5G PPP Architecture Working Group

View on 5G Architecture (Version 2.0)

Date: 2017-07-18 Version: 2.0

Executive Summary

Table of Contents

List of Acronyms and Abbreviations

5G PPP	5G Infrastructure Public Private Partnership
AaSE	AIV (Air Interface Variant) agnostic Slice Enabler
ADC	Analogue Digital Conversion
AF	Adaptation Function
AI	Air Interface
AI	Artificial Intelligence
AIV	Air Interface Variant
AL	Abstraction Layer
AMF	Access and Mobility Management Function
AN	Access Network
AP	Access Point
API	Application Programming Interface
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat reQuest
B2B	Business to Business
B2C	Business to Consumer
BB	Base Band
BBU	Base Band Unit
BH	Backhaul
BS	Base Station
BSS	Business Support System
B-TAG	Backbone VLAN Tag
BTS	Base Transceiver Station
C3	Central Controller and Coordinator
CA	Certification Authority
CAPEX	Capital Expenditure
CDF	Cumulative-Distribution-Function
CDN	Content Distribution Network
CH	Cluster Head
CM	Connection Manager
CMDB	Configuration Management Database
CMS	Catalogue Management System
CN	Core Network
COTS	Common Off The Shelve

CP	Control Plane
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CQI	Channel quality Indicator
C-RAN	Cloud Radio Access Network
cRRM	Centralised Radio Resource Management
CSAT	Carrier Sense Adaptive Transmission
CSE	Circuit Switching Element
CSE	Cognitive Smart Engine
CSI-RS	Channel State Information-Reference Signal
CU	Centralised Unit
DAC	Digital Analogue Conversion
DC	Data Centre
DC	Dual Connectivity
DEI	Drop Eligible Indicator
DL	Download
DPDK	Data Plane Development Kit - a Linux Foundation Project
DPI	Deep Packet Inspection
dRRM	Distributed Radio Resource Management
DRX	Discontinuous Reception
DTX	Discontinuous Transmission
DU	Distributed Unit
E2E	End to end
eDSA	Extended Dynamic Spectrum Access
EM	Element Manager
eMBB	Enhanced Mobile Broadband / Extreme Mobile Broadband
EMS	Element Management System
EN	External Network
EPC	Evolved Packet Core
ERO	Explicit Routing Object
ETSI	European Telecommunications Standards Institute
E-UTRA	Evolved Universal Terrestrial Radio Access
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FBMC	Filter-Bank Multicarrier
FCAPS	Fault Configuration Accounting Performance Security
FDMA	Frequency-Division Multiple Access

FFT	Fast Fourier Transform
FG	Focus Group (ITU-T)
FG	Forwarding Graph
FH	Fronthaul
FMC	Fixed Mobile Convergence
FS	Fast Switching
GENI	Global Environment for Networking Innovations
GPON	Gigabit Passive Optical Network
GUI	Graphical User Interface
H-ARQ	Hybrid Automatic Repeat reQuest
HM	Home Network
HMAC	Higher Media Access Control
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secured
HW	Hardware
IA	Infrastructure Association
ICIC	Inter-site/air Interface Resource Coordination
ICT	Information and Communication Technologies
ID	Identifier
IETF	Internet Engineering Task Force
IM	Identity Management
IMoS	Intelligent Monitoring Subsystem
IMT2020	International Mobile Telecommunications 2020 (ITU)
InP	Infrastructure Provider
IP	Internet Protocol
IP	Infrastructure Provider
IPS	Internet Protocol Service
ISG	Industry Standards Group (ETSI)
ISO	International Standards Organisation
IT	Information Technology
I-TAG	Backbone Service Instance Tag
ITU-R	International Telecommunication Union Radiocommunication Sector
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
JSON	JavaScript Object Notation
JWT	JSON (JavaScript Object Notation) Web Token

K-DR	Key Design Recommendation
KPI	Key Performance Indicator
LBT	Listen Before Talk
LCM	Lifecycle Management
LCSE	Lightweight Cognitive Smart Engine
LLA	Licensed Assisted Access
LMAC	Lower Media Access Control
LTE	Long Term Evolution
LWA	LTE/Wi-Fi Aggregation
MAC	Media Access Control
MADM	Multiple Attribute Decision Making
MANA	Management and service-Aware Networking Architecture
MANO	Management and Network Orchestration
MAPE	Monitor, Analyse, Plan, Execute
MBB	Mobile Broadband
Mbps	Megabits per second
MC	Multi-Connectivity
MCS	Modulation Coding Scheme
MdO	Multi domain Orchestration
ME	Mobile Equipment
MEC	Mobile Edge Computing
MEHW	Mobile Equipment Hardware
MeNB	Master eNB
MIMO	Multiple-Input and Multiple-Output
MM	Mobility Management
MME	Mobility Management Entity
mMTC	Massive Machine Type Communications
MNO	Mobile Network Operator
MPLS	Multiprotocol Label Switching
MPLS-TP	Multiprotocol Label Switching – Transport Profile
MSP	Mobile Service Provider
MTA	Multi-tenancy Application
MVNO	Mobile Virtual Network Operator
MWC	Mobile World Congress
NaaS	Network as a Service
NAT	Network Address Translation

NE	Network Element
NF	Network Function
NFV	Network Function Virtualisation
NFVI	Network Function Virtualisation Infrastructure
NFVIaaS	Network Function Virtualisation Infrastructure as a Service
NFVO	Network Function Virtualisation Orchestration
NGFI	Next Generation Fronthaul Interface
NGMN	Next Generation Mobile Networks
NGS-3GPP	Next Generation System – 3rd Generation Partnership Project
NM	Network Management
NOMA	Non-Orthogonal Multiple Access
NR	New Radio
NRE	Near Realtime Engine
NS	Network Slice
NSI	Network Service Information
NSO	Network Service Orchestration
OAM	Operation, Administration and Management
OASIS	Organization for the Advancement of Structured Information Standards
OF	OpenFlow
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
OGF	Open Grid forum
ON	Operator Network
ONF	Open Network Foundation
OPEX	Operating Expenditure
OS	Operating System
OSI	Open Systems Interconnection
OSS	Operations Support System
OSS	Open Source Software
OTT	Over The Top (service provider)
OVS	Open vSwitch
OWASP	Open Web Application Security Project
PaaS	Platform as a Service
PBB-TE	Provider Backbone Bridge Traffic Engineering
PBM	Policy Based Management
PCP	Priority Code Point

PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PFE	Packet Forwarding Element
PGW	Packet Gateway/PDN-Gateway
PHY	Physical layer
PNF	Physical Network Function
PON	Passive Optical Network
PPP	Public Private Partnership
PRB	Physical Resource Block
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RCM	RAN Configuration Mode
RCM	Radio Connection Manager
REE	RFB (Reusable Functional Block) Execution Environment
RFB	Reusable Functional Block
RLC	Radio Link Control
RM	Resource Management
RRC	Radio Resource Control
RRM	Radio Resource Management
RRU	Radio Resource Unit
RT	Radio Transceiver
RTC	Real-Time Controller
RU	Radio Unit
SBI	Southbound Interface
SCC	Security Control Class
SC-FDMA	Single Carrier FDMA (Frequency-Division Multiple Access)
SCTP	Stream Control Transmission Protocol
SDK	Software Development Kit
SDM	Software Defined Mobile Network
SDMA	Space Division Multiple Access
SDMC	Software-Defined Mobile network Control
SDM-O	Software Defined Mobile Network Orchestration
SDN	Software Defined Networks
SDO	Standards Developing Organisation

SeNB	Secondary eNB
SFC	Service Function Chaining
SGW	Serving Gateway
SINR	Signal to Interference plus Noise Ratio
SLA	Service Level Agreement
SLO	Service Level Objective
SMF	Service Management Function
SMS	Short Message Service
SN	Serving Network
SNR	Signal to Noise Ratio
SON	Self-Organised Networks
SON	Self-Organising Network
SoTA	State of The Art
SR	Security Realm
SRIOV	Single Root Input/Output Virtualization
STP	Service Termination Point
SW	Software
TA	Trust Anchor
TADS	Topology Abstraction and Discovery Subsystem
TAL	Tactical Autonomic Language
TAU	Tracking Area Update
TDMA	Time Division Multiple Access
TEE	Trusted Execution Environment
TLS	Transport Layer Security
TN	Transit Network
TOSCA	Text and Office Systems Content Architecture
TR	Technical Report (ETSI)
TS	Technical Specification (ETSI)
TSON	Time Shared Optical Network
TTI	Transmission Time Interval
UC	Use Case
UCA	Use Customer Address
UDP	User Datagram Protocol
UE	User Equipment
UICC	Universal Integrated Circuit Card
UL	Upload

UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable and Low Latency Communications
USIM	Universal Subscriber Identity Module
V2X	Vehicle to Anything
vBBU	Virtual Base Band Unit
VI	Virtual Infrastructure
VIM	Virtual Infrastructure Manager
VLAN	Virtual Local Area Network
VM	Virtual Machine
VNE	Virtual Network Element
VNF	Virtual Network Function
VNFaaS	Virtual Network Function as a Service
VNFM	Virtual Network Function Manager
WAF	Web Application Firewall
WDM	Wavelength Division Multiplexing
WG	Work Group
WLAN	Wireless Local Area Network
WP	White Paper
XCF	Xhaul Common Frame
XCI	Xhaul Control Infrastructure
XPU	Xhaul Processing Unit

1 Introduction

The development of the fifth generation (5G) mobile and wireless networks has progressed at a rapid pace. The initial non-standalone release of 5G is already set to be completed till the end of 2017 by third generation partnership project (3GPP). Since mid of 2015, the European Union (EU) funded 5G Public Private Partnership (5GPPP) Phase 1 projects[1] have played an important role in establishing a pre-standardization consensus on areas ranging from physical layer to overall architecture, network management and software networks. Various technologies and innovations from these projects have substantially contributed to the progress in standards developing organizations (SDOs). With the aim of consolidating the outcome of 5GPPP projects into an overall architecture vision and responding to the diverse requirements of 5G use cases and services, the 5G Architecture Working Group has been active since the start of 5GPPP initiative. To this end, the first version of the white paper was released in July 2016, which captured novel trends and key technological enablers for the realization of the 5G architecture along with harmonized architectural concepts from projects and initiatives. Capitalizing on the architectural vision and framework set by the first version of the white paper, this version of the white paper presents the latest findings and analyses with a particular focus on the concept evaluations.
Various 5GPPP Phase 1 projects have been finalized by June 2017 and Phase 2 projects[2] are already kicked-off to be aligned with the accelerated 5G development. The current white paper highlights the **key design recommendations** identified by the Phase 1 projects toward the 5G architecture design. Another goal is to provide a **baseline architecture** to be facilitated by the new Phase 2 projects to assist further development.
The 5G system has the ambition of responding to widest range of service and applications in the history of mobile and wireless communications categorized under enhanced mobile broadband (eMBB), massive machine-type communications (mMTC) and ultra-reliable and low-latency communications (URLLC). In responding to the requirements of these services and application, the 5G system aims to provide a flexible platform to enable new business cases and models to integrate vertical industries, such as, automotive, manufacturing, and entertainment. On this basis, **network slicing emerges as a promising future-proof framework** to adhere by the technological and business needs of different industries. To achieve this goal, network slicing needs to designed from an end-to-end perspective spanning over different technical domains (e.g., core, transport and access networks) and administrative domains (e.g., different mobile network operators) including management and orchestration plane. Furthermore, **security architecture shall be natively integrated into the overall architecture**, e.g., to ensure the requirements of the enhanced applications and services pertaining to the safety-critical use cases.

The white paper is organized as follows. In Chapter 2, the overall 5G architecture is presented highlighting the aforementioned key attributes. Radio access domain is described from functional and protocol stack perspectives in Chapter 3, where numerical evaluations are outlined supporting these perspectives. Various aspects related to the physical deployment in the edge and transport networks are discussed in Chapter 4. In Chapter 5, the design of the 5G management and orchestration plane is detailed. Chapter 6 presents potential impacts on standardization[3] and Chapter 7 concludes the white paper with outlook.

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

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

[3] The 5GPPP pre-standard WG is working on this chapter

2 Overall architecture

5G networks will meet the requirements of a highly mobile and fully connected society. The proliferation of connected objects and devices will pave the way to a wide range of new services and associated business models enabling automation in various industry sectors and vertical markets (e.g. energy, e-health, smart city, connected cars, industrial manufacturing, etc.). In addition to more pervasive human centric applications, e.g., virtual and augmented reality augmentation, 4k video streaming, etc., 5G networks will support the communication needs of machine-to-machine and machine-to-human type applications for making our life safer and more convenient. Autonomously communicating devices will create mobile traffic with significantly different characteristics than today's dominantly human-to-human traffic. The coexistence of human-centric and machine type applications will impose very diverse functional and KPI/performance requirements that 5G networks will have to support.

The vision of network slicing will therefore satisfy the demand of vertical sectors that request dedicated telecommunication services by providing "customer-facing" on-demand network slice requirement descriptions to operators as depicted in Figure 2-1. The need for mapping such customer-centric SLAs to resource-facing network slice descriptions, which facilitate the instantiation and activation of slice instances, becomes evident. In the past, operators executed such mapping in a manual manner on a limited number of service/slice types (mainly MBB, voice service, and SMS). With an increased number of such customer requests and according network slices, a mobile network management and control framework will therefore have to exhibit a significantly increased level of automation for the entire lifecycle management of network slice instances.

More specifically, slice lifecycle automation must be realized by an architecture and comprising functions and tools that implement cognitive procedures for all lifecycle phases: preparation phase, instantiation, configuration and activation phase, run-time phase, and decommissioning phase. Two fundamental technological enablers include softwarization, e.g., virtualisation of network functions, as well as software-defined, programmable network functions and infrastructure resources. Further key elements constitute efficient management & orchestration procedures and protocols. Finally, scalable, service-centric data analytics algorithms that exploit multi-domain data sources, complemented with reliable security mechanisms, will pave the way for deploying customised network services with different virtualised NFs (VNF) on a common infrastructure in a trustworthy manner.

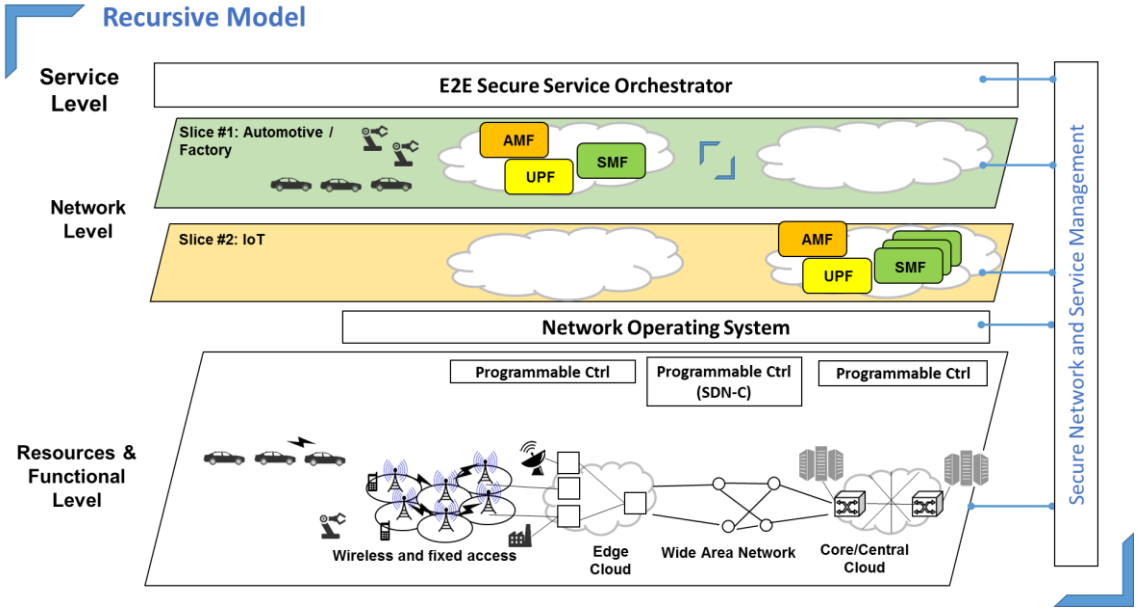


Figure 2-1: Overall Architecture

2.1 5G Services, Applications and Use Cases

In the course of identifying the requirements for the 5G network infrastructure a large number of use cases have been described and analysed in the context of standards bodies, such as 3GPP and ITU-T, industry forums such as NGMN and last but not least the projects of phase 1 of the 5G Public Private Partnership (5G-PPP). Those projects described use cases that guide the research and innovation in these projects towards demonstrating their scientific and technological achievements.

Through the interaction with the community of the industry verticals a number of additional use cases have been defined. Many available use cases are variations of a small set of basic 5G service classes, which have been consolidated and agreed in the context of 5G-PPP and different SDOs as follows:

- Enhanced Mobile Broadband (eMBB) – also called Extreme Mobile Broadband
- Ultra-Reliable and Low Latency Communications (URLLC), and
- Massive Machine Type Communications (mMTC)

Additional use cases are likely to emerge and which are not foreseen today. For future 5G systems, flexibility is necessary to adapt to new use cases with a wide range of requirements.

Currently five vertical industries have described their requirements in their respective white papers [2-1]. The requirements have been expressed in the form of vertical industry use cases, which have been further analysed in the white paper *5G empowering vertical industries* of the 5G vision and societal challenges work group [2-2], In this white paper the verticals use cases are mapped to technical capabilities of 5G that correspond to the main key performance indicators of the 5G-PPP programme, identified in the 5G-PPP contractual arrangement and extended in the *5G Vision* document [2-3].

From a technical architecture perspective version 1 of the View on 5G Architecture [2-4] by the 5G-PPP architecture work group introduces the key requirements for 5G networks and presents the design objectives for the architecture.

The document on 5G-PPP use cases and performance evaluation [2-5] provides an overview of the use cases that are used for evaluation of different 5G radio access network concepts. It refines the use case classes provided above, by defining use case family groups in order to better reflect their use in 5G-PPP phase 1 projects. The identified groups are:

- Dense urban
- Broadband (50+Mbps) everywhere
- Connected vehicles
- Future smart offices
- Low bandwidth IoT
- Tactile internet / automation

The grouping is based on stated ranges for the metrics for each of the KPIs relevant for the service experience of the customer, namely:

- Device density
- Mobility
- Infrastructure (related to topology)
- Traffic type
- User data rate
- Latency
- Reliability
- Availability (related to coverage)
- 5G service type (eMBB, URLLC, mMTC)

Additional use cases and related KPIs are identified that are relevant from the deployment and network operational perspective, namely:

- Network slicing, which considers the ability to create end-to-end slices on the same infrastructure for heterogeneous services

- Multi-tenancy, which considers the ability to offer connectivity services to multiple tenants and to combine resources from different operators
- Flexibility, which considers the possibility to dynamically configure networks in time and space, depending on foreseen or unforeseen events

Finally the following capabilities, although expressed from the vertical sectors perspective, are key for a successful commercialisation:

- Service deployment time, defined as the duration required for setting up end-to-end logical network slices characterised by respective network level guarantees.
- Data Volume, defined as the quantity of information transferred per time interval over a dedicated area.
- Autonomy, defined as the time duration for a component to be operational without power being supplied. It relates to battery lifetime, battery load capacity and energy efficiency.
- Security, defined as a system characteristic ensuring globally the protection of resources and encompassing several dimensions such as, among others, authentication, data confidentiality, data integrity, access control and non-repudiation.
- Identity, defined as the characteristic to identify sources of content and recognise entities in the system.

In the following table the KPIs and capabilities identified above are assessed with respect to 5G architecture relevance and mapped to the architecture mechanisms presented in this document.

KPI/capability	**Architecture relevance**	**Reference to mechanism**
Device density	High, RAN level	Sec. 3
Mobility	High, system level	Sec. 3
Infrastructure (related to topology)	High, system level	Sec. 4
Traffic type	Medium, RAN and system level	Sec. 3
User data rate	Medium, RAN and back-/fronthaul level	Sec. 3, Sec. 4
Latency	High, RAN and system level	Sec. 3, Sec. 4
Reliability	High, management level	Sec. 5
Availability (related to coverage)	Low, system level	Sec. 4
Network slicing	Fundamental concept	Sec. 2
Multi-tenancy	Fundamental concept	Sec. 2
Flexibility	Fundamental requirement implemented through complete softwarization	Sec. 5
Service deployment time	High, system level	Sec. 5
Data Volume	Medium, system level	Sec. 4
Autonomy	N/A	
Security	High, system level	Sec. 2.5
Identity	High, system level	Sec. 2.5

2.2 Network Slicing

The industry consensus is that by 2020, 5G network of the future will involve the integration of several cross-domain networks, and the 5G systems will be built to enable logical network slices across multiple domains and technologies to create tenant- or service-specific networks. The network slicing shall realize end-to-end (E2E) vision starting from the mobile edge, continuing through the mobile transport including fronthaul (FH) and backhaul (BH) segments, and up until the core network (CN). This will enable operators to provide networks on an as-a-service basis and meet the wide range of use cases that the 2020 timeframe will demand. In the same context,

a profound relationship is considered between the concept of network slices and 5G integrated environments.

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

In the following, further elaboration on the slicing definition and motivation is provided. The functional layers for the implementation of network slicing are highlighted, and the lifecycle management of network slices is discussed. Inter-slice and intra-slice control mechanisms are depicted along with implementation examples on the protocol stack. The business realization and possible extensions to the current network slicing context are captured, as well.

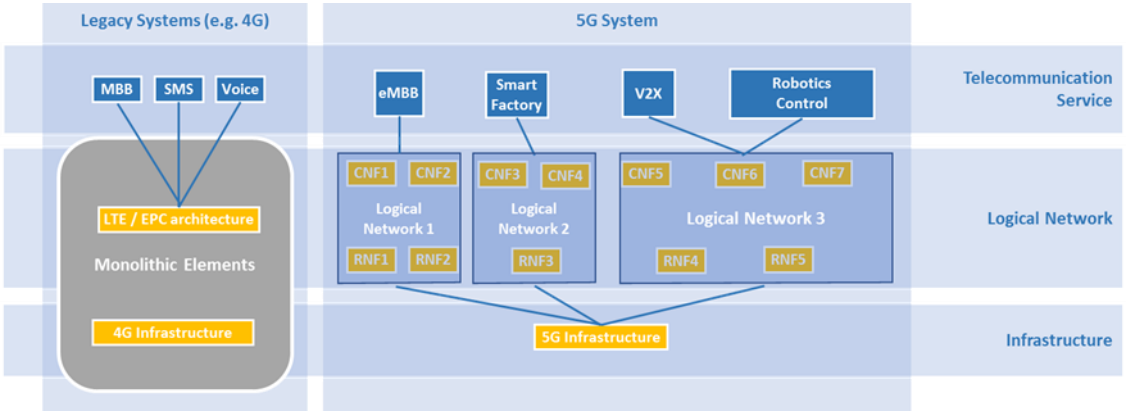


Figure 2-2: Multi-tenancy in legacy networks and slicing-enabled networks [2-21]

2.2.1 Network Slicing Context, Definition and Motivation

A number of definitions slicing as partitions of connectivity resources were used in the last ten years within the context of research into distributed and federated testbeds and in future internet research [2-6]. More recently in research and SDOs revised definitions were used [2-7][2-8][2-9][2-10][2-11][2-12].

The network slice is a composition of adequately configured network functions, network applications, and the underlying cloud infrastructure (physical, virtual or even emulated resources, RAN resources etc.), that are bundled together to meet the requirements of a specific use case, e.g., bandwidth, latency, processing, and resiliency, coupled with a business purpose. Following the 5G verticals paradigm [2-13], an infrastructure provider will assign the required resources for a network slice, that in turn realizes each service of a service provider portfolio (e.g., the vehicular URLLC network slice, the factory of the future URLLC network slice, the health network mMTC network slice, see Figure 2-3. Hence, a network slice comprises a subset of virtual network infrastructure resources and the logical mobile network instance with

the associated functions using these resources. It is dedicated to a specific tenant that, in turn, uses it to provide a specific telecommunication service (e.g. eMBB). The decoupling between the virtualised and the physical infrastructure allows for the efficient scaling-in/out/up/down of the slices, suggesting hence the economic viability of this approach that can adapt the used resources on demand. The network slices will span the whole protocol stack from the underlying (virtualised) hardware resources up to network services and applications running on top of them. This approach is aligned with the industry and telecom perspective, towards 5G [2-14], in order to meet the demands of extremely diverse use cases. Although, the infrastructure resources could be shared among several parallel network slices, every provider may use a specific control framework or/and a specific cloud management system and, in addition, all the configuration effort and fine-tuning of the components may be left to users. Advanced orchestration and automation is required to release the configuration burden from users and to enable an integrated end-to-end solution. **Network Slicing is an end-to-end concept covering all network segments including radio networks, wire access, core, transport and edge networks.** It enables the concurrent deployment of multiple end-to-end logical, self-contained and independent shared or partitioned networks on a common infrastructure platform.

From a business point of view, a slice includes a combination of all the relevant network resources, network functions, service functions and enablers required to fulfill a specific business case or service, including OSS and BSS.

The behaviour of the network slice realized via network slice instance(s). From the network infrastructure point of view, network slice instances require the partitioning and assignment of a set of resources that can be used in an isolated, disjunctive or non- disjunctive manner for that slice.

Network slicing considerably transforms the networking perspective by abstracting, isolating, orchestrating, softwarizing, and separating logical network components from the underlying physical network resources and as such they enhance the network architecture principles and capabilities.

To support network slicing, the management plane creates a group of network resources, it connects with the physical and virtual network and service functions as appropriate, and it instantiates all of the network and service functions assigned to the slice. For slice operations, the control plane takes over governing of all the network resources, network functions, and service functions assigned to the slice. It (re-) configures them as appropriate and as per elasticity needs, in order to provide an end-to-end service. In particular, ingress routers are configured so that the appropriate traffic is bound to the relevant slice.

The establishment of slices is both business-driven as slices are the support for different types and service characteristics and business cases, and technology-driven as slices are a grouping of physical or virtual resources (network, compute, storage) which can act as a sub network and/or a cloud. A slice can accommodate service components and network functions (physical or virtual) in all of the network segments: access, core, and edge / enterprise networks.

Network operators can use network slicing to enable different services to receive different treatment and to allow the allocation and release of network resources according to the context and contention policy of the operators. Such an approach using network slicing would allow a significant reduction of the operations expenditure. In addition, network slicing makes possible softwarization, programmability and allows for the innovation necessary to enrich the offered services. Network softwarization techniques may be used to realise and manage network slicing. Network slicing provides the means by which the network operators can provide network programmable capabilities to both OTT providers and other market players without changing their physical infrastructure. Slices may support dynamic multiple services, multi-tenancy, and the integration means for vertical market players (such as, the automotive industry, energy industry, healthcare industry, media and entertainment industry).

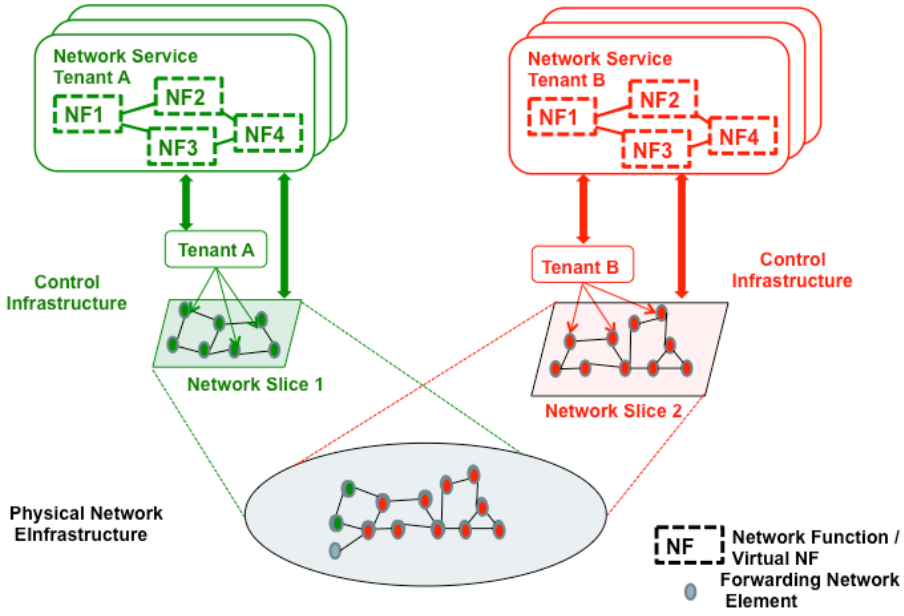


Figure 2-3: Network Slicing Representation

2.2.2 5G Functional Layers

In order to serve all aspects of network slicing, the 5G architecture is divided into different layers [2-15] as shown in Figure 2-4:

- The **Service layer** comprises Business Support Systems (BSSs) and business-level Policy and Decision functions as well as applications and services operated by the tenant. This includes the end-to-end orchestration system.
- The **Management and Orchestration layer** includes ETSI NFV MANO functions, i.e., the VIM, the VNF Manager and the NFVO. An Inter-slice Broker handles cross-slice resource allocation and interacts with the Service Management function. Further, the MANO layer accommodates domain-specific application management functions. E.g., in the case of 3GPP, this comprises Element Managers (EM) and Network Management (NM) functions, including Network (Sub-)Slice Management Function (N(S)SMF). Those functions would also implement ETSI NFV MANO interfaces to the VNF Manager and the NFVO. The Service Management is an intermediary function between the service layer and the Inter-slice Broker. It transforms consumer-facing service descriptions into resource-facing service descriptions and vice versa.
- The **Control layer** accommodates the two main controllers, SDM-X and SDM-C, as well as other control applications. Following the SDN principles, SDM-X and SDM-C translate decisions of the control applications into commands to VNFs and PNFs. SDM-X and SDM-C as well as other control applications can be executed as VNFs or PNFs themselves.
- The **Multi-Domain Network Operating System Facilities** which includes different adaptors and network abstractions above the networks and clouds heterogeneous fabrics. It is responsible for allocation of (virtual) network resources and maintain network state ensure network reliability in a multi domain environment.
- The **Data layer** comprises the VNFs and PNFs needed to carry and process the user data traffic.

In addition changes to all functional layers is realised via native softwarisation all network elements part of network segments: Radio networks, wire access, core, transmission and edge networks effective integration of communication and computation.

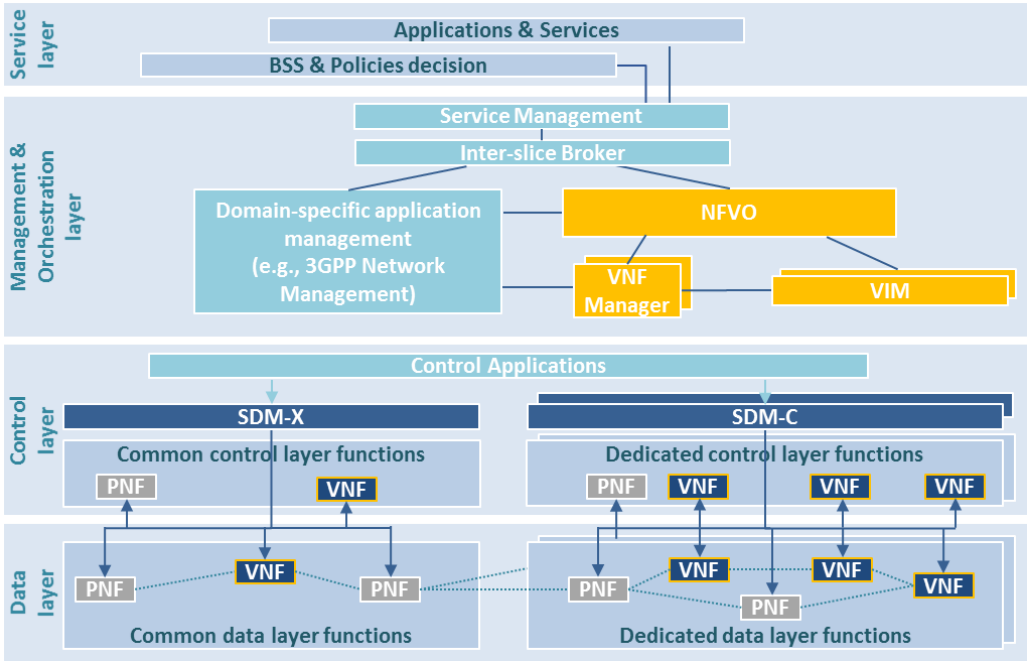


Figure 2-4: Architecture functional layers

Two main network slicing services can be considered that enable different degrees of explicit control and are characterized by different levels of automation of the mobile network slices management:

(1) the provisioning of Virtual Infrastructures (VI) under the control and operation of different tenants – in line with an Infrastructure-as-a-Service (IaaS) model, i.e., creation of a Network Slice Instance;

(2) the provisioning of tenant's owned Network Services (NS) as defined by the ETSI NFV architecture [2-16], i.e., creation of a Service Instance.

In the former service, the deployment of a mobile network deals with the allocation and de-allocation of VIs. A VI is defined as a logical construct composed of virtual links and nodes, which, as a whole, "behaves as" and "can be operated-as" a physical infrastructure. The logical entities within a VI encompassing a set of compute and storage resources are interconnected by a virtual and logical network. The VIs can be operated by the tenant via different SDN control models, enabling different degrees of internal control. This service involves dynamic allocation of a VI, its operation and de-allocation. The actual realization of a VI combines many aspects like partitioning and book-keeping of resources or the instantiation of connections supporting virtual links. The provisioning of a VI commonly requires direct hardware element support or its emulation via software for multiplexing over the shared infrastructure.

In the latter, Network Services (NS) are instantiated directly over a shared infrastructure, and as a set of interrelated Virtual Network Functions (VNFs). A NS corresponds to a set of endpoints connected through one or more VNF Forwarding Graphs (VNF-FGs). The allocation of a NS extends and complements the concept of VI deployment to deliver isolated chains of virtual services composed of specific VNFs, in an automated manner and exploiting the sharing of a common physical infrastructure with computing, storage and network resources. The tenant request usually specifies the type of VNFs (i.e. the desired virtual application components) in the NS Descriptor, their capabilities and dimensions through one or more VNF Descriptors and how they must be interconnected through a VNF-FG Descriptor. Templates for the unified description of these information elements are currently under standardization process in the ETSI NFV ISG and in OASIS TOSCA standards [2-17].

To enable both services providing different degree of control of network slices, a set of APIs can be defined:

- Network Service Allocation / Modification / De-allocation API,
- Virtual Infrastructure Allocation /Modification / De-allocation API,
- Virtual infrastructure control API with limited control, and
- Virtual infrastructure control API with full control.

2.2.3 Network Slicing Characteristics and Life-cycle management

Network slicing enables the operator to create logically partitioned networks at a given time customized to provide optimized services for different market scenarios. These scenarios demand diverse requirements in terms of service characteristics, required customized network and virtual network functionality (at the data, control, management planes), required network resources, performance, isolation, elasticity and QoS issues. A network slice is created only with the necessary network functions and network resources at a given time. They are gathered from a complete set of resources and network /virtual network functions and orchestrated for the particular services and purposes.

The network slicing reference framework is represented by two distinct levels:

- Network slice life-cycle management level, i.e. the series of state of functional activities through which a network slice passes: creation, operation, deletion, and
- Network slice instances level, i.e. activated network slice level, as shown in next figure.

Functions for creating and managing network slice instances and the functions instantiated in the network slice instance are mapped to respective framework level.

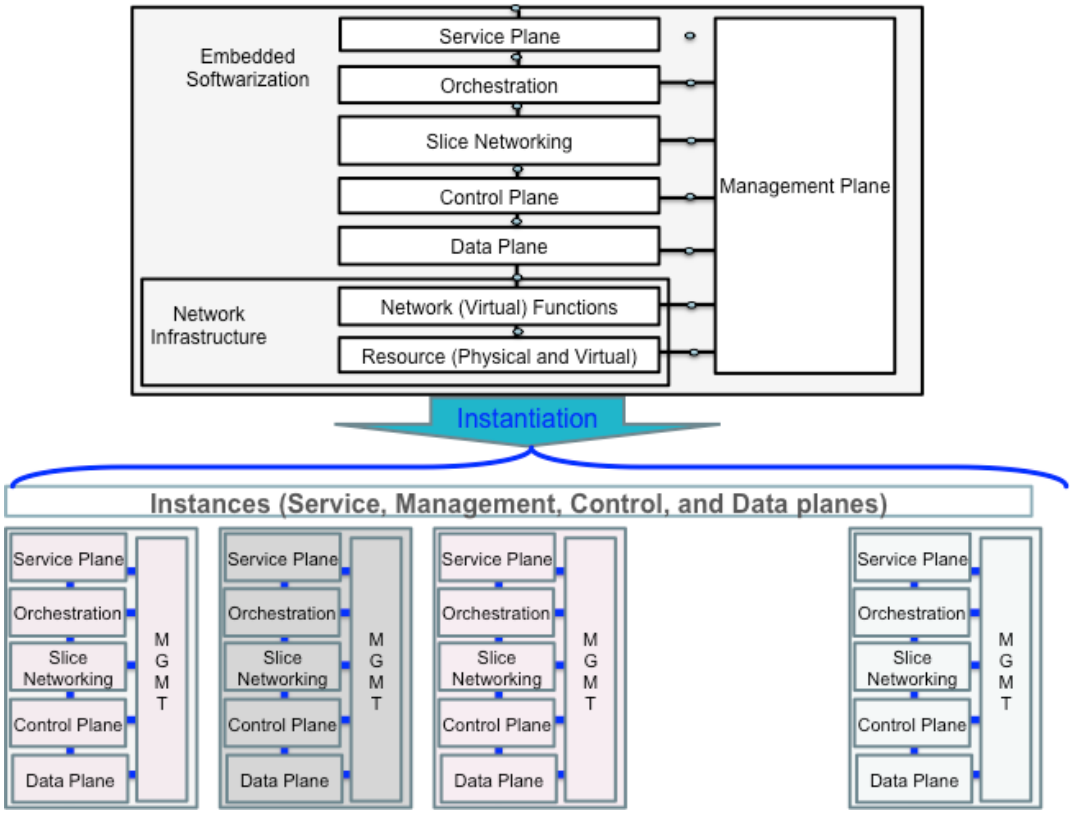


Figure 2-5 - Network Slicing Life-Cycle

In order to implement and use network slice functions and operations, there is a clear need to look at the complete life-cycle management characteristics of network slicing solutions based on the following architectural tenets:

- Governance tenet: A logically centralized authority for all the network slices in a domain.
- Separation tenet: Slices may be independent of each other and have an appropriate degree of isolation from each other.
- Capability exposure tenet: Allow each slice to present information regarding services provided by the slice (e.g., connectivity information, mobility, and autonomicity) to third parties, via dedicated interfaces and /or APIs, within the limits set by the operator.

In pursuit of solutions for the above tenets with the relevant characteristics within the context of 5G Networking, the followings are expected Network Slicing characteristics and challenges:

Network Slice Capabilities:

- Guarantees for isolation in each of the Data / Control / Management / Service planes. Having enablers for safe, secure and efficient multi-tenancy in slices.
- Methods to enable diverse service requirements for NS, including guarantees for the end-to-end QoS of a service within a slice.
- Recursion, namely methods for NS segmentation allowing a slicing hierarchy with parent–child relationships.
- Methods and policies to manage the trade-offs between flexibility and efficiency in slicing.
- Resources and network functions Optimisation, namely methods for automatic selection of network resources and functions.
- Monitoring the status and behaviour of NS in a single and/or muti-domain environment; monitoring of NS interconnection.
- Capability exposure for NS with APIs for slice specification and interaction.
- Programmability and control of Network Slices.

Network slice Operations

- Slice management including creation, activation / deactivation, protection, elasticity, extensibility, safety, sizing and scalability of the slicing model per network for slices in radio networks and wire access, core, transport and edge networks.
- Autonomic slice management and operation, namely self-configuration, self-composition, self-monitoring, self-optimisation, self-elasticity for slices that will be supported as part of the slice protocols.
- Slice stitching / composition by having enablers and methods for efficient stitching / composition / decomposition of slices: vertically (through service + management + control planes); horizontally (between different domains as part of access, core, edge segments); or a combination of vertically + horizontally.
- End-to-end network segments orchestration of slices.
- Service Mapping- dynamic and automatic mapping of services to network slices.
- Efficient enablers and methods for integration of the above capabilities and operations.

2.2.4 Slice Moderation: Inter-Slice/Intra-Slice Control and Management

Methods for resource sharing involve multitasking, virtualization, and multiplexing. All three enable the sharing of resources between multiple users by i) decoupling the functionality from the resources needed to execute this functionality, and ii) partitioning of resources into isolated execution environments.

This joint property suggests combining hypervisors, multiplexers and multitasking mechanisms in a common abstraction layer. While hypervisors manage the resources of x86-based servers in

the central cloud and the network edge cloud, multiplexers and multitasking mechanism perform the same task for other components, like DSPs and accelerators in the base stations and PNF nodes. In this way, partitioning can be applied to all components in the network, resulting in a network slicing from end-to-end.

As pointed out in [2-18], multiplexing and multitasking can be applied on different levels of the ISO-OSI protocol stack as shown in Figure 2-6.

