each NF to discover which instance of another NF can be accessed to receive a required service. The AN CP is connected to the Access and Mobility Function (AMF) of 5GC in case of 3GPP access and is connected to the Interworking Function (N3IWF) in case of non-3GPP Access.

Compared to the traditional functional based network architecture design, SBA is expected to have the advantage of short roll-out time for new network features, extensibility, modularity, reusability and openness [NGMN18]. **This reference architecture, as envisioned 5GC architecture for 5G-MoNArch, allows the definition and instantiation of flexible E2E networks, which can be customised by network operators' or vertical industries' requirements, in terms of performance, capabilities, isolation etc.** In other words, 5GC reference architecture allows the support of network slices, i.e., independent logical networks, either sharing partly/entirely the infrastructure they are instantiated on, or isolated and deployed over separate infrastructures. 5G devices will be able to access 5GC and requiring services from a number of supported network slices. The Network Slice Selection Function (NSSF) is an emerging NF dedicated to select the proper NSI for the 5G devices. The reference architecture provides multiple options to customise network slices capabilities. For example, the Session Management Function (SMF) may allow the support of different UP protocol models, such as IPv4/IPV6, Ethernet, or unstructured data format. The Policy Control Function (PCF) may allow customising the policy framework on network slice basis. Finally, the Unified Data Management function (UDM) may enable different authorisation, authentication, and subscription management mechanisms upon network slice tenant needs. It should also be noted that, thanks to SBI, the reference architecture also provides third parties with the possibility to influence the network behaviour, extend and customise network slices capabilities via the inclusion in the system of proprietary non-standard Application Functions (AFs). Using the SBI, the AF is possible to access services provided by other NFs, as well as to expose theirs services to other NFs, e.g., via NEF.

Despite the foundations for 5GC have been successfully established, the general framework still appears not entirely mature and seems to be still susceptible to significant technical and conceptual enhancements. Some key examples of issues still offering a large number of design options and room for further improvements are:

- The instantiation and selection of NFs for different slices in the infrastructure;
- The specific functional customisation of NFs to address requirements of specific use cases;
- The functional interaction among different network slices.

The 5G-MoNArch core network architecture adopted the 3GPP SBA based architecture and NFs currently defined in Release 15 as a baseline. Beyond current state of the art, several novel enhancements of the NFs e.g., AMF, PCF, NSSF, are proposed to address a set of key gaps listed in D2.2 [5GM-D2.2].

A separate distinguishing feature of 5GC, compared to previous generation networks, is network analytics capability embedded in the general framework, via the definition of the Network Data Analytics function (NWDAF). In short, as per 3GPP Release 15, NWDAF provides 5GC with the ability to collect and analyse *per slice aggregated data*, and to aid network optimisation via interaction with PCF. Albeit included in 3GPP release 15 specification, NWDAF description and capabilities are extremely rudimental. The exploitation of the full potentials of network analytics and Big Data technologies requires the clarification and investigation of a number of questions, including:

- What data should be collected by NWDAF and what feedback is expected from NWDAF;
- From which entities and how should the NWDAF collect the data;
- How NWDAF shall collect data on per PLMN and/or per slice and/or per user basis and/or per session basis;
- How NWDAF shall expose its services, and which NFs may benefit from them;
- How can NWDAF get the data from the NFs/NEs which are not connected to the SBI;
- The granularity of the optimisation to be enabled by NWDAF services, options being:
    - Per session basis;
    - Per user basis;
    - Per slice basis;
    - On Inter-slice basis.

3GPP SA2 WG started a study item FS_eNA and a subsequent work item (eNA) in Rel. 16, to study enablers for Network Automation for 5G and to further clarify the usage of data analytics capability in the network layer. FS_eNA /eNA envision improving NWDAF scope via the study of different use cases, Key Issues and related solutions on data analytics. It specifies a general framework (as shown in Figure B-3) to collect data from/provide data analytics to different Network Functions (NFs), application function (i.e., Application layer) and management layer (i.e., OAM).

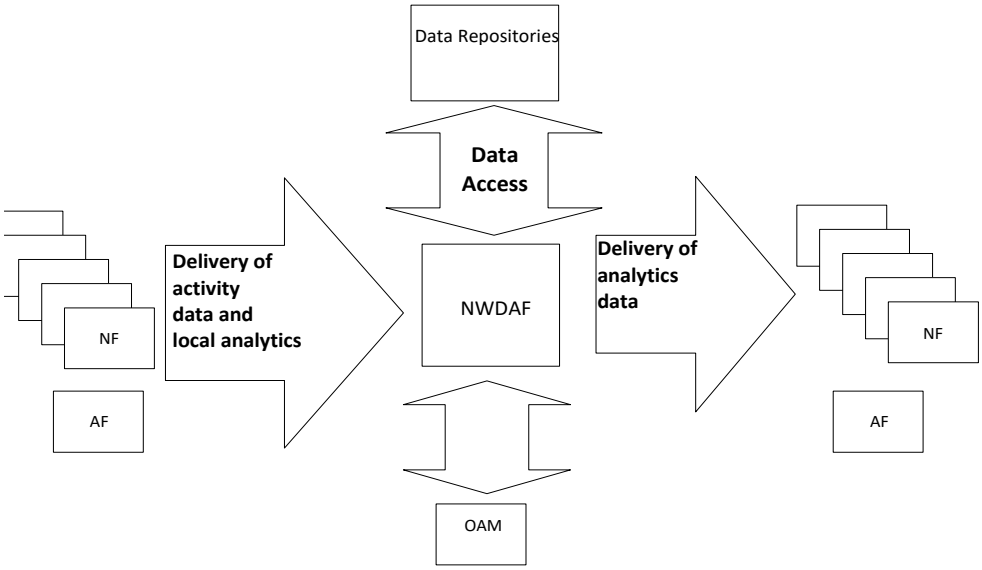


Figure B-3: General framework for 5G network automation (TR23.786)

Some key studied use cases included are: NWDA-assisted QoS provision, traffic handling, customised mobility management, policy determination, QoS adjustment, 5G edge computing, load (re-)balancing of NFs, determination of areas with oscillation of network conditions, slice SLA assurance or predictable network performance. NWDAF is also used for information retrieving from the application function, performance improvement and supervision of mIoT terminals, prevention of various security attacks and UE driven analytics sharing.

Some key issues were also identified with high priority such as the data collection from/analytic information exposure to 5GS NF/AF/OAM. Other key issues were derived from the specific use cases such as NWDAF assisted QoS profile provisioning, traffic routing or mobility management. The current conclusion for these key issues is summarised as follows.

NWDAF reuses similar service exposure mechanisms as other 5G NFs for data collection and data analytics exposure from / to other NFs. This includes both the subscription module and request-response model. Certain data structure needs to be followed for the data (data analytics) which should be included

in the content of the subscription, request or response. There are also proposals to define a new service for unified data collection and analytics exposure from/to NFs/AFs.

For the interaction with OAM, the data collection from OAM may reuse the existing SA5 services. And how NWDAF provides the data analytics to OAM is still under discussion together with SA5. **5G-MoNArch has directly contributed to FS_eNA/ eNA studies regarding the data collection and analytics exposure key issues. More details are described in Section 3.2.1.**

5G-MoNArch has investigated the enhancement of the CP/UP procedure (e.g., slice alignment procedure between RAN and CN) and the architecture (e.g., interfaces and functionality extension of NWDAF, new functionality to support inter-slice coordination) to address the above issues and questions. The related innovation elements are described in Section 2.2.2 as InE#2 Inter-slice coordination and InE#3 Inter-slice context sharing and optimisation. The detailed solution and analysis of the innovation elements in the core network are included in Sections 3.2.1, 3.2.2, and 3.2.3, respectively.