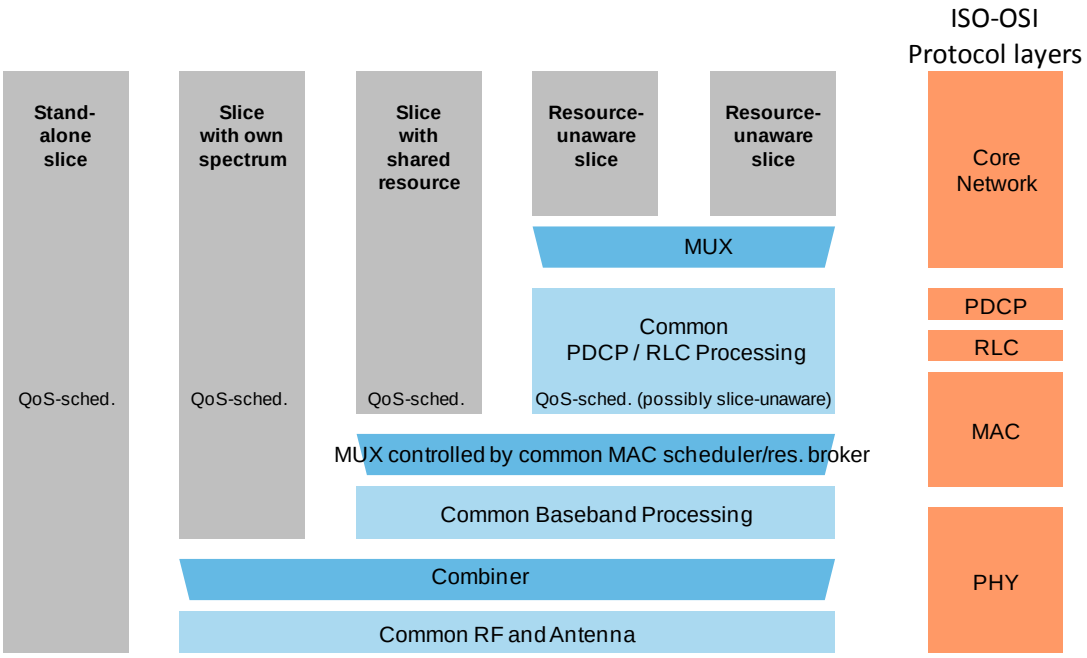


Figure 2-6: Options for slice multiplexing and their relation to the OSI protocol stack [2-21]

This yields several options for the design of network slices, ranging from standalone slices with own HW and spectrum, to slices that are completely unaware of the resources they are using and hence have no (direct) control on the resource scheduling. The differences between the slice variants will be reflected by the templates of the respective network slices.

From the RAN support of network slicing perspective, slicing can be realized as a limited number of different RAN functions that can serve a specific use case, e.g., uMTC. Different use cases can use the same combinations of RAN functions. We define this combination of RAN functions as "RAN configuration mode" (RCM). An RCM can be statically defined or fully flexible, and this is up to the implementation and the requirements for flexibility and future-proofness (i.e., in case a totally new use case arises with new unforeseen requirements).

The generic considerations for the RCMs have been presented in details in [2-19] and are captured Figure 2-7. In brief, it can be foreseen that:

- the different RCMs share an RRM (Radio Resource Management) function for ensuring the sharing of the common radio resources; also, this function can ensure that, in the case of the RCMs sharing the lower layer functions the slice isolation can be ensured at least using QoS classes. However, each slice anyway can apply its own RRM strategies according the slice specific characteristics.
- At least a common RRC part for all slices will be present, as it is seen there is a shared part which enables the slice selection. Each slice can have its own RRC functions and configurations as well so as to tackle the special UC requirements when it comes to particular functions (e.g., DRX, DTX, measurements reporting, TAU periodicity, cell selection strategies, etc.) when particular shavings can be achieved. One alternative

implementation of the common part of the RRC could be a common slice which will provide information for slice selection

- For PDCP and the RLC, depending on the message size, or the delay requirements certain functions can be either omitted (e.g., header compression, ciphering) or modified (e.g., segmentation, re-ordering, ciphering).
- The RCMs that share the lower layers (PHY, MAC, etc.) should have a joint "Unified Scheduler" for enabling them to share the resources more dynamically.

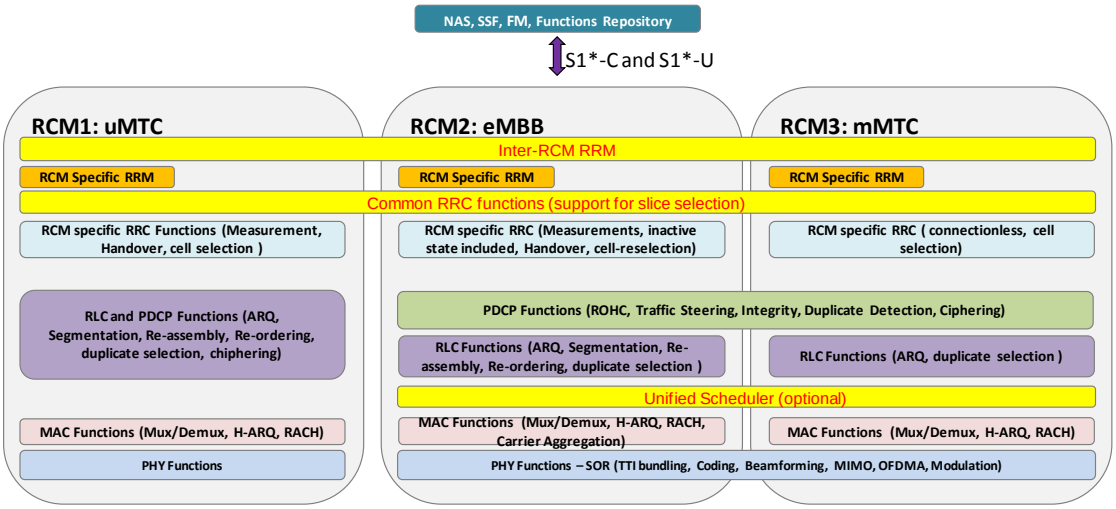


Figure 2-7: Example of RCMs with shared and independent functions

On this basis, inter-RCM RRM is a key aspect to fulfil the business-driven service-level agreements (SLAs) by exploiting the slice-specific QoS enforcement. At RAN level, an efficient sharing of scarce radio resources among the network slices is the key challenge. **The efficient multi-slice RRM is realized with the help of the AIV agnostic Slice Enabler (AaSE), which is responsible for monitoring and enforcing SLAs for individual slices by mapping the abstract slice specific SLA definition to the QoS policies,** see Figure 2-8a**.** It monitors the status of the SLAs and adapts QoS parameters accordingly. It could, for example, in case of a network slice with a latency guarantee, assign to all corresponding data flows that are part of it, a certain QoS class. Using Allocation and Retention Priority (ARP), the importance of individual data streams can be configured. It is then a task of the multi-AIV resource mapping, interference management, and real-time resource mapping to realize the corresponding QoS. More details on the proposed solution as well as simulation results can be found in [2-20].

Furthermore, a key functionality of AaSE can be the adaptive placement of intra-slice RRM functionalities to the RAN nodes, assuming that schedulers can coordinate clusters of APs. By taking into account the slice requirements, the backhaul/access channel conditions and the traffic load, AaSE can assign schedulers to BSs for pre-defined clusters of nodes, as well as RRM functionalities with different levels of centralization in order to meet the per slice SLAs (in terms of throughput, reliability, latency).

The simulation results in Figure 2-8b show a comparison of two RANs (subnetworks) with best effort traffic in terms of user throughput. In the first case (red curves), two dedicated networks with 10 MHz system bandwidth each are operated for independent businesses. The dedicated network 1 serves hundred users with a low demand, such that the network is low loaded. In contrast, the dedicated network 2 serves 710 users causing a fully loaded system with lower performance per user. In the second case (blue curves), a common RAN for both networks is operated on 20 MHz system bandwidth. The detailed simulation assumptions can be found in Annex A.10 of [2-20]. The pooling of resources enables a gain in user throughput as can be depicted from Figure 2-8 showing that the probability for users in both slices to miss a certain throughput figure is always well below that of users in both networks (solid curves). By means of

an SLA, it is targeted that users of the virtual network 1 (network slice 1) reach a similar capacity as in the case of dedicated networks. As the dedicated network 1 reached a mean network throughput (averaged over time) of 218 Mbps, an SLA was used to a guaranteed network capacity of 220 Mbps. Network slice 1 achieves a network throughput of 209 Mbps. This is slightly below the guaranteed capacity due to variations in the traffic pattern that cause a demand of less than 220 Mbps at some time instances. Consequently, the simulation results show that network slicing can achieve performance gains due to pooling of resources while protecting the performance of individual network slices.

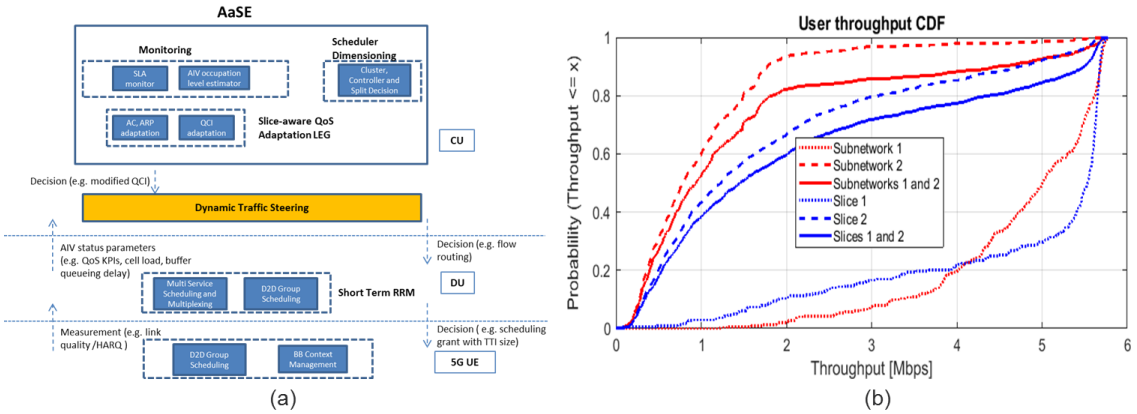


Figure 2-8: AaSE for multi-slice RRM (a) and simulation results (b)

In Figure 2-9, one additional evaluation study is shown for the scheduler dimensioning and placement of RRM functionalities. For different slices we may have different requirements for spectral efficiency and different RRM centralization requirement. For the example shown in a practical scenario (see Annex A.11 of [2-20]), for uMTC (termed also as URLLC) more than 1bps/Hz is an acceptable level, while for eMBB more than 2.5bps/Hz spectral efficiency is required. Thus, we select the level of centralization considering these requirements and the interference levels (e.g., for cell edge users we might need centralization to benefit from multi-connectivity at cell edges). The per-AP Spectral Efficiency for this particular simulation setup can be seen in Figure 2-9. As we can observe from the CDF of spectral efficiency, for the uMTC slice we do not need to centralize RRM, unless the users are near the cell-edge (e.g., 5 percentile), since the spectral efficiency KPI is fulfilled. On the other hand, for eMBB the higher the centralization the higher gain we can achieve.

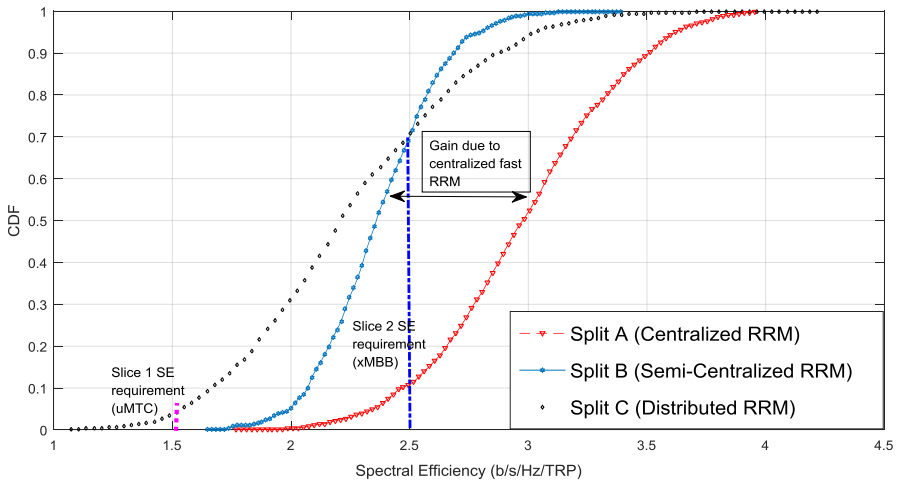


Figure 2-9: CDF of Spectral Efficiency – Comparison of different splits

The novel concept of network [2-21] control extends the software-defined routing (switching) approach to all kinds of mobile NFs from both data and control layer, with a focus on wireless control functions, such as, scheduling or interference control. For this purpose, controllers apply the split between the *logic* of the network function and the part that can be controlled (*agent*), implemented by a network function. As illustrated in Figure 2-10, Software-Defined Mobile Network Controller (SDM-C) and Software Defined Mobile Network Coordinator (SDM-X) take care of dedicated and shared NFs, respectively. In addition, SDM-O can set up slices and merge them properly at the described multiplexing point using the network slice templates of slice variants.

Each network slice has an SDM-C, responsible for managing the network slice resources and building the paths to join the network functions taking into account the received requirements and constraints which are being gathered by the QoS/QoE Monitoring and Mapping module. The SDMC, based on slice performance reports received by QoS/QoE Monitoring and Mapping module, may adjust the network slice configuration either by reconfiguring some of the VNFs in a network slice or by reconfiguring data paths in a SDN-like style. The QoS/QoE module along with the SDM-C constitute the intra-slice management If the requirements cannot be met by aforementioned reconfigurations of VNFs or data paths the SDM-O can perform a slice reshaping e.g., by adding more resources to the given network slice.

The SDM-O has a complete knowledge of the network managing the resources needed by all the slices of all tenants. This enables the SDM-O to perform the required optimal configuration in order to adjust the amount of used resources. While the SDM-C directly interfaces with dedicated NFs, the SDM-X controls shared NFs. Together with the SDM-O the SDM-X constitutes the inter-slice management. The inter-slice management and orchestration is a key-feature of the novel 5G architecture as it fosters and supports multi-service and multi-tenancy systems.

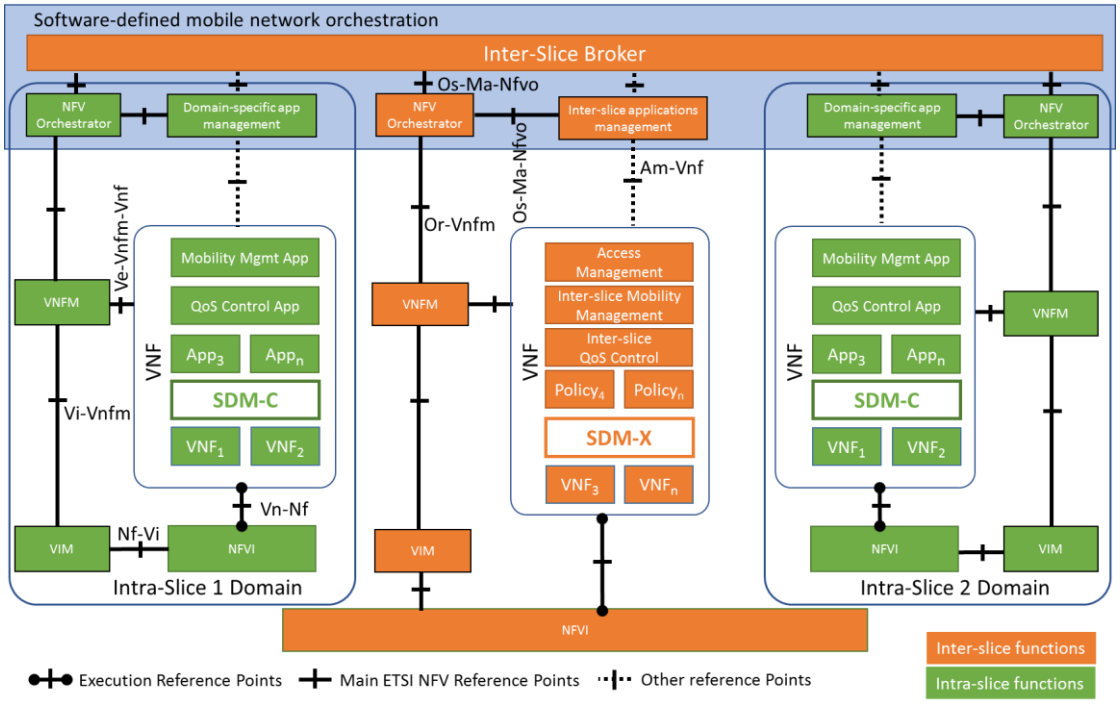


Figure 2-10: Inter- and intra-slice MANO framework [2-22]

2.2.5 Business Realization and Stakeholders

In 5G, the mobile service provider (MSP)'s role is central, as illustrated in Figure 2-11 [2-21]. The MSP is intersecting between tenant and InP. There is no direct relationship between InPs and tenants and the MSP is actually brokering the resources from possibly multiple InPs. The MSP's role is to acquire the necessary resources from one or more InPs to build an end-to-end virtual

network (slice) instance according to the needs of the tenant, i.e. a collection of (mobile) network function instances including their required resources necessary to operate an end-to-end (self-contained) logical mobile network. The MSP has to ensure that the SLAs he has with the tenants are satisfied, while being constrained by the availability of resources rented (bought) from possibly multiple InPs as presented in the figure below. In addition, when also owning the required resources, i.e., (parts of) the infrastructure (e.g. RAN), the MSP acts as an MNO.

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

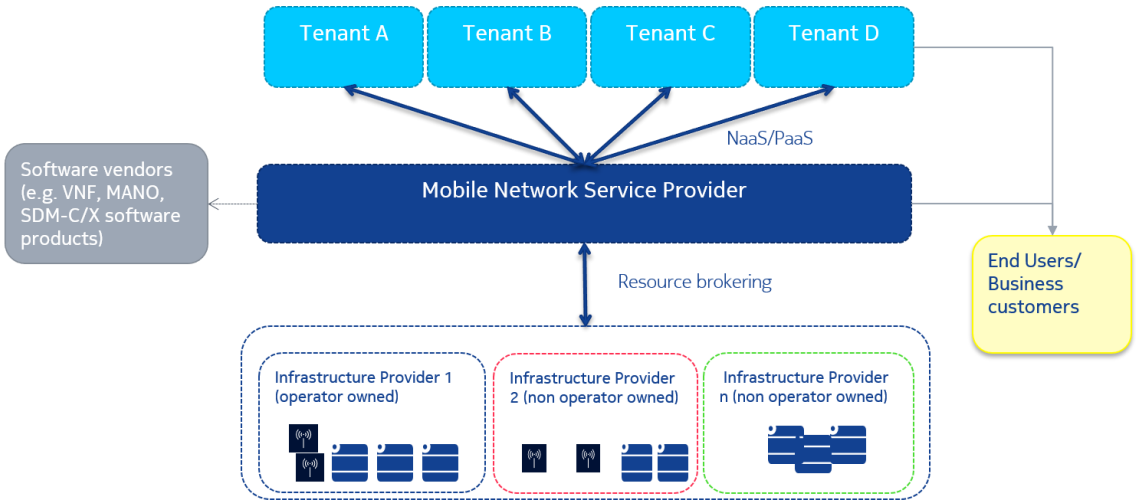


Figure 2-11: Relationship between stakeholders and Mobile Service Provider in the core place

In addition to the support sharing of the common transport infrastructure by multiple tenants, 5G architecture shall also allow each tenant to own and deploy and have a full degree of control of its slice. The designed needs to be designed to not only. This case is referred to as recursive architecture, building a hierarchy of tenants operating on top of slices of virtual infrastructure. This concept requires support for recursion of the *Management and Network Orchestration* (MANO) system to allow multiple instances of the system operating on top of the set of services provided by the MANO instance below.

Figure 2-12 shows the layered recursive architecture. In the lower layer, the owner of the physical resources (MNO) instantiates its MANO. On top of the physical infrastructure, different tenants request the MNO to allocate virtual infrastructures composed of a network subset with virtual nodes and links (i.e., a slice) through a Multi-tenancy Application (MTA), which orchestrates the assignment of the available resources. The MTA requests to the Virtual Infrastructure Manager (VIM) for the creation of a virtual topology according to tenant's demand.

Each tenant signs a Service Level Agreement (SLA) with the MNO, now the provider must take care of managing the available resources to meet the individual tenant requirements. The management on top of the virtual infrastructure is done through an API offered by the MTA with some defined operations and policies. In a recursive and hierarchical manner each tenant can operate its virtual infrastructure as the MNO operates on the physical, allocating and reselling part of the resources to other MVNOs in a transparent way to the MNO.

Figure 2-12 shows this practice between Tenant#1 and Tenant#2, the infrastructure of MVNO #2 operates over the virtual network offered by the MVNO #1 which operates on top of the MNO infrastructure (the physical one). In case of deploying an OTT tenant over an infrastructure, the

MTA is required to provide the tenant identification while the mapping of the virtual to physical resource will be done by the VIM through the NFVO.

This architecture also supports deployment of the network services of the OTT tenants on top, by extending the functions of the NFVOs such as Open Source Mano and OpenBaton to have the tenant separation and identify the mapping of a tenant to a network service consisting of a set of VNFs connected in a forwarding graph. Note that this recursive architecture also follows the recursion principles of the ONF architecture [2-23].

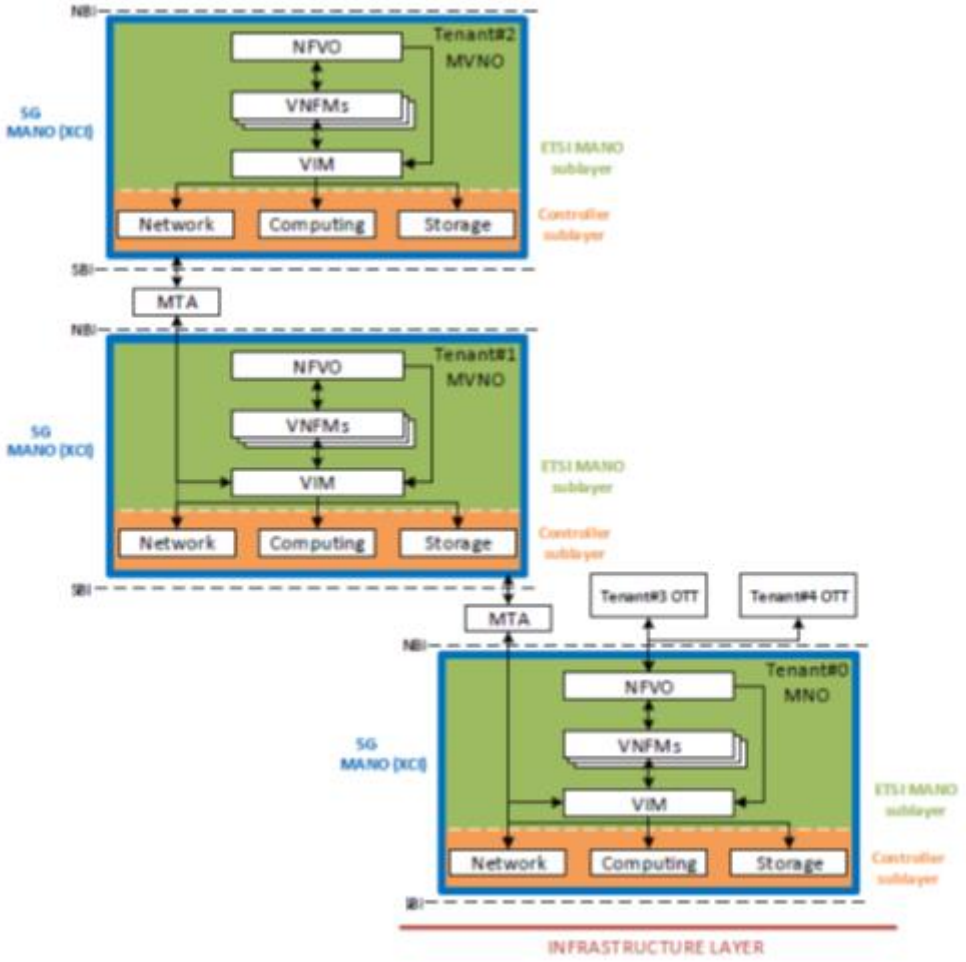


Figure 2-12: Recursive architecture

The business relationships between MSP, InPs, and tenant have impact on the architecture. Depending on the situation, the entities of the architecture belong to distinct administrative and technical domains as depicted in Figure 2-13. Thus, it is important that the cross domain functional interfaces are carefully designed for possible standardization. Moreover, security aspects at these interfaces must be covered.

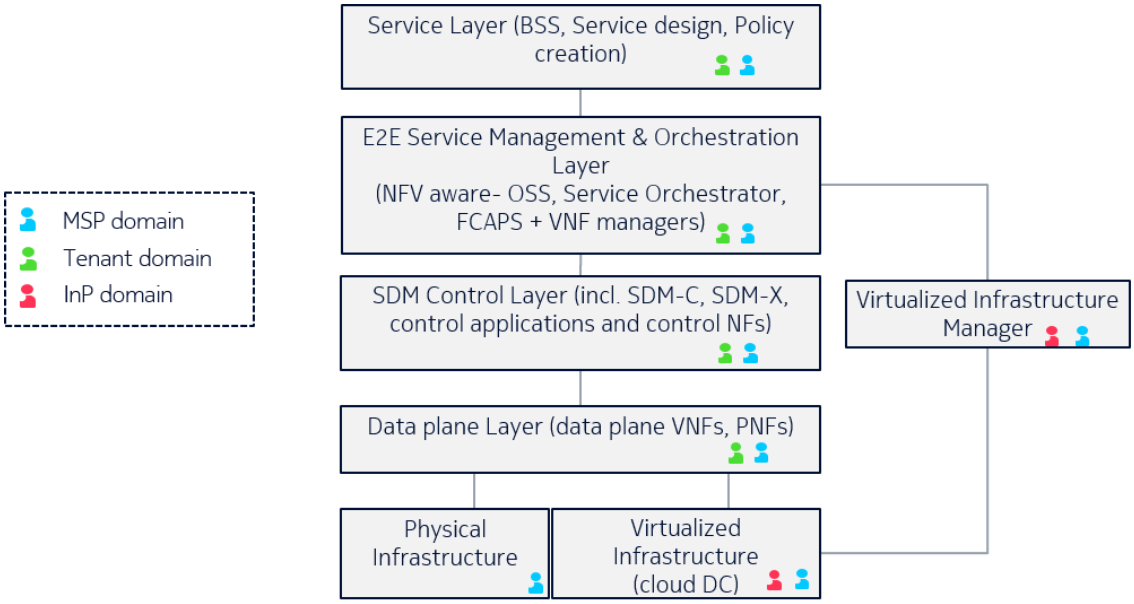


Figure 2-13: Possible domain ownerships of main functional blocks using the layers introduced before [2-21]

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

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

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

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

2.2.6 Possible Extensions

The considerations so far have been fairly specific to concrete incarnations of the slicing concept, for example, as it applies to mobile networks. The idea of slicing is, however, powerful enough to admit a broader consideration of further options, both technological and business-model ones. This section briefly touches upon such options.

The first notion to contemplate is that of recursive slicing (also often called "*slice-in-slice*"). In principle, it is a straightforward idea: slicing separates and isolates resources of a physical infrastructure (multiplexing the actual resources between slices, e.g., by time-division multiplexing a computational resource, aka multitasking), with each slicing offering an interface that is, ideally, indistinguishable from the actual underlying physical infrastructure. This allows to install any kind of services inside a slice (e.g., network functions, SDN controllers, cloud-like services, etc.). If the analogy between the physical infrastructure's interfaces and the slices' interfaces are close enough, nothing prevents one from applying the slicing a slice again. This slice in a slice needs to install its own functions (e.g., network functions, SDN controller) etc. and can operate inside the resource envelope of its encompassing slice. This idea is illustrated in the previous Figure 2-12.

This form of recursive slicing has, by now, been considered in various forms by multiple projects; it is conducive to a number of business models (e.g., reselling, MVNO). It is, however, a limited model in that the sub-slices only obtain resources from their parent slice but no functionality – in fact, they are deliberately isolated from their parents in that sense. Practically speaking, that would mean that any functionality beyond mere resources that is available in a parent slice is invisible and inaccessible to a sub-slice. For example, a particular mobility management mechanism used by the parent would *not* be available to the sub-slice, but the sub-slice would have to deploy and use its own mobility management mechanism. This raises the interesting question of feature interaction, yet with proper resource and functional isolation between slices, this should be manageable.

A more advanced concept of recursive slicing could allow to explicitly expose functionality of the parent towards the sub-slice. A typical example could be radio resource management: it could make a lot of sense to keep to the same radio resource management regime within a slice and (all of) its sub-slices, but there might be good reason to deviate from other behavioural aspects inside the core network. Such a concept allows sub-slices to *inherit* functions (with deliberate analogy to the object-oriented programming paradigm) from parent slices (with proper accounting for consumed resources) and eases the deployment and configuration burden on the operator of a sub-slice. This could turn into a very attractive business model (easier market introduction, higher reselling price as a value-added sub-slice is sold, etc.). However, up to now, this concept has not investigated in any detail so far and provides opportunities for further research.

A related, yet different concept from recursive slicing is slice composition (also called "*slice-cum-slice*"): how to build a slice out of individual slices? This makes obvious sense if the slices are functionally equivalent (i.e., an identical or at least compatible architecture is deployed inside them) yet have restricted scope (limited geographically, topologically, by user base, or some other means). Then, a composed slice naturally can be created by merging these scopes. It is also relatively easy to imagine slice-cum-slice if the scopes are limited to different functional parts of a network. For example, one slice could provide specific radio access functionality, another one has suitable user management functions. Out of them, a new slice can be created with the joint set of functions. Inside this new slice, additional functions can then be added to commoditize the new slice further. This model also allows more sophisticated approaches where a slice can be combined with multiple other slices, creating separate new slices in each case; the first slice has to be "multi-composable" (as an extension of multi-tenancy) to admit this model. Various concepts for accounting for the shared resource usage are conceivable here and will have to be defined in future. In general, the slice-cum-slice model is currently still in its infancy and requires considerable more research.

2.3 Programmability & Softwarization

In the recent moves of network soft re-architecture, significant attention is currently given to 5G networks are conceived as extremely flexible and highly programmable with native softwarisation E2E connect-and-compute infrastructures that are application- and service-aware, as well as time- location- and context-aware. They represent an evolution of native flexibility and programmability conversion in all radio and non-radio 5G network segments. They will in some

cases result in the decomposition of current monolithic network entities into network functions. These functions will constitute the unit of networking for next generation systems, and should be able to be composed in an "on-demand", "on-the-fly" basis.

Network Softwarization and Enablers which are aiming at providing networking and communication functionality through programmable software that is separable from hardware and that would not be restricted to run only as part of a firmware image. Important examples of such technology include Software-Defined Networking, Network Function Virtualization, Service Function Chaining, Network Slicing and Network Virtualization and transition from today's "network of entities" towards a "network of functions". Indeed, this "network of (virtual) functions", resulting, in some cases, in the decomposition of current monolithic network entities will constitute the unit of networking for next generation systems. These functions should be able to be composed on an "on-demand", "on-the-fly" basis. In fact, a research challenge consists in designing solutions which identify a set of elementary functions or blocks to compose network functions, while today they implemented as monolithic. More generally, it includes any networking and communications technology that features open programmable interfaces accessible to third parties, extensibility through software, software development kits, and separation of data forwarding, control, and management planes.

In addition Programmability in Networks and Enablers allow the functionality of some of their network elements to be dynamically changed. These networks aim to provide easy introduction of new network services by adding dynamic programmability to network devices such as routers, switches, and applications servers. Network Programmability empowers the fast, flexible, and dynamic deployment of new network and management services executed as groups of virtual machines in the data plane, control plane, management plane and service plane in all segments of the network. Dynamic programming refers to executable code that is injected into the execution environments of network elements in order to create the new functionality at run time. The basic approach is to enable trusted third parties (end users, operators, and service providers) to inject application-specific services (in the form of code) into the network. Applications may utilize this network support in terms of optimized network resources and, as such, they are becoming network aware. The behaviour of network resources can then be customized and changed through a standardized programming interface for network control, management and servicing functionality.

The following Figure 2-14 summarizes the 5G softwarisation and programmability ***abstractions*** in the form of a high-level logical design at the network element specific functionality.

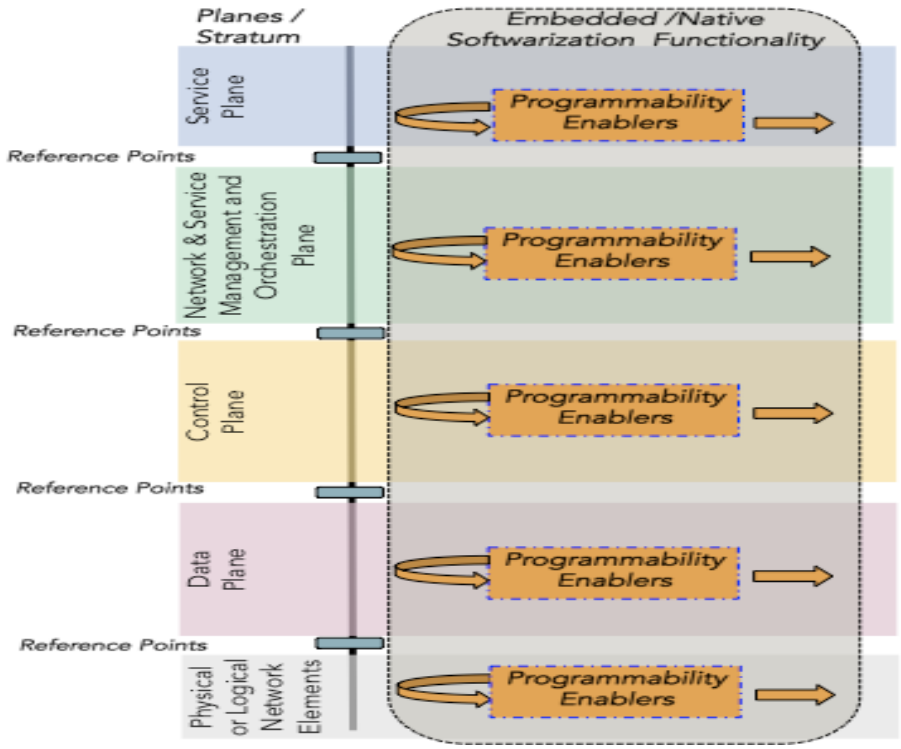


Figure 2-14: 5G Network Element Softwarisation & Programmability Viewpoint

The following Figure 2-15 summarizes the 5G softwarisation and programmability abstractions in the form of a high-level logical design at the network level.

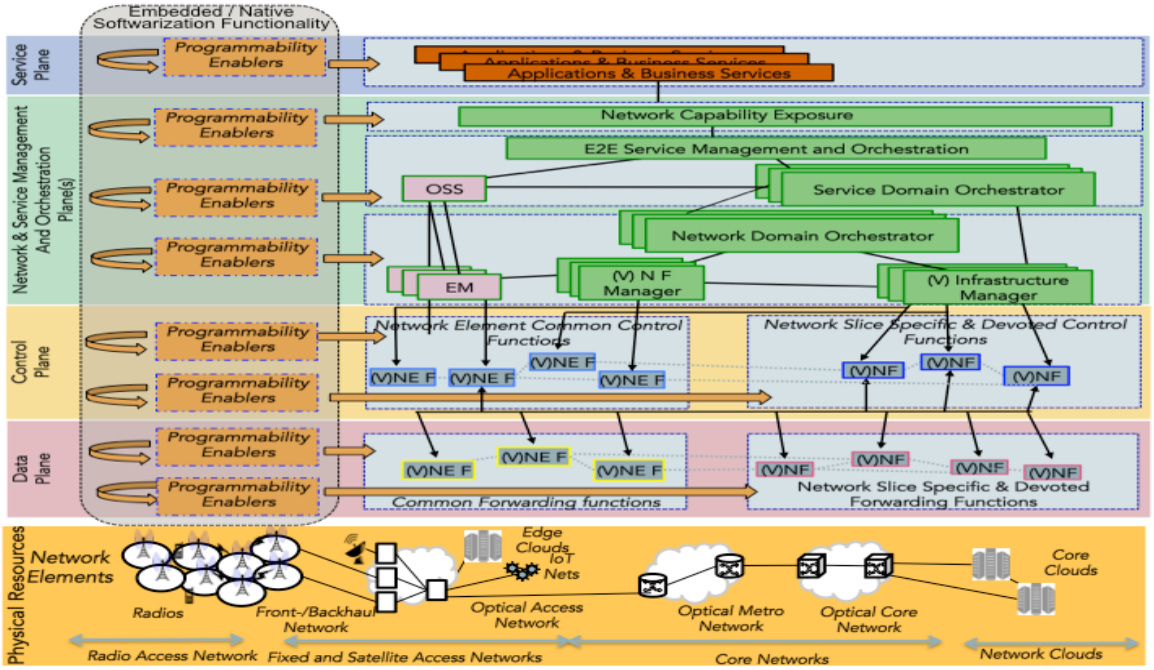


Figure 2-15: 5G Logical and Physical Network Softwarisation and Programmability Viewpoint

2.4 Management and Orchestration

Driven by the need to serve multiple customers with differing requirements operator infrastructures in 5G will be programmable to dynamically create typically virtualized network architectures over the same physical infrastructure. In 5G terminology this concept is referred to as network slicing. This change from a static architecture in LTE to on-demand dynamically provisioned architectures in 5G adds complexities in the design of the management plane. As mentioned in Section 2.3 vendor equipment now is expected to be programmable to support this scenario. This exposure of programmable interfaces for creation of those network slices contrasts with the traditional vendor approach of proprietary management interfaces for management of the network infrastructure.

To be able to uniformly manage the multiple slices that an operator is expected to host, ***programming interfaces to the virtualization infrastructures*** need to be exposed. The infrastructure needs to be capable of ***hosting multiple tenants*** and to be able to distinguish between the various types of ***flexibility and control required in the virtualized cloud resources.*** Chapter 5 covers these advances in the underlying infrastructure in detail.

Once the underlying infrastructure programmable, there need to exist components to use that programmability ***to manage and orchestrate those resources within a domain, across domains, as well as across providers.*** Providers in the new business ecosystem are not only limited to operators but could be specific technology providers, e.g. factory floors wanting to rent out their resources.

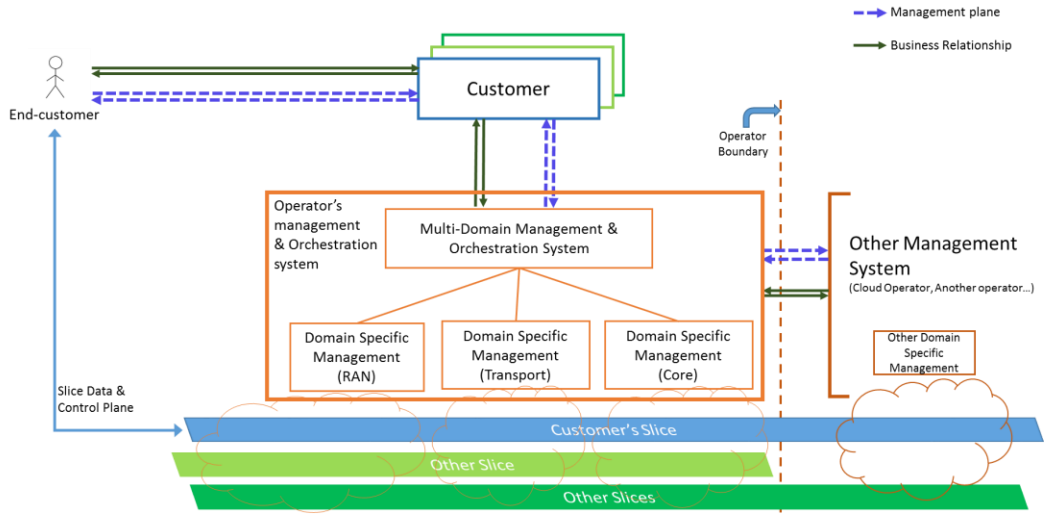


Figure 2-16: The perceived 5G Business and Management Ecosystem

Figure 2-16 shows the currently proposed architecture of the orchestration system as an extension of the ETSI NFV MANO architecture. The architecture is recursively stackable and consists at the lowest layer of domain-specific management and orchestration entities. The next higher level in the recursion combines these multiple domain-specific management and orchestration entities to create a multi-domain orchestration and management entity to coordinate end-to-end service and slice creation. The abstraction layer at each level exposes a generic set of interfaces while hiding specific technical and implementation details. The business and management interfaces are then exposed both northwards to customers as well as in the east-west direction to other operators.

Leaving the business models aside there is considerable work ongoing within the 5GPPP to realize the view presented in Figure 2-16. Over interface 1 between the customer and the operator there is the need of describing the network slices and the services they host accurately. Work done in 5GPPP has reviewed as well as contributed to the development of not only such a service description such as in [2-17], it has also developed newer mechanisms to compose those services

extending the principles of service function chaining. Furthermore, simplifying the management of services has been investigated in terms of verification mechanisms for the deployed service where the key question relates to the safety of instantiating of the network functions is adressed as well as applying cognitive machine learning techniques for predicting service characteristics.
Once the operator receives the service request the operators' management and orchestration system must then parse the description and decide how to realize this service both over interface 2 and 3 in Figure 2-16. In particular, it must place the NFs composing the service or the slice and the assign appropriate resources or their execution. This is referred to as **placement** or **embedding** algorithms. Within 5GPPP considerable work has been done on placement algorithms including ILPs, multi-provider embedding, greedy heuristics with backtracking, applying multiple algorithms simultaneously as well as considerations of future scale-in or scale-out operations for placing the virtual network architectures. Once the decision has been made on how to place the service the underlying programmable interface to the architecture are used to provision the service. This can also happen across operators. The challenges in multi-operator deployment stem mostly from the fact that one operator may only want to show a limited set of functionalities, resources and capabilities to another operator. In here 5GPPP considers the ***cooperative ecosystem*** for service deployment wherein the operators are enabled by abstraction of their internal capabilities to compete as well as cooperate with each other to realize a service. An additional challenge in inter-operator interface exposure is security over the exposed interfaces. Once the service is provisioned it enters service assurance phase wherein the service should be monitored for SLA compliance. This is must be done both within a single domain as well as across domains and providers.
Finally, 5GPPP work has looked into automated management for simplifying the management overload that comes with the introduction of network slicing. This is done by the introduction of long term machine learning mechanisms which act as are intelligent agents to perceive the network state and its external environment, and use these insights to assist the network management of the system. The agents analyze on the information collected during the monitoring phase and (re)configure policies to enact a desired alteration in the network infrastructure. Consequently, managed network resources, such as the VNFs forming the network slice or service are realigned by the orchestration based on those policy changes.

2.5 5G Security Architecture

As stated in the introduction to this chapter, 5G is being developed with new architectural concepts and capabilities to enable new business models and to provide enhanced applications and services to network subscribers. In order to ensure that 5G fulfils its promise, all security matters accompanying the 5G architecture need to be addressed. The security architecture presented here has been developed in the 5G-ENSURE project [2-24] and can be seen as an evolution based on the existing security architectures for 3G and 4G [2-25][2-26]. The basic concepts, e.g. domains and strata, remain but have been adapted and extended to fit and cover the 5G environment.

In [2-27] and the 5G-PPP (Phase I) Security Landscape white paper [2-28] the need for a new security architecture for 5G is discussed and motivated and an initial draft is presented. The work covered here, backed by the 5G-PPP WG on Security, leveraged on what has been described in both of these documents.

The most important architectural aspects missing in the earlier 3G/4G security architectures are those of softwarization, virtualization, trust models covering all players in the 5G ecosystem, multi-domain/multi-tenant (also slice concept) management and orchestration and new mission critical cyber threats.

In Section 2.5.1, the security characteristics of 5G are described and requirements on the security architecture are discussed. Then the security architecture is presented in Section 2.5.2. In section

2.5.3 some observations on the security architecture are discussed and in particular its relation to the overall 5G architecture.

2.5.1 Security Characteristics of 5G and new requirements

In the design of the new security architecture special care has been taken to ensure that the security architecture can embrace all new technologies and concepts used in 5G which are significant evolution steps from 4G.

As stated above, 5G will rely heavily on softwarization and virtualization of the network to enhance flexibility and scalability. One important development to achieve this is to refine the network slicing technologies and mobile edge computing (MEC) to more dynamically offer various kinds of services to different tenants. Virtualisation is the underlying enabling technology for this. Another development of equal importance is the large scale introduction of SDN allowing flexible and dynamic (re-)defining of the networking infrastructure.

The incorporation of many new target use cases [2-29][2-30][2-31][2-32][2-33][2-34] in 5G brings new business models with many new actors taking part in the provisioning of services. This means that new, multi-party trust relations will be present and have to be modelled and handled from a security perspective.

In an environment where many different actors independently manage resources that they own or lease, all aspects of management become critical from a security point of view. In particular, 5G has to handle multiparty management of security (e.g. provisioning of keys and credentials) and security management (e.g. ensuring that secure services are available and runs securely). Orchestration of virtualized environments, services and SDN's will also require secure management.

5G networks will be more complex and dynamic than earlier generations of mobile networks as e.g. new (virtualized) network nodes and slices can be added to and removed from the network at any time. To identify and model attack vectors in this dynamic environment and to be able to offer strong network protection, security control points have to be defined based one establish boundaries between different actors' network functions and slices and their interfaces.

With the increasing number of target use cases, the new multi-party trust relations and the new technologies employed, 5G will comprise an ever increasing number of security and non-security protocols and network functions. To identify and keep track of threats and attack vectors, and required protection mechanisms and their coverage, 5G networks have to be modelled in a structured way together with the security controls that need to be deployed to offer the necessary trust and confidence. Below we describe the concepts we use for this purpose.

2.5.2 The security architecture

2.5.2.1 Underlying concepts

The basic concepts in our security architecture are domains, strata, security realms, and security control classes. The definitions of these concepts are:

- A **Domain** is a grouping of network entities according to physical or logical aspects that are relevant for a 5G network. This concept is leveraged from TS 23.101[2-26].
- A **Stratum** is a grouping of protocols, data, and functions related to one aspect of the services provided by one or several domains. This concept is leveraged from TS 23.101 [2-26],
- A S**ecurity Control Class (SCC)** is a new concept introduced that refers to a collection of security functions (including safeguards and countermeasures) to avoid, detect, deter, counteract, or minimize security risks to 5G networks, in particular, risks to a network's physical and logical infrastructure, its services, the user equipment, signalling, and data.

- A **Security Realm (SR)** captures security needs of one or more strata or domains. As such it is similar to the Security Feature group concept as defined in TS 33.401[2-25]

The framework of the security architecture is flexible as it can be extended with definitions of new domains, strata, etc. This makes it possible to adapt the framework to future network solutions with new functionality and services.

Domains

The basis for the security architecture is the use of domains. The Figure 2-17 illustrates an instance of a 5G network and depicts the domains defined so far.

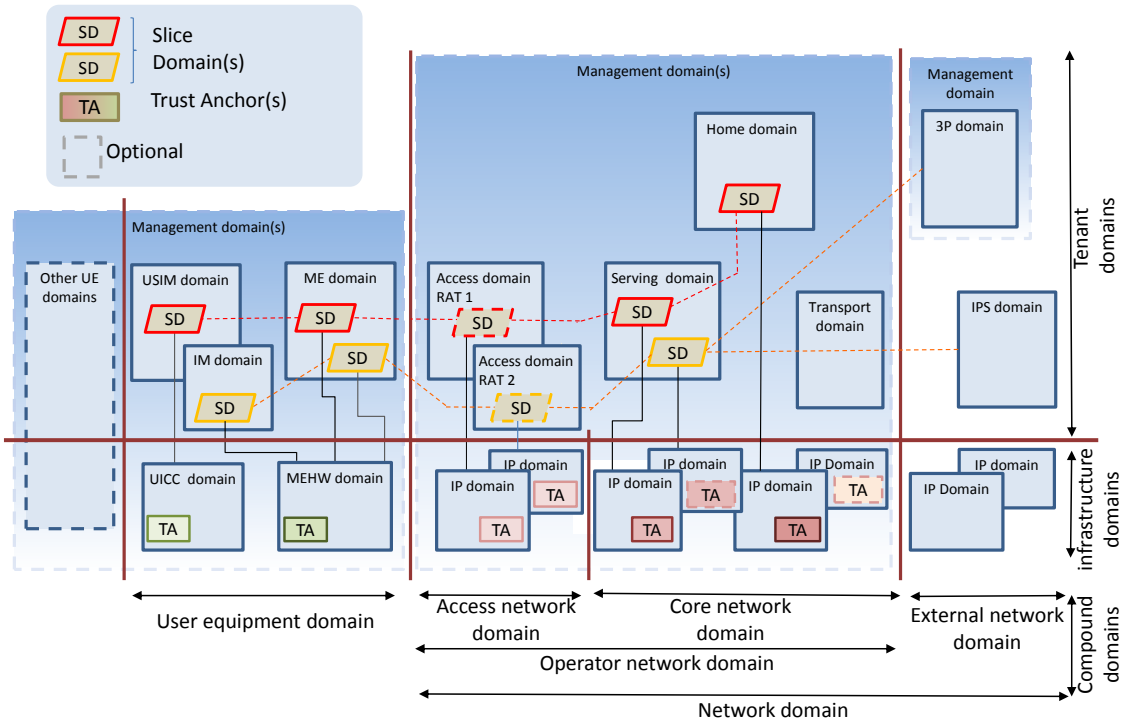


Figure 2-17 Domains in the 5G security architecture. NB: Lines from slice domains down to Infrastructure Domains denote that VNFs allocated to a slice make use of certain physical resources.

The domain concept is a cornerstone in the security architecture as it enables the definition of different types of domains used to represent a 5G network's different functionalities, services and actors. The defined domains may occur in multiple instances, and belong to different actors taking on different roles and responsibilities in the network, which provides flexibility for the modelling of different 5G network configurations and describing their inherent multi-party trust aspects. By observing interdependencies and required interactions between domains it becomes a relatively straightforward task to analyse and model their trust relations and their need for different security controls.

We use three different types of domains. First there is the **Infrastructure Domains** focussing on the relevant physical network aspects, i.e. they contain the "hardware" in the network. Then there are **Tenant Domains** which are logical domains executing in infrastructure domains. By this division into infrastructure and tenant domains, it is easy to map and handle virtualized environments onto the architecture as these types of domains give a clear division between the physical platform offering an execution environment and the logical functions and services in the tenant domain.

To capture higher order groupings of entities and/or functionality we have defined a third type of domains, namely **Compound Domains.** Such domains consist of a collection of other domains, grouped together according to some 5G relevant aspects, e.g. ownership, joint administration or the like. With this concept we can map 3G/4G defined domains onto our security architecture in a simple way.

Of particular importance are **Slice Domains**. They are compound domains used to capture network slicing aspects. A slice can cover only some parts of the network, e.g. parts of the Core Network domain, but are in general defined end-to-end. In this way slicing is explicitly handled. The use of slice domains also highlights the trust issues appearing between actors controlling a domain and other actors controlling concurrently operating slices in that domain. The requirement on strict isolation between domains and slices belonging to different actors is also made clear. We note that slicing may be implemented without relying on a virtualized system, but in most 5G systems it is.

The domain figure depicted above also shows so called trust anchors in the infrastructure domains. These trust anchors are used to capture trust issues appearing in virtualized systems, e.g. how to get assurance of tenant domain integrity and that a tenant domain executes on a designated and trusted infrastructure. The trust anchors can also be used to verify infrastructure domains' integrity and to bind tenant domains to infrastructure domains.

We only have three types of infrastructure domains. They are

1. **UICC Domains** containing the conventional tamper-resistant module offering protected storage and processing of long-term subscriber credentials and other security critical information.
2. **Mobile Equipment Hardware (MEHW) Domains** containing the hardware support for the Mobile Equipment (ME). The MEHW domain may include trusted execution environments (TEE) supporting e.g. other forms of credentials such as certificates.
3. **Infrastructure Provider (IP) Domains** containing the hardware platforms for the compute, storage, and networking resources required by both the network/telecom functionality and the access (radio) specific hardware.

At present, there are 10 tenant domains defined. They are

1. **Mobile Equipment (ME) Domains** containing the logical functionality required for using access to network services, for the operation of access protocols by users and for user applications.
2. **USIM Domains** containing the logical functionality for USIM operation together with other hosted security services (it is analogous to the USIM domain of TS23.101 but only contains the logical functionality).
3. **Identity Management (IM) Domains** containing functionality to support alternatives to USIM-based authentication, i.e. for industry automation use cases (the IM Domain may contain for example public key certificates. The IM domain preferable obtains security support from a UICC or from a TEE in the ME HW as discussed above).
4. **Access (A) Domains** containing the logical functionality which manages the resources of the access network and provides users with mechanisms to access the core network domain.
5. **Serving (S) Domains** containing the logical functionality which is local to the user's access point. It also routes calls and transports user data/information from source to destination. It has the ability to interact with the home domain to cater for user specific data/services and with the transit domain for non-user specific data/services purposes.
6. **Home (H) Domains** contains the logical functionality conducted at a permanent location regardless of the location of the user's access point. The USIM is related by subscription to the home network domain. The home network domain therefore contains at least permanently user specific data and is responsible for management of subscription information. It may also handle home specific services, potentially not offered by the serving network domain.

7. **Transit (T) Domains** containing the logical core network functionality in the communication path between the serving network domain and external remote parties.
8. **3rd Party (3P) Domains** containing functionality for use cases where a (semi-)trusted third party such as a factory/industry vertical provides its own authentication services for e.g. its M2M devices like industry robots and IoT-devices.
9. **Internet Protocol Service (IPS) Domains** representing operator-external IP networks such as the public Internet and/or various corporate networks. Such networks may be partially or fully non-trusted.
10. **Management Domains** containing the logical functionality required for management of specific aspects of a 5G network. Management domains may cover security management, management of security, traditional network management, orchestration of SDN and virtualized environments, and management of user equipment domains etc.

Finally, the compound domains defined are

1. Slice domains (described above).
2. **User Equipment (UE) Domains** defined by MEHW, ME, UICC, USIM and IM domains included, i.e. it consists of the equipment used by a user to access network services. The "Additional UE Domain" in Figure 2-17 is added to capture the so called direct-mode, UE-to-UE communication.
3. **Access Network (AN) Domain** defined by the A and IP domains included, i.e. it consists of the entities that manage the resources of the access network and provides the user with a mechanism to access the network. It may comprise of different types of accesses, e.g. both WLAN and 5G-radio accesses.
4. – 6. **Serving Network (SN), Home Network (HN) and Transit Network (TN) Domains** include the S, H, T and corresponding IP domains respectively and correspond to the same concepts in TS 23.101 [2-26].
7. **Core Network (CN) Domain** defined by the HN, SN, TN and IP domains included, i.e. it consists of the entities that provide support for the network features and telecommunication services. The support provided includes functionality such as user location information, control of network features and services, the transfer (switching and transmission) mechanisms for signalling and for user generated information.
8. **Operator Network (ON) Domain** defined by the AN and CN domains included, i.e. it consists of the physical nodes together with their various functions required to terminate the radio interface and to support the telecommunication services requirements of the users.
9. **External Network (EN) Domain** defined by the 3P, IPS and IP domains included and
10. **Network (N) Domain** defined by the ON and EN domains included

Strata

Figure 2-18 shows the strata of 5G security architecture. The definitions of the strata are analogous to the definitions given in TS 23.101 [2-26] except for the management stratum which is added in the 5G security architecture. The management stratum is graphically drawn to be behind and cover all other strata because the management stratum will perform management operations on network functions in all of the other strata. For instance, it will comprise protocols like OpenFlow for configuring network components. Obviously, there will also be dedicated protocols, data, and functions related to managing NFVs and network slices.

The strata provide a high-level view of protocols, data and functions that are related in the sense that they are exposed to a common threat environment and exhibit similar security requirements. The use of strata thus helps in structuring for which purpose and where different security controls are needed.

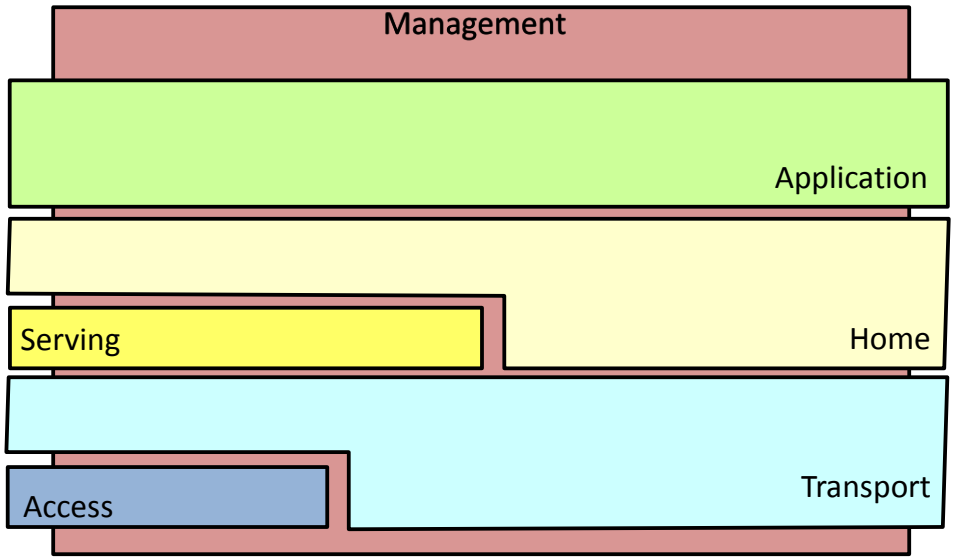


Figure 2-18: Strata of 5G Security Architecture

The **Application Stratum** represents the application process itself, provided to the end-user. It includes end-to-end protocols and functions which make use of services provided by the home, serving and transport strata and infrastructure to support services and/or value added services. End-to-end functions are applications which are consumed by users at the edge of/outside the overall network.
The **Home Stratum** contains the protocols and functions related to the handling and storage of subscription data and home network specific services. It also includes functions to allow domains other than the home network domain to act on behalf of the home network. Functions related to subscription data management, customer care, including billing and charging, mobility management and authentication are located in this stratum when end-users are at home network. When end-users are roaming, then serving network domain is allowed to do mobility management at serving network level.

The **Serving Stratum** consists of protocols and functions to route and forward data/information, user or network generated, from source to destination. The source and destination may be within the same or different networks. Functions related to telecommunication services are located in this stratum.

The **Transport Stratum** supports the transport of user data and network control signalling from other strata through the network. It includes consideration of the physical transmission, e.g., physical transmission format, error correction/recovery, data encryption, resource allocation, etc.

The **Access Stratum** is a sub-stratum of the Transport Stratum. It is located between the edge node of the serving network domain and the UE Domain. It provides services related to the transmission of data over the radio interface and the management of the radio interface.

The **Management Stratum** comprises aspects related to conventional network management (configuration, software upgrades, user account management, log collection/analysis, etc.) and, in particular, *security management* aspects (security monitoring audit, key and certificate management, etc.). In addition, aspects related to *management of virtualization* and service creation/composition (orchestration, network slice management, isolation and VM management, etc.) belong to this stratum.

Security control classes

The different security control classes can be found in Figure 2-19. The structure of the Security Control classes was inspired by the security dimensions found in ITU X.805 [2-35]. Several of the X.805 security dimensions were adopted, some with minor modifications, and then

complemented with a few new Security Control classes relevant for 5G networks. The exact functions and mechanisms to enforce a specific security control are left for consideration in the detailed design phase.

Security Control Class	Description
Identity & Access Management	A collection of security functions and mechanisms addressing access control (authorization), management of credentials and roles, etc.
Authentication	A collection of security functions and mechanisms serving to verify the validity of an attribute, e.g. a claimed identity.
Non-repudiation	A collection of security functions and mechanisms serving to protect against false denial of involvement in a particular action.
Confidentiality	A collection of security functions and mechanisms protecting data against unauthorized disclosure.
Integrity	A collection of security functions and mechanisms protecting data against unauthorized creation or modification.
Availability	A collection of security functions and mechanisms serving to ensure availability of resources, even in the presence of attacks. Disaster recovery solutions are included in this category.
Privacy	A collection of security functions and mechanisms serving to the right of an entity (normally a person), acting in its own behalf, to determine the degree to which it will interact and share it personal information with its environment.
Audit	A collection of security functions and mechanisms providing review and examination of a system's records and activities to determine the adequacy of system controls and detect breaches in system security services and controls. The necessary data collection to enable audit (e.g. logging) is also included.
Trust & Assurance	A collection of security functions and mechanisms serving to convey information about the trustworthiness of a system. For a trusted party such information constitutes a claim which may or may not persuade them to trust the system, while a trustee would see such information as evidence of the security level achieved.
Compliance	A collection of security functions and mechanisms provided to allow an entity or system to fulfil contractual or legal obligations.

Figure 2-19: Table of Security Control Classes (SCC)

Security Realms

The domains and slices in the security architecture provide boundaries between different network functions and services and the strata provide information on required security needs for domain

interaction and communication. A joint analysis of domains and strata will thus enable identification of required security control points for groups of protocols.

The Access Network (AN) SR captures security needs of the access Network domain and access stratum as part of the transport stratum - in particular aspects related to users securely accessing 5G services over 3GPP (5G radio) and certain non-3GPP (e.g. WLAN) access technologies.

The **Application (App) SR** captures security needs of the application stratum. That is, end-user applications/services provided over the 5G network, either as operator provided services (from HN or SN Domain), or provided from External Network Domains (3P or IPS Domain with associated IP domains). Note that when the service is hosted by an External Network Domain, the services may not always be fully trusted by 5G network operators. Examples of applications/services include: VoIP, VoLTE, V2X, ProSe, HTTP-based services, etc.

The Management (Mgmt) SR captures security needs of the Management Stratum and Management Domains, including secure management (secure upgrades, secure orchestration etc.) and management of security (monitoring, key and access management, etc.). Thus, Management Security is either a concern related to communication between a Management Domain and some other (semi-)trusted Domain, or, related to security of the Management Domain itself.

The **User Equipment (UE) SR** captures security needs of user equipment (UE) domain comprising the ME, ME HW, UICC, USIM, and IM domains and other UE domains, e.g. visibility and configurability and security aspects related to communication between these domains.

The Network SR captures security needs of communication in core network domains and between the core network domains and external network domains - including aspects related to securely exchanging signalling and user data between nodes in the operator and external network domain.

The **Infrastructure and virtualization (I&V) SR** captures security needs of IP Domains, e.g. for attestation, secure slicing/isolation, and trust issues between tenant domains and tenant domains and infrastructure domains.

2.5.2.2 Security methods

The defined security control classes provide a structured way to prevent or answer to a risk identified regarding specific data, functions and services in a network. The defined security realms capture needs of one or more strata or domains and are there to group different network aspects with different but area specific security concerns. Bringing these two concepts together by analysing which security controls that are required in a given security realm will provide a detailed and structured view of the required security mechanisms to ensure that security requirements are fulfilled. The security realms should be subdivided with respect to functionality, domains, strata and end-points of protocols and for each such subdivision a security control mechanism should be selected. In this way it is possible get a detailed overview of the security mechanisms needed in a 5G network. Figure 2-20 illustrates how such a structured view of required security controls per security realms could be recorded.

Security Realms	Security Control Classes									
	Identity & Access Mgmt	Authentication	Non-repudiation	Confidentiality	Integrity	Availability	Privacy	Audit	Trust & Assurance	Compliance
Access Network										
Application										
Management										
UE										
Network										
Infrastructure and Virtualization										

Figure 2-20: Security Realms with Controls

2.5.3 Mapping to overall 5G architecture

In the following we present some considerations regarding the implementation and enforcement of key security aspects in the 5G architecture. We note that for a security architecture to be useful beyond a mere "abstract thought experiment", it must be reflected in the design of real operational networks. To this end, it must be possible to map the components in the 5G security architecture on the architectural model of 5G systems and in particular to the high level 5G architecture presented in this document.

The focus of the high-level architecture depicted in Figure 2-1 is to illustrate the softwarization of 5G systems allowing efficient management and orchestration procedures providing a flexible sliced service platform for different types of service verticals running on a common infrastructure. We see two major features; the first is softwarization and the second is slicing and they are straight forward to map onto the security architecture. The softwarization functionality is in essence management functionality (orchestration, slice control, etc.) and should as such be mapped onto management domains and the management stratum, i.e. the management security realm. Slicing relies to a large extent on virtualization and use of shared infrastructure services for compute and storage. The security needs here would then be covered by the virtualization and infrastructure security realm.

Architecture enforcement of security objectives (incl. privacy) should be based on implementation and use of mandatory security controls. The required security controls must be implemented in all relevant domains, i.e. in all relevant 5G entities, networks and network functions and platforms on top of which the controls are implemented. Their implementation should provide strong assurance that the security controls cannot be circumvented. The security

controls should also ensure that 5G entities, networks and network functions are legitimate, can be authenticated and integrity verified and additionally provide privacy and confidentiality of data and users. Furthermore, the architecture enforcement should define how infrastructure as well as tenant domains are securely deployed, integrity verified, protected from outside threats and from threats coming from other domains.

In general, domains should be isolated and only have well defined entry points where controls can be implemented so that only legitimate traffic and signalling can take place. However, as domains in many cases are virtualized the isolation properties and controlled entry points must be enforced by logical means. This means that infrastructure domains used for network function virtualisation must provide strong isolation between tenants. Similarly, slice aware tenant domains must provide strong isolation of and between slices. Furthermore, it is essential that tenant domains are bound to the infrastructure domain on which they are deployed. The required form of binding may vary depending on the characteristics of the tenant domain as well as the infrastructure domain employed. Typically mutual authentication between a deployed tenant domain and the underlying infrastructure domain would be a relevant requirement. As is clear from the above, most of the security requirements for support of virtualisation, slicing, and mobile edge computing have bearings on the Infrastructure provider domain. The domain must fulfil a number of new security features with the most prominent ones being to provide proof of platform integrity, isolation between slices and control of services deployment, execution and migration.

The 5G security architecture will certainly reuse components of the existing 4G architecture when appropriate. Examples of such solutions are the 4G security features developed to cope with threats to radio base stations in physically exposed locations (e.g. when the AN Domain is EUTRA) and tampering threats to user credentials in devices (the USIM Domain is protected by the UICC). Other areas will as discussed above exhibit new aspects of security e.g. less emphasis on protection at physical and logical domain borders and more on defence-in-depth, the new need for “roots of trust” in virtualized settings to ensure legitimate use of resources as well as authenticated points of deployment.

2.6 References

[2-1] https://5g-ppp.eu/white-papers/
[2-2] https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf
[2-3] http://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf
[2-4] https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-July-2016.pdf
[2-5] 5G-PPP, Living Document on 5G PPP use cases and performance evaluation models, https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf
[2-6] GENI Key Concepts - Global Environment for Network Innovations (GENI) http://groups.geni.net/geni/wiki/GENIConcepts.
[2-7] Galis, A. et all - "Management and Service-aware Networking Architectures (MANA) for Future Internet" - Invited paper IEEE 2009 Fourth International Conference on Communications and Networking in China (ChinaCom09) 26-28 August 2009, Xi'an, China, www.chinacom.org/2009/index.html.
[2-8] Hedmar, P., Mschner, K., et all - NGMN Alliance document "Description of Network Slicing Concept", January 2016; www.ngmn.org/uploads/ media/160113_Network_Slicing_v1_0.pdf
[2-9] Study on Architecture for Next Generation System - September 2016; www.3gpp.org/ftp/tsg_sa/WG2_Arch/Latest_SA2_Specs/Latest_draft_S2_Specs.
[2-10] Paul, M., Schallen, S., Betts, M., Hood, D., Shirazipor, M., Lopes, D., Kaippallimalit, J., Open Network Fundation document "Applying SDN Architecture to 5G Slicing", April 2016; www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/Applying_SDN_Architecture_to_5G_Slicing_TR-526.pdf

[2-11] Technical Report Application of network softwarization to IMT-2020, ITU-T FG IMT2020, December 2016, http://www.itu.int/en/ITU-T/focusgroups/ imt-2020/Pages/default.aspx
[2-12] Galis, A., et all "network Slicing Problems" - https://tools.ietf.org/html/draft-galis-netslices-revised-problem-statement-03; "Network Slicing Architecture" - https://tools.ietf.org/html/draft-geng-netslices-architecture-01
[2-13] 5G Infrastructure Association – Vertical Sectors White Papers - https://5g-ppp.eu/white-papers/
[2-14] Next Generation Mobile Networks (NGMN) Alliance, "5G White Paper", Feb. 2015, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf.
[2-15] 5G NORMA D3.2 Network Architecture – Intermediate Report, https://5gnorma.5g-ppp.eu/dissemination/public-deliverables/
[2-16] ETSI, Network Functions Virtualisation, Network Functions Virtualisation (NFV); Management and Orchestration, December 2014. [Online]. Available: http://www.etsi.org/deliver/etsi gs/NFVMAN/001 099/001/01.01.01 60/gs nfv-man001v010101p.pdf
[2-17] OASIS, Organization for the Advancement of Structured Information Standards, TOSCA Simple Profile for Network Functions Virtualization (NFV) Version 1.0, March 2016. [Online]. Available: http://docs.oasisopen.org/tosca/tosca-nfv/v1.0/tosca-nfv-v1.0.pdf
[2-18] 5G NORMA D4.1 RAN architecture components – preliminary concepts, https://5gnorma.5g-ppp.eu/dissemination/public-deliverables/
[2-19] METIS-II, "D6.2 Final asynchronous control functions and overall control plane design", April 2017.
[2-20] METIS-II, Deliverable D5.2, "Final Considerations on Synchronous Control Functions and Agile Resource Management for 5G", March 2017.
[2-21] 5G NORMA D3.2 5G NORMA network architecture – Intermediate report, https://5gnorma.5g-ppp.eu/dissemination/public-deliverables/
[2-22] 5G NORMA D5.2 Definition and specification of connectivity and QoE/QoS management mechanism – final report, https://5gnorma.5g-ppp.eu/dissemination/public-deliverables/
[2-23] Open Networking Foundation (ONF), "SDN Architecture", https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR_SDN_ARCH_1.0_06062014.pdf
[2-24] 5G-ENSURE, http://www.5gensure.eu/
[2-25] 3G-PPP TS 33.401, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2296.
[2-26] 3GPP TS 23.101, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=782.
[2-27] 5G-ENSURE D2.4 Security Architecture Draft, http://www.5gensure.eu/sites/default/files/Deliverables/5G-ENSURE_D2.4-SecurityArchitectureDraft.pdf.
[2-28] https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP_White-Paper_Phase-1-Security-Landscape_June-2017.pdf.
[2-29] https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf.
[2-30] https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Factories-of-the-Future-Vertical-Sector.pdf.
[2-31] https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White_Paper-on-Energy-Vertical-Sector.pdf.
[2-32] https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf.

[2-33] http://5g-ppp.eu/wp-content/uploads/2015/06/Specialized-Services-Network-Management-and-5G.pdf.
[2-34] https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-Media-Entertainment-Vertical-Sector.pdf.
[2-35] X.805, https://www.itu.int/rec/T-REC-X.805-200310-I/en

3 Radio Access

To address 5G challenges a combination and integration of new radio access technologies with existing technologies is anticipated. New types of frequency bands like micro- and millimeter waves are expected to be used. These will make small cells even smaller and denser than in current setups. Also, the adoption of massive MIMO systems will necessitate more efficient interference management schemes, e.g., by coordinated multi-point (CoMP) techniques. However, interference between heterogeneous macro and small cells is only an exemplary aspect that has to be coordinated tightly. Another example is the co-deployment of tightly integrated LTE-A and novel 5G radio via e.g., multi-connectivity.

While these goals could be reached in conventional D-RANs, C-RAN concepts have been also proposed. C-RAN promises significant CAPEX and OPEX advantages for operators by centralizing hardware and by significantly reducing energy consumption, but a key feature of 5G RAN will be the flexibility to adopt intermediate solutions suitable for the scenario at hand. The different deployment options are expected to set limits on what RAN technologies can offer in real deployments. The ability to resolve these dependencies in a real system, by using different deployment scenarios over a single infrastructure, is a core advance of 5G over previous architectures.

In this section of the white paper, we cover key architectural aspects of 5G RAN, including RAN softwarization and programmability concepts, control/user plane split options, protocol stack integration options, LTE/5G interworking, as well as centralized and distributed radio resource management approaches. We then present a selection of radio access functions that are enhanced or completely new in 5G. Table 3-1 presents a summary of those sample functions and sub-functions covered in this white paper, along with a brief description of each of them.

Table 3-1: Key radio access functions in 5G

Enhanced/new network access functions	Brief description
Multi-connectivity • LTE/5G tight integration	Multi-connectivity enabled at different network layers (micro/macro), spectrum (sub-6 GHz/mm-wave), user plane (MAC/RLC/PDCP), and technologies (WiFi/LTE/5G). Particularly relevant is the tight RAN interworking of LTE and 5G.
Initial access • Low-frequency assisted	Set of CP functions across multiple layers of the RAN protocol stack and, at some extent, the CN / RAN interface. Non-standalone deployments can benefit from low frequency RAT assistance.
Dynamic traffic steering	Unlike LTE, traffic steering may not be performed in RRC level but rather in a much more agile way and on a faster time scale on lower protocol stack layers.
RAN moderation	Energy-efficient network operation can be attained by moderation of the RAN access nodes, i.e., the optimal active-mode operation, with the help of QoS and channel quality awareness.

Mobility management • Mm-wave cell clustering	The 5G mobility framework consists of several new methods which need to cover use cases with uses in different RRC states. Furthermore, techniques such as cell clustering need to be employed to support mobility in mm-wave enabled RANs
Self-backhauling	Set of functions to provide technology- and topology-dependent coverage extension and capacity expansion utilizing same frequency band for both backhaul and access links,

3.1 SW controlled architecture definition

Software-defined mobile network control (SDMC) extends the SDN paradigm beyond mere packet forwarding to put the majority of mobile network functions under a centralized control. In addition, it provides an interface for full network programmability instead of configuration as it is the case in currently deployed mobile networks. Decomposition into function blocks enables sharing of network functions among slices for reuse and consistency among slices, or where common resources must be shared. A slice may be partly composed of a set of common function blocks to be shared across slices and a set of dedicated function blocks that implement customized and optimized functionality of a slice.

Accordingly, the control plane is reworked to add network programmability and RAN slicing: Most control functions run as SDMC-enabled control applications on top of a controller for slice-dedicated and another coordinating controller for shared functionality. Only time critical and frequent radio resource control (RRC) and management (RRM) functionalities are kept as "legacy" distributed network functions.

The 5G control and user plane functions are based on the current 3GPP LTE to equally support LTE and novel 5G RATs. For 5G, the data plane is modified and extended specifically at MAC and PDCP layer for integration of multi-tenancy, multi-service, multi-connectivity and multi-RAT support.

Future 5G heterogeneous radio access networks need a programmable control and coordination that offers fine-grain, real-time control without sacrificing scalability. The programmable control and coordination is driven by a key characteristic, namely abstraction. The control and coordination plane capitalises abstraction of low-layer resources in radio access networks. The abstracted resources should allow any RAN operations desired by the network slices while hiding the configuration details of RAN hardware. More specifically, abstracting RAN resources manages the complexity and greatly simplifies the implementation and deployment of advanced control and coordination functions in the RAN, while leveraging on a variety of physical layer technologies.

Furthermore, functional decomposition enables the function blocks to be (i) selected and (ii) placed according to its service needs and the concrete deployment scenario, i. e., the available execution environments such as distributed (edge) or centralized resources, including centralized RAN and distributed RAN as the two extremes of possible RAN functional splits, as well as beyond by additional integration of selected core functions to provide e.g. low latency and local breakout.

The optimal resource utilisation in heterogeneous mobile networks is a challenge. Solutions with distributed control are scalable and flexible, but often yield sub-optimal results far away from the expectation. Therefore, new RAN control architecture is needed to adopt a centralised solution,

SDN in RAN, which could achieve the global optimisation. Furthermore, applying SDN in 5G RAN enables the programmability. The benefit of global optimisation and programmability comes at the expense of scalability and latency. To address the scalability and latency issues, the hierarchy control structure is needed: the centralised controller for network-wide control and coordination, and the local controller for network functions requiring real-time operation.

The logically centralized controller, designed by the SDN principles, can provide network-wide control and coordination. For overcoming scalability issue in a large and dense RAN deployment, or for performance/reliability reasons, the logically centralised control and coordination entity can be implemented with distributed physical control instances sharing network information with each other. The distribution of abstraction shields higher layer from state dissemination and collection, making the distributed control problem a logically centralised one.

One challenge in creating such a software defined radio access network is the inherent delay between any centralised control entity and individual radio elements. For example, RAN operations require for real-time control when it comes in Resource Block (RB) scheduling with the time scale of 1ms. The controller operations for such a network function (RB scheduling) must be extremely fast and supporting hard real-time applications.

Under such limitation of the inherent latency, the local controller can offer real-time control. The local controller should be close to the physical radio elements so that it could adjust to rapidly varying wireless networks. The local controllers in the RAN do not coordinate with each other and therefore perform distributed control in the RAN.

By separating control functionalities between the centralized controller and the local controller, the centralized controller makes decisions that affect the logically centralised network states, while the local controller handles control decisions for latency-sensitive network functionalities in low layers without coordinating with other local controllers. Moreover, different network slices contain different network applications and configuration settings. Some application modules in network slices may be latency-sensitive. For such a slice, these modules are located in the local controller.

Based on the software defined control architecture, a set of generic RAN APIs are introduced to allow the control plane to interact with the data plane of the underlying network. These API calls can be invoked either by the centralized controller or the local controller through the control protocol. Table 3-2 provides a list of RAN-specific API call [FLEXRAN]. It can be seen that different type of network applications can be developed including (a) monitoring for a better decision making (e.g. adaptive video transcoding), (b) control and coordination to optimize network performance (e.g. joint scheduling, interference management, beamforming, and HO), and (c) programmability for a higher adaptability and flexibility (e.g. changing functional split, or even disabling/enabling network functions, slicing).

Table 3-2: RAN-specific API Calls

API	Target	Direction	Example	Applications
Configuration (synchronous)	eNB, UE, Slice	Centralized controller → Local controller	UL/DL cell bandwidth, Reconfigure DRB, RSRP/RSRQ/TA	Monitoring, Reconfiguration, SON
Statistic, Measurement, Metering (Asynchronous)	List of eNB, UE, Slice	Local controller → Centralized controller	CQI measurements, SINR measurements, UL/DL performance	Monitoring, Optimization, SON
Commands	Local controller	Centralized controller →	Scheduling decisions,	Realtime Control,

(synchronous)		Local controller	Admission control Handover initiation	SON
Event Trigger	Centralized controller	Local controller → Centralized controller	TTI, UE attachment, Scheduling request, Slice created/destroyed	Monitoring, Control actions
Control delegation	Local controller	Centralized controller → Local controller	Update DL/UL scheduling, Update HO algorithm,	Programmability, Multi-service, Slicing

3.2 Control/User Plane Split

This section analyses how the different functional split options can be best mapped to different deployment scenarios (physical architectures). Centralization of RAN NFs on the one hand provides gains in terms of centralized scheduling and flow control, etc. but on the other hand increases the x-haul (back-/mid-/fronthaul) requirements in terms of bandwidth and latency.

In order to simplify the CP/UP NFs have been structured into 3 parts with respect to their position in the radio protocol stack (-H: high; -M: medium; -L: low) with the following meaning:

- Control Plane network functions:
  - CP-H: High-level inter-site/air interface resource coordination like ICIC or slow load balancing
  - CP-M: User and network specific NFs (e.g. RRC, RAN mobility, admission control)
  - CP-L: Cell configurations, Short-term scheduling, PHY layer control
- User Plane network functions:
  - UP-H: QoS/Slice enforcement, PDCP
  - UP-M: RLC[4], MAC, Higher PHY
  - UP-L: Lower PHY

Figure 3-1 shows the different split options. Five different deployment options are shown: Fully decentralized CP and UP, fully centralized CP and UP, CP/UP partially centralized, CP fully centralized and finally CP (almost) fully centralized. The benefits are:

- **Fully decentralized CP and UP**: simple and no delay and data rate requirements, coordination via X2 interface without any centralized options.
- **CP/UP fully centralized:** High degree of freedom for centralized algorithms between different RU but has very high requirements on the latency and data rate between the RU and UP/CP NFs.
- **CP/UP partially centralized**: Good possibility for centralized handling of asynchronous (slow) functions while still modest requirements on the latency and data rate.

[4] UPNF-H may also consider asynchronous functions of RLC, so only synchronous functions of RLC will remain in UPNF-M

- **CP (only) fully centralized**: Higher CP centralized, i.e. the RRC allows for limited centralized handling of asynchronous NFs. No UP traffic steering functions is possible in the RAN in this scenario
- **CP (almost) fully centralized**: correspond to the software defined networking (SDN) approach. The short-term scheduler can be implemented at the CU which poses strict latency requirements. No UP traffic steering functions is possible in the RAN in this scenario

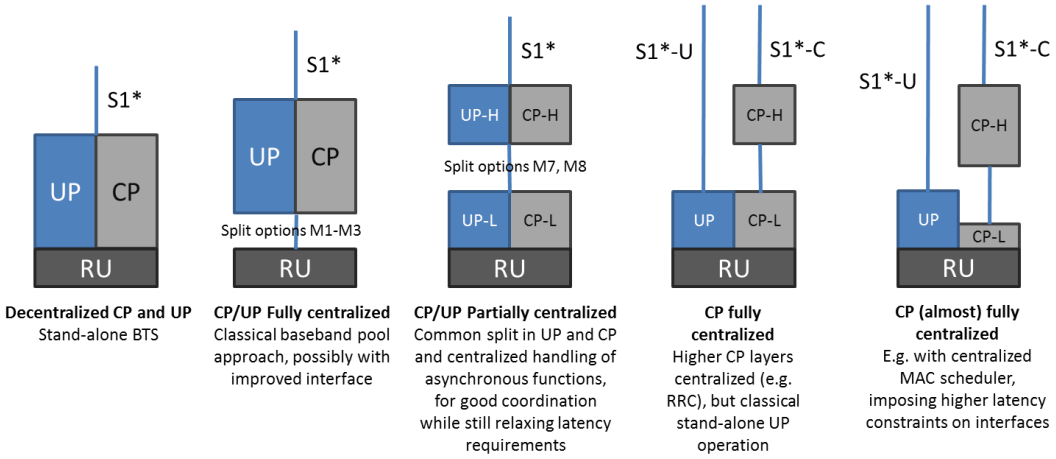


Figure 3-1: Considered UP/CP Split Constellations

Three options of RAN slicing are derived in order to support NW slicing also within the RAN. Each option targets different use cases and requirements such that the same network may support and instantiates different RAN slicing options at the same time.

1. In the first option, all cell-specific NW functions and above are customized by each NW slice but only the transmission point specific functions are shared. The multiplexing is coordinated by the SDM-X controller which makes use of flexible and efficient radio resource management, e.g., in -resource and user-centric control.
2. In the second option, all user-specific NW function and above are customized by each NW slice while all cell and transmission point specific functions are shared. The multiplexing would take place at MAC layer, which is then controlled by the SDM-X taking into account SLAs.
3. Finally, the third option regards a shared RAN where all NW slices use the RAN functions but apply individual QoS scheduling as well as individual CN slices.