One other aspect is the E2E slice view which needs the alignment of a network slice between the Network layer and M&O layer. The M&O layer looks at the network slice deployment in a longer time scale for one tenant/one group of services. The Network layer takes care of the individual user, connects them to the already deployed NSI by the M&O layer and controls the shorter time scale slice KPIs. Both layers need to work together to guarantee the Service Level Agreement (SLA).

In the real network deployment, not all network slices are supported over the complete PLMN network, especially when considering the E2E perspective. 3GPP defines the network slice availability as following:

"A Network Slice may be available in the whole PLMN or in one or more Tracking Areas of the PLMN. The availability of a Network Slice refers to the support of the NSSAI in the involved NFs. In addition, policies in the NSSF may further restrict from using certain Network Slices in a particular TA, e.g. depending on the HPLMN of the UE."

More specifically, the E2E slice availability is decided by the RAN slice capabilities, CN slice capabilities, the NSI management, network configuration, and also network policies. This brings up the following issues:

- Whether the current network layer slice selection mechanism in 3GPP is sufficient to address different deployment scenarios.
- How the Network layer interacts with management layer on individual network slices.
- How to map the customer services to the actual deployed network slice in the operator network.

Since end to end aspects of slice covers from service layer, management and orchestration and network layer, the enhancements are discussed in different sections e.g., context aware slice selection is covered in Section 3.2.3, CP/M&O layer per slice interaction is covered in Section 3.2.1, and service to slice mapping is captured in Section 3.2.2.

B.3 Management and orchestration

Main Principles

5G-MoNArch M&O system capitalises on 3GPP guidelines using virtualisation and slicing to fill the identified gaps D2.2 [5GM-D2.2]. VNF are aggregated into network slices and foresees automation and orchestration functions also considering Self Organising Network (SON) algorithms. E2E management and orchestration is performed at different levels in a coordinated manner. These levels are: service, network configuration, virtualisation, and transport. The **5G-MoNArch M&O layer** (see Figure B-4 **takes care of this job, interworking with Controller layer and Network layer, to deploy the required NFs and to configure the appropriate interconnections according to the service and network requirements.**

The 5G-MoNArch M&O layer complies with 3GPP specifications that foresee a management system that coordinates network and slice management and orchestration. Current 5G-MoNArch architecture explicitly considers the interaction with the 3GPP Management Entities dedicated to Network management and configuration. For slice management the NSMF will implement 3GPP standards for slice management and orchestration.

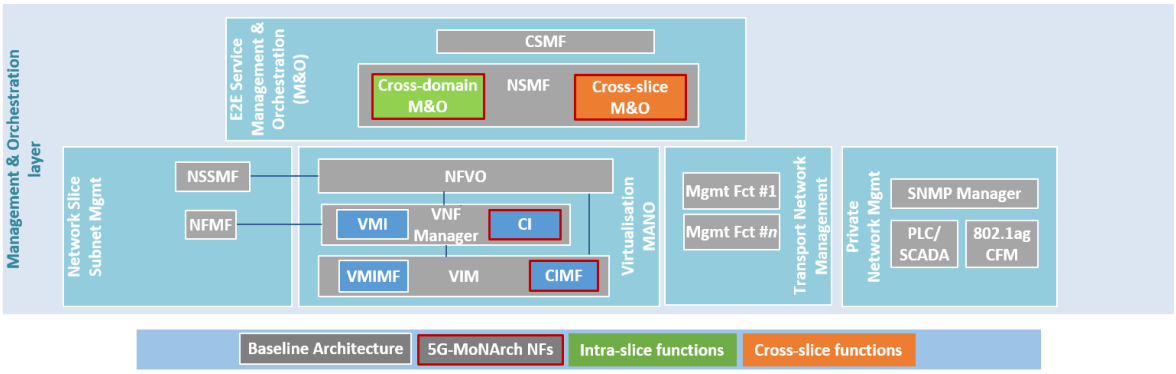


Figure B-4: 5G-MoNArch Management & Orchestration layer

5G-MoNArch M&O layer is coherent with the 5G management architecture framework [3GPP TS 28.533] adopted by 3GPP for its management layer. This framework is based on the SBA approach which foresees that each management function is decomposed in management services. Each management service is a consumer of other services (e.g. produced by NFs or by other management services) and it is a producer of management services. Any authorised service consumer can use any management service, this concept gives a great flexibility in the composition of the orchestration process.

In the E2E Service Management & Orchestration sublayer, service requirements are translated into network requirements by the CSMF. The obtained network requirements are forwarded to the NSMF which is composed by sub-entities or micro-services, that address the management and orchestration of each slice (Cross-domain M&O) and the management according to the possible interaction among slices in terms of resources and features sharing (Cross-slice M&O).

The Management Function defined into 5G-MoNArch E2E Service Management & Orchestration sub-layer are needed to support E2E cross-slice optimisation allowing simultaneous operation of multiple network slices. The Management layer, with the interaction of NSMF and NFVO, fits the specific requirements of each covered service supporting Orchestration-driven elasticity. Analysing performance and assurance data the management layer orchestrates action at slice level and cross-slice level to support telco cloud resilience

The two service-level sub-entities then interact with Domain-Specific Application Management (e.g., 3GPP Network Management and ETSI NFV MANO). To fill several of the identified gaps D2.2 [5GM-D2.2], the M&O layer has to:

(1) Identify the requested VNFs/PNFs that support the service requirements.
(2) Identify the forwarding graph that links the VNFs/PNFs.
(3) Identify the configuration and policies (e.g. for elasticity) to fulfil the required service and SLAs.
(4) Identify the most appropriate Network Slice Template (NST) (for network management) and Network Service Descriptor (NSD) (for VNF deployment).
(5) Identify KPIs for Performance Management (PM) to meet the requested SLAs.
(6) Orchestrate the deployment and activation of the NSI.
(7) Activate PM and Fault Management (FM).
(8) Run PM and FM comparing the data with the defined KPI for the slice.
(9) Activate orchestration to fulfil service changes requests or to meet the SLAs using FM and PM.
(10) Expose PM and FM data to the customer (if requested).
(11) Orchestration performs the LCM of VNFs and performs the requested action on the transport part.

The deployment and management of a network slice is performed to fulfil the request of a customer asking for a Communication Service, 5G-MoNArch M&O layer is coherent with some aspects studied

by 3GPP in [3GPP TR 28.801] and specified by 3GPP in [3GPP TS 28.530]. In the following are reposted the 3GPP principles that 5G-MoNArch is following from [3GPP TS 28.530].

The 5G-MoNArch M&O layer takes care of the LCM of an NSI working with all the other Domain Specific orchestrators. When providing a communication service, 5G-MoNArch M&O layer has to use non-3GPP parts (e.g. Transport Network) in addition to the 3GPP managed network components. Therefore, in order to ensure the performance of a communication service according to the business requirements of the customer.

5G-MoNArch M&O layer has coordinated with the management entities of the non-3GPP parts (e.g., ETSI MANO system) when preparing an NSI for this service. This coordination may include obtaining capabilities of the non-3GPP parts and providing the slice specific requirements and other resource requirements of the non-3GPP parts.