The key characteristic of 5G networks will be the increased degree of flexibility and adaptability enabling to cope with huge diversity of radio access technologies, application service requirements and deployment scenarios. In general, there is a trade-off between centralized processing (C-RAN) allowing throughput maximization by efficient radio resource and interference management in dense networks and distributed processing (D-RAN) enabling latency minimization by leveraging baseband processing at access point. As quality of service requirements in terms of resiliency, throughput and latency becomes more and more heterogeneous and application-specific, flexible split and distribution of network functions across processing resources is inevitable. This, however, pose challenge on design and deployment of baseband functions due to the highly heterogeneous RAN environment. TUD addresses this problem within Superfluidity by developing dataflow framework comprising, firstly, a highly configurable dataflow baseband application enabling flexible functional split while supporting multi-cell, multi-antenna, multi-band processing, and secondly, multi-platform dataflow runtime system with dynamic application transformation and resource assignment capability. Complemented by real-time monitoring system, this approach make use of runtime state for dynamic optimization of application structure and mapping to available processing resources according to the performance requirements.

The flexible implementation of 5G RAN will be impacted by application performance requirements and capabilities of computing platform. Unfortunately, the design space and the exploration effort for implementing application on different platforms explodes due to the multitude of possible application arrangements and diversity of computing platform characteristics. Therefore, the automatic application adaptability is of high importance. The proposed dataflow concept allows the application adaptability by leveraging three following principles:

- Dynamic (5G) dataflow baseband application composition according to specific operating conditions and use case scenario
- Dynamic dataflow graph transformation i.e. the optimization of the graph structure that minimizes the structural complexity according to available parallel computing resources
- Dynamic resource allocation and scheduling that maps the tasks to processing elements by exploiting inherent parallelism and data locality

3.3 Protocol stack integration options

There are pros and cons as to which protocol layer is chosen as the aggregation point of a multitude of RATs [3-1]. The general consensus in selecting aggregation point is to choose protocol layers which are time-agnostic to the radio. RRC and PDCP are two protocol layers that do not have to be time synchronised with the radio and therefore often seen as the preferred layers that should be harmonised in supporting different RATs. However, none of these is involved in real-time medium accessing and utilization of the radio resources, and therefore are less powerful in coordination of the different RATs which is a strength of integration at MAC level. RLC and MAC on the other hand, are two protocol layers that are not entirely time-agnostic to the radio and any loss of synchronisation could lead to packet and call drops.

In the case of the PHY, there is a general consensus that a full harmonisation between different air interface variants may not be possible at this layer due to the different requirements and delays of the different baseband processing chains. If the time synchronisation with the radio can be overcome though proper design concepts, then aggregation of RATs at the MAC layer has better gains with respect to radio resources utilization and RATs coordination. This is possible with the MAC framework and functions proposed further down in Section 3.3.2.

3.3.1 PDCP level integration

It is envisioned in 5G that overall air interface (AI) comprises different AI variants (AIVs) that are optimized, e.g., for the specific frequency bands of operation (below 6 GHz, millimeter wave, etc.) and for one or more target 5G use cases [3-2][3-3]. In this regard, a hierarchical RAN control plane (CP) design can effectively address the 5G requirements. Hence, **he higher CP and user plane (UP) protocol stack layers can be implemented in the access network-outer (AN-O), e.g., at a centralized unit (CU), and the lower CP and UP protocol stack layers can be implemented in the access network-inner (AN-I), e.g., at a distributed unit (DU)**. Each DU may support one or more AIVs. Considering the multi-connectivity (see Section 3.3), the service flows can be steered dynamically on the AN-O layer.
Considering the goal of tight interworking between 5G and legacy AIVs, this functional split option is preferred to be at packet data convergence protocol (PDCP) level not to impact the 5G specification with legacy AIV constraints, see Section 3.4. A specific protocol stack implementation along with resource management functions is depicted in Figure 3-2 [3-7], where two AIVs are exemplarily illustrated. Therein, AIV-overarching mechanisms are located at the AN-O while AIV-specific mechanisms are located at the AN-I.

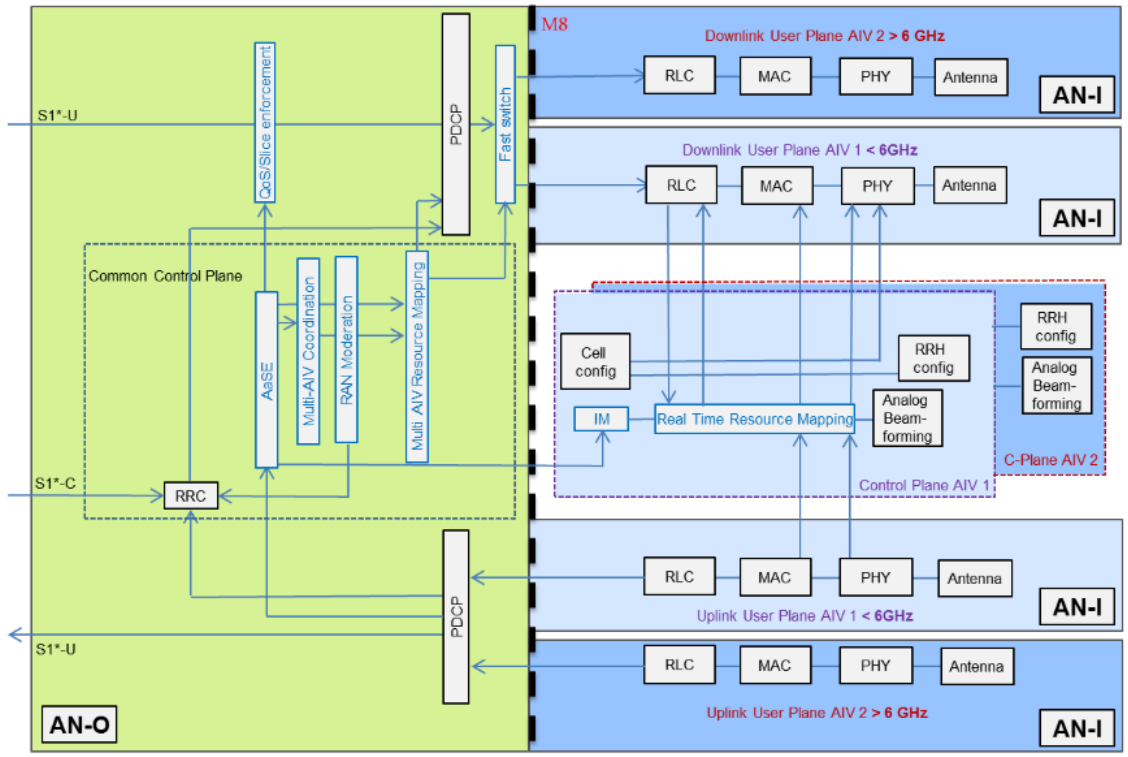


Figure 3-2: Hierarchical CP Design in 5G RAN

For non-standalone mm-wave, to minimise interruption times or ideally even to avoid interruptions and to guarantee reliability, we foresee that multi-connectivity (MC) will be an essential or rather a fundamental feature in a 5G-access network. Moreover, a UE must be able to detect and receive from multiple mmAPs to ensure the possibility of MC, link monitoring, and fast selection. To provide redundant coverage, multiple mmAPs are placed within the low-band 5G coverage area, building per UE "serving cluster". Under the consideration of backhaul capabilities in current and future low-band radio access nodes, i.e., to relax backhauling requirements with respect to 5G low band access points, we further propose a split between high and low data rate PDCP. We split the data layer in high data rate and low data rate flows and in accordance with that, and we define high rate PDCP-H typically hosted in an edge cloud and low rate PDCP-L hosted in the low band 5GAP. 5GAP additionally hosts RRC and manages the mm-wave cluster incl. traffic steering from PDCP-H to mmAPs.

3.3.2 MAC level integration

The Extended Dynamic Spectrum Access (eDSA) MAC framework is multi-RAT capable, and is able to support a multitude of RAT-specific MAC protocols. It is integrated into an architecture that facilitates this capability, see Figure 3-3. The novel multi-RAT-capable protocol architecture is developed with LTE-A protocol stack as the base.

Non-Access Stratum layers, such as Mobility Management (MM) etc. and access stratum layers, as e.g. RRC, PDCP and RLC are beyond the scope of this work on MAC level integration. However, valid assumptions have to be made for any form of interaction that involves them. The focus here is however only on MAC and PHY layers. Radio Resource Management (RRM), though not a visible entity in the LTE-A protocol stack architecture, is a key component with which the eDSA MAC has to interact in order to intelligently and efficiently utilize the available radio resources from different spectrum bands. A centralised RRM (cRRM) is proposed, which is a major paradigm shift from the current distributed RRMs deployed in LTE and 3G Systems. The cRRM, with the support of the Operations and Administration and Management (OAM) entity, Spectrum Manager, and other relevant operator-specific network elements, coordinates the spectrum and the radio resources utilizations among the supported RATs. A multitude of

frequency spectra from different spectrum regimes are expected to be supported by eDSA systems – a key requirement of 5G Radio Access Network (RAN) design.

One fundamental question related to the design of a multi-RAT 5G RAN is how the different RATs or air interfaces can be integrated (or aggregated) such that the complexity of implementation is minimised and the performance of the individual RAT is not hampered. This raises questions related to protocol harmonisation and on which layer aggregation should take place. Different approaches imply different benefits and challenges.

As shown in Figure 3-3, the MAC layer is chosen as the harmonisation point for the different RATs and air interfaces. This is in sharp contrast to the recently standardised LTE/Wi-Fi Aggregation (LWA) where the aggregation and/or split point is at the PDCP layer [3-4].

In order to support eDSA and overcome time synchronisation issue between the MAC layer and the radio, this layer is divided into two sub-layers: Higher MAC (HMAC) and Lower MAC (LMAC) (see Figure 3-3). There could be many instances of LMAC, each tied to a specific RAT/air interface, but only a single HMAC that coordinates the inter-RAT and coexistence functionalities of the different instances of RAT-specific schedulers at the LMAC. In essence, each LMAC instance which is tied to some RAT is responsible for the real-time scheduling of user and control traffic, whilst the HMAC coordinates, depending on the load distribution, which of the RATs should be employed and how much traffic is distributed to each. The LMAC instances may operate concurrently, each scheduling user traffic independently. *The entire eDSA MAC supports LTE-A and its evolution, WiFi and its evolution, as well as novel 5G air interfaces*. In [3-5], the eDSA MAC design was presented with initial simulation results.

In the proposed design, every LMAC instance interfaces with an instance of a PHY at the PHY layer, which may not be shared with other LMAC entities. For example, LMAC LTE-A uses a PHY instance that supports OFDM waveforms and the characteristics of the frequency-time grid resources, modulation schemes etc. may not apply to WiFi, if this PHY instance is used by a WiFi LMAC instance.

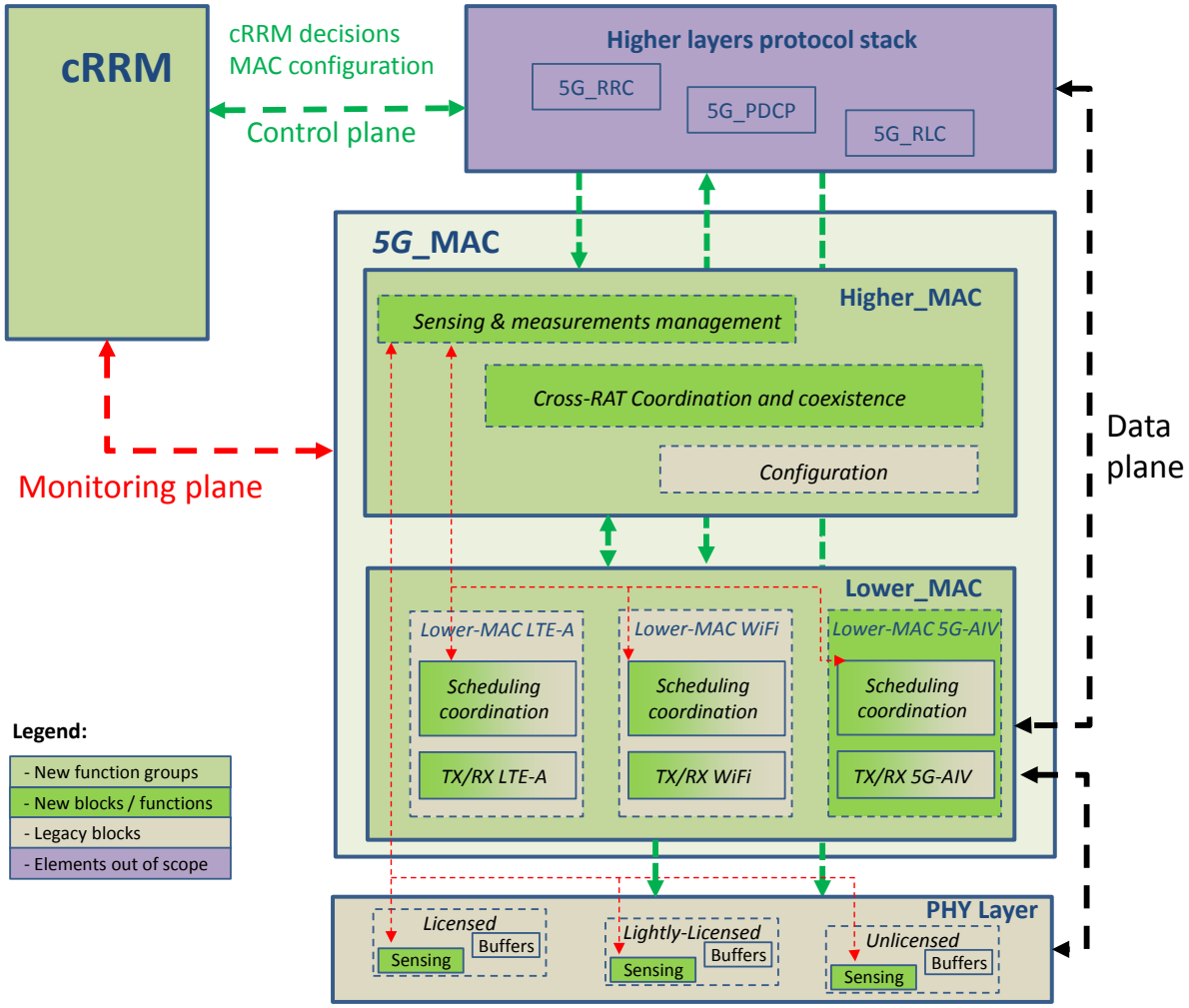


Figure 3-3: eDSA MAC Framework Architecture and functional blocks

Several Service Access Points (SAPs) and other interfaces have been defined between the eDSA MAC and cRRM and the existing protocol layers (e.g. RLC, PDCP and RRC), many of which are expected to evolve in 5G.

5G-RRC has a control interface with the HMAC. This is a configuration interface relaying RRM policies and UE configurations, radio bearer or logical channels and other configurations that are to be applied by the HMAC and the different LMAC instances. There is no 5G-RRC configuration interface to LMAC. This gives absolute control to the HMAC which interfaces with LMAC in deciding which RAT-specific LMAC gets configured and activated or deactivated, depending on the cRRM mid-to-long term policies of the spectrum utilization and the traffic load distribution.

The 5G-RLC interfaces with LMAC to handle the control and the data plane traffic respectively. A set of logical channels are expected to be statically mapped to each RAT and the traffic on these logical channels scheduled by the respective LMAC RAT-specific entities.

The LMAC interfaces with the PHY for control and data traffic respectively. As highlighted above, the LMAC has many instances, each tied to a specific RAT that facilitates adopting a PHY of any waveform which fits with the objective of the MAC protocol, implementing the 5G use-case [3-6] of interest. For example, one key waveform of interest is the Filter-Bank Multicarrier (FBMC). An instance of the LMAC_5G_AIV interfaces with a PHY that has FBMC waveform implementation and this is expected to operate in the unlicensed spectrum because of its reduced spectral leakage and high spectral localisation, which in turn leads to lower co-channel interference. With such flexible protocol architecture and the identified SAPs, higher layers such as RRC, PDCP and RLC can evolve to be RAT-independent and the eDSA MAC framework can

offer the flexibility to be enhanced with further air interfaces over time. Flexibility in integrating air interfaces is one key requirement of 5G RAN design [3-7].

The HMAC has the responsibility of ensuring coexistence of the supported RATs and existing wireless systems in shared spectrum with coexistence functionalities such as Listen-Before-Talk (LBT), Carrier Sense Adaptive Transmission (CSAT), Frame Format adaptation, etc. In addition, it has to intelligently coordinate the utilization of the different RATs with an encapsulation that makes higher layers RAT-independent, leading to leaner specification and design of those layers. Besides the control and user plane, a new plane has been introduced to support eDSA, and this is called the **Monitoring Plane.** This plane supports sensing and measurements and other relevant KPIs reporting to the cRRM. With the multi-RAT capable framework, an eDSA system is supposed to take quick decisions on handover, whilst providing non-interrupted transmission and reception of data to user equipment (UEs). HMAC is expected to complement 5G-RRC with this procedure via its KPI and measurements collection. HMAC can instruct measurement collection from the UE and cell's PHY and can relay that to the cRRM, so that cRRM can react promptly on the utilization of the radio resources with any imminent change in the radio environment. This is facilitated by the monitoring plane. Moreover, load balancing and traffic steering are further objectives of the HMAC, leveraging on the coordination capabilities due to the interfaces to the individual LMAC entities.

LMAC, on the other hand hosts a number of RAT-specific MAC protocols, each providing real-time scheduling of user and control traffic and operating autonomously when configured and activated by the HMAC.

The LMAC 5G_AIV shown in Figure 3-3 is an innovation that schedules traffic over a 5G air interface. It includes a KPI collection function, which can be invoked by the HMAC to collect statistics of some specific performance indicators to assist HMAC and cRRM to assess the overall behaviour of the cell and also the spectrum utilization. The LMAC 5G_AIV also supports Carrier Aggregation of radio frequencies from a heterogeneous set of spectrum bands and depending on the spectrum of usage, coexistence with incumbent same-tier users are strictly observed, thanks to LBT/CSAT and Frame Format adaptation functionalities. Besides the coexistence and carrier aggregation functionalities, the LMAC 5G_AIV also supports RACH, a 5G random access scheme, specific to the 5G air interface to which it operates and the other functions include configuration at real-time, radio resource grid settings, frame format for transmitting and receiving of control and user traffic. Techniques that allow multiple users to utilize same resource elements of the radio resource grid, often referred to as Non-Orthogonal Multiple Access (NOMA) [3-9], is also supported by the LMAC 5G_AIV.

With the diversity of the LMAC entities, QoS of cell edge users will be improved, as a multi-RAT capable UE could have user traffic and control messages steered on RATs experiencing lower inter-cell interference. The diversity of LMAC entities increases reliability of transmitting user traffic and control messages.

Full details on the proposed MAC Framework Architecture can be found in [3-10].

3.4 LTE and 5G RAN interworking

The interworking between different radio access technologies (RATs) has been on the core network level and based on inter-RAT hard handovers. In the era of 5G, **it is aimed that 5G and evolved LTE shall be integrated on the RAN level in the form of a tight interworking**. The tight interworking can be realized through a dual connectivity (DC), where a user equipment (UE) can connect to evolved LTE and 5G AIVs (UP and CP) simultaneously. As also depicted in Figure 3-2, the integration can be realized at the PDCP level. DC can increase the UE throughput due to UP aggregation and make the connection more reliable via CP diversity. One disadvantage of DC is that it may be less resource efficient than to be connected only to the best cell, since DC will

transmit packets also on links with relatively bad quality and, thereby, taking resources from users who could utilize the link better.
An alternative to DC is Fast UP Switching (FS), where the CP is connected to both AIVs at the same time but the UP is only transmitted via one of the AIVs. If the CP is connected to both the LTE access node and the 5G access node, no signaling is required when switching AIV, and the UP switch can be almost instantaneous. The need on the frequency of the radio link feedback between AN-I and AN-O, and accordingly the speed of FS depends on the radio frequency and speed of the UE in order to follow the fast fading. An example set of results are provided in Figure 3-4 [3-8]. It is shown that at low load DC performs better; yet, on the high load FS outperforms DC since DC is impacted by the increased interference due to the simultaneous connection to two AIVs.

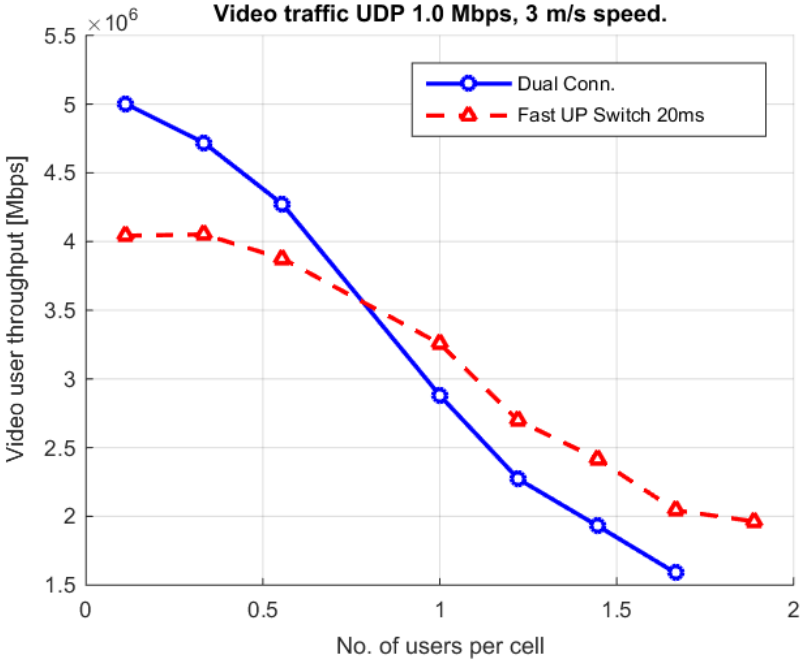


Figure 3-4: Hierarchical CP Design in 5G RAN

In order to fulfil the stringent 5G requirements, NR is being developed to be operational also at higher frequencies of up to 100 GHz (i.e., at mm-wave frequencies) where significant bandwidth resources are available. However, the high frequencies bring new challenges in terms of availability and reliability as a consequence of the reduced coverage due to increased path loss and signal blockage. The mm-wave system is expected to extend on the concept of dual connectivity by i) allowing more than two nodes to be connected, ii) enabling a new bearer type where the traffic is terminated in the SeNB and split into the MeNB and/or the SeNB, and iii) enabling a tight interworking between LTE and NR

Although the path loss can be mitigated through the use of narrowed directional beams, this may aggravate the poor coverage in case the directions of the beams are uncertain. Despite the fact that NR will eventually support standalone operations, where sufficient mm-wave coverage can be provided, a tight interworking between LTE and NR will allow a gradual deployment of the mm-wave RAT and leverage on the incumbent LTE installations infrastructure. For instance, users with good coverage to both LTE and NR can aggregate carriers from both RATs to obtain even higher throughputs, while edge users requiring large resources to obtain a modicum of performance can beneficially be served by LTE. To allow this dynamic RAT selection and utilization, a harmonized set of procedures in LTE and NR will allow an optimized selection and smooth transition between them.

To evaluate the impact of utilizing mm-wave access, a system evaluation was performed in a dense urban deployment consisting of 1442 buildings with heights varying between 16 and 148

m distributed in a 2x2 km2 area, with the taller building located in the centre, as shown Figure 3-5. The evaluations compare legacy LTE at 2.6 GHz with 40 MHz bandwidth to standalone NR at 15 GHz with 100 MHz bandwidth, as well as non-standalone deployments with co-located LTE and NR.in a dense urban deployment consisting of 1442 buildings with heights varying between 16 and 148 m distributed in a 2x2 km2 area, with the taller building located in the as shown Figure 3-5. The traffic is served by a macro network with an inner and an outer layer. The inner macro layer contains 7 sites deployed on the building rooftops with 200 m inter-site distance and the outer macro layer contains 28 sites deployed at 30 m height with 400 m inter-site distance. The traffic is generated by deploying 10000 users distributed uniformly randomly in the area with 80% of the traffic generated inside the buildings. Each user downloads a single packet with a size depending on the traffic load.

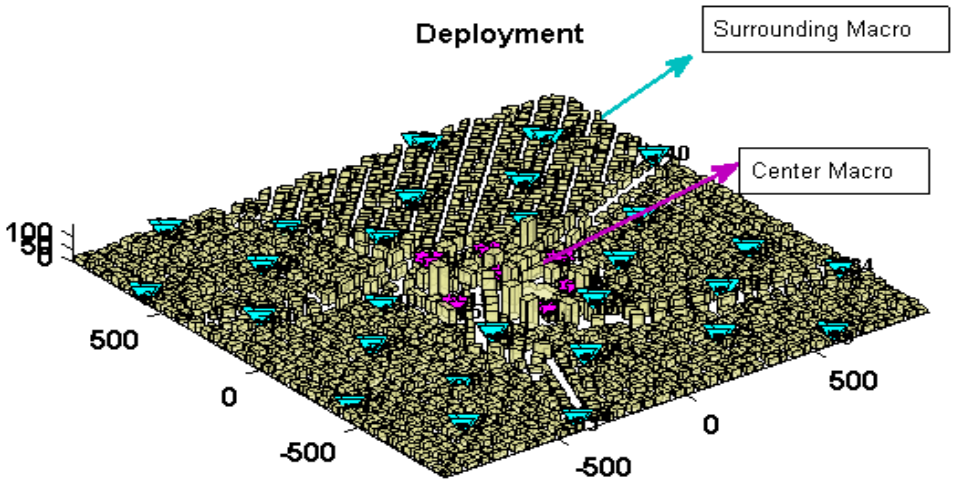


Figure 3-5: Network layout for LTE-NR tight interworking simulations

To avoid boundary issues from users located near the edge of the simulation area, the evaluation only considered the performance of users located within the central 1x1 km2. The users outside this area still generate traffic but are only considered as a source of interference. The end-user performance of the evaluated systems is illustrated in Figure 3-6.

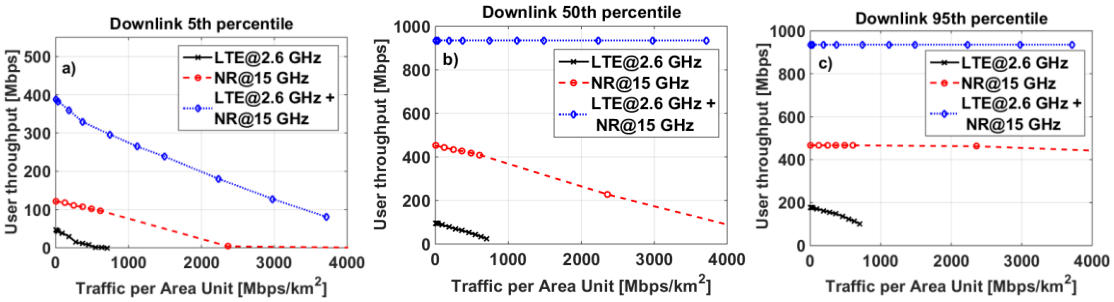


Figure 3-6: Simulation results for LTE at 2.6 GHz standalone, NR at 15 GHz standalone, and non-standalone deployment with LTE at 2.6 GHz and NR at 15 GHz. (a) 5th percentile; (b) median; (c) 95th percentile user throughput.

As the figure shows, LTE standalone cannot cope with the traffic load and cannot provide 50 Mbps for the 5th percentile users, even at very low loads. For NR standalone, the throughput even for the 5th percentile is increased thanks to the wider bandwidths. However, when employing both LTE and NR, the throughput is increased for all users to more than the sum of the two systems. An explanation for this synergy effect is that in the standalone deployment, all UEs compete for the same resources while for the non-standalone deployment, the UEs can be served by whichever system gives the best performance. For instance, for a UE in poor coverage, i.e. a low signal to noise ratio (SNR), require much more radio resource to transfer the same amount of data as a UE in good coverage as a more robust modulation-coding scheme (MCS) need to be

used. Consequently, since the UE in poor coverage require more resources, these resources cannot be used by other UEs whose throughputs are reduced. In addition, as the UEs with poor coverage have a reduced throughput, they also take longer time to finish their data transfer, meanwhile generating interference for the other UEs. This will reduce the overall performance as the signal to interference plus noise ratio (SINR) reduces. When both LTE and NR are utilized, the UEs with good coverage to both systems get the benefit of receiving the full capacity of both RATs while the UEs with poor NR coverage are only served by LTE. These edge users in turn also benefit from NR as a significant amount of traffic is offloaded from LTE to NR, thus reducing the traffic load as well as the interference.

3.5 Centralised and distributed RRM

The Radio Resource Management (RRM) is a key component in the 5G network architecture. The RRM framework proposed here consists of two types of components: the cRRM (centralised part) and dRRM (distributed part). An adaptation layer interfaces southbound to the upper MAC layer and northbound to OAM, spectrum manager, and KPI collector. A network operator will have one or more cRRM entities located typically in a gateway entity of the network, controlling several hundred cells, whereas the dRRM is located in every cell and will communicate with one or more cRRMs. The dRRMs can communicate with cRRMs from different operators and in this way the cell becomes multi-tenanted.

The proposed RRM design is capable of incorporating algorithms from multiple vendors whether those algorithms are centralized or distributed. Moreover, the proposed RRM framework fully decouples the underlying algorithms by the introduction of an abstraction layer (AL) and supports multiple interfaces transparently from the algorithmic point of view. This means that communications with the entities outside the RRM are the responsibility of the AL. The framework can also work in Cloud Computing environments, as well as in embedded equipment.

Figure 3-7 shows the proposed RRM architecture where dRRM and cRRM are deployed.

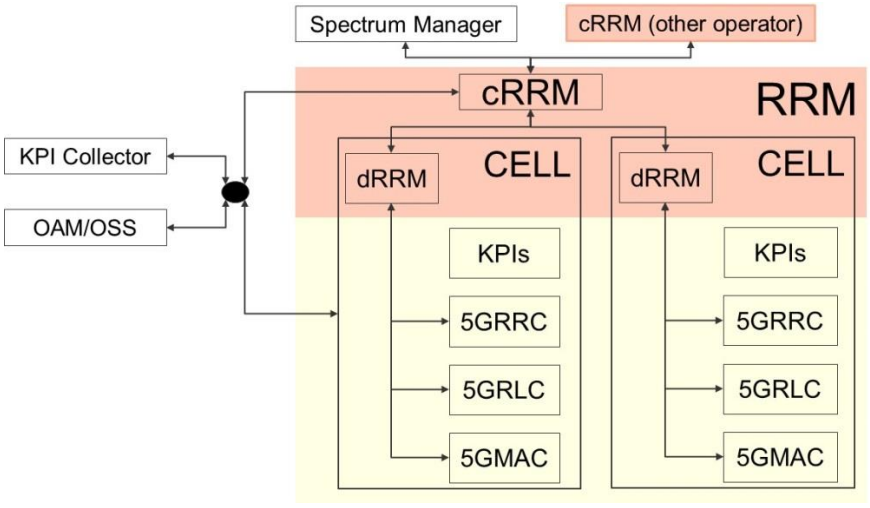


Figure 3-7: RRM overview

The most important design requirements of the framework are:

- The RRM is composed of one cRRM and one or more dRRMs.
- The cRRM is always deployed in a network, even if it is only a gateway between the Spectrum Manager and the cell.
- The dRRM is always deployed in a cell, even if it is only a gateway between cRRM and cell layers.
- Cells have only one corresponding cRRM per operator.
- dRRMs may be connected with more than one cRRM, but exactly one per operator.

- Decisions taken at cRRM have priority over those taken at dRRM. When the cRRM sends a new configuration, the dRRM resets its status with cRRM information.
- Only the cRRM is capable of requesting data from the Spectrum Manager. Any dRRM request goes through the cRRM.
- Cells support RAN-sharing.
- A new interface between cRRMs is provided, to support RAN-Sharing.
- SON and current 3GPP RRM procedures are part of the RRM algorithms.
- Ready for Cloud Computing, through its flexible degree of distributed processing.

High-level design of the RRM framework

The high-level design of the RRM framework is shown in Figure 3-8 below, where the main blocks are:

- RRM configuration: This module contains the entire RRM configuration required for the functioning of the system. The configuration module contains all the information related to the configuration, e.g. how temporal data is stored (in a database or RAM memory), the allowed protocol interfaces, or the authorized algorithms. This entity contains various parameters, for example, required IP addresses, algorithm order (when executed in series), different groups of algorithms to be executed in series, priority of the algorithm when more than one targets the same issue, etc. The information required for secure connection to the network is also stored in this module. The *RRM configuration* is managed by the OAM, which in turn is controlled by the network operator.

- Internal KPI collector: The RRM stores KPIs and reports them to the OSS. The network manager, usually the operator, will define the required KPIs and the algorithms responsible for providing the information. The proposed framework provides the mechanisms to allow algorithm to report their KPIs. The network manager (OAM) configures the periodicity of KPI reporting for each report as multiple reports with different periodicity may be requested by the operator.

- Message constructor: This entity is responsible for mapping the received data to internal and external format and forwards it to the algorithms or to other entities respectively. Main functionalities include validation of received data, avoiding errors into the framework, inspection if new inputs are relevant or not, and mapping internal messages into output message and vice versa.

- Abstraction layer: manages the physical interfaces and translates physical messages through its SW abstraction layer. The key idea here is to separate higher functions (especially cRRM, but also dRRM algorithmic operations) from lower layers (RLC, HMAC etc.). The AL translates all messages passing through it from a SW abstraction viewpoint, to HW-specific message primitives for the layers below it.

- Interfaces: This entity is in charge of managing the physical interfaces required by algorithms in the RRM. Every entity connected to RRM requires a specific interface based on as SCTP or TR-69 transport. All the required procedures to establish and maintain the interface connection are managed by the Message Dispatcher. *Interfaces* is also in charge of validating the received data, avoiding RRM malfunctions. It is important to note that the validation process over the Interfaces is different from the one that is done in the Message constructor. The Interface validates the data itself while the Message constructor validates the content of the message.

- Algorithms: In addition to supporting legacy RRM algorithms, new algorithms are proposed for an optimal spectrum usage. The grouped algorithms run sequentially and the final result is provided when all the algorithms terminate. Some of the algorithms may require more data from the system to do their calculations when they receive the start trigger. The algorithms may be grouped and the framework has the capability to run each

defined group in parallel. The groups of algorithms or the algorithms themselves are isolated and, as a consequence, each one will provide the best solution. Artificial Intelligence (AI) techniques can be deployed to make a final selection.

- Requester: is used by the algorithms to get data periodically or on-demand. It does so by scheduling its periodic request, or by issuing requests when the algorithm runs. The entity has to be configured and is the network operator the one which knows where the information is stored. On the other hand, it is the algorithm which knows which data is required. Therefore, an interface is required between the RRM and the OAM to configure the Requester.

More details can be found in [3-11].

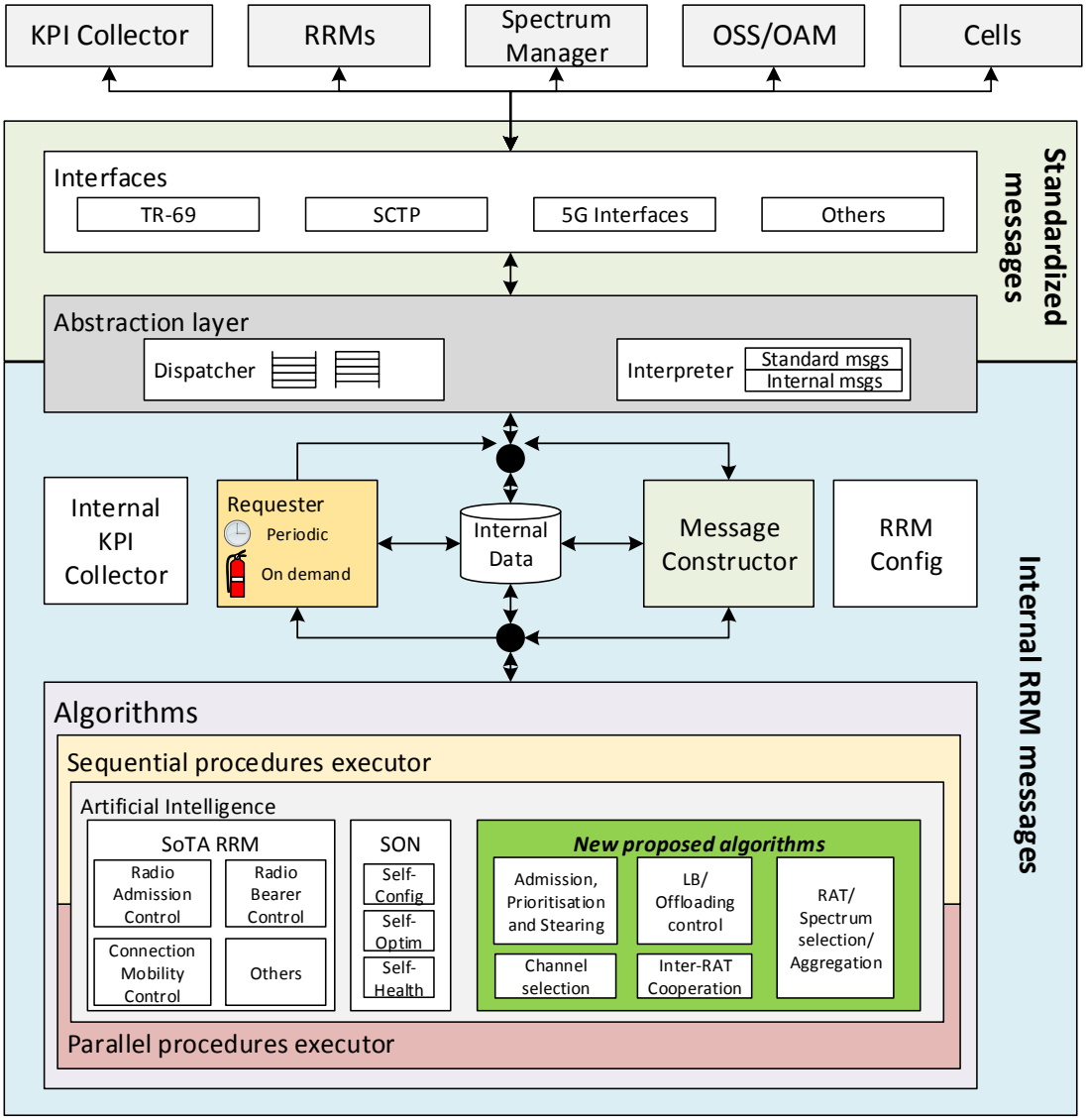


Figure 3-8: High-level RRM framework. Note the abstraction layer, a critical component of the design separating higher-level algorithms from internal components.

Finally, an overview of the new RRM algorithms is given.

- Algorithm 1 is designed for efficient licensed-assisted access (LAA) operation in small cells, based on reinforcement learning. Running in the dRRM, the algorithm chooses the best-unlicensed channels to use on the downlink based on spectrum availability and the QoS requirements from the cRRM. Simulations show that there is an improvement in the worst served UEs.

- Algorithm 2 is used for RAT/spectrum/channel selection based on hierarchical machine learning. Using dRRM it chooses the best option for the downlink taking into account a pool of bands and various licensing schemes (licensed/unlicensed/lightly-licensed) and the need to fulfil certain traffic requirements. The focus in the lightly licensed band of 3.5 GHz spectrum.
- Algorithm 3 is Radio resource allocation with aggregation for mixed traffic in a WiFi coexisted heterogeneous network. It performs load balancing across WiFi and licensed spectrum. Using knowledge of the available capacity on the unlicensed spectrum it decides which UEs can use WiFi.
- Algorithm 4 is a Fuzzy MADM strategy for spectrum management in multi-RAT environments. Working on the uplink, a connection manager (CM) is introduced on the UE side to collect the various components of the context and acts according to a policy that is remotely adjusted by the network manager. Based on this, a fuzzy multiple attribute decision making (MADM) implementation of the CM is developed to select the best RAT for a set of heterogeneous applications.
- Algorithm 5 is Co-primary spectrum sharing in uplink SC-FDMA networks. The algorithm takes into consideration the users' buffer status and real-time delay constraints, as well as the operator priorities and the constraints of a realistic LTE system in order to perform uplink resource allocation in a QoS and energy efficient manner.
- Algorithm 6 is Dynamic resource allocation algorithm for the coexistence of LTE-U and WiFi. The algorithm maximizes network throughput in the multi-operator scenario for 5G mobile systems by jointly considering a licensed & unlicensed band, user association and power allocation subject to minimum rate guarantee and co-channel interference threshold.

More details can be found in [3-11].

3.6 Enhanced/new network access functions

3.6.1 Initial access

Initial access refers to a set of CP functions across multiple layers of the RAN protocol stack (e.g. PHY, MAC and RRC) and, at some extent, the CN / RAN interface as in the case of paging and state handling. In LTE, some of these functions are synchronization (time and frequency, UL/DL), Cell Search, System information distribution and acquisition, Random access and Paging [3-12].

The initial access solutions aim at handling both the initial access bottlenecks due to massive connectivity and the service prioritization. The proposed components proposed are:

- **Group based RACH: The devices are being grouped by the network based on their mobility and their communication characteristics (e.g., data to be transmitted, packet delay requirements) for reducing the collision rate. The network schedules the cluster heads' transmission opportunities based on their transmission requirements.**
- **URLLC: Combine preambles transmissions from certain UEs to prioritize service requests in case of collisions.**
- **The RAN based paging concept benefits from the new Connected Inactive which allows the network be able to page the UE more accurately (fewer NBs broadcasting the paging).**
- **5G RAN lean design: With on-demand system information, avoiding always on reference symbols (self-contained with data) and longer time between**

synchronization signals, the 5G system may sustainably decrease the energy consumption compared to LTE.

From the evaluation results depicted in Figure 3-9 it is observed that using the group based system access reduces the collision rate significantly. Additionally, the average initial access delay (i.e., random access, random access response, terminal identification, and contention resolution) is reduced, as shown in Figure 3-9, since the devices are accessing the system with fewer collisions and thus experience fewer retransmissions.