5G-MoNArch M&O layer has identified the requirements for RAN, CN and non-3GPP parts of a slice by breaking down the customer requirements into different parts and sending them to the corresponding management systems, respectively. To support this capability, and according to 3GPP actors and roles, 5G-MoNArch M&O layer introduces the Communication Service Management Function (CSMF).

The coordination may also include related management data exchange between those management systems and 3GPP management system. As defined by 3GPP in [3GPP TR 28.801], 5G-MoNArch M&O layer manages NSIs using three new functions:

- Communication Service Management Function (CSMF): this function takes care of the management of the communication service and translates the requirements related to the communication service to network slice related requirements.
- 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 NSSMF and CSMF.
- Network Slice Subnet Management Function (NSSMF): responsible for management and orchestration of NSSI. Communicates with the NSMF. NSMF, according to 5G-MoNArch, could be useful to take care of specific management domains or to aggregate NF from a specific vendor.

5G-MoNArch approach on slice offering is coherent with the 3GPP definition of Network Slice as a Service (NSaaS) [3GPP TS 28.530]. NSaaS can be offered by a Communication Service Provider (CSP) to its Communication Service Customer (CSC) in the form of a communication service. As defined by 3GPP, 5G-MoNArch M&O comprises the option of exposing some management interface. For 5G-MoNArch this feature is important to let the customer to operate the slice applying custom LCM and optimisations.

5G-MoNArch approach on slice offering is also coherent with the 3GPP definition of "Network Slices as NOP internals" model. Network slices are not part of the CSP service offering and hence are not visible to CSCs. However, the NOP, to provide support to communication services, may decide to deploy network slices, e.g. for internal network optimisation purposes. 5G-MoNArch Deliverable D2.1 [5GM-D2.1] identified some gaps that require an improved management and orchestration (M&O) system in the 5G-MoNArch architecture. The compliancy and enhancement of what defined in 3GPP, for the management of 5G networks, is the chosen path to fill those gaps D2.2 [5GM-D2.2] related to management and orchestration.

The offering of slices implies the exposure of management services to the customer to let him partially control the slice. SA5 defines in [3GPP TS 28.530] a possible deployment scenario with a management function, Exposure Governance Management Function (EGMF), which intermediates the exposure of the management APIs to external consumers. EGMF is a service consumer of others Management Service producers with the aim to expose those management services to other consumers such as the management system of another Operator or to some Vertical industry. The level of service exposure can be different according to the different consumers and can be policy driven. **In 5G-MoNArch architecture, when it comes to the exposure of the management services, the 5G-MoNArch M&O layer is involved and the NSMF intermediates the exposure of the slice management APIs.**

To support management and orchestration, 3GPP SA5 also introduces in [3GPP TS 28.533] the Management Data Analytics Functions (MDAF) that exposes one or multiple Management Data

Analytics Service(s) (MDAS). Unlike an atomic function, an MDAS can exist at NF, network slice subnet, and network slice level. The MDAS at slice level consumes the service produced by the MDAS at subnet level which consumes the MDAS at NF level. Deployment options for MDAS comprise centralised deployment (e.g., at a PLMN level) and domain-level deployments (e.g., RAN, CN, and NSSI). Domain MDAS provides domain-specific analytics, e.g., resource usage prediction in a CN or failure prediction in a subnet, etc. A centralised MDAS can provide end-to-end or cross-domain analytics service, e.g., resource usage or failure prediction in a network slice, optimal CN node placement for ensuring lowest latency in the connected RAN.

5G-MoNArch ETSI MANO evolution

Network slicing, multi-tenancy and flexibility of supporting different services are the key requirements that novel 5G systems have to support. **5G-MoNArch architecture fulfils these requirements and provide mechanisms and framework that manage NFs that are shared between network slices or belonging to different management domains.**

Mapping 3GPP network slicing concepts to ETSI NFV framework

5G-MoNArch architecture embeds ETSI NFV MANO orchestration framework besides 3GPP compliant modules. This includes:

- VIM: Responsible for control and management of NFV Infrastructure (NFVI) compute, storage and network resources.
- VNFM: Responsible for LCM of VNF instances.
- NFVO: Responsible for the orchestration of NFVI resources and LCM of NSs.

This section briefly describes the ETSI MANO concepts that are used and enhanced in 5G-MoNArch and how it can coexist along 3GPP compliant M&O modules [3GPP TR 28.801] to support E2E network slicing.

ETSI NFV Architectural Framework [ETSI NFV13] introduces a concept of NS (network service) as a set of NFs connected according to one or more forwarding graphs [ETSI NFV16]; it additionally adds the concept of nested NSs. The NFVO would use the NSD as a template with information used to manage the lifecycle of an NS. VNF Descriptor (VNFD), on the other hand, is a template describing the requirements of VNF. It is used by the VNFM for VNF instantiation and by NFVO to orchestrate the virtualised resources.

[3GPP TR 28.801] describes a model where a network slice contains one or more network slice subnets. Each network slice subnet can be composed of one or more NFs. Therefore, a NS can be considered as a network slice subnet in case it contains at least one VNF. Similarly, the network slice blueprint described by [ETSI NFV13] could be associated with nested NFV NSDs. Additionally, [3GPP TR 28.801] describes three management functions dealing with network slicing management as described in Section 2.2.3, i.e. CSMF, NSMF and NSSMF. In reference to the 5G-MoNArch overall architecture, Cross-domain M&O could be mapped to NSMF while cross-slice M&O could be either NSMF or NSSMF as described in Section 3.4.2

Figure B-5 shows how these functions could match the NFV MANO model using the Os-Ma-Nfvo reference point as a way of interaction between 3GPP slicing related management functions and NFV-MANO. The role of the NSMF and/or NSSMF would be to determine the type of NS, VNF and PNF that can fulfil the requirements for an NSI or NSSI.

As described in [5GM-D2.1], there are several gaps that need to be addressed in order to properly interface with NFV-MANO while slice-related management functions are still under definition in 3GPP SA5 regarding the interaction with NFV MANO.

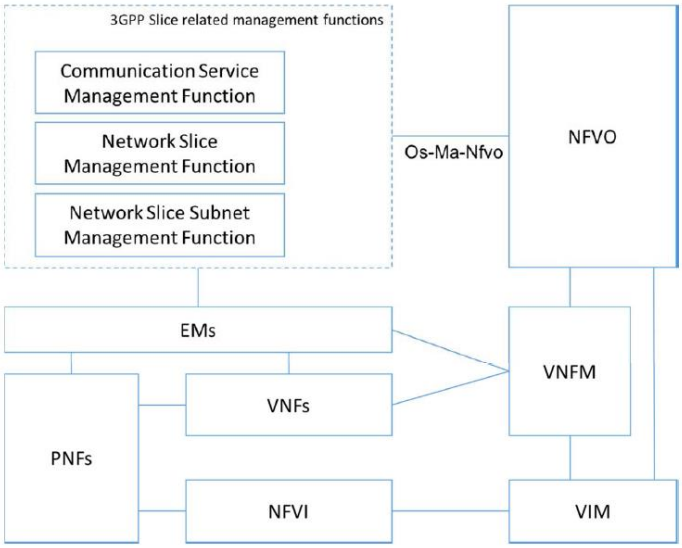


Figure B-5: Network slice management in an NFV framework [ETSI NFV17]

Role of ETSI NFV MANO in NSI management

According to [3GPP TR 28.801] the lifecycle of a network slice is comprised of the four following phases. This has been further discussed in Section 4.3:

- Preparation;
- Instantiation, Configuration and Activation;
- Run-time;
- Decommissioning.