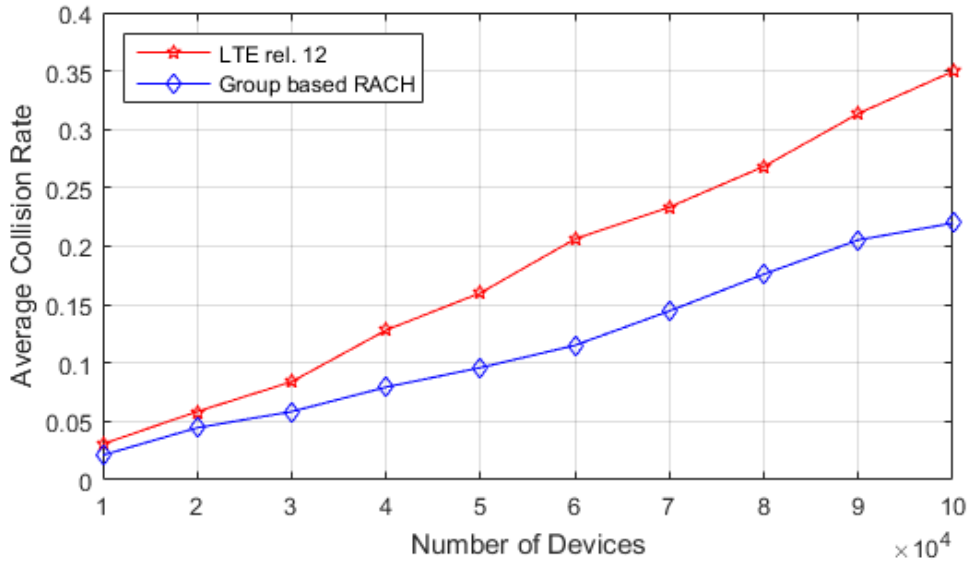


Figure 3-9: Group-based initial access gains in terms of average collision rate decrease

Low frequency assisted initial access

The initial access process comprises the three tasks of downlink timing and frequency synchronization, system information acquisition and uplink timing synchronization. During initial access a UE has to establish a RRC connection with the corresponding mm-wave AP. The performance of this procedure directly impacts the user experience. Therefore, on PHY layer a beam alignment must be achieved within short time. Exploitation of the limited a-priori information on the preferred transmission direction at both ends of the link will support this. In a non-standalone deployment, i.e., a heterogeneous network, as the one illustrated in Figure 3-10, where mm-wave small cells are located within the coverage area of a macro cell operating at low frequency, low frequency RAT assistance can improve initial access performance significantly. Especially UE power consumption and latency can be reduced.

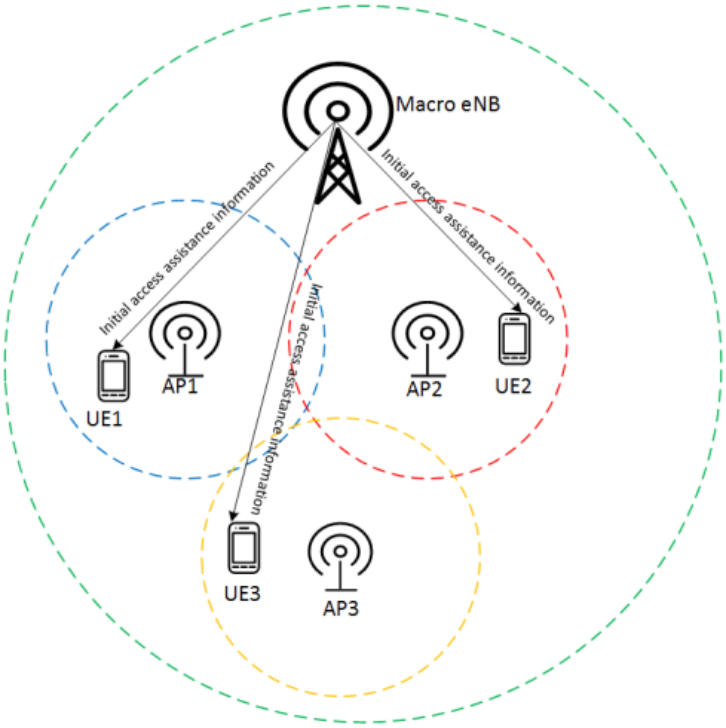


Figure 3-10: Low frequency-assisted initial access to a heterogeneous network.

In the following, the three mentioned tasks of low frequency RAT assistance are highlighted.

Downlink synchronization

For downlink synchronization the UE exploits synchronization signals transmitted by the AP. These are in particular time-frequency resources with a certain periodicity, which allow acquisition of symbol, slot and sub-frame timing. After achieving that, the UE is able to obtain the cell ID. If the UE is located in a low frequency RAT coverage area, the low frequency RAT can transmit information about frequency and cell IDs of mm-wave small cells within its coverage area, by e.g., dedicated signalling to the UE. With this signalling, the UE does not need to perform an exhaustive search over the whole small cell ID space, but it only tries to detect the signalled cell IDs. As a consequence, the UE power consumption for downlink synchronization is significantly reduced.

System information transmission

The second task of the initial access procedure is to acquire the system information which provides all the essential information for accessing the network to the UE. The coverage of the system information determines the coverage of the cell. Some of the system information components, e.g. the system frame number, are changing fast on the basis of one or several mm-wave RAT frames. Other system information components vary relatively slowly, so information about system bandwidth, random access resources, paging resources and scheduling of other system information components is typically semi-static. For this reason, it can be energy efficient to convey some of the slowly varying system information by exploiting the existing low frequency RAT. The fast changing system information components, however, need to be transmitted by the mm-wave RAT.

Uplink synchronization

It is important that efficient uplink (UL) data transmission in the mm-wave RAT is supported as well, especially for "UL data traffic dominant" use cases, e.g., uploading content, such as high-resolution videos to social media during sports events, concerts etc. UL synchronization needs to be achieved prior to any UL packet transmission to ensure that all the co-scheduled UEs' UL

signals are time-aligned at the eNB. A RACH procedure, similar to that standardized in LTE can be used. Based on the RACH preamble transmitted by the UE, the eNB can determine the timing advance value for the UE. The radio resources for the preamble transmission are typically part of the system information and such system information can be signalled by the low frequency RAT. This can be viewed as a basic assistance to the UL synchronization. To ensure a certain UL preamble coverage, if several preamble formats are supported by the system, the low frequency RAT can signal a particular preamble format to the UE in order to realize the network assisted preamble format selection. In case of contention free RACH, the low frequency RAT can signal the exact preamble sequence to be used by the UE.

During the LTE-like RACH procedure, the RACH response signal can be also transmitted by the low frequency RAT. In addition to the above mentioned options for UL synchronization assistance, the low frequency RAT may also offer assistance to the possible beam alignment operations during the initial UL synchronization procedure.

3.6.2 Dynamic Traffic Steering

The RAN design presented in Section 3.1 enables certain network functions to operate on a faster time scale in 5G than in legacy systems. On this basis, **traffic steering may not be performed in the form of hard handover on radio resource control (RRC) level between different RATs (e.g., between 3G and 4G) or access nodes, but could be done in a much more agile way and on a faster time scale on lower protocol stack layers**. The resultant dynamic traffic steering exploits multi-connectivity (see Section 3.3) and applies traffic flow adaptation on a faster and possibly synchronous time scale. To this end, QoS requirements of the traffic flows are fulfilled considering the real-time radio link conditions pertaining to different AIVs serving the UE. For instance, relative unpredictability of the radio links especially on the higher frequency bands needs to be taken into account when considering high-priority traffic flows, as depicted in Figure 3-11 [3-8]. Current LTE radio link failure detection and recovery mechanisms would take several seconds in order to re-establish the radio bearer, and since this is unacceptable for high-priority traffic, the dynamic traffic steering framework will ensure that the QoS policies received from the 5G core network (CN) are successfully enforced. Some example results are shown in Fig. 3.w, where normalized delay can be improved relative to hard handovers as in legacy (see subfigure b) and, by employing packet duplication, SINR can be increased particularly in the low regime, which can improve the cumulative link reliability (see subfigure c).

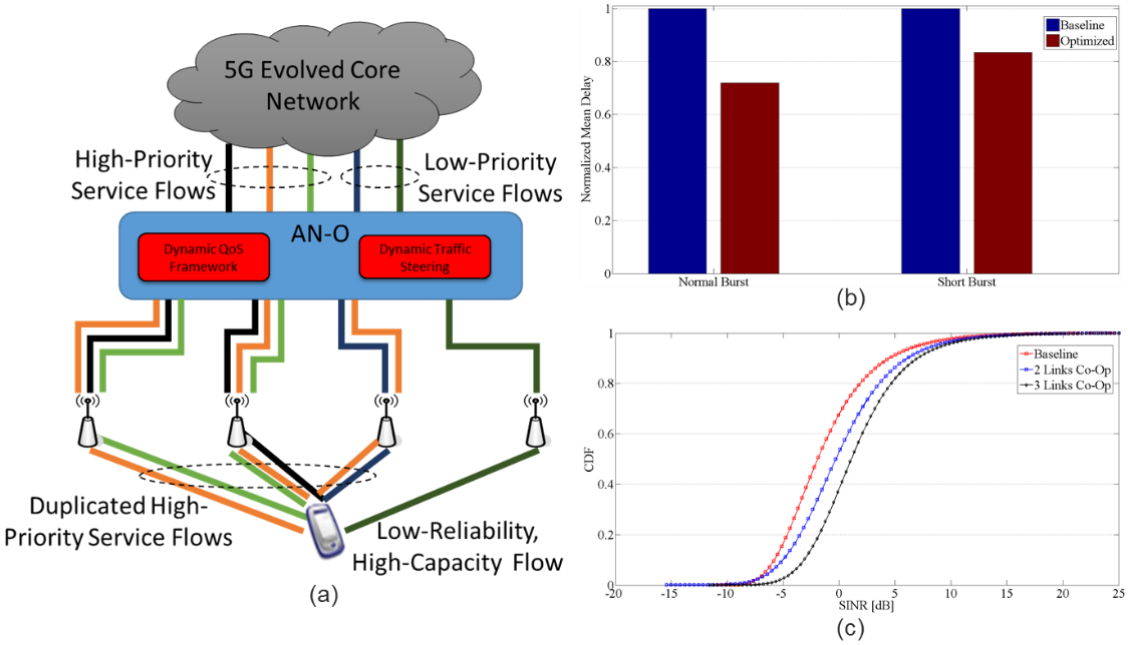


Figure 3-11: Dynamic Traffic Steering framework (a), delay gains (b) and reliability increase in terms of SINR gains (c).

3.6.3 RAN Moderation

While it is foreseen that traffic will increase massively in 5G networks, a key element to allow sustainability of future systems is that the absolute energy consumption in cellular networks should not increase as compared to today. Accordingly, **the energy-efficient network operation can be attained by moderation of RAN access nodes, i.e., the optimal active-mode operation, with the help of QoS and channel quality awareness.** The energy saving gains of the moderated network is obtained thanks to the envisioned lean system design of 5G, which enables the access nodes to enter sleep mode longer when there is no traffic to be served. An example consideration is illustrated in Figure 3-12 [3-10], where we consider the RAN network to also include self-backhauling (sBH) nodes where access and backhaul links are provided using the same 5G radio, possibly using the same frequency band for operation. The coordination is achieved when each sBH node informs the aggregation node (see AN-O in Section 3.1) about the intended sleep decision and about the time periods where it would be in sleep mode. The aggregation node can also enter sleep mode in a synchronized manner along with the sBH nodes. Clear power consumption gains are shown in Fig. 3w2. Therein, for LTE, we assume the presence of fiber access links without sleep modes. For the 5G-gNB with dedicated backhaul (BH) link, we assume similar BH architecture, with the Gigabit Passive Optical Network (GPON) sleep mode power consumption.

RAN moderation framework [3-8]also includes mechanisms exploiting coordination approaches like joint transmission so that the number of cells that should be activated to satisfy the existing traffic can be reduced. Furthermore, mobile access nodes, e.g., vehicular nomadic nodes integrated into cars, within a dynamic radio topology can be activated or deactivated on-demand to cope with changing traffic conditions over time and space.

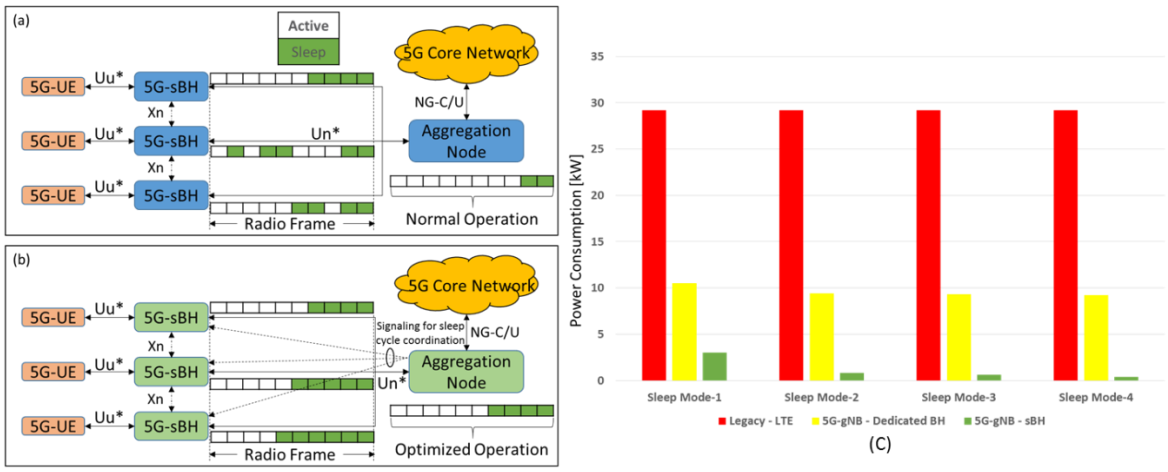


Figure 3-12: RAN moderation framework considering sBH nodes with normal operation (a), optimized coordinated operation (b), and achievable power consumption gains via optimized operation (c). Sleep modes 1 - 4 indicate the level of power savings that could be achieved at the BS, depending on the available sleep mode duration (0.071 – 1000 ms).

3.6.4 Cell clustering

During recent standardization efforts in 3GPP, it has been agreed that NR will support two different levels of mobility, namely RRC driven and mobility without RRC involvement [3-13]. For the RRC driven mobility, the handover procedure will be similar to LTE, where the UE reports measurement on serving and neighbour cells when the measurements fulfil certain criteria e.g. neighbour cells is offset better than serving cell. The network can then reconfigure RRC in the UE to perform a handover to a target cell. For the mobility without RRC involvement, the

measurements are handled at PHY or MAC layer based e.g. on channel-state information reference signals (CSI-RS) or channel quality indicators (CQI). The UE will then be configured to select a specific beam which may come from an AP different from the serving AP. 3GPP has recently agreed to standardize a protocol split with a centralized PDCP and distributed RLC/MAC/PHY [3-14], it would be possible to maintain a single PDCP entity in one gNB while switching the RLC/MAC and PHY from one transmission/reception point (TRxP) to another. This protocol split also allows for arranging UE-specific base station clusters [3-15]. The cluster will consist of at least two APs, where one of the APs is designated as cluster head (CH). The CH terminates the control plane and an interface to the CN, and can configure which other APs should be included in the cluster. All cells in the cluster transmit reference signals (e.g. CSI-RS) in beams for the UE to measure on. The UE then reports the quality of the measured beams, either directly to the CH, or via another AP which forwards the measurement reports to the CH. Based on the measurement reports, the CH can switch between beams from different AP, only relying on PHY layer beam management and beam refinement procedures.

Since the availability and configuration of the base station cluster is transparent to the UE, the UE will not be aware of which AP is the CH, or which neighbouring cells are included in the cluster and which are outside the cluster. Instead, the UE will measure on all the beams it detects of the configured measurement objects. The network can then decide whether multiple APs should be prepared to serve the UE, based on transport capacity between the nodes, as well as processing and storage capacity of the target node.

To ensure connectivity within the cluster, it may be necessary to rely on the wide area coverage of low-frequency RATs, e.g. LTE-A, when the mm-wave RAT has limited reliability e.g. due to signal blockage. The lower frequency can then relay traffic and control signals from the CH to the UE, and assist in intra-cluster mobility. This makes strong connection between inter-frequency multi-connectivity and mm-wave clustering. Additionally, the mm-wave access clustering is expected to work even with wireless self-backhauling, where the nodes may relay traffic using the mm-wave air interface. However, this may introduce additional latencies in the system which needs to be considered.

3.6.5 Mobility Management

Mobility in the 5G framework needs to cover use cases with active users in RRC Connected state and low activity users in RRC Connected Inactive state and in RRC Idle state. In addition, 5G must support the tight interworking between LTE and 5G (NR) for mobility users (see Section 3.4).

The 5G mobility framework consisting of several new methods including UE autonomous mobility, make-before-break handover and mobility concepts for URLLC.

Centralization of the UE AS context, RRM, and multi-connectivity have potential to improve the efficiency of various Connected Inactive and Connected state mobility and multi-connectivity procedures. **Inter-RAT mobility between LTE and 5G allows seamless inter-RAT mobility and multi-connectivity by anchoring the RAN/CN interface to the LTE or 5G network node. Inter-RAT mobility can support UEs in both RRC Light Connected in LTE and RRC Connected Inactive in 5G state thus allowing low energy dissipation and fast state transition to Connected state from either of the systems in inter-working scenarios**. Finally, **UE context information is used to enhance the accuracy of mobility prediction enabling a uniform service experience for the user, even in deep shadow regions or coverage holes.** User profiles are used for predicting the future network requirements. This approach reduces the signalling cost for context information transfer.

In addition, efficient mobility management in the mmWave RAT is required for a seamless service experience for users on the move. As the mmWave RAT supports multi-connectivity (between different mmWave nodes or inter-RAT between mmWave RAT and LTE-A), the mobility management should support the required coordination between the nodes even

when the transport network to which they are connected is capacity-limited or adds up latency. As previously outlined in Section 3.6.4, mmWave AP clustering is instrumental to support mobility in a mmWave RAT. Intra-AP beam switching from one beam to another within a cluster can be supported by UE measurement feedback and direct communication between UE and AP on the beam switch. In case of inter-AP beam switching, the report will be forwarded to the cluster head from the current serving AP. The cluster head will request the target AP for beam switching. If positive feedback is received from the target AP, the UE will be eventually informed (via cluster head and serving AP) to switch its beam and will be served via target AP after the switch. In case of inter-RAT handover, the maximum data rate supported by the technology and expected rate fluctuations (as a result of a handover) may be shared as cross-layer information with application level services to avoid frequent rate variations, impacting the end user QoE.

3.6.6 Self-backhauling

A further step of wireless backhauling is self-backhauling, which refers to a set of solutions to provide technology- and topology-dependent coverage extension and capacity expansion utilizing same frequency band for both backhaul and access links, as shown in Figure 3-13. Self-backhauling provides an efficient way to combat infrastructure constraints especially in dense network deployment, where access to fibre may be limited to only some APs. However, self-backhauling also brings challenges to the radio resource management (RRM) between backhaul and access links, which leads to joint backhaul and access RRM optimize system efficiency.

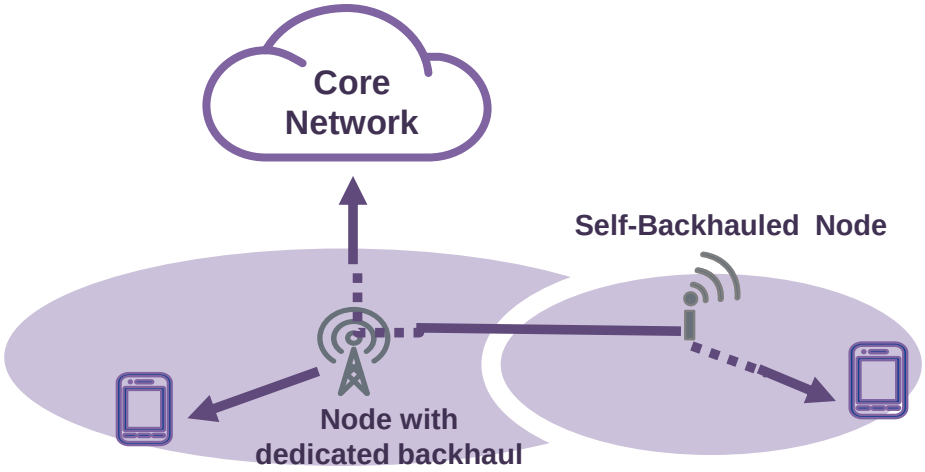


Figure 3-13: Concept of Self-backhauling

The joint optimization problem is mathematically decomposed into transmission link scheduling, transmission duration and power allocation governed by a set of constraints. The scheduling and resource allocation algorithm is further proposed to exploit space division multiple access (SDMA) that allows non-conflicting links to be transmitted simultaneously. The proposed solution exploits self-backhauling within a unified backhaul/access optimization framework. In the following, we describe the framework assuming non-standalone deployment. However, this can be also applicable to a standalone network with internode coordination.

Here, we assume backhaul and access links share the same air interface, and all network elements (including BS, APs and UEs) are equipped with directional steerable antennas and can direct their beams in specific directions. The BS processes transmission link scheduling and adjusts transmission duration and power on both backhaul and access links. Figure 3-14 shows an example of considered HetNet.

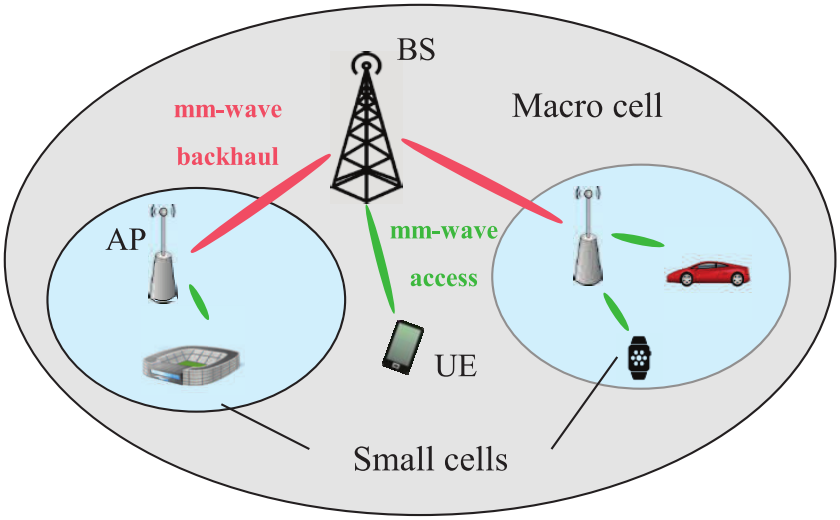


Figure 3-14: Illustration of a HetNet with mm-Wave wireless BH and access

Monte Carlo simulations are used to evaluate the efficiency of the proposed algorithms in enhancing user throughputs. For the evaluation, we consider a HetNet deployed under a single Manhattan Grid, where square blocks are surrounded by streets that are 200 meters long and 30 meters wide. One BS and four APs are located at the crossroads. 100 UEs are uniformly dropped in the streets.

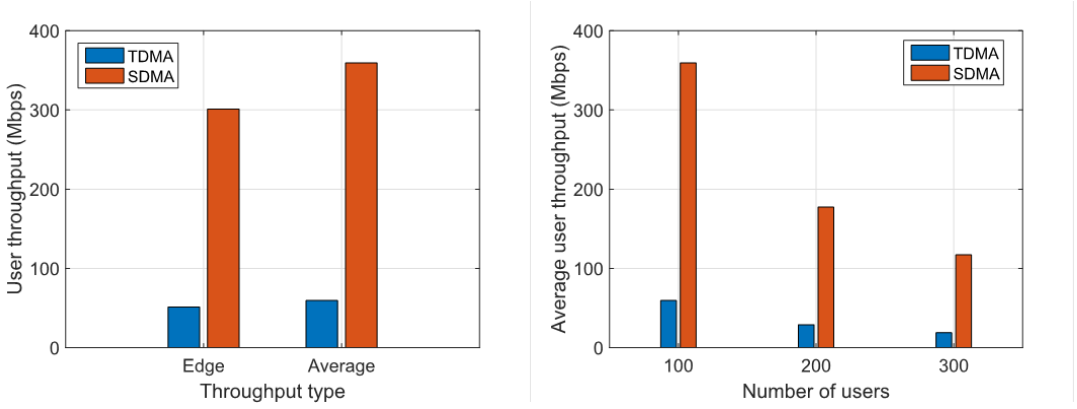


Figure 3-15: Comparison of user throughputs for carrier frequency of 28GHz and bandwidth of 1GHz: (a) edge/average user throughputs for 100 users; (b) average user throughputs for different number of users.

Figure 3-15(a) shows the simulation results of user throughput at carrier frequency of 28 GHz and bandwidth of 1 GHz. Here, cell edge user throughput is defined as 5-th percentile point of the cumulative distribution function of user throughputs. Compared to the benchmark TDMA scheme, our proposed algorithm provides considerable improvement in both edge user throughput and average user throughput due to exploiting spatial multiplexing that allocates more time resources to each link in the network by allowing multiple links to transmit concurrently. Figure 3-15(b) shows the simulation results of average user throughputs for different numbers of users in the network. On one hand, as expected, increasing the number of users reduces average user throughput due to limited bandwidth. However, enabling space dimension still achieves high user throughput in the case of 300 users, and provides significant improvement compared to the benchmark scheme. On the other hand, as user density increases, gain of proposed scheme to TDMA scheme also grows (604 percent, 614 percent and 623 percent by the proposed algorithm against TDMA for 100, 200 and 300 users, respectively). This is mainly because with the increasing number of users, allocable slots for each link in TDMA scheme are limited and become

dominant factor in determining user throughputs, consequently user throughputs benefit more from the spatial multiplexing gain.

3.7 References

[3-1] I. Da Silva et. al, "Tight integration of new 5G air interface and LTE to fulfil 5G requirements", 2015.

[3-2] 5G PPP 5G Architecture – (White Paper) Updated July 2016, https://5g-ppp.eu/white-papers/

[3-3] METIS II D2.2 Draft overall 5G RAN design, https://metis-ii.5g-ppp.eu/documents/deliverables

[3-4] LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (3GPP TS 36.300 version 13.3.0 Release 13)

[3-5] SPEED-5G Deliverable, "D5.1: MAC Approaches with FBMC and Simulation Results (First Version)", ICT-671705, H2020-ICT-2014-2, December 2016.

[3-6] SPEED-5G Deliverable, "D3.2: SPEED-5G enhanced functional and system architecture, scenarios and performance evaluation metrics", ICT-671705, H2020-ICT-2014-2, June 2016.

[3-7] METIS-II, 5G RAN Architecture and Functional Design White Paper, https://metis-ii.5g-ppp.eu/wp-content/uploads/5G-PPP-METIS-II-5G-RANArchitecture-White-Paper.pdf.

[3-8] METIS II D5.2 Final Considerations on Synchronous Control Functions and Agile Resource Management for 5G, https://metis-ii.5g-ppp.eu/documents/deliverables

[3-9] A. Benjebbour et.al. "Non-orthogonal Multiple Access (NOMA): Concept, Performance Evaluation and Experimental Trials", 2015

[3-10] SPEED-5G Deliverable, "D5.2: MAC approaches with FBMC (final)", June 2017, https://speed-5g.eu.

[3-11] SPEED-5G Deliverable, "D4.2: RM framework and modelling", ICT-671705, H2020-ICT-2014-2, April 2017, https://speed-5g.eu.

[3-12] 3GPP 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2

[3-13] 3GPPP 38.804, Study on new radio access technology Radio interface protocol aspects

[3-14] R3-171287, Status of Higher-Layer Functional split between Central and Distributed unit

[3-15] mmMAGIC, D5.1 Initial multi-node and antenna transmitter and receiver architectures and scheme, https://5g-mmmagic.eu/results/

4 Physical Infrastructure and Deployment

This section considers the physical infrastructure and deployment for 5G mobile radio networks. The physical infrastructure comprises the interconnection network between the different 5G clouds which is now expressed referring to the distance between the user and the place where the function is executed. The central cloud will be connected to many EDGE clouds via backhaul network. The EDGE clouds, placed at maximum 20Km to the user, will be connected to the antenna site via fronthaul network.

The Mobile access EDGE computing enhances the infrastructure through the deployment of functionality like Traffic Offload Function able to execute services in the EDGE and well adapted to the RAN measurements or situation (congestion, interference, etc). This functionality is dependent on the way the Cloud RAN is designed and split is implemented. This is described in Section 4.1.

The implemented split in the Cloud RAN depend on a set of parameters: supported services, services requirements, technology of the fronthaul and the backhaul, the used protocol in the fronthaul and backhaul, etc. In section 4.2, we describe the unified data plane architecture including FH/BH traffic and an evaluation of the benefits of a 5G infrastructure supporting FH and BH, i.e. end user, services.

The Physical infrastructure plays an important role in the deployment and execution of services. From this perspective, monitoring of platform and services is critical for 5G networks in order to provide and assure a desired service performance level during the lifecycle of the services. In section 4.3, the monitoring workload performance is described.

The 5G-PPP phase 1 results do not cover all the challenges related to the physical infrastructure and the deployment. A set of problem are in the track of many projects in phase 2. One can cite, the workload prioritization and network slicing, Lifecycle of infrastructure, Accountability/ Authorisation/Billing, Infrastructure security, etc.

4.1 Physical infrastructure improvements

Multi-access Edge Computing (MEC), which formerly stood for Mobile Edge Computing, offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the mobile network. This environment is characterized by ultra-low latency and high bandwidth, as well as real-time access to services that can be leveraged by applications. Examples of these services are, radio network information or location.
MEC provides a new ecosystem and value chain. Operators can open their networks edges, e.g. Radio Access Network (RAN), to authorized third-parties (e.g. app providers), allowing them to rapidly deploy innovative applications and services towards the subscribers, enterprises and vertical segments.
Multi-access Edge Computing will leverage new vertical business segments and services for consumers and enterprise customers. MEC Use cases include, among others:

- Video analytics
- Location services
- Internet-of-Things (IoT)
- Augmented reality
- Optimized local content distribution and
- Data caching

It uniquely allows software applications to tap into local content and real-time information about local-access network conditions. By deploying various services and caching content at the network edge, core networks are alleviated of further congestion and can efficiently serve local purposes.
One of the important features in the MEC is the Traffic Offloading Function (TOF) enhancing

the infrastructure to be able to re-route the traffic from a global view to a local view.
The TOF has three data-plane interfaces (BBU, Serving GW and Apps) and a single control interface (API). The data-plane carries user-data to/from the respective elements it connects. The BBU is the virtual radio access, the Serving GW is a data plane EPC element, and Apps are running at the edge cloud and require access to the data plane.
The data-plane interfaces should be seen as network adapters. Although there are three interfaces in the data-plane, the two that connect to the BBU and Serving GW can be merged into a single network adapter that connects to the S1 interface.

The tight relationship between the MEC and RAN is dependent on the way the RAN is deployed or splitted between the Remote Unit and the Central Unit called also EDGE cloud.
The split option impacts the fronthaul network which is analysed below.

The allocation of functions between the RU and CU, i.e., the functional split, has a major impact on the transport network and the corresponding NGFI requirements regarding data rate, latency and synchronization, and by the way to the information accessible by the MEC. Figure 4-1 depicts a generic RAN signal processing chain of a mobile network. In principle, the split between RU and CU may be between any of the blocks depicted. We focus on three functional splits, which are capturing the most relevant trade-offs, denoted as A, B, and C in Figure 4-1. In general, splits higher up in the processing chain offer less centralization gains in terms of RRU size and cooperative processing, while reducing the requirements in terms of fronthaul data rate, latency, and synchronization. In Figure 4-1, split A places antenna processing in the RU, thus achieving a scalable way to accommodate 5G technologies such as Massive-MIMO into current C-RAN architectures based on CPRI. Split B is based on the transport of frequency domain symbols, which, unlike CPRI, varies according to cell-load and enables statistical multiplexing gains. Finally, split C represents an upper-MAC, or PDCP-like split, and may represent backhaul type of interfaces. The interested reader can found a detailed analysis of these functional splits in [4-6].

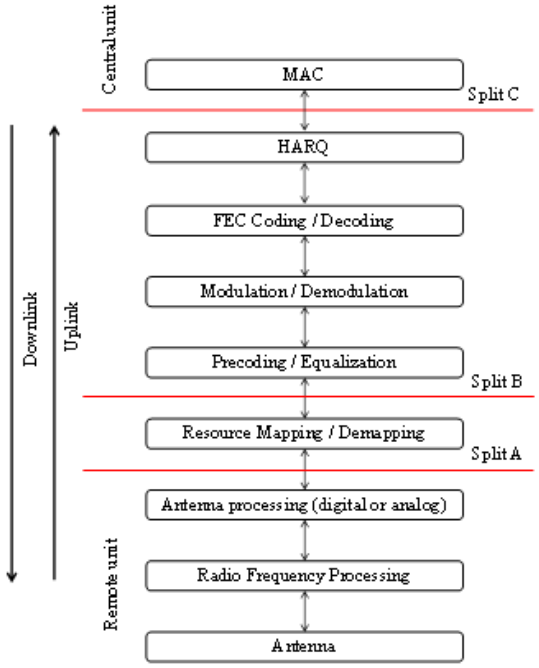


Figure 4-1: Functional split options

Parameter:	Symbol	Unit	4G	5G		
			LTE	Sub-6	Low mmWave	High mmWave
Carrier Frequency	f_C	GHz	2	2	30	70
Channel Size	BW	MHz	20	100	250	500
Sampling Rate	f_S	MHz	30.72	150	375	750
# Antennas	N_A	-	4	96	128	256
# ADC/DAC chains/ layers	N_L	-	4	16	12	10
Overhead (control, line coding)	γ	-	1.33	1.33	1.33	1.33
Quantizer resolution time domain	$N_{Q,T}$	Bit	15	15	12	10
Quantizer resolution frequency domain	$N_{Q,F}$	Bit	9	9	8	7
Modulation order	M_{MCS}	-	64	1024	256	64
Max. code rate	$R_{C,MCS}$	-	0.85	0.85	0.85	0.85
Frame duration	T_F	ms	1	1	1	1
FFT size	N_{FFT}	-	2048	2048	2048	2048
# Active subcarriers	$N_{SC,act}$	-	1200	1300	1300	1300
# Data symbols per frame	N_{Sy}	-	14	70	150	300
Peak utilization	μ	-	1	1	1	1
Resulting requirements:						
Channel coherence time at *v*=3 km/h	$T_{C,3}$	ms	76.17	76.17	5.08	2.18
Channel coherence time at *v*=50 km/h	$T_{C,50}$	ms	4.57	4.57	0.30	0.13
Channel coherence time at *v*=500 km/h	$T_{C,500}$	ms	0.46	0.46	0.03	0.01
Delay accuracy	T_j	ns	65	13.33	5.33	2.67
Peak data rate Split A	D_A	Gbps	4.9	95.8	143.6	199.5
Peak data rate Split B	D_B	Gbps	1.6	34.9	49.8	72.6
Peak data rate Split C	D_C	Gbps	0.46	16.5	21.2	26.5

Table 4-1: Parametrization of considered 5G RANs [4-6].

Dimensioning Peak Requirements

Following the parametrization of three considered 5G-RATs in Table 4-1, the peak requirements per cell of these 5G RATs are summarized at the bottom of Table 4-1 and are illustrated in Figure 4-2. Note that while the parameters of 5G RATs are currently not standardized, the numbers provided are a realistic outlook to the transport requirements if the RATs discussed are indeed implemented. Note for example that the data rate given for Split C is between 16 and 26 Gbps. Considering that Split C is close to the data rates experienced by the user, but still includes MAC and transport overhead, these data rates are in line with the capacity of up to 10 Gbps considered for 5G. Looking at the delay accuracy and channel coherence times, we can see that they differ by orders of magnitude depending on the different RATs. This already indicates that future transport networks will have to deal with a large variety of requirements.

For the lower functional splits, the data rates range from 35 to almost 200 Gbps per cell. However, these are peak requirements. We have already remarked that a key benefit of Splits B and C is that the associated transport data rates scale with the actual utilization of resources on the access link. Thus, when using these splits, dimensioning the transport network according to peak data rates is not efficient in most cases. Instead, the NGMN Alliance has derived guidelines for dimensioning transport networks based on busy hour utilization, which leverage the statistical multiplexing occurring in practical networks [4-4].

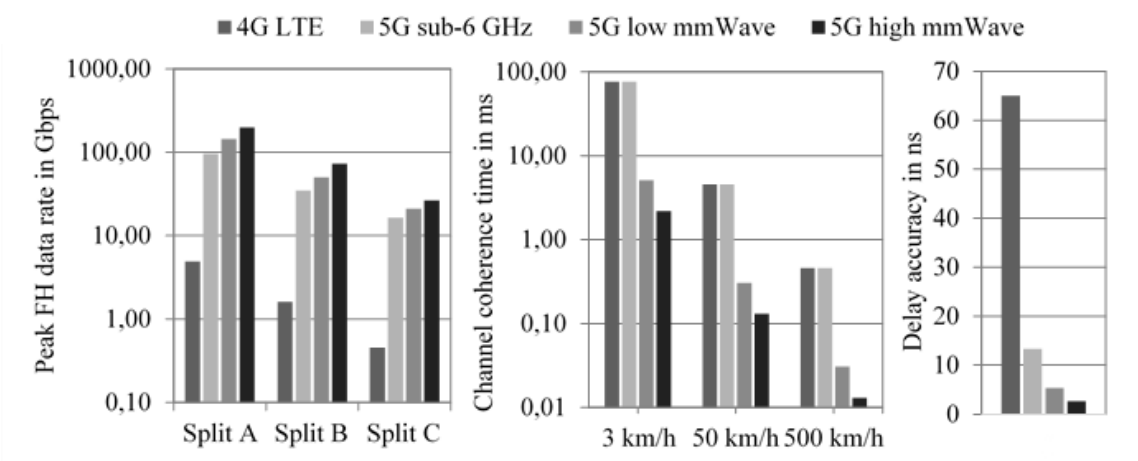


Figure 4-2: Peak 5G FH data rate, channel coherence time and delay accuracy requirements

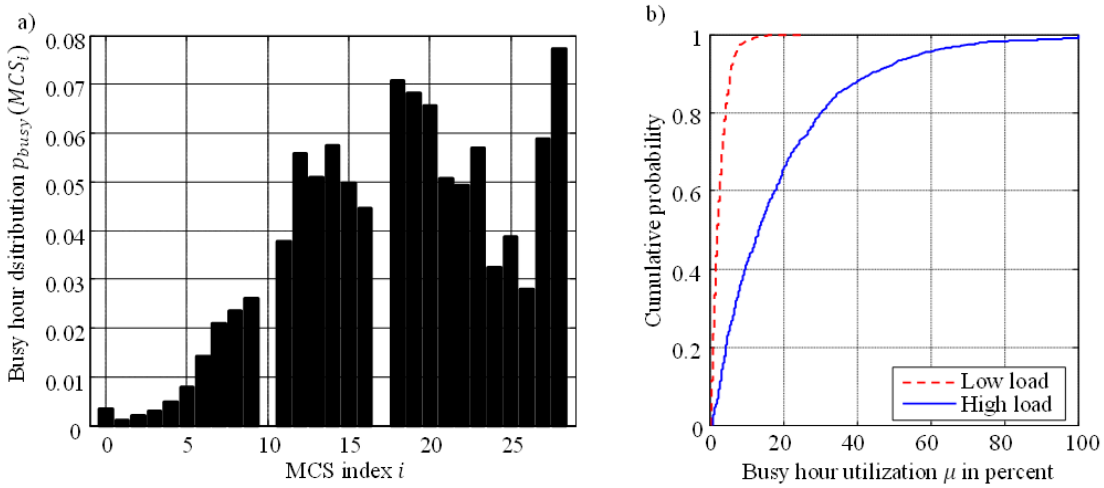
Dimensioning based on Live Network Measurements

The values reported in Figure 4-2 do not account for the statistical multiplexing gains often encountered in real networks, and which are the main motivation to use Splits B or C. To this effect, we combine the 5G RAT configurations from Table 4-1 with real downlink measurements from a real-life LTE network. The measurements were collected from 10 LTE sites with C=33 cells in total, covering an area of about 2.5 km^2 in downtown Athens, Greece. 27 of the 33 cells utilize 20 MHz bandwidth, while the other 6 cells utilize 10 MHz. The measurements were taken on a 15-minute-basis for a time-period of 15-days.

While these measurements are based on a 4G network, we assume that similar traffic patterns will be observed for the mobile broadband use-case of 5G. However, to model 5G we scaled the empirical load measured in the 4G network as described in the following. We assume that each user has an average traffic demand of D_5G=300 Mbps as recommended for the downlink by the NGMN alliance for broadband access in dense areas in [4-4] and that the spectral efficiency of a 5G network would be approximately 5 times higher than in a current LTE network. Accordingly, the load projection for a 5G network is defined as:

$$\mu'(c,t) = \mu(c,t)\frac{D_{5G}}{D_{\max}(c,t)/N_{\mathrm{UE}}(c,t)\cdot 5}.$$

Hereafter we will use two load scenarios to study the requirements of 5G networks. The measured utilization in the 4G network as μ ("Low load scenario"), and the scaled utilization μ' ("High load scenario"). We expect the real requirements of a 5G network to lie between these two scenarios.
In order to dimension the transport network appropriately, we need to consider the busy hour traffic, which in our live network is found to be from 12:15 h to 13:15 h. Consequently, a busy hour MCS distribution was calculated as the average MCS distribution in that hour, which is depicted in Figure 4-3(a). Note that MCS 10 and 17 are very close to their adjacent MCS in terms of spectral efficiency and are hence not utilized in this network. Finally, the loads of all cells observed in the busy hour were used to accumulate a discrete busy hour load distribution illustrated in Figure 4-3(b).