From an NFV perspective the role of NFVO in the preparation phase is to ensure the resource requirements for an NST. NFVO contains the NSDs that have been previously on-boarded and that can be used to create new NSTs that are created and verified in the preparation phase. The NSDs can be updated and created from the beginning if required, if a new NST is necessary.

During the instantiation phase the NFV MANO functions are only involved in the network slice configuration if parameters related to virtualisation are required for any VNF instance and can be called in the network slice activation step. During the activation the NSMF or the NSSMF functions can activate VNFs by means of Update NS sent towards NFVO. This operation could include adding, removing or modifying VNF instances in the NS instance.

During the run-time phase NFV MANO is responsible for PM, FM that could affect a VNF's functioning, and lifecycle of virtualised resources. This could include for example scaling of NS.

Use cases and impact on NFV architecture

[5GM-D2.1] described some of the M&O use cases from 3GPP perspective. [ETSI NFV17] additionally considers the NFV MANO architectural framework [ETSI NFV13] and evaluates the impact of network slicing, multi-tenant, and multi-domain scenarios on NFV architectural framework. Some of the evaluated use cases are:

- Single operator domain network slice.
- NSI creation.
- NSSI creation.
- NSI creation, configuration and activation with VNFs.
- NSI across multiple operators.

In case of single operator domain network slices, [ETSI NFV17] suggests that additional functionality may be needed to support configuring policies, access control, monitoring/SLA rules, and usage/charging consolidation rules. The specification proposes to add an external entity called Network Slice Manager that would be responsible of:

- Determining the requirements for NSIs from the description of applications and services by mapping appropriate features into NSD and VNFD.
- Management of network slice catalogue, network slice and/or sub-network blueprint, and lifecycle of network slices.

ETSI NFV-MANO system supports and manages the resources of the VNFs, as the NSI can be composed of VNFs and PNF. In NSI creation use case the MANO is responsible for management of virtualised resources while 3GPP application takes care of network applications. The NSD contains requirements for QoS and resources of a network slice. During the instantiation the deployment flavour is selected during the instantiation. Another use case is derived from [3GPP TR 28.801] and consists of NSSI creation that is done by NSSMF. This function specifies which NFs and resources are needed. The NFs can be either VNFs or PNFs. In this case, NFV MANO supports the management of the virtualised resources. If VNFs are included in NSSI, NSSMF triggers NFV-MANO to instantiate or configure the VNFs that are needed.

NFV in multi-tenant and multi-domain environment

5G-MoNArch M&O layer supports multi-tenancy and flexible E2E network slicing. The network slices are isolated between each other and are capable to run on shared infrastructure without affecting each other.

Tenants manage the slices in their operative domains by means of NFVO. Each tenant has its own NFVO that is responsible for resource scheduling in the tenant domain. The resources can belong to different administrative domains in the infrastructure, so NFVO has to be able to orchestrate resources across different administrative domains. This is the role of Cross-domain M&O function in 5G-MoNArch. Cross-domain M&O function is in charge of managing and coordinating NSs between different management domains. On the other hand, Cross-slice M&O is responsible for common functions between different slices.

5G-MoNArch ETSI Experiential Network Intelligence Extension

In response to the industry demand for automatic networks based on AI principles, ETSI has created the Experiential Network Intelligence (ENI) workgroup [WF18]. The goal of this group is to improve the network cost efficiency and add value to the Telco provided services, by assisting in decision making. Specifically, ENI aims to define an architecture that uses AI techniques and context-aware, metadata-driven policies, in order to adapt service functionalities and parameters based on changes in user needs, environmental conditions, and business goals, by using an "observe-orient-decide-act" control loop model [GDD18]

ENI has designed a modularised system architecture [GDD18], whose main modules are defined below:

- The Policy Management module provides decisions to ensure that the operator goals and that the broader system policies, goals and objectives are met.
- The Context Awareness module describes the state and environment in which a set of the assisted system entities exists or has existed. For example, an operator may have a business rule that prevents 5G from a specific type of a network slice in a given location.
- The Situational Awareness module enables ENI to understand how information, events, and recommended commands that it may provide to the assisted system, may impact its actions and ability to meet its operational goals.
- The Cognition Management module operates at the higher level and enables ENI as a whole to meet its end to end goals.
- The Knowledge Management is used to represent information about ENI and the assisted system, differentiating between known facts, axioms, and inferences.

The interactions and interoperability of ENI with an assisted system is supported by the ENI Reference Points.

Network slicing for 5G can serve as a prime example to demonstrate ENI's architecture and the operator's benefits it provides, especially around computational resources efficiencies, while preserving the user requested SLA. In WP4, 5G-MoNArch is adapting the ENI architecture to embrace the elasticity concepts by design, provided the required algorithms to allow the elastic network operation.

As discussed above, the ETSI ENI architecture introduce the needed modules for the cognitive management of the network. These concepts, in the context of network elasticity are at the basis of the "Touristic City testbed" described in Section 4.3.2.7. The overall testbed architecture is depicted in Figure B-6. All the enablers implemented in the context of this architecture were developed in WP4 [D4.2], still the overall architecture is built on top of the one defined in WP2. WP4 is also defining the different interfaces needed for the demo operation.

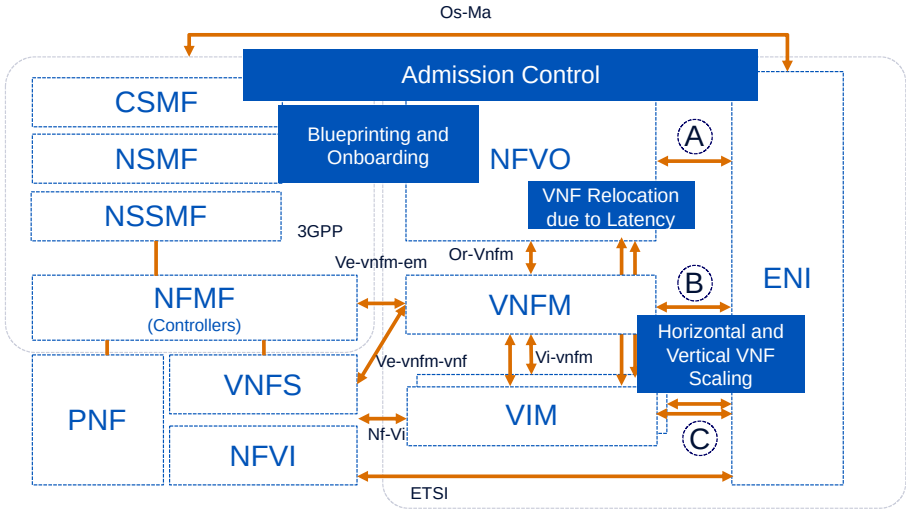


Figure B-6: Touristic City Testbed ESTI ENI aligned architecture

5G-MoNArch Relation with the Open Network Automation Platform (ONAP)

The ONAP initiative was launched in 2017 with the goal of providing a common platform to deliver differentiated network services on a shared infrastructure, cf. Figure B-7. As the main objective of ONAP is generality, in the latest version of their architecture, the ONAP consortium propose a clear split between the general, abstract models that tackle the problem of service design and the specific modules that control the lifecycle management of such services. More specifically, they define the Service Design and Creation (SDC) and the Runtime Framework realms. In a nutshell, they perform tasks that are commonly categorised under Network Management (SDC) and Orchestration (Run Time). Therefore, all the tasks related to the abstraction of resources and the high-level deployment of network services are performed within the Design Time Framework, while all the others related to the lifecycle management and the actual representation of those resources, are performed by the run time execution framework. The full specification of the architecture modules is depicted in Figure B-8.