(a) (b)

Figure 4-3: MCS distribution in the busy hour, and resulting load CDF distributions in the high load and low load scenarios

To fully benefit from the statistical multiplexing that occurs when multiple cells are aggregated under a single transport link, we need to consider a certain outage rate. For example, transport networks are typically dimensioned to guarantee a certain percentile, e.g. the 95th percentile Q_95, i.e., the offered traffic can be transported without queueing with a probability of 95%.
Figure 4-4 illustrates the resulting capacity to be deployed in aggregation networks for the different functional splits, RATs and for both the high load and low load scenarios. We give the capacity of up to 1000 cells, which we consider to be the approximate order of magnitude for the number of cells in a single large city.
First, note that according to the overall increase in data rates considered for 5G networks, the transport data rates are also expected to increase by order of magnitude and can easily reach the Tbps range. However, the introduction of the lower functional Splits B and C can mitigate this effect, and statistical multiplexing can further reduce the required capacity by almost 10 times as seen in Figure 4-4. The degree of statistical multiplexing gain depends on the number of aggregated cells but shows already high values for a few dozens of cells. In addition, Split C, which exhibits the highest variability of per-cell traffic due to its dependence on both the overall load and the utilized MCS, shows the highest gains, while the static rates of split A do not offer any statistical multiplexing. These factors need to be considered in the design of future transport networks in order to make them economically feasible.

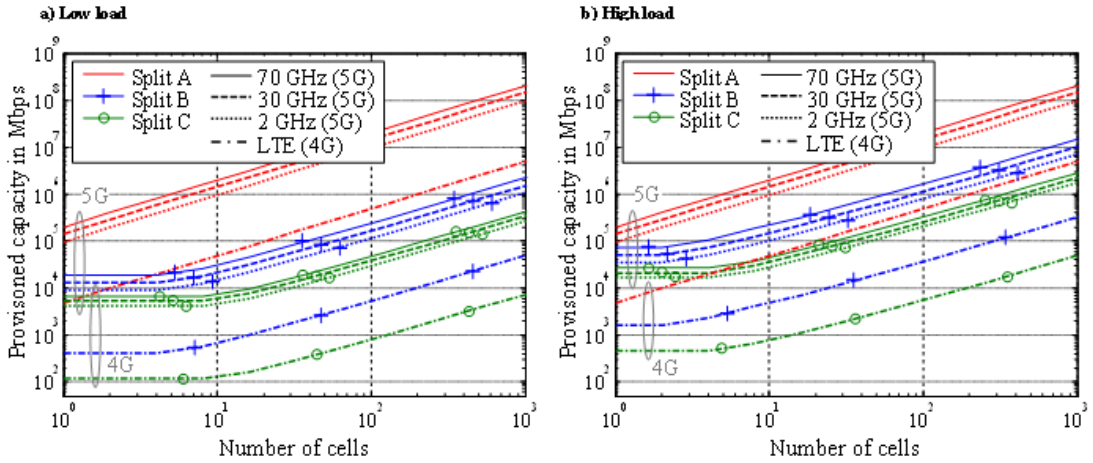


Figure 4-4: 5G Transport data requirements for the low and high load scenarios, functional splits and 5G RANs

4.2 Physical access network

In order to achieve the required 5G data rates, extensive support for novel air interface technologies such as Cooperative Multipoint (CoMP), Carrier Aggregation (CA) and Massive MIMO will be needed. Such technologies require processing of information from multiple base stations simultaneously at a common centralized entity and also tight synchronization of different radio sites. Hence, backhaul and fronthaul will have to meet the most stringent requirements not only in terms of data rates but also in terms of latency, jitter and bit error rates. The first subsection presents the unified data plane architecture including FH/BH traffic. The second subsection evaluate the benefits of a 5G infrastructure supporting FH and BH, i.e. end user, services

Integrated fronthaul and backhaul networks

Figure 4-5 shows a unified data plane architecture, including circuit- and a packet- switched paths, for the integrated fronthaul and backhaul networks, namely Crosshaul networks. This two-paths switching architecture is able to combine bandwidth efficiency (through statistical multiplexing in the packet switch) while deterministic latency is ensured by the circuit switch.

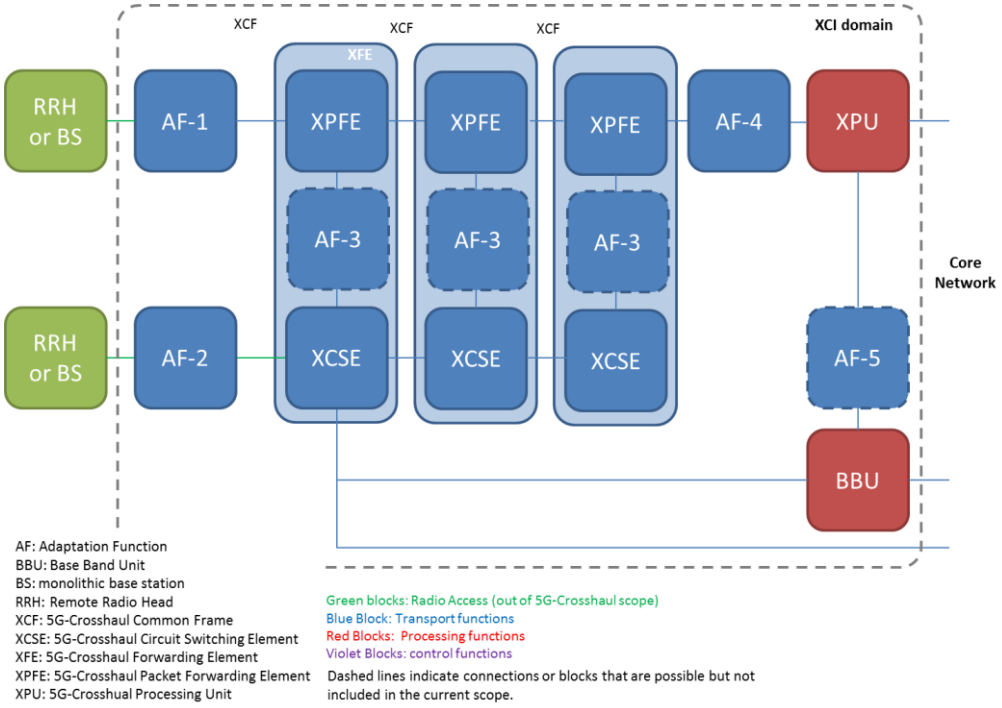


Figure 4-5: Data plane architecture [4-1]

The fundamental block of the data plane architecture is the XFE (Crosshaul Forwarding Element) that, in the most general implementation, is a multi-layer switch, made up of a packet switch called the Crosshaul Packet Forwarding Element (XPFE) and a circuit switch called the Crosshaul Circuit Switching Element (XCSE).

The packet switching path is the primary path for the transport of most delay-tolerant fronthaul and backhaul traffic, whereas the circuit switching path is there to complement the packet switching path for those particular traffic profiles that are not suited for packet-based transporting (e.g. legacy CPRI or traffic with extremely low delay tolerance) or just for capacity offloading. The modular structure of the switch, where layers may be added and removed, enables various deployment scenarios with traffic segregation at multiple levels, from dedicated wavelengths to VPN, which is particularly desirable for multi-tenancy support.

The radio access units (the green blocks in Figure 4-5) support flexible functional splits, with some functions of the access interface virtualized and placed at the cell site, and their complementary functions virtualized and pushed to the baseband processing nodes. The fronthaul interface between the remote access unit and the central access unit for a base station can be any existing or new interface, such as CPRI or future packet based fronthaul interfaces.

Although it is not made explicit in Figure 4-5, some aggregation may be performed by the radio units before interfacing the XFE in order to decrease the number of data flows, increase the bit rate on the XFE ports and simplify the XFE implementation. The first example is provided by a cascade of RRHs, where CPRI traffic is added and time-multiplexed at each RRH. In the second example, client signals at different radio carrier frequencies are multiplexed in a RoF system and the aggregate signal is converted from analogue to digital.

The radio units are connected to the XFE by means of adaptation functions (AFs) that perform media and protocol adaptation. The purpose of the adaptation function AF-1 is media adaptation (e.g. from air to fibre) and translation of the radio interface (CPRI, new 5G fronthaul packet interfaces, Ethernet used in backhaul links, mmWave/802.11ad frames, analogue radio over fibre, etc.) into a Crosshaul Common Frame (XCF), that interface the XPFE. Similar to AF-1, the

adaptation function AF-2 maps the radio interface into the protocol used by the XCSE, e.g. OTN or the simpler circuit framing protocol. The XCF is a packet interface based on an evolution of the Ethernet MAC-in-MAC standard, namely Provider Backbone Bridge Traffic Engineering (PBB-TE) [4-2], adding mechanisms to deal with time sensitive applications. The XPFEs talk to each other using the XCF. XCF is also the interface between XPFE and Crosshaul Processing Unit (XPU), the virtualized unit in charge of hosting baseband processing and other virtual functions. An adaptation function, AF-4, could be needed.

The XPFE may be connected to the XCSE, through the adaptation function AF-3 that maps the XCF into the protocol used by the XCSE. As a further advantage, this connection can be used to offload the XPFE, avoiding overload situations and therefore decreasing the probability of discarded packets. Another adaptation function, AF-5, will also be necessary if pre-existing BBUs needs to be interfaced with 5G XPUs.

The XPFE consists of a common switching layer that supports the XCF format in the packet switching path across the various traffic types (of fronthaul and backhaul) and the various link technologies in the forwarding network [4-9] The XCF is a frame format which defines how the frames in the common switching layer look like. The corresponding control of the forwarding, i.e. which frame is forwarded by an XPFE in which way is defined in the SDN controller. Especially when using an existing frame format such as Ethernet or MPLS, this does not imply that the XFEs have to understand all the related control protocols, the control of forwarding remains is managed by control plane. This is in line with the SDN approach, where all control aspects are moved to a logically centralized controller.

Different protocol stacks for deploying networks for LTE backhaul traffic have been described in [4-10]. Depending on the physical topology different protocols can be used for deployment of the backhaul network. In case of LTE, the backhaul traffic is IP-based. Depending on the functional split and its implementation, fronthaul traffic can be also IP-based, see [4-11]. Therefore, the XCF has to support a larger variety of services to transport different fronthaul and backhaul traffic types.

The XCF has to contain enough information in the frame headers to enable the XPFEs to fulfill their task as a common switching layer for both fronthaul and backhaul traffic. Obviously, different functional splits, as well as multiple tenants, have to be supported. To enable the migration to the future 5G transport network, it must be possible to interact with legacy devices. A frame format based on MPLS, specifically on MPLS-TP, is a viable alternative as XCF format. A comparison of them is studied in [4-12], however no clear advantages or disadvantages of MAC-in-MAC or MPLS-TP as XCF have been identified.

The MAC-in-MAC header contains a new Ethernet header with MAC addresses, a B-TAG (Backbone VLAN Tag) to support VLANs, and an I-TAG (Backbone Service Instance Tag) to support further service differentiation. The outer MAC addresses are used to address the XPFEs. The destination B-MAC address is the MAC address of the XFE to which the tenant device, identified by C-Dest address, is connected. The B-VLAN tag contains the VLAN-ID in the provider network as well as the priority code points (PCP) used to prioritize the packets appropriately. The PCP and DEI values of the I-Tag are redundant to the ones in the B-Tag and are not used in 5G-Crosshaul. The UCA (Use Customer Address) is used to indicate whether the addresses in the inner header are actual client addresses or whether the frame is an OAM frame.

Benefits of a 5G infrastructure supporting FH and BH

A suitable data-plane for the 5G transport involves converged optical and wireless network domains that can support both transport and access. In the wireless domain, a dense layer of small cells can be wirelessly backhauled to macro-cells through mm-Wave and sub-6 technologies. Alternatively, the proposed architecture allows connecting small cells with a CU through an active

optical network. This adopts a hybrid approach combining a dynamic and elastic frame based optical network solution with enhanced capacity WDM PONs, to support the increased transport requirements of 5G environments in terms of granularity, capacity and flexibility. It should be noted, that this design is featured in the 5G PPP version 1 architectural vision [4-5].

Although the adoption of the C-RAN concept can overcome traditional RAN limitations, it introduces the need to support new operational network services (FH) over the transport network. These emerge from the need to connect densely distributed RUs with compute resources handling the BB processing, meeting very tight latency and synchronization requirements. To maximize coordination and resource sharing gains it is proposed to support BH and FH jointly in a common infrastructure. Thus, efficiency improvement and management simplification can be achieved leading to measurable benefits in terms of cost, scalability and sustainability. Aiming to relax the C-RAN challenges described flexible split options that can relax the tight transport requirements in terms of capacity, delay and synchronization can be adopted. "Optimal split" options, span between the "traditional distributed RAN" case where "all processing is performed locally at the AP" to the "fully-centralized C-RAN" case where "all processing is allocated to a CU". All other options allow allocating some processing functions at the RU, while the remaining processing functions are performed remotely at the CU. The optimal allocation of processing functions to be executed locally or remotely i.e. the optimal "split", can be decided dynamically based on a number of factors such as transport network characteristics, network topology and scale as well as type and volume of services that need to be supported.

Given the technology heterogeneity supported by the envisioned 5G data plane, a critical function of the converged infrastructure is interfacing between technology domains. The required interfaces are responsible for handling protocol adaptation as well as mapping and aggregation/de-aggregation of the traffic across different domains. Different domains (wireless/optical) may adopt different protocol implementations and provide very diverse levels of capacity, granularity etc. A key challenge also addressed by these interfaces, is mapping of different Quality of Service (QoS) classes across different domains as well as flexible scheduling enabling QoS differentiation mechanisms. More specifically, according to the approach put forward by the xHaul project [4-7], a dense layer of small cells interfaces first with a passive optical network based on WDM-PON technology, which bring the traffic until the Central Office (CO), where it interfaces with an active optical network based on Time Shared Optical Networks (TSON), at the metro level. Thus, at the optical network ingress point the interfaces receive traffic frames generated by fixed and mobile users and arrange them to different buffers. The incoming traffic is aggregated into optical frames and is assigned to suitable time-slots and wavelengths according to the adopted queuing policy, before transmission in the TSON domain. For FH traffic a modified version of the CPRI protocol supporting the concept of functional split (eCPRI) has been adopted. It should be noted that due to the large variety of technologies involved in 5G, these interfaces need to support a wide range of protocols and technology solutions and execute traffic forwarding decisions at wire-speed. This requires the development of programmable network interfaces combining hardware level performance with software flexibility. The reverse function is performed at the egress points.

In the case of C-RAN with vBBUs, where optimal split options are deployed, two types of servers have been considered in phase 1 (xHaul project): a) small scale commodity servers (cloudlets) close to the APs and, b) commodity servers hosted by large scale DCs. Although both types of servers can provide the necessary processing power for C-RAN and CDN services, large scale DCs provide superior performance per Watt, compared to cloudlets. The figure below shows that significant energy savings (ranging between 60-75%) can be achieved when adopting the C-RAN approach using the integrated wireless-optical infrastructure, compared to traditional RAN. However, due to sharing of network resources between BH services and high priority FH services, C-RAN leads in increased BH service delays that remain below 25 ms. On the other hand, traditional RAN provides minimum end-to-end BH service delays, as no sharing with FH services is required, but at the expense of increased energy consumption due to the lack of BBU sharing.

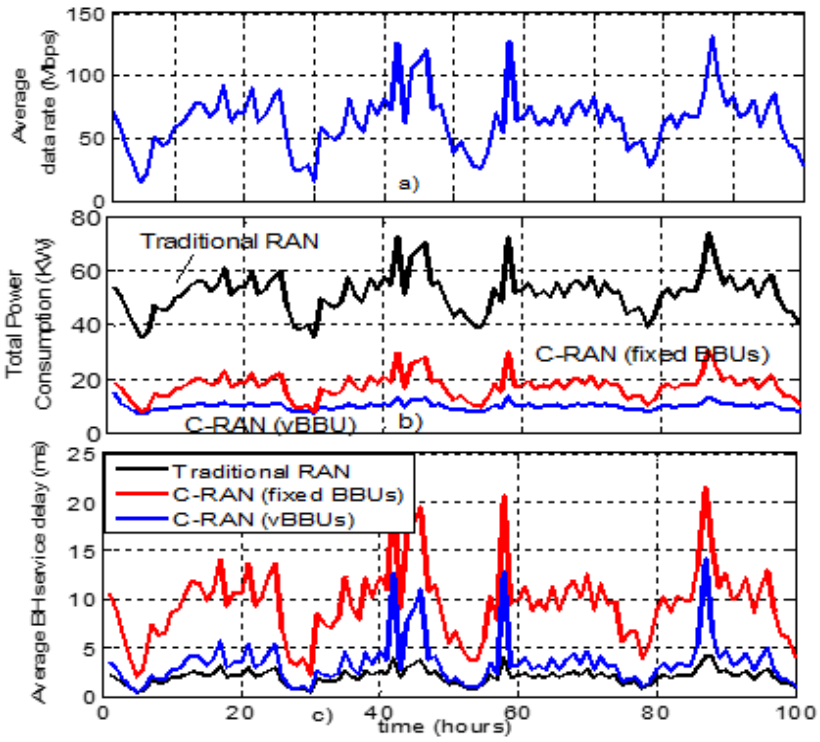


Figure 4-6: a) Bristol 5G city network topology with mmWave backhauling, b) Snapshot of spatial traffic load and c) average traffic/BS based on the dataset [10] during 8/2012, d)-e) Total power consumption and total service delay over time for the traditional RAN.

Figure 4-7 illustrates the impact of service requirements in terms of network and compute resources on the optimal split option adopted. Services with high network-to-compute ratios require significant network resources to operate, leading to overutilization of transport capacity. This effect is counter-balanced by the selection of higher split options that require lower bandwidth for the interconnection of RUs with CUs compared to the bandwidth requirements of lower split options (i.e. options 1, 2). Our modelling results have shown that for higher traffic load, the total power consumption increases.

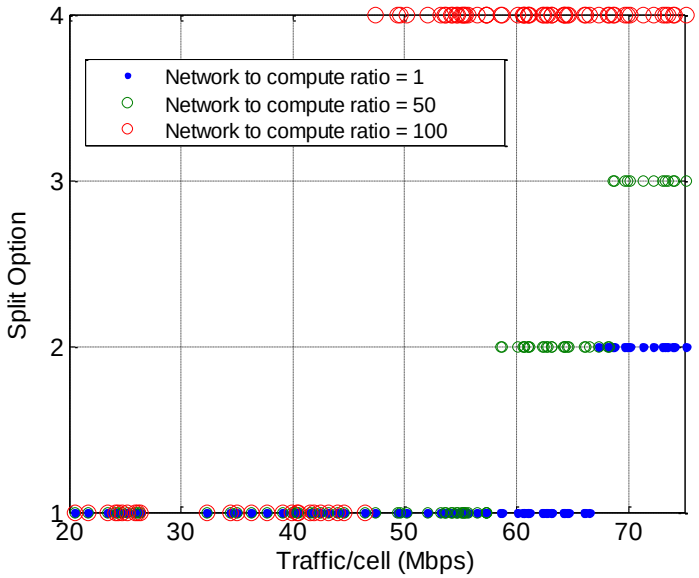


Figure 4-7: Split option as a function of load for different Compute to network ratios

The interested reader is referred to [4-8] for an in-depth evaluation of an joint FH-BH 5G network infrastructure.

4.3 Monitoring of workload performance

One of the most important requirements for a fully operating 5G Infrastructure is the deployment and execution of services in efficient, fast and scalable manner. From this perspective, monitoring of platform and services is critical for 5G networks in order to provide and assure a desired service performance level during the lifecycle of the services.

The problem requires to be addressed from at least two different perspectives: white-box perspective, assuring the internal configuration of each service component supports the desired level of performance, making the best use of service functionalities and avoiding the creation of bottlenecks; black-box perspective, observing and analysing the usage of the infrastructure resources and the relation that they have with the service performance.

For this reason an empirical methodology called TALE has been introduced as technique to profile services to be deployed on a 5G infrastructure and improve their performance. The methodology is based on full stack telemetry to collect metrics related to Throughput, Anomalies, Latency and Entropy (TALE) related to the service under investigation stressed under dynamic load conditions. This methodology aims at gaining understanding from the inside of the service to identify potential issues or available improvements that depend on the internal configuration of the service. An analytical approach has then been added to identify key server side platform performance metrics empirically that strongly correlate with client performance indicators (i.e. throughput, latency, etc.), called KPI Mapping. The combination of the two methodologies has been applied on Unified Origin service [4-3] deployed on an OpenStack environment and stress-tested using Citrix Hammer packet generator.

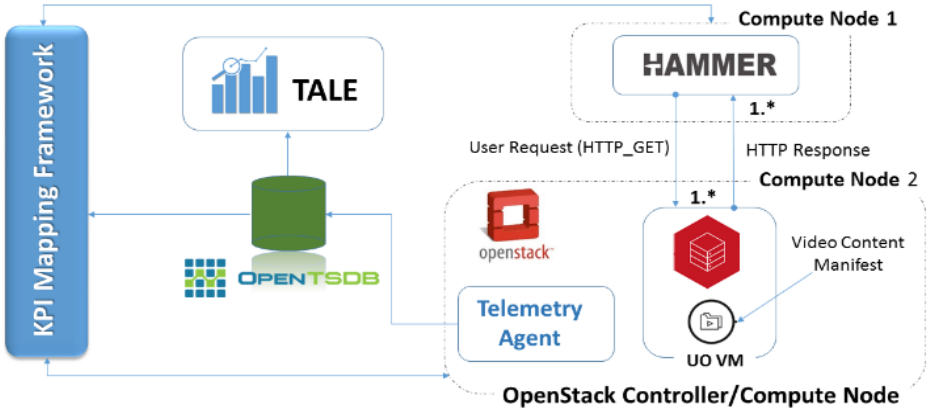


Figure 4-8: TALE and KPI mapping methodology

The structured approach of TALE codifies experiential knowledge of cloud workload characterization by focusing on the system areas and associated metrics which will mostly likely yield opportunities for bottleneck identification or performance optimization opportunities. In this context it important to consider the layered constructs of virtualized cloud workloads and the explicit relationships between the physical and virtual resources. In order to elucidate these relationship, TALE uses the automated collection of targeted system metrics coupled with analytics and visualization.

The execution of the KPI Mapping approach aims at identifying the most relevant platform metrics related to a Key Platform Indicator (KPI) for a service from an infrastructure perspective.

This opens the door to a simpler and more efficient monitoring of complex network systems, reducing the size of the telemetry to look at for service assurance. The methodology is based on a full end-to-end analytics pipeline that processes telemetry data and applies different features selection approaches to select those metrics of relevance for the satisfaction of Service Level Objectives.

4.4 References

[4-1] 5G-Crosshaul, "Detailed analysis of the technologies to be integrated in the XFE based on previous internal reports from WP2/3", Deliverable 2.1, July 2016.

[4-2] IEEE 802.1 Task Group, IEEE 802.1ah-2008 - IEEE Standard for Local and metropolitan area networks – Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges.

[4-3] http://www.unified-streaming.com/products/unified-origin

[4-4] NGMN Alliance, Guidelines for LTE backhaul traffic estimation, white paper, 2011. http://www.ngmn.de/uploads/media/NGMN_Whitepaper_Guideline_ for_LTE_Backhaul_Traffic_Estimation.pdf. Accessed 1 Aug 2016

[4-5] Chapter 5 of 5G PPP Architecture Working Group, "View on 5G Architecture, version 1.0", white Paper, July 2016.

[4-6] Bartelt, J., Vucic, N., Camps-Mur, D., Garcia-Villegas, E., Demirkol, I., Fehske, A., Grieger, M., Tzanakaki, A., Gutiérrez, J., Grass, E. and Lyberopoulos, G. 5G transport network requirements for the next generation fronthaul interface, EURASIP Journal on Wireless Communications and Networking, 2017(1), p.89.

[4-7] 5G-XHaul Project, Deliverable 2.2, "System Architecture Definition," 1 July 2016. [Online]. Available: http://www.5g-xhaulproject.eu/download/5G-XHaul_D_22.pdf.

[4-8] A. Tzanakaki, M. Anastasopoulos, I. Berberana, D. Syrivelis, P. Flegkas,T. Korakis, D. Camps-Mur, I. Demirkol, J. Gutiérrez, E. Grass, Q. Wei, E. Pateromichelakis, N. Vucic, A. Fehske, M. Grieger, M. Eiselt, J., G. Fettweis, G. Lyberopoulos, E. Theodoropoulou, and D. Simeonidou, Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services , Accepted at IEEE Communications Magazine

[4-9] 5G-Crosshaul, Deliverable 2.1

[4-10] NGMN, LTE backhauling deployment scenarios, white paper, 2011.

[4-11] IEEE 1904.3, Draft Standard Radio over Ethernet Encapsulations and Mappings.

[4-12] R. Vaishampayan et al., "Application Driven Comparison of T-MPLS/MPLS-TP and PBB-TE — Driver Choices for Carrier Ethernet," IEEE INFOCOM Wksps. 2009, Apr. 2009.

5 Softwarization and 5G Service Management and Orchestration

In this chapter we look at the advances in the design of the 5G management and orchestration plane to support the concept of slicing. As we mention in Section 2.4 one of the main focus areas of 5G architecture is its support for slicing. Network slicing refers to the existence of multiple, possibly isolated, service and network architectures to support different usage scenarios, in particular services hosted by different verticals. Network slicing has evolved as a fundamental feature of the emerging 5G systems enabling dynamic multi-service support, multi-tenancy and the integration means for vertical market players. This move from a static, well understood architecture in the previous generation of mobile networks to dynamic, multiple use case-based architectures in 5G is driven by operator business considerations. This has a significant impact ***on the management plane*** as shown in Figure 2-16. In the figure the operator is enabled to ***sell parts of its infrastructure*** to customers including verticals, factories and other operators. Each customer has a different set of requirements and is expected, typically, to resell a service to its end customers. Every point in the ecosystem that specifies a business relationship also must specify a previously agreed upon management interface to activate and configure the service that is being sold. Note that the operator can be its own customer.

The exposure of management interfaces contrasts with the traditional vendor approach of proprietary management interfaces for management of the network infrastructure. This change is driven by network slicing. At a high level, the network slicing concept is about operating virtual network architectures on top of physical infrastructures, possibly with virtual resource isolation and virtual network performance guarantees. The separation of different functions by programmable abstractions (e.g., radio resources from packet processing) simplifies the integration challenges especially for applications supporting vertical industries beyond telecommunications. Note that the notion of resources in 5G network slicing includes network, compute and storage capacity resources; virtualised network functions; shared physical resources; and radio resources. To be able to uniformly manage the multiple slices that an operator is expected to host, ***programming interfaces to the virtualization infrastructures*** need to be exposed. The infrastructure needs to be capable of ***hosting multiple tenants*** and to be able to distinguish between the various types of ***flexibility and control required in the virtualized cloud resources.*** This chapter covers these advances in Section 5.1 under ***advances in supporting technologies***.

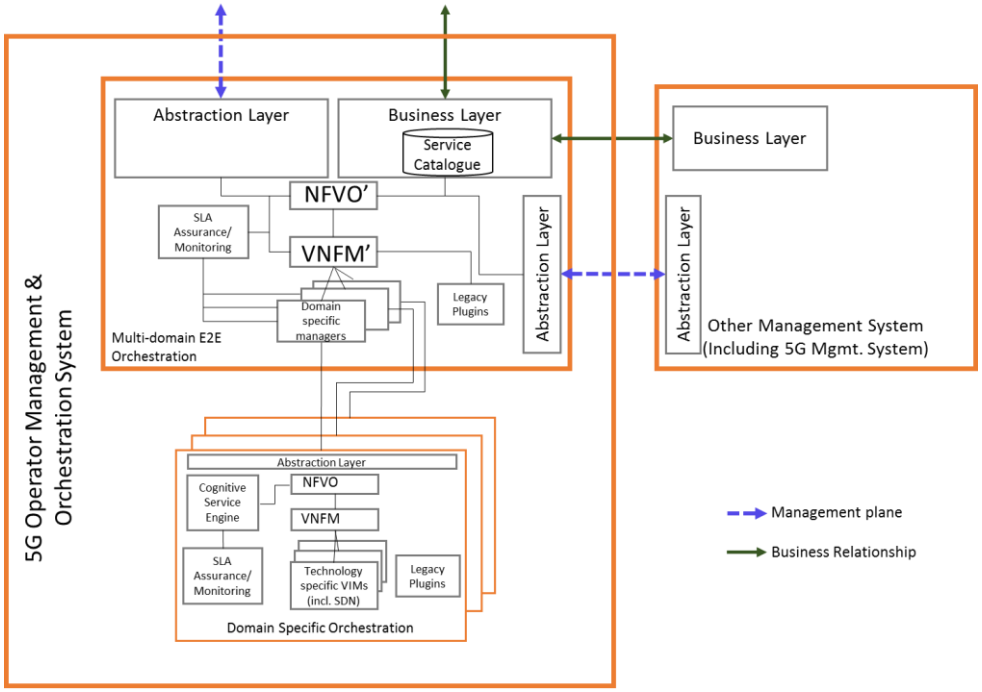


Figure 5-1: Internal architecture is a recursive extension of the MANO Framework

Once the underlying infrastructure is able to support the requirements stated above, there needs to be a way ***to manage and orchestrate those resources in an abstract manner, within a domain, across domains, as well as across providers.*** Providers in the new business ecosystem are not only limited to operators but could be specific technology providers, e.g. factory floors wanting to rent out their resources. Figure 5-1 shows the currently proposed architecture of the orchestration system as an extension of the ETSI NFV MANO architecture. The architecture is recursively stackable and consists at the lowest layer of domain-specific management and orchestration entities. The next higher level in the recursion combines these multiple domain-specific management and orchestration entities to create a multi-domain orchestration and management entity to coordinate end-to-end service and slice creation. The abstraction layer at each level exposes a generic set of interfaces while hiding specific technical and implementation details. The business and management interfaces are then exposed both northwards to customers as well as in the east-west direction to other operators. In Section 5.2, we look into the details of the management and orchestration framework.

To enable automation in such orchestration and management for service/slice deployment and (re)configuration functionalities, the management plane must be able to ***receive from the customer and describe the services and their requirements*** in sufficient detail to reflect the requirements of the customer.

All this brings immense complexities to the management systems underlining the need to simplify administrator load by including ***self organizing management*** mechanisms.

5.1 Enabling Technologies

Programmable virtualized infrastructure is considered a main supporting technology to realize slicing. In principle slicing could be achieved without such an infrastructure, however, it would lack the considerable benefit of installing a custom-tailored architectures dynamically. To this end, the 5GPP effort has made significant advances in the field of virtualization and cloud technologies. In particular, work has been done to incorporate plug-n-play deployment of clouds over physical infrastructure. This is described in Section 5.1.1.

Towards programmability, SDN has been extended to add support for multi-tenancy. Multi-tenancy support in the infrastructure is key to be able to support multiple possibly isolated slices at the same time as described in Section 5.1.2.

5.1.1 Multi-Tenancy Support

Multi-tenancy support is at the heart of the slicing-based business model for providing services. Without multi-tenancy, multiple simultaneously existing services cannot be provided by the 5G operator and rendering the concept of slicing moot. The platform of Figure 5-1 is able to execute service and function-specific management/orchestration code separately for each tenant (or, if done properly, even separately for each service or function) and in isolation. Functions/services are enabled to provide information about themselves as to whether they are multi-tenant capable and can be reused across tenants.

The system architecture shown in Figure 5-1 has been designed not only to cater to end-to-end (mobile) network services across multiple domains but it also makes provisions to provide multi-tenancy support. While multi-tenancy support in the cloud environment already exists, for 5G the following three aspects of the multi-tenancy with respect to the Network Slicing design paradigm are considered [5-1][5-2]:

- **Infrastructure Sharing:** although cloud computing can be used to meet traditional challenges, like scalability concerns and provide for fast resource provisioning times, an analysis is required when it comes in multi-operator environments with time-critical applications and services. Towards building infrastructure sharing mechanisms for LTE, the evolutionary approach applies the SDN concept into a part of the traditional core network architecture. This evolutionary approach analyses the traditional mobile network functions, such as PGW, SGW, MME, etc., and decides which functions should be implemented in the controller and which should be implemented in traditional, dedicated hardware. Work in this direction focuses on slicing techniques to create multiple virtual core networks shared among multiple network operators.
- **Spectrum Sharing:** Dynamic spectrum access is a promising approach to increase spectrum efficiency and alleviate spectrum scarcity. Many works investigate issues like cooperative spectrum sharing under incomplete information. Other approaches are both incentive-compatible and individually rational and are used to determine the assigned frequency bands and prices for them. The idea is to find policies that guarantee the largest expected profits by selling frequency bands jointly.
- **RAN Sharing:** This class relies on Hypervisor-based solutions to create the virtual eNB, which uses the physical infrastructure and resources of another eNB, depending on requests from the MNO.
- **Network Sharing:** With reference to the architecture framework shown in Figure 2-2 and Figure 2-4, the underlying SDN technology has been extended by defining and integrating SDN-based controller paradigms in the form of SDM-O, SDM-C and SDM-X functional components. For example, SDM-X runs applications that exclusively control shared network functions and resources – multi-tenancy scheduler being such an application that coordinates resource sharing among multiple tenants. Moreover, the Inter-slice Resource Broker at the MANO layer is also specifically designed to manage and orchestrate resources allocation for network services and functions across different slices and tenants.The paradigm of software-defined mobile network orchestration (SDM-O) enables the inter-slice management and orchestration, and thus it has a fundamental role in realizing *multi-service and multi-tenant* aspects of 5G NORMA network. The responsibilities of SDM-O is to map the slice templates representing the slice requirements along with the corresponding tenants' SLAs with the available network resources. The decision upon which network functions can be shared among slices/tenants as well as their placement in the network will be carried out by the software-defined mobile network orchestration (SDM-O). E.g. for a V2X slice with a stringent latency requirements the SDM-O might tend to deploy the network functions closer to the network edge. On the other hand, for eMBB slice with relaxed latency requirements

the network functions might be placed in the central cloud. The SDM-O has a complete (inter-slice) view on different slices and tenants' requirements as well as corresponding resources for slice realization. Moreover, the SDM-O incorporates the domain-specific knowledge, i.e. the logic of network functions that it orchestrates. Having such cross-slice knowledge the SDM-O can efficiently decide on rules/instructions that need to be conveyed to other MANO entities, control applications and SDM-X and SDM-C in order to properly orchestrate, manage and control the network functions and resources of slices. The network functions and resources can be shared among different network slices and tenants. Such multiplexing improves the network resource utilization and reduces the cost of network service deployment. Inter-slice control of 5G NORMA enables efficient sharing of such resources and network functions. The sharing rules are derived by SDM-O from slice templates and SLAs and implemented on top of inter-slice controller (SDM-X) in the form of policies. E.g. different Virtual Mobile Network Operators (vMNOs), which are the tenants of the same infrastructure can have dedicated implementations of virtualized EPCs (vEPCs) but the (S)Gi-LAN functions such as Firewall, Parental Control, DPI, etc. can be common (shared) among all vMNOs. The control of such shared (S)Gi-LAN will be done by SDM-X as depicted in Figure 5-2 below.

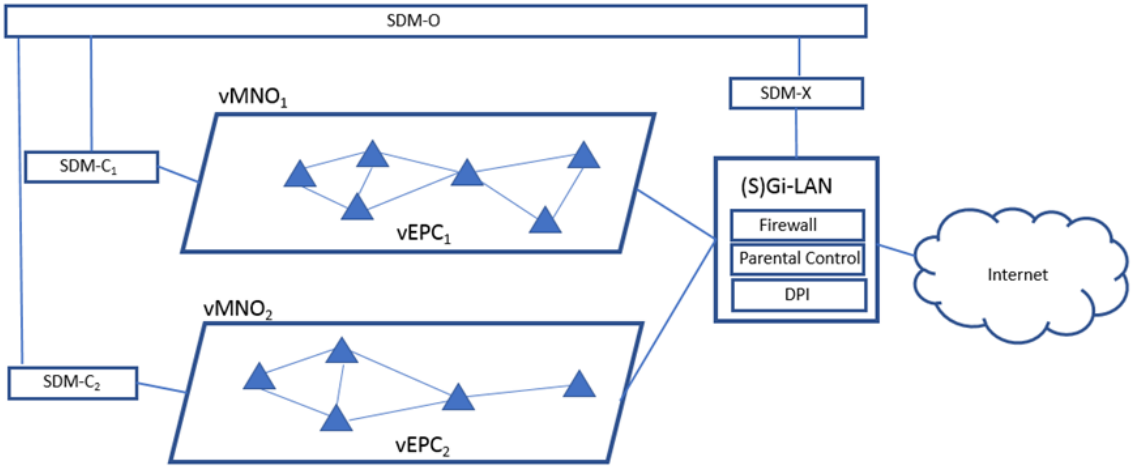


Figure 5-2: An example of shared network functions (Firewall, Parental Control, DPI, etc.) under the control of SDM-X

In what follows, we look at the concepts of RAN sharing and network sharing in further detail

5.1.1.1 Multi-tenancy in the RAN

RAN sharing for 5G can be implemented by adding advanced control features to the current 3GPP LTE. LTE uses Orthogonal Frequency Division Multiplexing (OFDM) for the downlink and Single Carrier Frequency Division Multiple Access (SC-FDMA) in the uplink. The Physical Resource Block (PRB) is the smallest element assigned by the base station scheduler. Transmission Time Interval (TTI) is the duration of a transmission on the radio link. A scheduler can determine to which user the shared resources (time and frequencies) for each TTI should be allocated. As shown in Figure 5-3, the RAN sharing problem is related to the design and implementation of policies that are able to effectively schedule Resource Blocks effectively between different MVNOs with respect to specific differentiation objectives and with isolation guarantees.

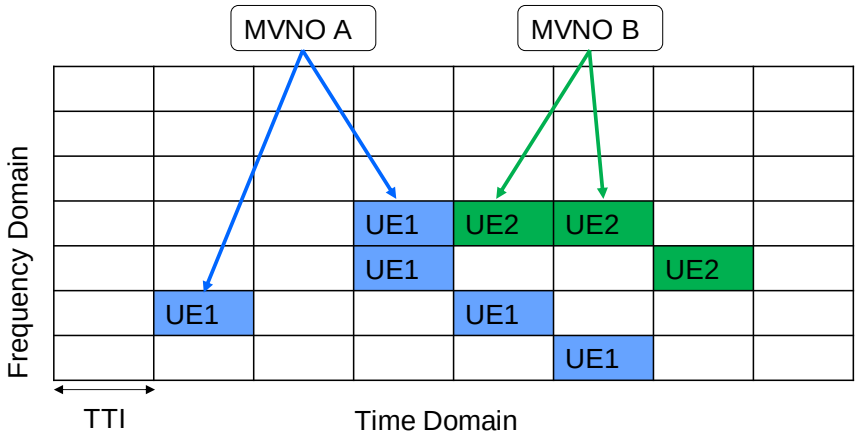


Figure 5-3 RAN sharing

Figure 5-4 presents a high-level representation of the control plane interaction with different tenants sharing the RAN. The local real time controller (RTC) component of the architecture facilitates efficient policies that share RAN resources. A local real time controller keeps information consistent within its regional and real-time scope of control, meaning that it has to synchronise information from various infrastructure sources to produce a complete local network view. In Section 3.1, the logically centralized controller and the hierarchy control framework from RAN are introduced. Together with the logically centralized controller (C3) integration and a fully programmable underlay, it is able to support the Infrastructure Sharing, the Spectrum sharing and the RAN sharing by coordinating the RAN resource from different abstraction levels by the hierarchy control structure.

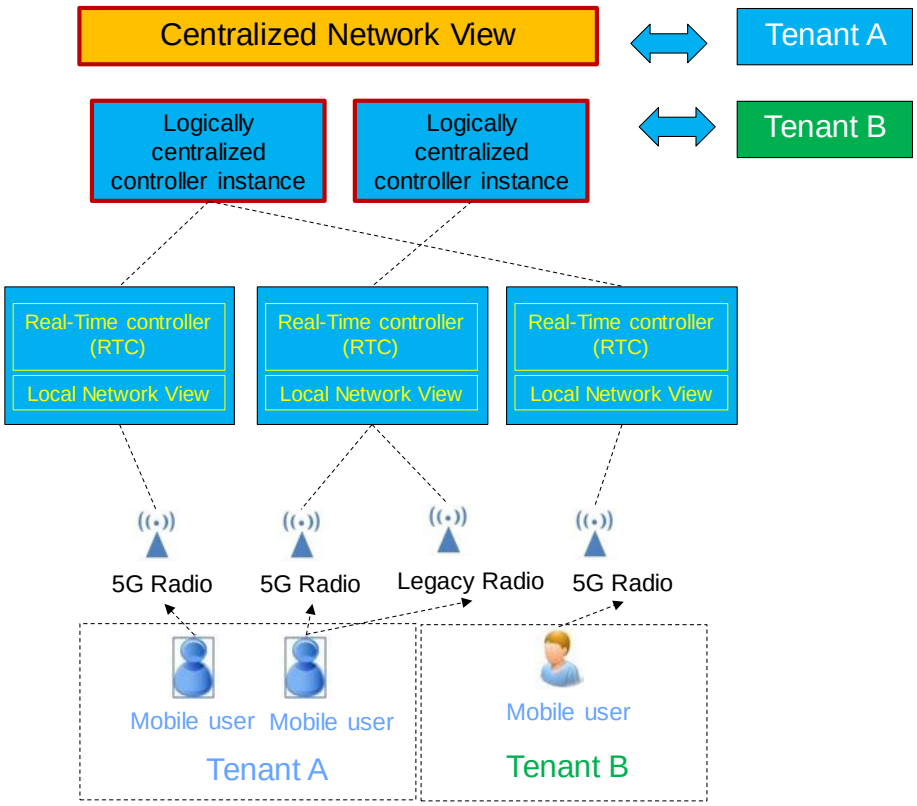


Figure 5-4 Multi-tenancy operation and RAN sharing

5.1.2 Cloud and Virtualization Technologies

The primary technology for slice deployment is currently considered to be NFV, supported by SDN. NFV consists of the network functions that form a slice or a service to be implemented in a virtualized manner as Virtual Network Functions. A traditional NFV Infrastructure, based on full VMs implementing VNFs may not be adequate to cover all the requirements of 5G networks. Different types of virtualization technologies will need to coexist to fully exploit the softwarization potential. This is already happening with the coexistence of full VMs and container-based virtualization. In addition, Unikernels are a very lightweight virtualization technology that is gaining momentum for some specific application scenarios. Unikernels offer very good performance in terms of low memory footprint and instantiation time and they have very good isolation and security properties (better than containers). The recent advances conducted in superfluidity show that the Unikernels, Xen based hypervisor, can reach a small footprint (around 5 MB of memory when running), can be instantiated within around 30 milliseconds, processes up to 10Gb/s of traffic and does not need a persistent disk drive to work. These three technologies (full VMs, containers and Unikernels) have different properties and should coexist in the same infrastructure, to be used opportunistically depending on the requirements of the scenarios and workloads. An example of such usage is applications where workloads need rapid instantiation and high performance with minimal overhead. In such cases, Unikernels, whose deployment requires expert optimization, will run inside containers to take the advantage of the easy portability and life-cycle management of containers, unlike old-fashioned bare metal installations and configurations. Another example is where applications may require higher network performance (latency, throughput) but still retain the flexibility given by containers or VMs, thus requiring a VM with special forms of performance acceleration such as SRIOV or DPDK.