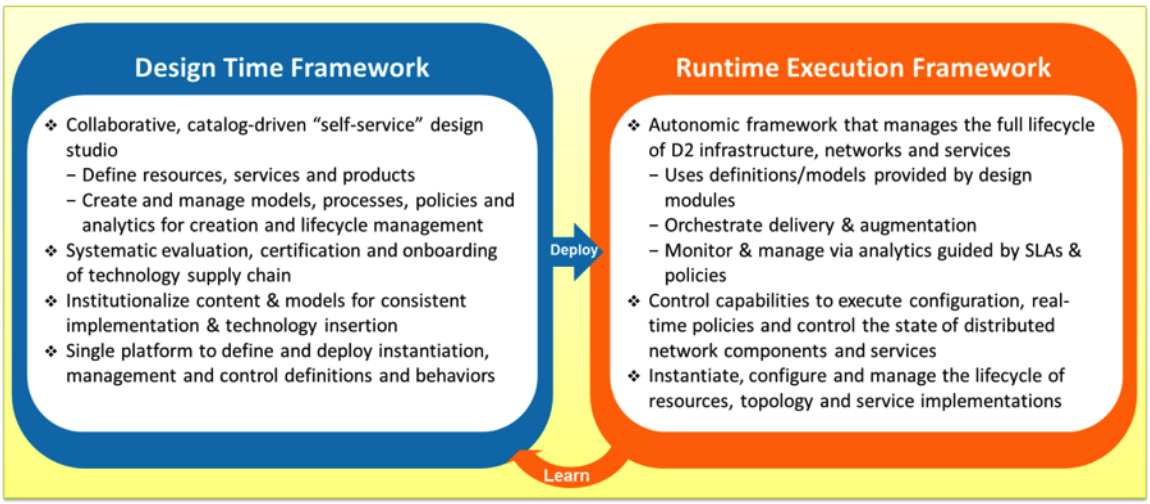


Figure B-7: ONAP architecture principles

Within ONAP, a network service is thus defined as a collection of recipes that specify the behaviour of a specific service which are deployed in the ONAP Operation Manager Portal. Recipes detail, among other things, factors such as the VNF deployment, the metric that have to be analysed and the self-healing of the network. These aspects are then enforced in the run-time of the network, as depicted in the specific part of the architecture (see Figure B-9). The ONAP and the 5G-MoNArch architecture share the same field of operation (i.e., the management, orchestration and operation of a multi-service network) although from a very different standpoint. ONAP is very much code oriented and submodule driven, while the 5G-MoNArch builds on top of the ETSI NFV framework and tackles the same problem with a top-down approach. In the following, we describe how the different modules of the two architectures relate among them.

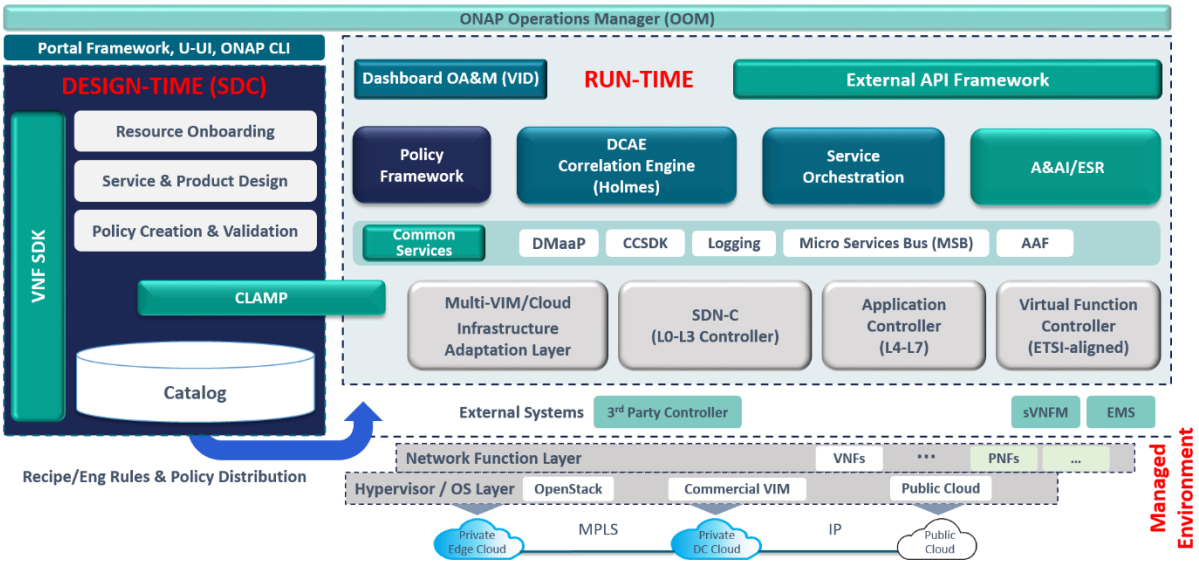


Figure B-8: ONAP architecture modules

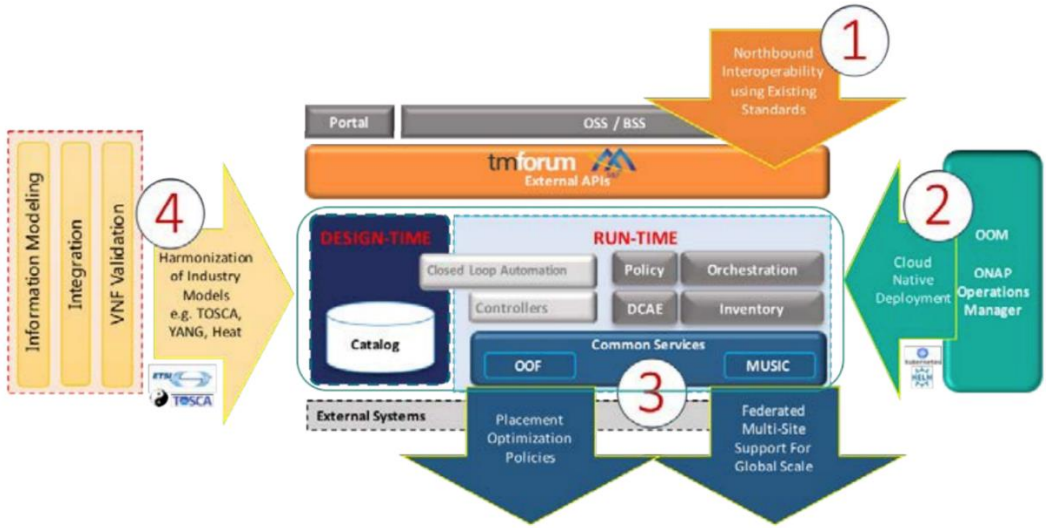


Figure B-9: ONAP architecture functional representation

Management and Service Orchestration: 5G-MoNArch relies on and extends the 3GPP management modules (CSMF, NSMF, and NSSMF) and defines the interface towards the NFV-O for the subsequent resource orchestration procedure. Within ONAP, this functionality is performed by the SDC framework that then interfaces towards the run-time modules for the lifecycle management.

Resource Orchestration and Lifecycle Management: 5G-MoNArch specifies procedures for the network slice lifecycle management by leveraging on the ETSI NFV Architecture modules and extending them

as for the case of the Intelligent Elastic Lifecycle Management [5GM-IR4.2] that also builds on the work done for ETSI ENI. ONAP adapts also a similar approach, being the Virtual Function Controller a replacement of the ETSI ENI Orchestration Stack.

From the above discussion, we can recognise one major difference between the two proposals. While in the ONAP architecture the concepts of network slicing and multi tenancy are left open and possibly enforced through the ONAP Operations Manager, in our architecture we clearly define specific roles for the involved stakeholders. We believe that this tighter definition of the interaction between the roles of the tenants / service providers and infrastructure providers as done within our proposed architecture will eventually lead to a better and clearer interaction of concurrent services provided on the same infrastructure.

B.4 Data Analytics in 5GS

This sub-section provides the current requirements on data analytics in 3GPP 5GS as well as the current architecture focusing more on CN and OAM.

B.4.1 Requirements for Data Analytics in 5GS

Data analytics can be used to serve different purposes, depending on the time granularity of the derived statistics and the defined parameters it may support. This section aims to decompose the requirements for prediction functionalities in the 5GS based on different optimisation objectives or expected benefits. This analysis considers different types of prediction models over different domains, e.g., CP, management plane, and service plane. The following sections explicitly describe the key requirements for employing data analytics in the 5GS.

Analytics for Service Enhancement

3GPP SA1 has provided the 5G requirement (SMARTER) specification [3GPP TS 22.261], where the key vertical use cases include V2X communication and vertical industry automation. For industrial automation, [3GPP TR 22.804] has studied communications for factories of the future in 3GPP Release 16 including application areas and mapped applications (e.g., motion control, massive wireless sensor networks, augmented reality, process automation, connectivity for the factory floor, and inbound logistics for manufacturing). For some use cases, data analytics can be useful for ensuring network availability or for providing predictive maintenance features. Furthermore, for enhanced 5G-V2X scenarios such as extended sensor sharing and automated cooperative driving [3GPP TR 22.886], 3GPP SA1 introduced the notion of network prediction, by enabling 5GS to notify a V2X application that the QoS of a UE's ongoing communication might need to be downgraded, e.g., due to predicted bad network conditions, change of radio technology, and radio congestion.

Analytics for 3GPP Network Enhancement

5G has introduced network slicing support to allow for dedicated network configuration and optimisation for individual scenario and services. However, due to the heterogeneous network (HetNet) deployment, the variation of network conditions as well as the changing of the traffic demand at different locations and times, the service assurance for each network slice may require complex network operation and management. This may have impact in RAN and CN domains:

- RAN: With its inherent characteristics, namely scarce resources (e.g., spectrum and computational resources), dynamic wireless channel conditions (e.g., fading and user mobility), necessity for low complexity in RAN deployments, along with the wide-spread utilisation of higher frequency bands that are even more susceptible to radio conditions, the RAN may greatly benefit from data analytics in the 5G era. In addition, the RAN can be shared by a multitude of network slices, where the essential slicing objectives, such as slice isolation, SLA guarantee, and service continuity, shall be fulfilled.
- CN: When it comes to 5G, there is significant room to put predictive and perspective analytics in usage as they enable an operator to predict an event (e.g., network overload and an upcoming outage or failure) earlier ahead to adopt suitable pre-emptive actions to ensure smooth network operation. The current usage of data analytics is limited to the individual NF/entity at intra slice-level. However, E2E service assurance requires joint consideration on intra and inter-slice coordination of CP and management plane information, as well as the feedback from the

application layer (e.g., 3rd party or PLMN-owned AFs authorised to closely interact with the 5GS). Based on these observations, a data analytics module should jointly consider the data from different NFs and different layers in a cross-slice manner.

Analytics for Network Management Enhancement

The slice requirements, along with the requirement of increasing the flexibility of the network to ensure homogeneous SLA across the slice coverage area, may present management and operational challenges and complexities when it comes to slice configuration and optimisation. Therefore, the design of analytics should follow the below principles in the slice management domain:

- Network slice management shall be driven by complex data that are the outcome of aggregation / elaboration of signals coming from multiple network resources or slice subnets. Hence, the MDAF should be responsible for providing these processed data.
- Management data analytics operates at NF level, at network slice subnet level, and network slice level:
    - Analytics at a NF level requires the collection of NF's load related performance data, e.g., resource usage status of the NF. This analysis could recommend appropriate configuration and lifecycle management actions, e.g., scaling of resources, admission control, load balancing of traffic.
    - Analytics at network slice subnet level, shall provide information for closed-loop management of the subnet and information for the overall network slice management. The analytics service may further classify or shape the data in different useful categories and analyse them for different network slice subnet management, needs, e.g., scaling and admission control of the constituent NFs.
    - Analytics at network slice level shall consume the analytics services exposed at slice subnet level to manage and orchestrate the slice life cycle in real time providing assurance management to the different communication services that are leveraging on the same shared network slice.
- The management data analytic service (MDAS) should utilise the network management data collected from the network (including e.g., service, slicing and/or NFs related data) and make the corresponding analytics based on the collected information to improve networks slice configuration and optimisation. For example, the information provided by performance MDASs can be used to optimise network performance, and the information provided by fault MDASs can be used to predict and prevent failures at network slice level.
- Network slice configuration and optimisation have to deal with the complexity of the management of shared resources and has to fit the different requirements and optimisation needs coming from all the communication services that the slice has to support. In this regard, the management system may consider what is happening in the network using real time data analytics to decide the optimised configuration parameters which are used to create a new slice or to maintain a deployed one.

B.4.2 5GS Architecture

5G-CN Architecture

The key technological components of the 5G Core Architecture rely upon principles of *architecture modularisation, CP and UP separation, and* SBI. These are reflected in the SBA as mentioned earlier (specified since 3GPP Release 15) where the CP NFs are interconnected via a common SBI. Compared to the traditional functional based network architecture design, SBA is expected to have the advantage of short role out time for new network features, extensibility, modularity, reusability, and openness. As outlined earlier, NWDAF is one key function within SBA, facilitating access to network data analytics. In Release 16, 3GPP SA2 Working Group (WG) started a new study Item, FS_eNA [3GPP TR 23.791, to study enablers for Network Automation for 5G to further clarify the usage of data analytics capability in the network layer. FS_eNA envisions improving NWDAF scope via introducing use cases and solutions for supporting network automation deployment and the related framework, to collect/provide data analytics in relation to different NFs, AFs, and Management Functions (MF), i.e., OAM. NWDAF

reuses similar service exposure mechanisms as other 5G NFs (as described in 3GPP Release 15) for data collection and data analytics exposure from / to other NFs. There are also ongoing proposals to define a new service for unified data collection/analytics exposure from/to NFs/AFs.

In addition, a new use case on "UE-driven analytics sharing" has been proposed in this study item for enhancement of NWDAF and/or other NFs. In this use case, UEs are natural data collection points to gather more localised analytics within the network. Examples of data that a UE can provide are positioning information (e.g., collected from inertial or other sensors of a UE) or user profiling info (e.g., when a UE changes environment from outdoor to indoor or from vehicular to pedestrian mode). Such information may help the NWDAF to make more intelligent decisions on slice selection (e.g., to switch from a slice with more flexible resources to a resilient one or vice versa). As UEs can simultaneously connect to or switch across different slices (e.g. in case of mobility), they can have more prominent role for data preparation for the network to provide relevant localised contextual information and to identify earlier any changes in the network compared to the past intra-slice and/or inter-slice. Some key issues proposed and currently investigated for this use case include "How the NWDAF collects the UE's information" or "How the NWDAF uses the data provided by the UE to do analytics and provides the analytics information to other NFs".