In the path toward network softwarization, the decomposition and modularity concepts will become increasingly relevant and the granularity of the decomposition will be finer. Software routers are an example: elementary packet processing functions can be composed to build a complex node function (e.g. a router/NAT/firewall). Hardware accelerators on NICs can execute packet processing functions designed in software. The architectural model for 5G should be flexible enough to support and take advantage from this heterogeneity. The notion of heterogeneous execution environments should be taken into account. The concept of VNF should be generalized to cover different type of functions (not only networking functions, but any kind of processing) and to cover a wide range of granularity in the definition of a VNF (from full VMs implementing “big” node functions, like a S-GW or a P-GW, to micro-components implementing packet level operations like address rewriting). The concept of Reusable Functional Block (RFB) (Figure 5-5) as a generalization of VNFs, and the concept of RFB Execution Environment (REE) to cover the heterogeneous infrastructure that can support the service execution. REE can be nested, for example containers can run on VMs, a software router can run inside a VM or inside a container, so that a hierarchy of REE is actually building up a 5G infrastructure.

The RFB/REE principles allow an abstraction of the heterogeneous platform and hardware which is more and more faced in 5G deployment where the EDGE clouds could have different hardware configuration and capabilities.

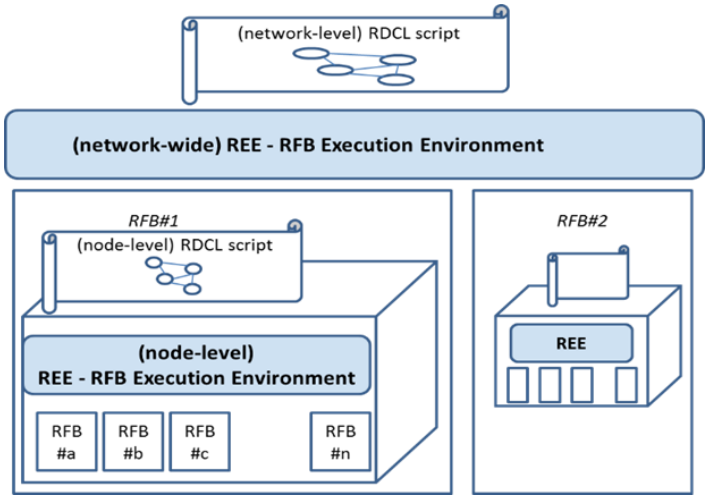


Figure 5-5: The concept of RFB and REE that generalize the execution environment

5.1.3 Network Programmability

Programmability is a key supporting technology for the realization of ***dynamism in the 5G service and slice architecture***. The main ***advantage of slicing*** is the ability to manage the infrastructure to the needs of the service while hosting multiple services over the same infrastructure to enable some degree of multiplexing gains, striking a balance between actually separate infrastructures for each service and running all services on the same, non-sliced infrastructure which would have perfect multiplexing gains. This implies that the underlying infrastructure must be programmable. In Figure 5-1 the internal architecture for the 5G operator management system relies on interfaces exposed by the technology-specific VIMs to the VNFM and the NFVO to achieve domain specific orchestration and management. Furthermore the domain specific orchestration in itself exposes a north bound interface to the end-to-end orchestration level to be able to create service and slices that extend technology and provider domains. In this section we look at the objectives of introducing programmability in a particular technology domain: the Radio Access Network (RAN) to create programmable abstracted network models for use at the higher layers in the orchestration framework.

RAN coordination and programmability are central concepts in 5G that are aimed to improve service quality, resource usage, and management efficiency, while addressing the limitations of the current LTE and WLAN systems caused by distributed control among them. A coherent representation of the network state and infrastructure resources is crucial for effective RAN coordination and control of programmable infrastructures and services. Moreover, programmable infrastructures require programmability constructs that provide means to observe and manipulate virtual and physical Network Functions (NFs) and their behaviour via high-level abstractions. One example of RAN programmability model is shown in Figure 5-6.

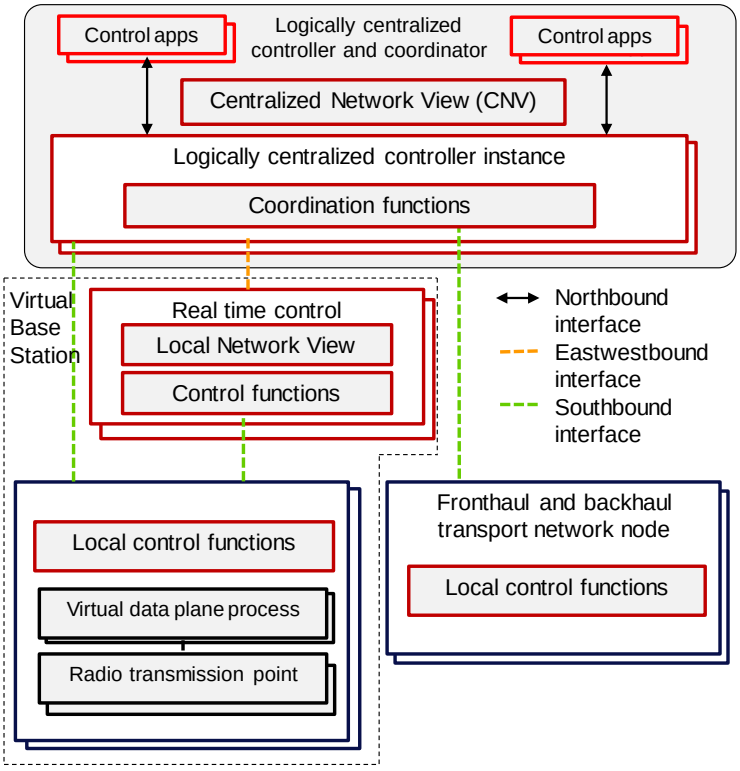


Figure 5-6 RAN programmability model

At the low layers, the status information is abstracted and fed to the higher control layers to generate network views. Abstractions encompass representations and models of time-frequency resources, spatial capabilities (i.e. number of transmit and receive antennas), as well as throughput per network slice or per allocated resources. In principle any data structure can be used for storing and accessing abstracted representation of the network state (e.g., CQI defined in LTE). However, for unified large-scale coordination of infrastructure resources, structuring network information into network graphs in a systematic way offers effective representation of physical and virtual infrastructures. In Figure 5-6, network views is presented by network graphs. Network views can be applied to RAN coordination and wide range infrastructure coordination and control operations. Essentially, network graphs enable the possibility to apply mathematical models and algorithms, which often leads to highly efficient solutions in terms of convergence and network performance. Graph-based abstractions can for example model LTE resource allocation problems that are solved efficiently using constraint satisfaction and local search algorithms [5-3].

The concept of network graph abstractions maps horizontally and vertically at different levels of the proposed architecture. The elements of network graphs (i.e. the vertices and edges) are created from distributed data sources such as raw metrics provided by the infrastructure or more sophisticated monitoring functions operating in a decentralised and centralised manner. A network graph is created by collecting information accessible directly from infrastructure entities or stored in some dedicated storage (e.g., storage networks and databases). At the level of local real-time controller entities, a network graph may represent the network state relative to a certain RAT infrastructure and associated nodes (e.g., WiFi or LTE). Network graphs at the centralised coordination level represent the state of a defined part of the network, such as a smaller region or domain. High-level network graphs (e.g., for centralised coordination and control) can be aggregated based on selected sets of regional network graphs for the purpose of multi-domain coordination.

In the hierarchical architecture of Figure 5-6, controller instances implement capabilities for creating network graphs for performing control operations and for providing regional or logically centralised views of the infrastructure via the technology specific VIMs. The domain specific orchestration can then include functionality for: 1) gathering network information from

distributed data sources for the purpose of creating a network graph; 2) aggregating existing network graphs; 3) processing of network graphs for the purpose of coordination and control operations; 4) disseminating network graphs and results to other controllers and network entities upon request or as part of a coordination and control operation, synchronously or asynchronously, including providing it to the end-to-end orchestration.

5.2 Services and Service Design

The ability to support multiple services simultaneously and dynamically is expected to be the driving force behind 5G success. This implies that from the perspective of the management plane, the ability to design, describe and manage services in an automated manner are expected to be the key trends when looking towards the evolution to 5G. Since 5G is based on an NFV architecture where a service is expected to be composed of a set of network functions the composition of these network functions to create a final service is an interesting topic presented in Section 5.2.2. The ability to then verify these deployed services is then presented in Section 5.2.3. Finally, supporting diverse services at the same time can create an administrative nightmare for the operator. Hence machine learning based mechanisms for service planning are investigated in Section 5.2.4

5.2.1 Service description

For the description of both individual network functions as well as entire services, a couple of de facto standards are emerging. The de facto standards start from the ETSI model and make some straightforward extensions. Relevant extensions are: (1) providing quantitative statements relating offered load, required resources, and provided performance for both individual functions and services, (2) extending a service description by more powerful and flexible annotations how a service should be managed (with regard to lifecycle, scaling, placement, e.g.), going beyond simple threshold-based policies that are commonly used to do. These extensions enable a far broader range of automated optimizations, e.g., automated placement and scaling in a service-specific manner rather than using one-size-fits-all, simplistic approaches. Overall, service quality will be improved by extended descriptions.

The following subsections present a brief survey of the current service description standards performed within the 5GPPP.

5.2.1.1 ETSI NFV

ETSI NFV is actively working on the definition of Network Service Templates that could allow the instantiation of services, understanding them as a composition of several linked VNFs, including some information regarding the (virtual) links connecting them in the forwarding graph. Specification [5-4] provides more detail on the fields and the structure of such templates. The Service templates are not only for requesting service but also for service lifecycle, then consistency with the functionality in Ix-F interface should be ensured, in this case.

Another relevant effort in ETSI is the open source initiative for releasing an Open Source MANO component. As part of this effort there is a parallel specification of descriptors for service request. The latest specification can be found here [5-5].

Such descriptors are basically YANG models based on [5-4], ensuring alignment with NFV framework.

5.2.1.2 OASIS TOSCA

TOSCA, as defined by OASIS, serves as language and metamodel to describe services, its components, relationships and management procedures. The main components for TOSCA are:

- A topology template which defines the structure of a service as a set of node templates and relationship templates that together define the topology model as a (not necessarily connected) directed graph.
- Node and relationship templates that specify the properties and the operations (via interfaces) available to manipulate the component

There is already a specific profile in TOSCA defined for NFV [5-6] that is being used by some commercial solutions in the market as in the case of Ciena Blue Planet). Figure 5-7 shows the relationship between TOSCA and NFV.

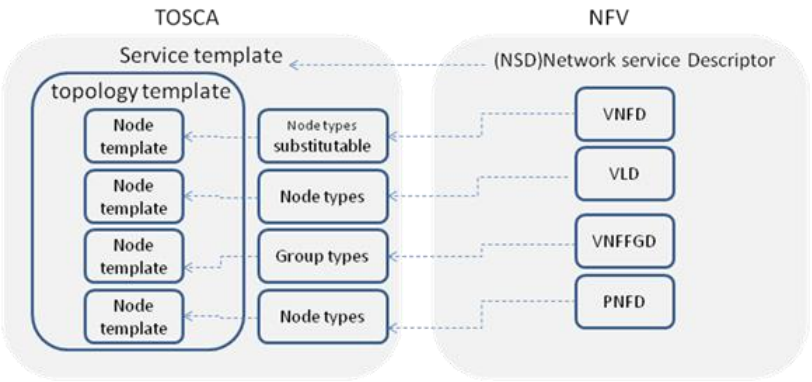


Figure 5-7. TOSCA relation to NFV

In addition to that, ETSI NFV has started to work on a new specification document on NFV descriptors based on TOSCA Specification [5-7], with the expected result of alignment with the previous reported effort. By now, there are no actual templates yet drafted, being work in progress.

5.2.1.3 IETF Service Function Chaining

The Service Function Chaining (SFC) working group [5-8] in IETF defines a service architecture around three main components to be deployed in an SFC-enabled domain, used to compose the service end-to-end:

- Service functions (SFs), which are the equivalent to the VNFs
- Service function forwarders (SFFs), which are the elements in charge of delivering the traffic among the SFs, ensuring connectivity and reachability
- Service classification functions (SCFs), which are the components devoted to classify the traffic entering the SFC domain, associating it to the proper SFC based on certain characteristics previously configured.

In IETF SFC the definition of service description is not addressed, therefore requiring extra logic for accomplishing that.

5.2.1.4 OGF NSI

The Network Service Information (NSI) [5-8] proposal of OGF allows for requesting Connectivity services including service-specific information as follows:

- Ingress and egress Service Termination Points (STPs)
- Explicit Routing Object (ERO)
- Capacity of the Connection
- Framing information

Figure 5-8 represents the way in which NSI facilitates the request of connectivity services among providers.

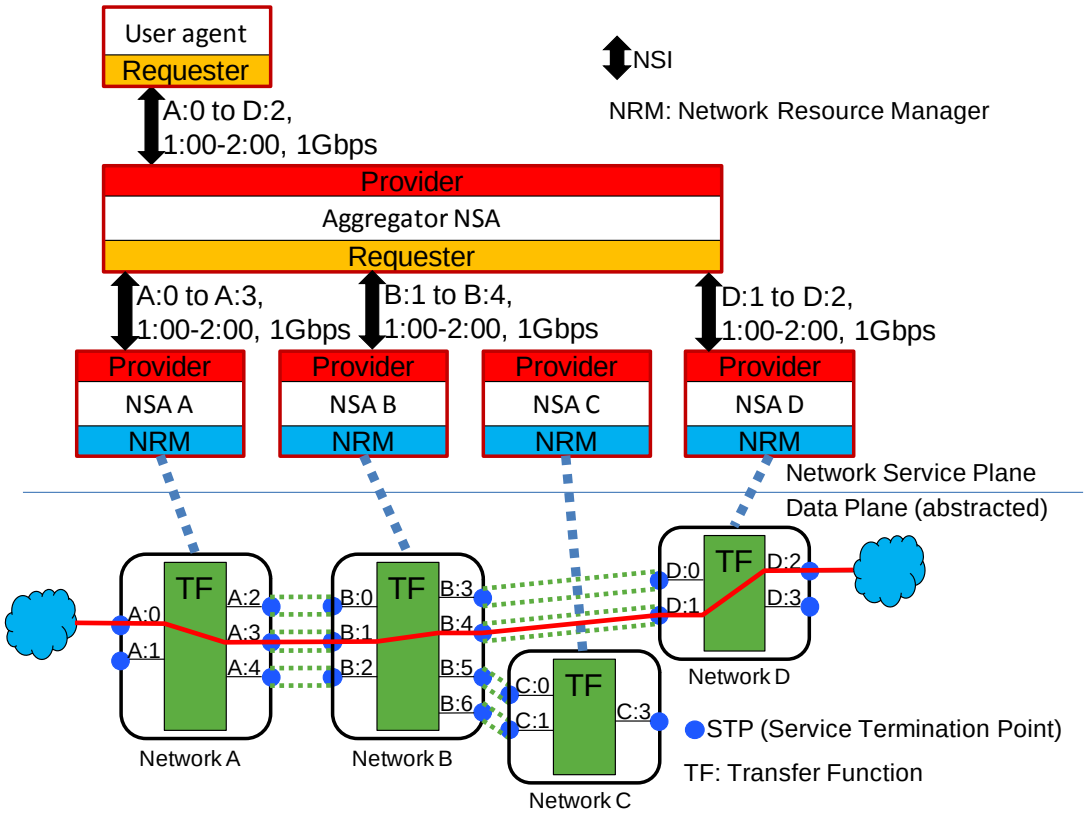


Figure 5-8. OGF NSI relation among providers

Despite connectivity services are covered by NSI, it does not allow by now any capabilities for requesting more advanced services as VNFaaS or NFVIaaS.

5.2.2 On-demand composition

The composition of VNFs or sub-services into new services is supported in the management plane in a straightforward and extensive manner. Explicit support for this is provided by an orchestrator's repositories, keeping track of which VNFs/sub-services are already available or already deployed with annotations expressing whether a service is capable to work for different users in different contexts (multi-tenancy). Similar repositories exist in multiple catalogues and can be queried and interacted with via an SDK. Via the SDK and the orchestrator's API, services can be reconfigured or new service versions can be created.
As mechanism to achieve the mapping of the conventional and new telco services to the new Cloud-enabled infrastructures, Service Function Chaining techniques are being applied. Software Defined Networking (SDN) mechanisms for Service Function Chaining (SFC) are the most prominent candidates to enforce traffic steering through a logical network graph and to achieve certain service functionality among the virtualized components. Extending the concept of SFC in the context of the 5G ecosystem requires deeper understanding on the NFV concepts in such a scenario. Carefully identifying the requirements of the specific setup is important in order to choose the most suitable mechanisms and protocols to establish the desired functionality.

The advantage of using OpenFlow (OF) protocol as basis for SFC is that routing can be steered over a specific networking path by programmatically applying OF based rules (flows) on the SDN controller or the virtual switch (OVS) inside the VM hosting the VNF. From a single data center point of view, the SDN controller is placed between components such as the Orchestrator. The steering and rule enforcement policy are kept within the controller application logic and enforced over the network that hosts the Light DC specific NFVs. If the routing happens across different Light DCs, then the SFC approach may alter depending on the placement of the controller within

the given architecture. This has to be in accordance with the networking protocols to be adopted in that case, as used today for intra data-center routing. Alternatively a combination of SDN controllers can be employed, one to manage the traffic rules within the cloud, and other for the radio specific topology.

5.2.3 Verification of deployed services

The 5G network management architecture of Figure 5-1 will allow operators and third parties to quickly instantiate services composed of network functions. A key question relates to the safety of instantiating these network functions. Essentially, network functions provide traffic processing in one way or another and can be considered as processing blocks. How does the operator know that the deployed network function or processing block is functioning correctly and according to its developer's specification? It is well known that router misconfiguration still causes havoc even in traditional networks running simple destination-based forwarding and tunnelling. In this context, installing custom processing functionality, as 5G proposes, creates potential risks compared to faulty configuration changes today, such as outlined in the following examples:

- A misconfiguration in a firewall VNF may allow attack traffic to reach internal operator services
- An application gateway may wrongly change packets leading to loss of connectivity
- Two VNFs that work correctly in isolation may misbehave when deployed together. For instance, a firewall filtering on source address will work incorrectly when placed after a NAT that changes the source address of packets.

These examples, and many others, show the need for formal verification before instantiation and teardown: every time the network configuration changes, it is possible that the new configuration breaks the operator's policy, and therefore would be incorrect after the change. To ensure that we can provide high levels of reliability we need to have tools to assess the correctness of configurations using model checking.

Network data plane verification tools based on formal software verification techniques are an alternative and very promising approach. Such tools work by using a snapshot of the network data plane, including router FIBs, tunnelling configurations and VNF processing specifications and simulate what happens when a generic packet (with wildcard fields) is injected at one of the network ports. By examining the way the packet is handled by the network, one can verify if the policies of the network operator hold. The limitations of existing works relate to their scalability (the size of networks that can be verified in "real-time", as network configuration updates arrive) and expressivity (the type of processing they capture).

In particular, symbolic network execution works in the following way. Each network box' processing is described in the SEFL language as a sequence of {port: instruction_list}, where ports can be external ports connected to other boxes, or internal ports. SEFL has an instruction called forward that allows processing to continue at a specified port. The instruction set of SEFL is limited to header field assignments (and simple arithmetic), encapsulation, decapsulation, as well as per-packet metadata set/get operations. It has the if instruction that all programming languages do, but it also includes a fork instruction that duplicates the current packet (similar to the process fork in Linux).

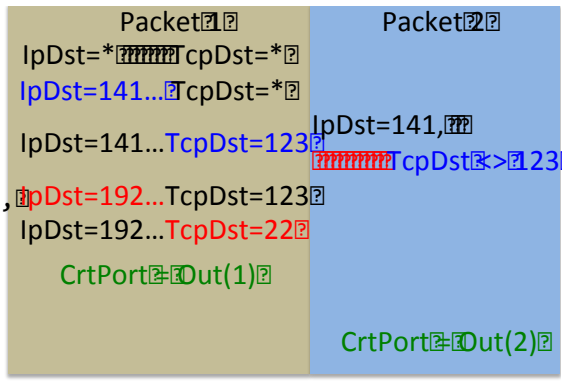
Symnet, our symbolic execution tool, takes a network described in SEFL, together with links between boxes and a port where a symbolic packet should be injected. A symbolic packet is a packet that contains one or more fields whose value is unspecified (i.e. symbolic), and can take any value. The job of Symnet is to simulate how all the possible concrete packets will be treated by the given network: which path they will take, how their fields will be changed, when and whether they will be dropped, and so forth. The trick in this simulation is to get the results without actually trying out all possible concrete packets, which is infeasible.

Symbolic execution works by tracking the allowed values for a given header field for every

possible packet (or path, in symbolic execution parlance). Execution starts with a single symbolic packet and continues until a firewall instruction is found (Constrain) – in this case, the constraint is applied to the current packet and, if it is satisfiable (as verified by a SMT solver), the constraint is added to the packet and execution continues; otherwise the packet stops. When an if instruction is reached, symbolic execution first checks the if condition for satisfiability; if so, the then path is explored. The else path is also explored, with the constraint that the negated if condition must hold. Thus, after an if instruction we have two potential execution paths (or symbolic packets) which will be explored separately by Symnet. The same applies for the Fork instruction.

Below we provide an example of Symnet symbolic execution for a box that does port forwarding. Execution begins with a packet with symbolic IP destination address and TCP port, and the box first checks the packet is destined for it, then checks whether the destination port should be forwarded, and if so will overwrite the destination address and port, otherwise will send the packet as is to another output port.

```
Element A model
InputPort(*):
  Constrain(IPDst==141.85.37.1),
  If (Constrain(TcpDst===123),
     InstructionBlock(
       Assign(IPDst,192.168.1.100),
       Assign(TcpDst,22),
       Forward(OutputPort(1))
     ),
     Forward(OutputPort(2)),
```



Symnet is very expressive as it can track header changes, tunnelling, stateful firewalls as well as basic reachability. The 5GPPP effort has extended Symnet to:

- Include a language that allows operators to express their policies easily and succinctly.
- include a verification tool that performs symbolic execution guided by the operator policy, in order to reduce the number of explored paths.
- include provably correct transformations from our SEFL language (used by Symnet) to dataplane languages like P4, ensuring that the verification results are accurate.
- Integrate Symnet verification in Openstack Neutron (ongoing)

A related issue is that of performance verification: Does a VNF actual conform with its stated load/resources/quality parameters, or are more resources needed than claimed to ensure a required quality? For such non-functional properties, symbolic approaches are often less well suited and have to be enhanced and complemented by (carefully steered) testing approaches.

5.2.4 Machine learning in Service Planning

Management in 5G is going to be an increasingly difficult task for operators due to the rapid increase in network demand. Machine learning services are proposed to be used to provide more efficient network management. The work presented on SLA management and monitoring as presented in Section 5.3 focuses on FCAPS (fault, configuration, accounting, performance and security). This work provide rule based solutions for network management for the five FCAPS functionalities. 5G networks need to leverage on such functionalities and yet provide more dynamic solutions that use machine learning for effective and automated network management.

There is considerable work that demonstrates the usefulness of machine learning for problems related to various functionalities of FCAPS in the literature. Zander et al.and Williams et al. proposed models for automated network traffic classification for machine learning. Network traffic classification can help the accounting and performance functionalities in FCAPS. Zender et al. clustered the network traffic traces using EM to find the traffic patterns that belong to the same class. Williams et al. compared various performance impact of feature set reduction, using Consistency-based and Correlation-based feature selection, is demonstrated on Naïve Bayes, C4.5, Bayesian NetworkandNaïveBayesTree algorithms for the network traffic classification problem. In both studies the findings are found to outperform solutions based on static set of rules. Towards security, Sinclair et al. proposed a machine learning model to differentiate normal network activity and anomalous network activity and used genetic algorithms and a decision tree based learner to evolve predefined anomalous network activity rules over time dynamically. In their work, Sinclair et al. used only traffic source, destination IPS and ports. Lastly, dynamic resource management related to the configuration. The authors used reinforcement learning for dynamic resource management in virtual networks by to improve the quality of service by reducing packet drop rate and virtual link delay. In an experiment based on simulated events the model showed promising results.

There is therefore a clear advantage and thereby a market need for integrated services based on machine learning to leverage the output provided by researchers for network management. The work done in 5GPPP phase 1 offers an integrated set of services that can be selected by operators based on the individual needs of an operator. The service portfolio can also be used to complement the tools used for network management to handle FCAPS functionalities (Fault, Configuration, Accounting, Performance and Security).

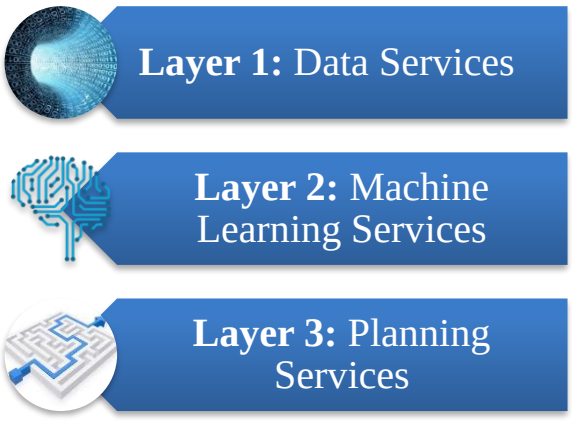


Figure 5-9 The Service Layers for Machine Learning

The services are offered in three layers, as illustrated in Figure 5-9. Data services are used to import and process the data required by the machine learning modules. Multiple services can be selected and integrated from these three layers based on the requirements of an operator and the availability of data. Machine learning services provide the core predictive functionality and the planning services orchestrate the predictive services for action recommendation and policy implementation. The machine learning models can also be used with other tools that gather network management data.

Figure 5-10 Service Portfolio

Figure 5-10 presents the service portfolio. Data services and planning services complement the core machine learning services. The services are divided in five categories. Data services and planning services are used to complement machine learning services offered in the portfolio. The categories are explained briefly as follows:

1. *Data Services:* Data services provide the input for all the machine learning services in the CogNet services portfolio. Data gathering service extracts data from all the third-party data sources. Data preparation service removes noise and processes the data for the active machine learning modules. Data preparation service is required to process the input of some of the machine learning services. Data dimensionality reduction service reduces the dimension of the input data (i.e., the number of variables considered for the analysis).
2. *Quality Assurance Services:* These services tackle the problem of anomaly detection as follows: SLO breach detection, Noisy neighbour detection and anomaly detection. These services use various machine learning algorithms, such as classification (e.g., in the network threats/noisy neighbour detection) and regression (e.g., in SLO breach).
3. *Network Demand Prediction Services:* Network traffic classification service classifies network traffic for various use cases, such as the Collaborative Resource Management, where the classification is applied only on the header of the traffic packets, to enable classification on encrypted traffic.
4. *Location-based Services:* The recurrent crowd mobility and functional regions detection service can be used to discover the different crowd mobility patterns and functional regions of an urban area and associate them with different network demand. The large-scale events detection service deals with processing external data to identify large scale events which reflect abnormal or unexpected events (e.g., a concert) that might disrupt the recurrent demand patterns, needing adjustments based on the situational context.

Finally, the vehicular mobility patterns recognition service predicts the mobility of cars to adjust the mirror adaptation for optimum coverage within the considered area. All these services use geo-tagged data, such as social media data (e.g., Foursquare check-ins with location/venue information).

5. *Planning Services:* This category comprises of the action recommendation service. Action recommendation service recommends different actions depending on the events and data sent by the Cognitive Smart Engine. For instance, in the case of an anomaly being detected, a corrective action should take place to mitigate potential damaging effects. Distributed security enablement service enforces security policies based on the warnings from the quality assurance services.

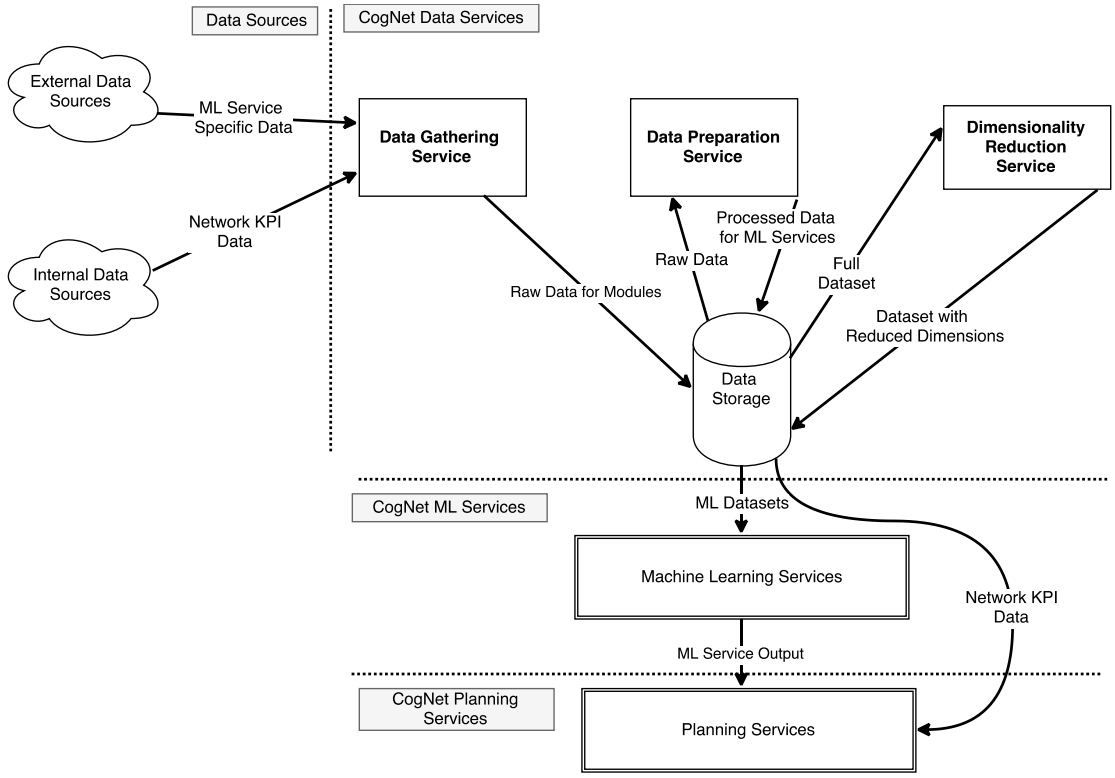


Figure 5-11: Data Services Interface with Input and Output of Each Service.

Data services are used to provide the data that are used by the machine learning services. These services are used to create the input of every machine learning service in the portfolio by importing raw data and processing it. In Figure 5-10, we show the summary of the data services that are offered.

Data gathering service extracts data from internal data sources such (network KPI monitors) and external sources that are applicable for a given machine learning service. The service can be configured based on which machine learning services that are actively used and caches the data it gathers in a database for later usage. Data gathering service can import raw data in batch mode or streaming mode depending on the service requirements and scheduled for periodical data imports.

Data preparation service processes raw data stored by the data gathering service based on the requirements of each machine learning service. Data preparation service provides the following functionality: (1) Aggregation of data, (2) Cleaning of data by detecting noise, (3) Splitting data for machine learning experiments.

The output of the data preparation service is processed datasets that can be used by the machine learning services directly or after dimensionality reduction. The output can be sent in batch mode or streaming mode based on the machine learning service requirements.

The services from the portfolio can be combined to form higher level services as illustrated in Figure 5-11. For example, crowd mobility and functional region detection can be combined with network traffic classification to increase the information content of the classification output by using functional regions and crowd mobility patterns as additional input.

5.3 Management and Orchestration

A generic architecture of the management and orchestration system based on an extension of the ETSI/NFV Management and Orchestration (MANO) architecture is an underlying trend in the evolution of the management plane architecture. Figure 5-1 already presented this evolution in brief. In this section we look at work done in some of the core functionalities that will support the functional blocks of Figure 5-1.

5.3.1 Embedding of Virtual Functions

Embedding or placing virtual network architecture that form a service is one of the most important aspects of 5G. To efficiently implement slicing, the placement of the VNFs and their interconnections in an efficient manner that maximizes the usage of the operator's resources is required. Considerable work has already been done in the research community in this regard as presented in [5-10][5-11].

The work done by 5GPPP has extended the placement concept in [5-12] to include several additional features and capabilities, particularly considering heuristic approaches. The heuristic mapping algorithm searches for (possibly a number of) embedding of service chains, considering a greedy step as the united mapping of a network function and an adjacent service graph link. The greedy mapping is guided by parameterize-able preference metrics to identify the locally best steps. If the greedy search fails, a bounded backtracking procedure is responsible for exploring a subset of the state space by trying locally less preferred steps. The preference metrics consider various parameters such as, load balancing on nodes, minimizing link resource usage, and they guide the mapping process to terminate as early as possible while complying to end-to-end delay requirements.

In addition, work has been done to not have a single, fixed placement/scaling/lifecycle management approach but rather by allowing individual algorithms to be used for each chain separately. Orchestrator then deals with resource conflicts and arbitrates, if necessary. Isolation of these service-specific placement/algorithms is ensured, as is information hiding.

Moreover, it is important to realize that looking at the placement problem only as a network embedding problem falls short of the mark. Unlike the conventional virtual network embedding problem (investigated, e.g., in the context of experimental testbeds over the internet like PlanetLab), a service graph is typically a flexible entity, modified by scaling operations. Hence, VNE algorithms don't immediately apply but have to be modified to incorporate the scaling step as well, enhancing their performance [5-13] .

5.3.2 Service Assurance and Monitoring

The 5G deployment envisions that a number of inter-operating virtualisation environments are used, some of which provide integrated telemetry (monitoring) solutions such as Celiometer in OpenStack. The type, quantity, quality and configuration of the exposed metrics can be limited. In addition, the metrics are typically tied to a specific type of virtualisation implementation approach such as virtual machines (VMs) and have scalability challenges. A specific case in 5G

is the telemetry platform that can be used to support diverse virtualised environments and bare metal (non-virtualised environments). To address this issue the project has adopted standalone telemetry agents to provide a wide range of metrics across different virtualisation methods and for different use cases i.e. operation service monitoring and service characterisation.

The characterisation methodology relies on a structured experimental approach, leveraging a full stack monitoring approach to collect metrics. The collected metrics relate to both the physical and virtualised compute/storage/network environments and the actual service under test in an operational context. The adopted telemetry platforms are currently an Intel proprietary CMDB, Cimmaron and the open-source Snap telemetry framework. The telemetry data collected can be used to:

1. Identify configuration optimisation opportunities;
2. Establish breakpoints for services and their underlying failure mechanisms;
3. Identify appropriate metrics for operational monitoring.

Although there are other data monitoring solutions available, Snap is more suitable for 5G environments because it allows for dynamic reconfiguration without interruption, can be used in tribes / swarms for monitoring similar information on large scales of similar machines and can be used with many types of plugins for data ingress and egress.

5.3.3 Life-Cycle Management

5G will be built upon softwarisation and virtualization technologies, particularly SDN and NFV. The 5GPPP effort has designed and prototyped a complete management of SDN and NFV Apps, thereby paving the way for truly dynamic, on-demand, flexible and automated/autonomic SDN/NFV service deployment through an orchestrator. This Apps Management subsystem has the following advantageous features. Firstly, it provides a fully automated lifecycle management of NFV and SDN applications, from Apps encapsulation, onboarding, and instantiation to deployment, configuration, update/modification and termination. In particular, it enables one-click automated Apps onboarding. New Apps are made available in the system (and ready for deployment) with a single action from the GUI.

Secondly, it provides common Apps lifecycle mechanisms and procedures for various kinds of Apps including VNFs, SDN Apps, SDN controller Apps, and Physical Network Functions (PNFs) for backward compatibility. Thirdly, it offers flexibility and extensibility by design. The design has taken a plugin-based approach for the architected components, and the prototypes are easily extensible, as shown in Figure 5-12. For instance, the encapsulation of NFV applications translates into two major enhancements to the ETSI NFV and MANO architectures. The first is the enhancement of the ETSI MANO VNFM functionalities towards a multi-tenant aware management of sensor and actuator VNFs lifecycle. The VNFM is the component responsible for the enforcement of VNF lifecycle management actions as requested by the upper layer Orchestration components. It directly interacts with the actuator and sensor VNFs by means of unified procedures and interfaces. Secondly, it employs containerization of VNF and EMS into encapsulated VNFs. The actuator and sensor VNFs expose a common interface that allows the VNFM to operate basic lifecycle management actions, like VNF configuration, re-configuration, start, stop irrespectively of the given VNF type. This is achieved by embedding a common EMS software layer within each sensor and actuator VNF.

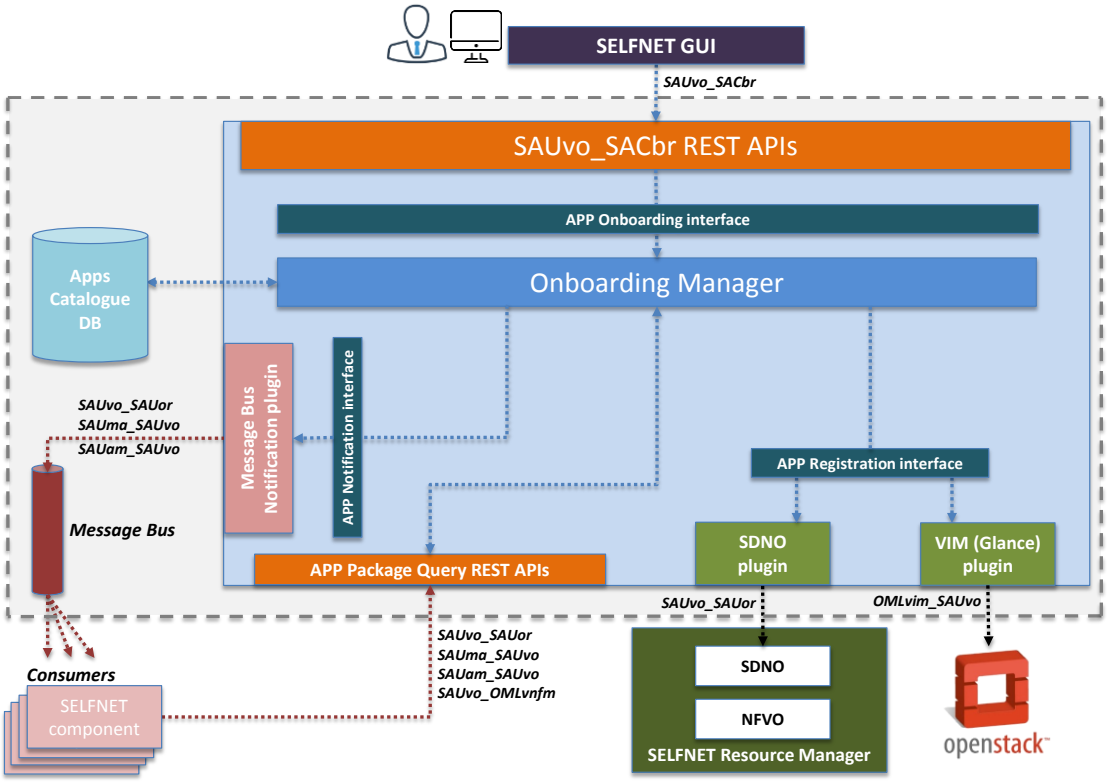


Figure 5-12: VNFs Onboarding Service and Plug-ins

The Figure shows the high level VNFM functional split and highlights its main functional entities and the interactions with the other components. This functional split includes those lifecycle management entities needed at the VNF side in support of the VNFM operations. Here, following the ETSI NFV principles, each VNF is generally considered as composed by a set of VNF components (VNF-C). Indeed, VNF developers and vendors may structure their VNFs into software components, and each one is deployed within an independent Virtual Machine. A VNF-C is therefore associated with a VNF software component, and a 1:1 correspondence exists between VNF-Cs and Virtual Machines. The way VNFs are structured into VNF-Cs generally depends on specific VNF vendor and developer factors. For each VNF-C, a lifecycle management (LCM) agent is embedded in the correspondent Virtual Machine to enable the communication with the VNFM and implement the lifecycle actions on the VNF. This agent enables the containerization of VNF and ETSI MANO EMS components into encapsulated VNFs, providing a common interface towards the VNFM.