One key consideration for allowing the inter-domain interaction of data analytics in 5GS is discussed in 3GPP SA5. For the interaction of NWDAF with OAM and RAN, the data collection from OAM may reuse the existing SA5 services. And how NWDAF provides the data analytics to OAM is still under discussion. Further details can be found in [3GPP S5-186486]. Figure B-10 provides an overview of how analytics can be used across CN, RAN, and OAM to enable network automation.

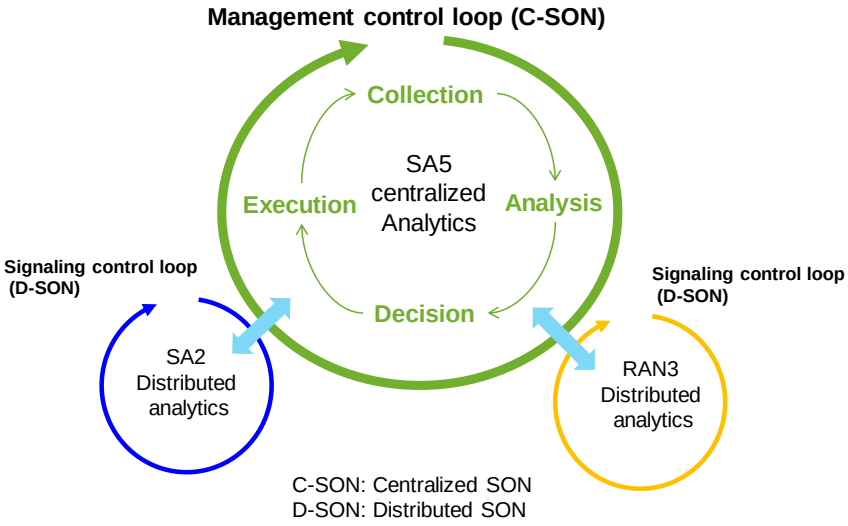


Figure B-10: General framework for 5G network automation

5G Management Architecture

For Release 15, 3GPP SA5 WG has specified an architectural framework for telecom management that realises an SBA approach. In this framework, a management service offers management capabilities that can be accessed by service consumers via a standardised service interface. Such management services include, for example, the performance management services, configuration management services, and fault supervision services. Consuming services may in turn produce (expose) these services to other consumers. Service producer and consumer may interact in a synchronous ("request-response") or asynchronous ("subscribe-notify") manner [3GPP TS 28.533]. Within this framework, SA5 introduced the MDAF that exposes one or multiple MDAS(s). Unlike an atomic function, an MDAS can exist at NF, network slice subnet, and network slice level. Deployment options for MDAS comprise centralised deployment (e.g., at a PLMN level) and domain-level deployments (e.g., RAN and CN network slice subnet instances, NSSIs). Domain MDAS provides domain-specific analytics, e.g., resource usage prediction in a CN or failure prediction in a subnet. A domain MDAF produces domain MDAS that is

consumed by the centralised MDAF or another authorised MDAS consumers (e.g., infrastructure manager, network manager, slice manager, slice subnet manger, other 3rd party OSS). A centralised MDAS can provide E2E or cross-domain analytics service, e.g., resource usage or failure prediction in a network slice, optimal CN node placement for ensuring lowest latency in the connected RAN. A centralised MDAF produces centralised MDAS, and it is consumed by different authorised MDAS consumers. SA5 is also looking at the open source scenario for the data analytic topic. Specifically, it started a study [3GPP TR 28.900] on the Open Networking Automation Platform (ONAP) Data Collection, Analytics, and Events (DCAE), the module for data collection and analytics. DCAE together with other ONAP components, gathers performance, usage, and configuration data from the managed environment. This data is then fed to various analytic applications, and if anomalies or significant events are detected, the results trigger appropriate actions. As a part of this study, 3GPP SA5 is comparing the data analytic approach and implementation of SA5 and ONAP to define new requirements for 3GPP SA5 MDAS or to give requirement to the ONAP consortium on DCAE.

Appendix C Further Analyses and Evaluation Results for 5G-MoNArch Enabling Innovations

For the sake of brevity not all analyses and evaluation results are provided in the main body of the report. In the following, more detailed analyses and results are provided for the innovation elements/enablers discussed in Chapter 3.

Context-aware relaying mode selection

The assumptions provided in [BRR+09] are taken as the basis for the relaying options, where a single DSC is considered either in the AF mode or the DF mode. Considering the links illustrated in Figure 3-26, the signal-to-noise-plus-interference ratio (SINR) on the E2E AF link at UE is given in terms of link signal-to-noise ratios (SNRs) on the individual links as

$$SINR_{\mathrm{AF}} = \frac{SNR_{\mathrm{BL}} \cdot SNR_{\mathrm{AL}} + SNR_{\mathrm{DL}}(1+SNR_{\mathrm{TI}}+SNR_{\mathrm{BL}})}{SNR_{\mathrm{BL}} + (1+SNR_{\mathrm{AL}})(1+SNR_{\mathrm{TI}})}, \quad (1)$$

where BL, DL, and AL correspond to backhaul ink, direct link, and access link, respectively. TI marks the total amplified interference in case of the AF mode, which factors in the effect of loop-back interference and co-channel interference collectively. The spectral efficiency (SE) of a link is calculated based on the Shannon approximation, i.e.,

$$SE_{\mathrm{link}} = A \cdot \log_2(1 + B \cdot SNR_{\mathrm{link}}), \quad (2)$$

where A=0.88 and B=1/1.25 are the bandwidth and SNR efficiency factors, respectively. Assuming an optimal resource split between backhaul and access links in case of DF mode, the E2E SE of the DF mode is given as [BRH+10]

$$SE_{\mathrm{DF}} = \left(\frac{1}{SE_{\mathrm{BL}}} + \frac{1}{SE_{\mathrm{AL}}}\right)^{-1}, \quad (3)$$

where BL and AL correspond to backhaul link and access link, respectively.

In the MATLAB simulation set-up, the DSC is placed at the different locations inside the cell, i.e., cell edge, cell middle, or cell centre as illustrated in Figure 3-26. During the operation, the DSC is static, e.g., while the vehicle is parked. Access link (between DSC and UE) SNR, i.e., SNR_{AL}, is varied from 0 dB to 20 dB. The SE of the direct link, i.e., SE_{DL}, is set as 0.7 bps/Hz at the cell edge, 1.2 bps/Hz at the cell middle, and 2.4 bps/Hz and the cell centre. The backhaul link is assumed to be 5 dB better than the direct link, e.g., thanks to better antenna installations and lower noise figure [3GPP TR 36.814]. Accordingly, given SE_{DL}, one can determine SNR_{DL} based on (2) and then SNR_{BL}, where $SNR_{\mathrm{BL}}=SNR_{\mathrm{DL}}+5$ dB. In order to account for the impact of amplified interference in case of the AF mode, the total interference levels (i.e., loop-back interference plus co-channel interference) are assumed as SNR_{TI}= [3, 2, 0] dB for cell-edge, cell-middle, and cell-centre operations, respectively.

In Figure C-11, an example case study is illustrated. In particular, the figure illustrates an example E2E spectral efficiency performance (BS-DSC and DSC-UE link) of DF half-duplex mode and AF full-duplex mode versus the signal to noise ratio (SNR) on the access link. A direct link performance is also exemplified, where the direct link performance indicates the relative position of the UE with respect to the serving BS. In Figure C-11, the performance breaking point between AF and DF modes is illustrated when the DSC is located closer to the cell edge. Due to the aforementioned co-channel interference, on most of the access link SNR values, DF mode outperforms the AF mode.

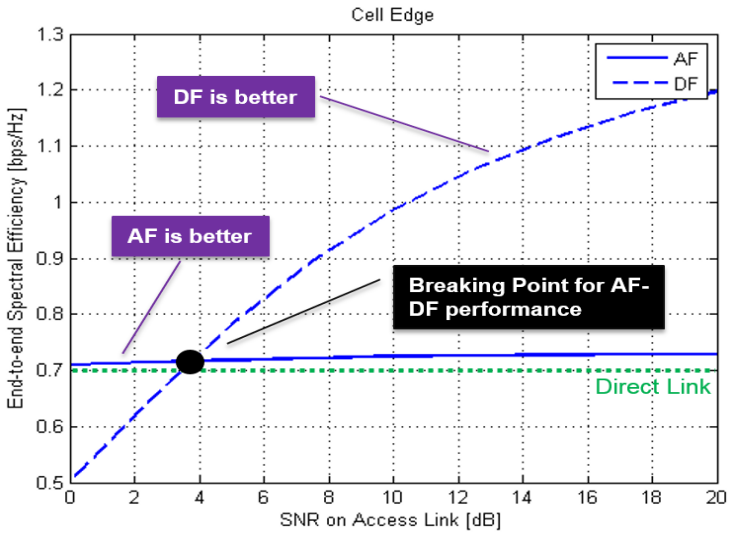


Figure C-11: Example performance comparison between AF and DF modes; breaking point shows the operation point where AF and DF performances are the same

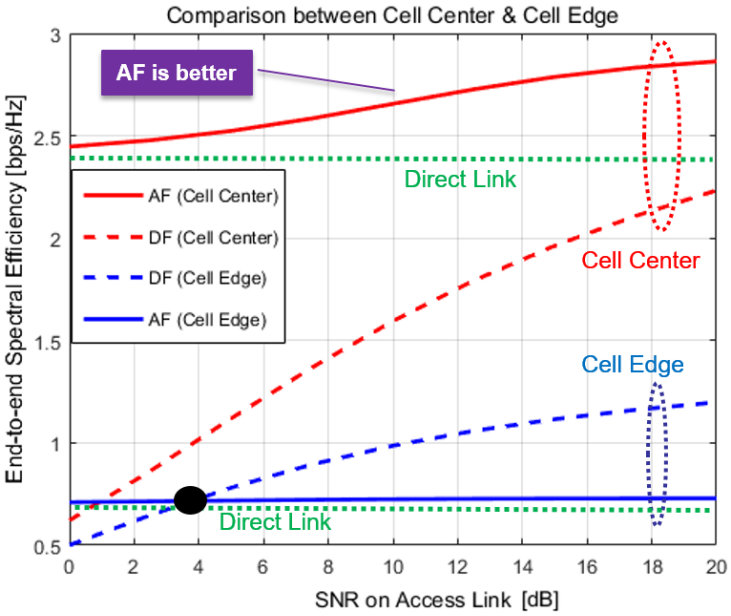


Figure C-12: Example performance comparison between AF and DF mode; breaking point shows the operation point where AF and DF performances are the same

Besides, Figure C-12 shows the performance comparisons considering both cell edge and cell centre locations. It is shown that DF mode outperforms the AF mode at the cell edge, where co-channel interference is high and thus AF undergoes interference amplification. In the cell centre, however, the AF mode outperforms the DF mode since the DF mode experiences performance loss due to the half-duplex operation [BRR+09].

Framework for slice admission results

This paragraph presents the process used to create the synthetic data set utilised in the simulation presented in Section 3.4.1: Values for the CPU and Transmit Power requirements were randomly picked from a predefined range assuming a truncated normal distribution with a mean value $\mu = \frac{\max - min}{2}$ and $\sigma^2 = 0.05 * \mu$. CPU consumption is assumed to have a linear relation with memory (CPU * 0.8), data-centre power consumption (CPU * 0.005) and cost (CPU * 0.003) while transmit power is assumed to

have a linear relation with bandwidth (Transmit power * 150) as presented in Table 3-7. Additionally, a flowchart of the slice admission process control used in the examined scenario is show in Figure C-13.

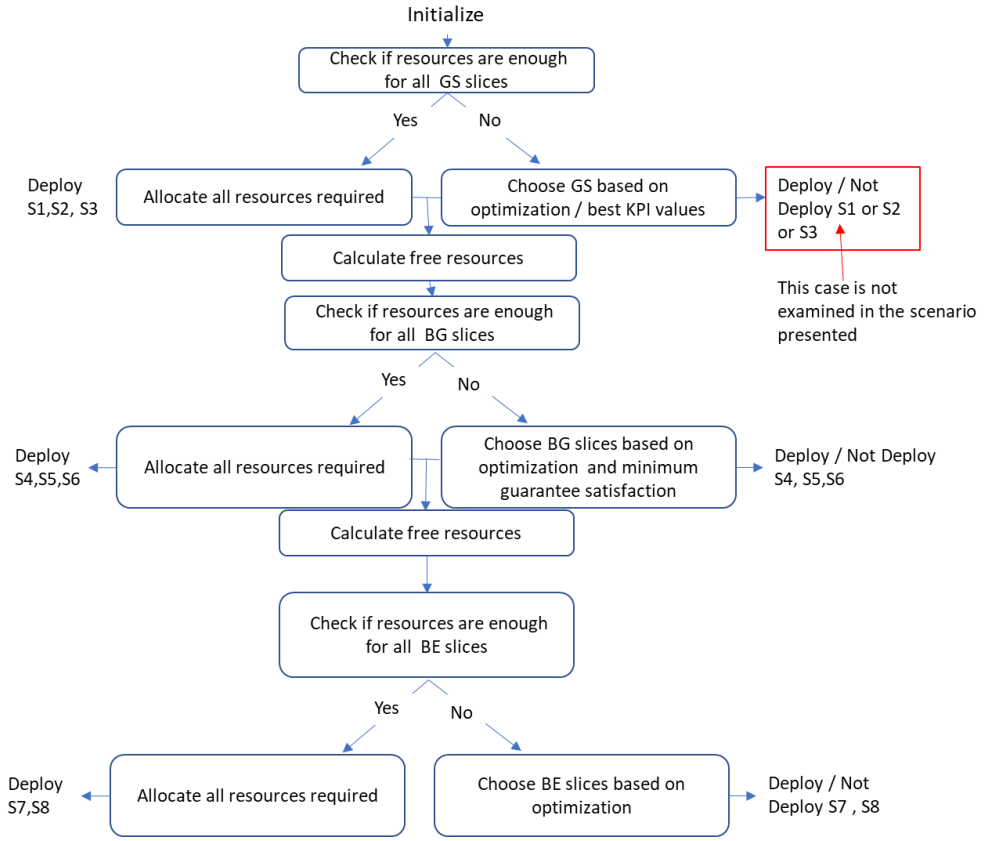


Figure C-13: Flowchart of slice admission control in the examined scenario