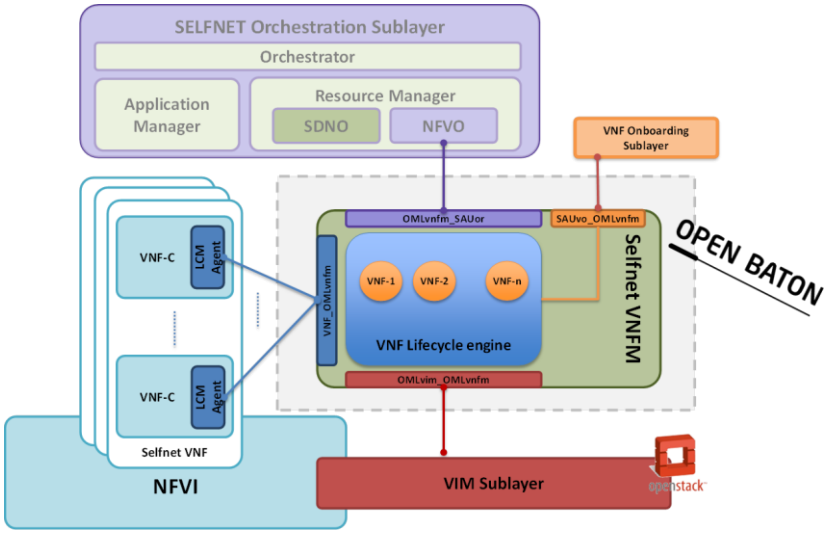


Figure 5-13: SELFNET VNFs Lifecycle Management

5.3.3.1 Automated deployment of physical and virtual infrastructures

5G architectures will include both physical and virtual infrastructures. The 5GPPP effort has achieved integrated management of physical and virtual infrastructures, which enables automated deployment of 5G infrastructures and services running on top of them, including virtualization services, cloud computing, Multi-access Edge Computing (MEC), SDN/NFV services and value-added services such as Service Function Chaining (SFC). Consequently, the creation and deployment time for infrastructures and their services are greatly reduced, from days to minutes. This achievement significantly contributes to realizing the service deployment KPI envisioned by 5G PPP. In addition, a 5G topology viewer allows visualizing correlated physical and virtual infrastructure elements and mobile users' connectivity in real time.

Management of the physical layer has been achieved by using the latest version of the MaaS software package provided by Ubuntu. MaaS provides the following primary functions: automatic deployment of Operating Systems in multiple zones, automatic configuration of network interfaces and storage layouts and support for on-demand acquisition of computing resources by multiple users. When a new computing resource is plugged into the infrastructure for the first time, it is automatically discovered by MaaS, and at this point the resource status is marked as New. This allows the administrator to commission the node by installing an OS in a reserved partition and thus gain the control and governance of the node. This action moves the node to Commissioned status, which in turn allows a user to acquire the computing resource, moving it to Acquired status. Once a node has been acquired, the user can configure the network, layout of the disks, and select the OS he/she wishes to deploy. By installation and configuration of these user-selected options, the node is moved to Deployed status, and at this point, the node is fully operational. Furthermore, network-facing SDN/NFV services or facilities such as OpenStack based VIM and OpenDayLight based SDN controller can be automatically installed. For the Cloud-RAN in a 5G network, these services can be pushed to the RRH and the BBU as part of the automated installation process. Figure 5-14 illustrates this scenario, using LTE evolution based approach as an example, where OpenAirInterface is employed using LXC (Linux Containers).

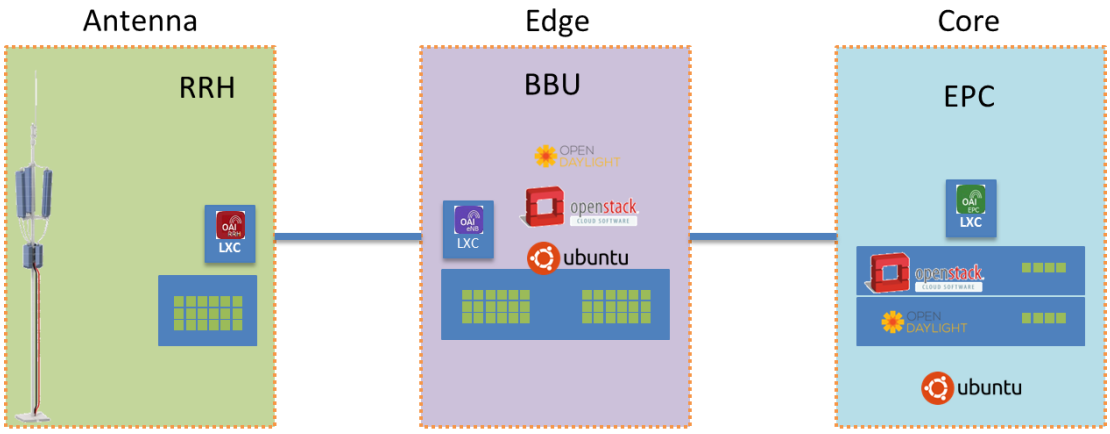


Figure 5-14: Automated Physical and Virtual Infrastructure Deployment

5.3.4 Multi-Domain and Multi-Operator Operation

The key component of 5G is going to be the inclusion of verticals into the telecommunication evolution. However, verticals like automotive and industry companies have sometime footprints that do not match the telecommunication coverage area. Furthermore, the specialized services for the verticals may require technical skills or components not available at a single operator. In such a scenario, operators would have to work together to create a *coopetative* ecosystem where while the operators compete with each other they also cooperate to serve their end customers, the verticals.

To achieve multi-operator interaction work has been done in defining the interfaces and components between the operators' multi-domain orchestrators (MdOs), see Figure 5-15. The architecture maps to the overall architecture picture shown in Figure 5-1. The Topology abstraction and discovery system implements the abstraction components of Figure 5-1. While the business level relationships such as pricing are not yet clarified the catalogue management system implements the service repository of Figure 5-1. The detail interfaces are shown showing the different types of functionality that exists between the operators. The section provides a brief overview of the functionalities exposed between the operators.

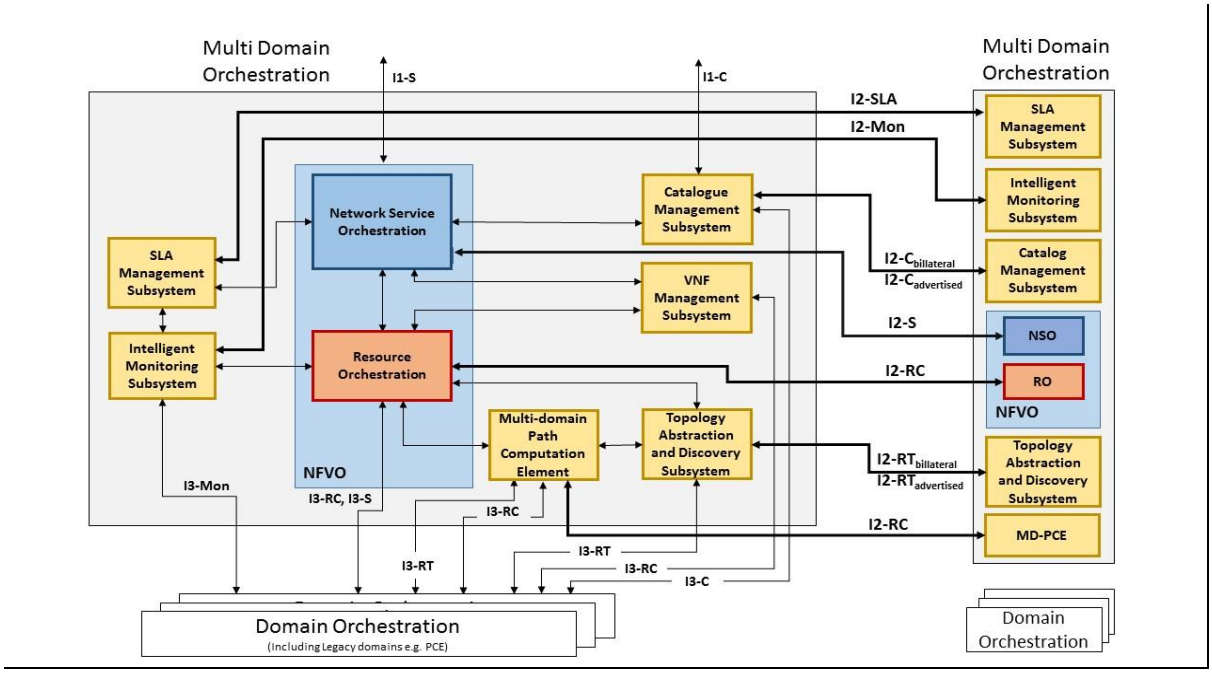


Figure 5-15: Implementation Design Architecture for multi-domain orchestration

The components of the architecture for inter-operator orchestration can be classified into three main groups, according to the core functionalities of the MdO. The components belonging to the first group are responsible for capabilities and resource acquisition and these ones can be seen on the right part of the MdO architecture. The components belonging to the second class are responsible for service request deployment (e.g., orchestration) and can be seen in the middle of the MdO. Furthermore, the third group contains components that are in charge of service assurance and SLA management; these components are placed on the left side of the MdO.

The short description of the architecture components and the relations among them is summarized according to the above-mentioned classification:

1. *Components that gather resource and service related information*: Each MdO needs to be aware of the service capabilities and resources of other administrative domains in order to make service and resource orchestration decisions. Two types of interfaces are distinguished in the architecture, namely advertisement-based and bilateral ones. With the help of the advertisement-based interfaces (such as I2-$C_{advertised}$, i2-$RT_{advertised}$) an MdO can announce its available resource types and capabilities on the service level. By using bilateral interfaces, the MdOs can exchange e.g. topology, resource, service, pricing information. The components responsible for information gathering are the Catalogue Management Subsystem (CMS) and the Topology Abstraction and Discovery Subsystem (TADS).
   The Catalogue Management Subsystem is responsible for maintaining the VNF Catalogue (list of available function a Service Provider can utilise to compose Network Services), Network Service Catalogue (list of services that a certain MdO provides for a

Customer) and takes care of the catalogue synchronization among MdOs via I2-C. In P2, a pricing model was introduced as a subcomponent of the Catalogue Management Subsystem.
Topology Abstraction and Discovery Subsystem (TADS) is in charge of maintaining a database about the networking, compute and storage resources available for the RO and NSO. There is also a topology information exchange between the TADS and the Multi-Domain Path Computation Element (MD-PCE); TADS acquires the intra-domain connectivity topology from MD-PCE, while informs MD-PCE about the Multi-domain topology gathered from the TADS components of other MdOs.

2. *Components that relate to the orchestration procedure*: When the MdO receives a request from the customer the deployment operation is carried out by the Network Service Orchestration (NSO), Resource Orchestration (RO) and the MD-PCE. The NSO is responsible for handling the Network Services (NS), requested by the customers or NSO of another MdO. RO is in charge of embedding the resource requests to the available resources offered by the TADS. The MD-PCE can be considered as a legacy component to support connectivity services, in charge of provisioning the network connectivity services using the information provided by TADS. MD-PCE has several interfaces to other components in the same MdO, as well as to MD-PCE in other MdOs.

3. *Assurance components for the deployed service*: When the service deployed, the NSO/RO passes the SLA and resource entities to be monitored to the assurance components of the architecture, namely SLA Management Subsystem and Intelligent Monitoring Subsystem (IMoS). IMOS is responsible for the coordinated deployment and management of probe-based measurement methods for different domains within a provider or across multiple providers by interworking with IMOS component of other MdOs. The SLA Management Subsystem is responsible for evaluating service KPIs belonging to a certain running instance by using the performance measurement reports provided by the IMoS. I2-Mon is defined as an interface between the IMoS components of different MdOs, dedicated for exchanging performance monitoring related actions.

The above description includes the main parts of the orchestration functionality that needs to be supported across operators. However, the main challenge of multi-domain operations comes from the fact that each operator exposes a very limited/abstracted view of the topology to other operators. Each of the other functionalities work on this abstracted view of the topology and resources. The host MdO then needs to translate the requests for deployment as well as monitoring and management over the abstracted components to actual orchestration and management decisions over the real topology as perceived by the host MdO. The other challenge of exposing access to such programmable interfaces is on how to actually secure them. This is explained in the following section.

5.3.4.1 Secure Multi Domain Interfaces

The multi-domain scenario is definitely a new case in the 5G arena, and requires special attention from security perspective. The most relevant security recommendations have been found in [5-14][5-15][5-16][5-17].

The security of such a multi-operator instantiation entails multiple actors who are using different MdO interfaces, and a trust model needs to be defined for each of them. Obviously, any security guideline relevant to single operator system is a pre-requisite for a multi-operator scenario, for instance, a Firewall is assumed to be present to protect I1-* interfaces from external attacks, specifically a WAF (Web Application Firewall) might be suitable for this case.

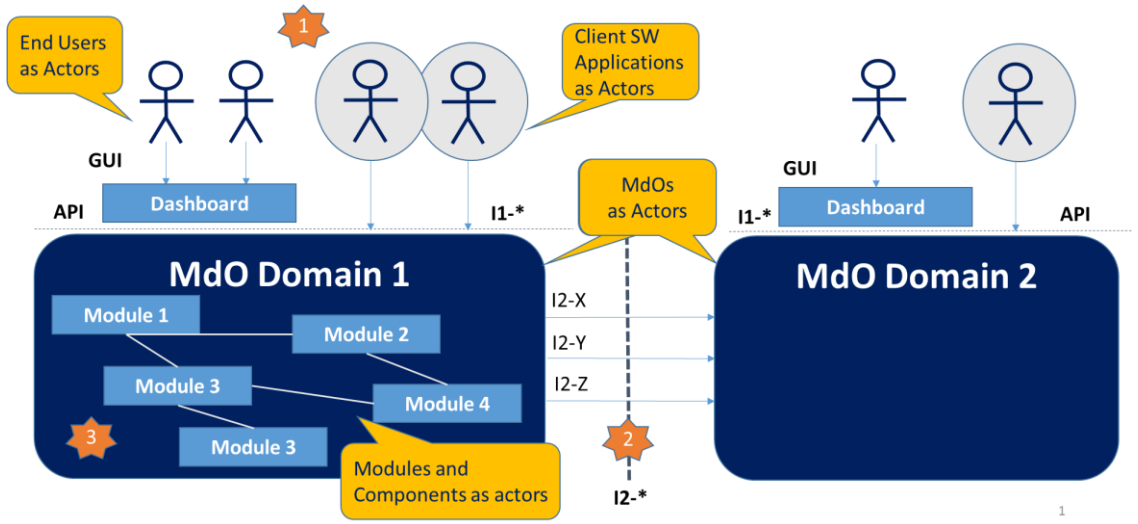


Figure 5-16 Security in the Multi-Operator Scenario

With reference to the interactions depicted inside Figure 5-16, the following actors inside MdO security trust model have been considered (the numbering of the next paragraphs matches the digit inside the orange star in the figure):

1. Customers (tenants) interact with Multi Domain Orchestrator (MdO) through interface I1 (B2C interaction), that provides a set of services to operate with VNFs and NSs. These customers are 'private' to the MdO they're interacting with and therefore:
   - They need to be identified, authenticated, authorized and accounted by the specific MdO (provider)
   - Credentials used by customers are valid only for the MdO who provided it to their clients.
   - Their identity, for business privacy reasons, should never be forwarded to any partner provider (other MdOs) in any traceable way

   Any software using interface I1 to provide client access to the MdO functionalities on behalf of a customer shall use a specific MdO service to authenticate the customer (login) that will return a JWT token (JSON Web Token [5-18]). This JWT shall be inserted in any other I1 service invocation header, so that the MdO will be able to authenticate, authorize and account the customer on any invoked service. Moreover each customer has specific roles (e.g. VNF provider, NS provider, NS consumer) that will be used to perform a further step of role-based authorization on certain I1 services.

   The connection between the client SW (e.g. the Dashboard or any other client custom developed SW using I1) needs to use a secure channel to the MdO endpoint (i.e. HTTPS with TLS, that provides data privacy and server authentication).

2. MdOs collaborate one with another (B2B interaction) using I2 services. These services are exclusive for the MdO to MdO communication, and can't directly be used by other actors, and therefore:
   - All collaborating MdOs need to mutually identify, authenticate, authorize and be accounted
   - X509 certificates will be used as MdOs credentials, an each MdO will use a single certificate for all services of I2 interfaces.
   - These certificate can be provided either by an external public Certification Authority (CA) – and in this case each MdO will need only one certificate - or directly by each provider private CA to its own 5GEx MdO partners - in this case an MdO will have one certificate for each 5GEx partner)

The service requests are responses need to support non-repudiation; therefore, appropriate footprint digital signing schemes need to be applied (under discussion at ETSI NFV ISG Security WG).

3. The SW modules and components of an implementation of MdO are another relevant actor for the security, and need to be protected inside an MdO specific trusted zone of the provider space:
   - The module exposing the I1 interface (e.g. MdO customer endpoint) needs to be protected as any other provider external service (for instance by a Firewall), still having external visibility on the customer accessible network (very likely Internet)
   - No other MdO module shall be directly reachable from the external customer accessible network
   - Modules of 5GEX implementation can communicate with each other inside the specific provider trusted zone (with IP whitelisting enforced)
   - All the modules/components implementing communications to/from other MdOs need to comply to the previously mentioned inter-MdO trust model. To simplify the implementation of a secure inter-MdO communication, it's highly desirable to consider the usage of a proxy/gateway also for all the services of I2 interfaces, i.e. a single module that receives requests from all other internal modules, and consistently applies the security/trust model for all inter-MdO communications
   - From operational perspective, the implementation of MdO based on virtualized components, like virtual machines or containers, shall follow all the typical guidelines for such deployments in trusted zones scenarios, and – in addition - consider digital signatures of the virtual image as an additional measure to guarantee SW integrity.

5.4 Self -Organizing Networks and Services

5G is expected to support a large number of slices, each specialized to support a particular vertical or a particular usage scenario. As such, these slices and the services they host will need to be designed and managed by specialists in the respective fields. Such a scenario would be strongly assisted by autonomic management of the network for the operator on a day to day basis. Hence, 5G will include intelligence-based autonomic network management to improve network performance whilst reducing OPEX.

5.4.1 Automated/Cognitive network management

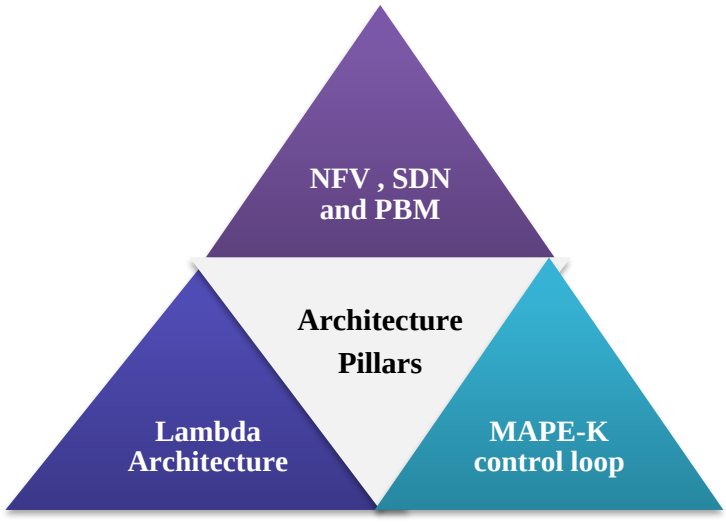


Figure 5-17 Architecture Pillars for Cognitive Management.

The Cognitive Smart Engine (CSE) is responsible for delivering the Machine Learning capabilities towards an autonomic network management. This component receives as input monitoring information from the managed environment and optionally external inputs such as contextual information, and it will further process this information and send events to the Policy Engine based on the machine learning modules outputs.

The architecture of the CSE component in Figure 5-1 relies mainly on three pillars depicted in Figure 5-17 and is further reviewed in the following subsections: (1) It assumes an environment based on Network Functions Virtualisation (NFV) and Software-Defined Networking (SDN).. (2) The machine learning components of the CSE, the key focus for automation, are able to take advantage of both batch and streaming processing methods, are inspired by the lambda architecture. The lambda architecture is a framework for designing big data applications that serves a variety of use cases with different latency requirements. The CSE is inspired from the lambda architecture but adapted to support machine learning functionalities rather than querying purposes which is the original aim of the lambda architecture. (3) CSE aims to deliver an autonomic network management solution based on machine learning, hence components and workflow were designed to support the *M*onitor, *A*nalyse, *P*lan, and *E*xecute parts that share Knowledge (MAPE-K) autonomic loop. To this end the CSE re-uses the monitoring data collected by the monitoring engine in Figure 5-1.

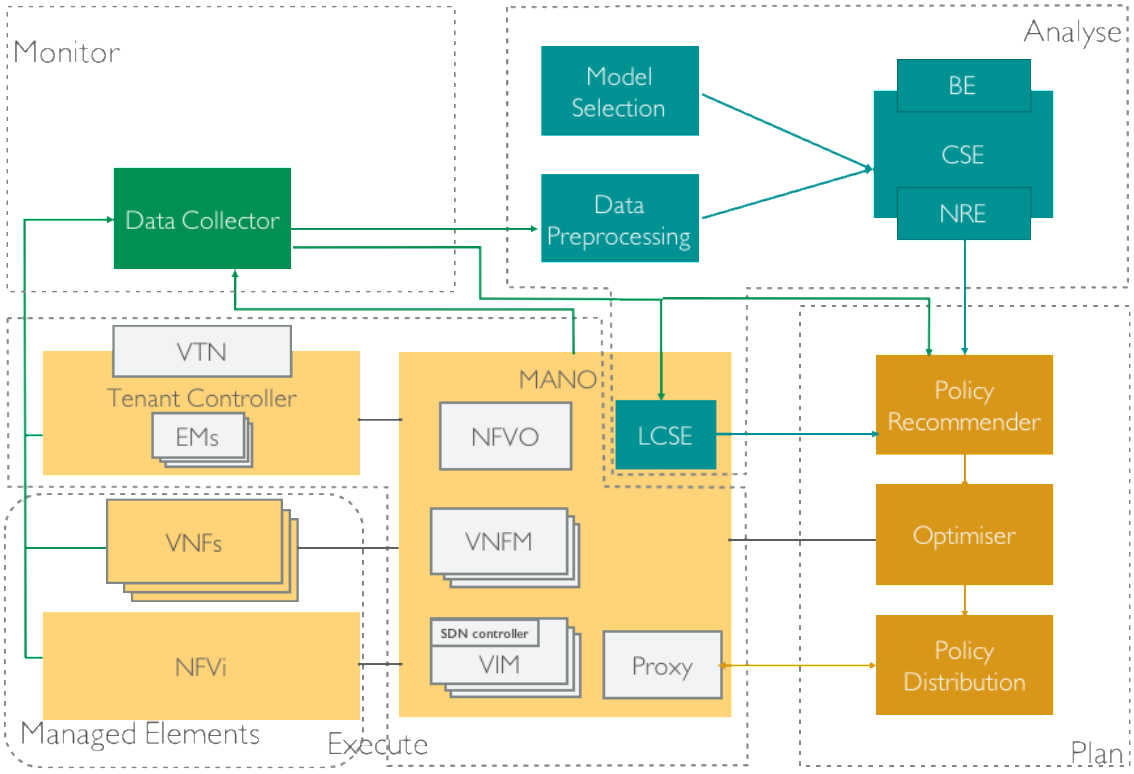


Figure 5-18 CSE in depth functioning architecture mapped to MAPE autonomic loop: Data Sources (Light Orange), Data/Collection (Green), Cognitive Elements (Blue), Policy Engine Elements (Dark Orange).

One of the main features of CSE is the ability to dynamically adapt to changes by combining machine learning models and network management policies. The architecture can be described by the MAPE autonomic loop, as presented in Figure 5-17 which executes the following actions in a loop manner:

1) Monitor: The Data Collector works gathers the monitoring details from the monitoring element as well as details of the required SLA from the SLA management element.

2) Analyse: The collected information is forwarded for analysis to the CSE or Light-weight CSE (LCSE)[5] which are intelligent agents to perceive the network state and its external environment, and use these insights to assist the network management of the system.

3) Plan: The output of the (L)CSE is sent to the Policy Engine, which recommends network policies to enact a desired alternation in the network infrastructure.

4 on 5) Execute on Managed Elements: Consequently, managed network resources, such as VNFs and NFVi, are adjusted though request to the NFVO and other controllers based on the recommended actions.

The (CSE) is responsible for receiving the state and resource consumption records, pre-processing the records, selecting suitable algorithms, and then applying selected models to further process the received data. The CSE is enhanced by a Batch Engine (BE) that processes data in batches, and by a (Near) Real-time Engine (NRE) that processes data in lower latency manner. The CSE supports various machine learning modules that in turn help deliver different functionalities for the slices. These include data gathering functionality, forecasting and prediction functionality, anomalies and fault recognition functionality, and action recommendation functionality for the policy engine.

The Policy Engine is mainly responsible for mapping insights from the LCSE/CSE into appropriate policy actions that can be directly understood by related components in the Management and Orchestration functions.

[5] Note that LCSE comes with less computational resources but is deployed closer to data, thus it can offer the results of the analysis in a lower latency manner. Contrarily, CSE is equipped with more resources but is deployed away from data. Both of them can work together to compensate each other and offer a more flexible and efficient solution.

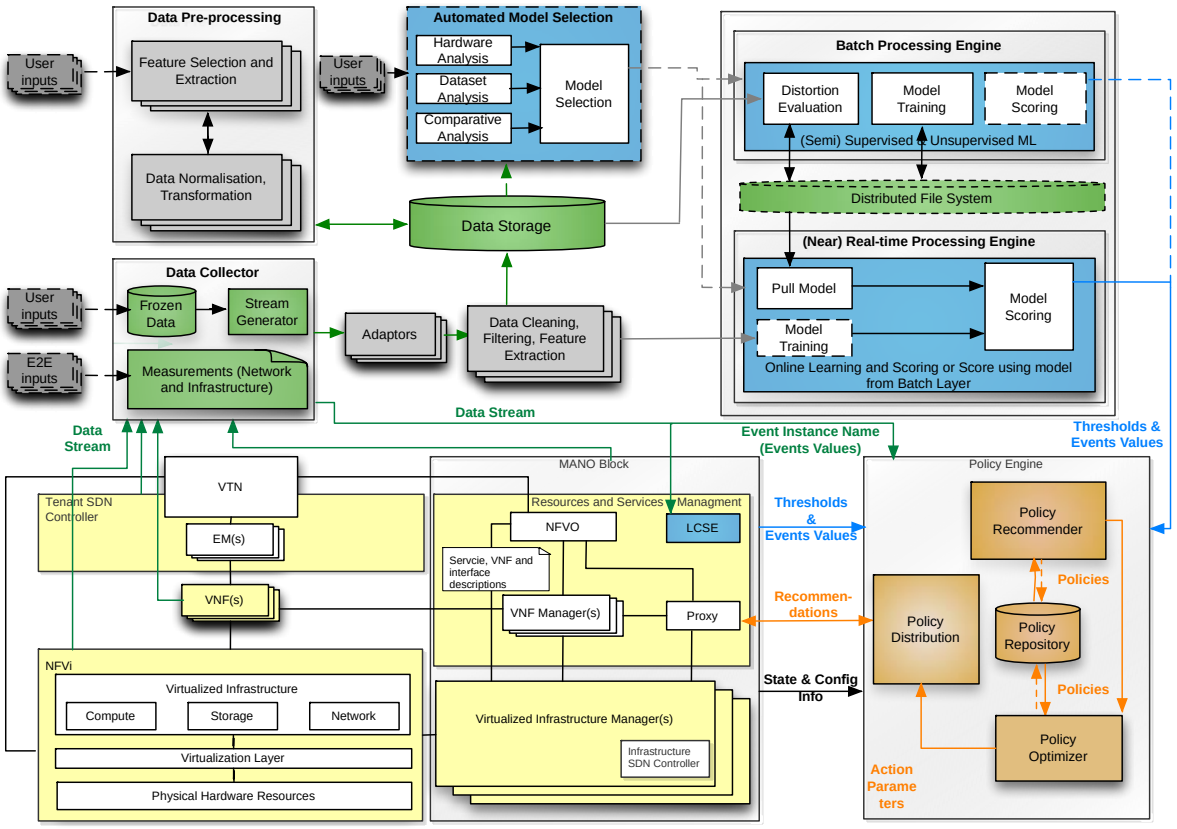


Figure 5-19 CSE Operational Architecture.

The CSE, depicted in the upper half of the operational architecture from Figure 5-19 is an enabler for autonomic network management. This component is responsible for receiving the state and resource consumption records, pre-processing the records, selecting suitable algorithms, and then applying selected models to further process the received data. The objective of the CSE is to support the various Machine Learning modules that in turn contribute to the delivery of various functionalities These functionalities in turn have associated policies in the Policy Repository of the Policy Engine.

The input of the CSE will be a data stream on the relevant events from both resource provider-side and consumer-side, whilst its output will be scored on the states of given components in the architecture. This is intended to increase the openness and transparency of services delivered by 5G networks, and subsequently provide better user experience. Moreover, the output of the CSE are prediction scores, which can be divided into:

1. Thresholds for specific policies, such as the maximum CPU utilisation before reaching performance degradations, or
2. Metric data at timestamp *t*, such as predicting the: (i) %CPU (*numerical*), or (ii) presence of an anomaly or fault (*categorical*).

The CSE consists of following sub-components:

- *Data Collection & Adaptors*: it collects data from multiple resources, and maps collected data into those that can be processed directly by the following components.
- *Data Cleaning & Filtering*: it cleans and refines received data, and then stores it into the Data Storage or forwards it to the (Near) Real-time Processing Engine.
- *Data Storage*: it stores historical data, and makes them available for multiple components constituting the CSE.
- *Data Pre-processing*: it can work in either automatic or manual mode to pre-process collected data stored in the Data Storage and make them ready for the Batch Processing Engine. Feature extraction can be achieved by Deep Neural Networks, which can

generate highly informative features automatically. Such functionality is essential to support the overall flexibility of the architecture and to keep it adjustable to constantly changing environment. It controls the noise and reduces the processing time of analytic works in the big data context.

- *Algorithm Selection*: similar to the Data Pre-processing, this component is able to identify the model(s) that will be deployed on both processing engines automatically or based on customer requirements. In the automatic mode, it will take account into the features of given datasets, the performance of candidate algorithms, and available resources of the processing engine.
- *Batch Processing Engine*: it retrieves consumption and state data from the Data Storage, and applies these data to train a model or generate scores. In the former case, the Batch Processing Engine will evaluate the distortion of current model. If the model has become *stale* or no model is available, it will generate a new model from scratch to facilitate the work of the (Near) Real-time Processing Engine. In the latter case, this engine works independently to analyse collected records in a more accurate but higher latency manner. Note that the scoring in both the Batch Processing Engine and (Near) Real-time Processing Engine is not to simply apply one machine learning model but may involve a sequence of models associated with post-processing. For example, to detect network anomaly, we may need to score a number of records and then make a conclusion based on a linear combination of generated scores.
- *Distributed File System*: it stores models generated by the Batch Processing Engine that will be deployed on the (Near) Real-time Processing Engine. Note that this component is optional since the Batch Process Engine may forward generated models directly, such as through message queues/RESTful Web Services or the two processing engines may not share data between each other without writing it to an external storage system if they are implemented and deployed in some cluster computing systems, such as Apache Spark.
- *(Near) Real-time Processing Engine*: it consumes the data from the sources directly, and scores the received data within a short period of time. This can be achieved by applying the model generated by the Batch Processing Engine, or light-weight on-line learning approaches directly, such as some on-line clustering algorithms. Depending on the latency requirement of a network, this component can generate scores either in real-time if it is implemented and deployed in a distributed real-time computation system, such as Apache Storm, or near real-time if it is powered by a mini-batch system, such as Spark Streaming.

This section presented architectural pillars and design towards delivering autonomic network management. It explored the integration of cutting-edge technologies in related areas, such as NFV and machine learning. By leveraging feature extraction and model selection, this architecture work applies machine learning to predict and evaluate states of network elements and responses to events that can potentially affect network performance and SLA. This turns networks into flexible, programmable platforms with intelligence to scale up and down. Subsequently, this can result in a significant reduction in CAPEX, OPEX, as well as service deployment time, and large savings in energy efficiency.

5.4.2 Autonomic network management framework

SELFNET has designed an autonomic network management framework based on an innovative Tactical Autonomic Language (TAL) specification, which facilitates a formal definition of the autonomic behaviours in the SELFNET framework and logically links a Self-Organising Network (SON) engine, the monitoring subsystem, the orchestrator, Apps management and other relevant components in the architecture, as shown in Figure 5-20. This creates an autonomic control loop to significantly facilitate network management operations in an automated and autonomic way. The investigation of machine-learning based algorithms is underway to improve the prediction and self-learning capabilities of the framework. It is expected that the SON autonomic management engine will extend the current 4G SON concept in the physical layer of 4G networks

to the management plane of both physical and virtual networks in 5G networks, and significantly simplify network management tasks and minimise human intervention and labour in diagnosing complex network problems and then determining corresponding reactive or pro-active tactics to resolve existing or even avoid potential issues. Representative use cases include Self-healing against network failures or vulnerabilities for improved reliability, Self-protection against cyber-attack threats for improved security, and Self-optimization against network constraints for improved users' QoE.

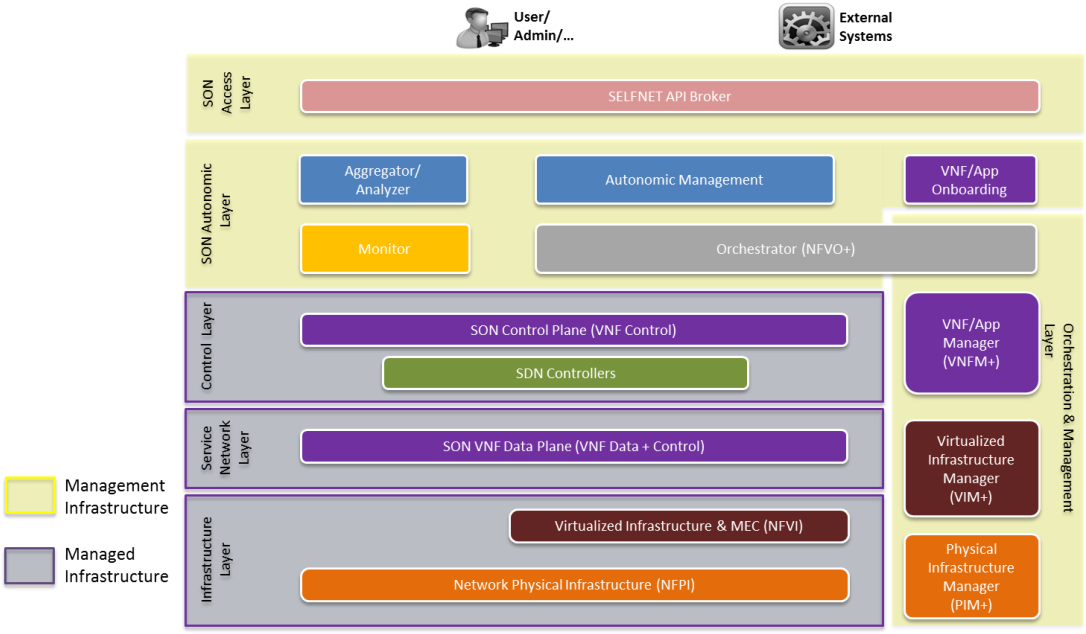


Figure 5-20: Overall Self-Optimized Architecture

The concept of 5G SON is addressed in a twofold manner within the SON-based autonomic management engine, as illustrated in Figure 5-21. At first, a set of tactical strategies is expressed by the use of the TAL. Their aim is to specify the human-based tactical approaches regarding the reactions of the system in view or detection of certain events and anomalies in the system being controlled. Secondly, an algorithmic processing of the indicated strategies along the gradually built (machine learning based) artificial intelligence is expected to produce optimal decisions. In the context of this autonomic management framework, TAL constitutes the static definition of the intelligence, usually incorporating experience of network management personnel, that at least, provides the initial starting point where to the common sense intended behaviour is expressed and that may thereafter be optimized by the artificial intelligence based processing.

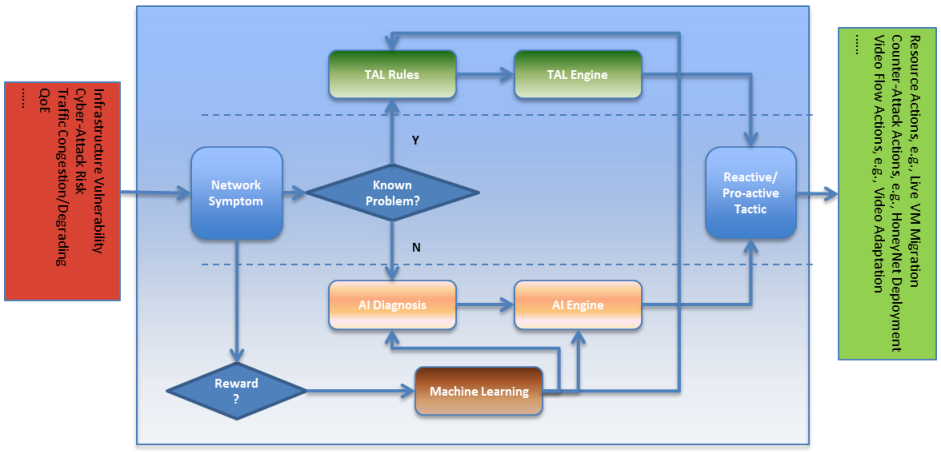


Figure 5-21: Autonomic Network Management Framework: SON Engine

5.4.3 Balancing autonomy against explicit control: A pragmatic approach

The previous paragraphs have outlined options to do autonomic or cognitive management. In principle, these techniques are applicable on a wide range of fields, and will be applied to NFV and slice management as well. Nevertheless, there is a balancing desire to extend explicit control to both network operators as well as service developers, if they choose to exert such controls and offer the opportunity for such control.

This can take place on multiple levels. On the level of an individual (virtual network) function, it is difficult to imagine entirely autonomous, self-managed operation without taking into account the corresponding actions of the peer functions inside the same service; this could easily produce counterproductive behavior or race conditions. Hence, some coordination across functions in a service is needed. This coordination could take place in a one-size-fits-all fashion, attempting to regulate and manage all services (and their constituting functions) in a single manner. While this would be easy to realize and implement, it obviously contradicts the ideas of autonomic actions defined by the service.

When embracing autonomy, we hence need a way to balance the needs of explicit control, aligned with typical MANO frameworks, with the desire to delegate as much functionality as possible to services. The architectures outlined above could fulfill this task. Another approach could be to endow a more conventional MANO framework with the ability to incorporate such autonomic management approach and give services the power to regulate themselves, while having access to relevant parts of infrastructure, load, topology, descriptions and so on. Clearly, such information needs to be carefully shepherded, the interactions of services with each other have to carefully controlled, and the control actions of a service's self-management functionality has to be carefully curried to ensure stable operation. Once a MANO framework provides such functionality, it could be a both pragmatic and far-reaching option to combine autonomic approaches with the desire for more fine-grained control in an actual network.

Once we accept such an approach, it also opens a new avenue to thinking about the relationship of slices and services. In the simplest case, a "slice" could be just a single service along with an SLA, executed in (conceptually) perfect resource isolation. In fact, inside such a conceptually perfect slice, it does not really matter much how a service is managed as it runs in isolation. The picture changes somewhat if the notion of a slice is widened, with multiple services living inside a single slice (an easy to imagine, plausible extension of a simple reference model). Then, it does in fact make sense to start thinking about the way these services are managed or manage themselves with respect to each other. In this sense, autonomic service management opens a door to a more powerful, useful slicing model.

5.5 References

[5-1] Navid Nikaein, et al., "Network store: Exploring slicing in future 5g networks," in Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture, ACM, 2015.

[5-2] Katsalis et al., "Architectural Design Patterns for the RAN," in IEEE ICC, 3rd International Workshop on 5G Architecture, 2016.

[5-3] F. Ahmed et al., "Distributed Graph Coloring for Self-Organization in LTE Networks," Journal of Electrical and Computer Engineering, 2010.

[5-4] ETSI - GS NFV-IFA 014, Network Functions Virtualisation (NFV); Management and Orchestration; Network Service Templates Specification, v2.1.1, October 2016.

[5-5] https://osm.etsi.org/wikipub/images/0/0c/Osm-r1-information-model-descriptors.pdf

[5-6] http://docs.oasis-open.org/tosca/tosca-nfv/v1.0/tosca-nfv-v1.0.html

[5-7] Draft ETSI GS NFV-SOL 001, Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; NFV Descriptors based on TOSCA; TOSCA-based NFV descriptors, v 0.0.2, July 2016.
[5-8] Open Grid Forum NSI-WG, "Network Services Framework v2.0", available at: https://www.ogf.org/documents/GFD.213.pdf
[5-9] Open Grid Forum NSI-WG, "Network Services Framework v2.0", available at: https://www.ogf.org/documents/GFD.213.pdf
[5-10] Ishan Vaishnavi, Riccardo Guerzoni, Riccardo Trivisonno, "Recursive hierarchical embedding of virtual infrastructure in multi-domain substrates", Network Softwarization (NetSoft) 2015 1st IEEE Conference on, pp. 1-9, 2015.
[5-11] Guerzoni, Riccardo, et al. "A novel approach to virtual networks embedding for SDN management and orchestration." Network Operations and Management Symposium (NOMS), 2014 IEEE. IEEE, 2014.
[5-12] B. Németh,B. Sonkoly, M. Rost, S. Schmid, "Efficient service graph embedding: A practical approach," in IEEE NFV-SDN, 2016.
[5-13] Dräxler, H. Karl, Z. Á. Mann: Joint Optimization of Scaling and Placement of Virtual Network Services. Accepted in 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2017).
[5-14] http://www.etsi.org/deliver/etsi_gs/NFV-SEC/001_099/001/01.01.01_60/gs_NFV-SEC001v010101p.pdf
[5-15] http://www.5gensure.eu/
[5-16] https://5g-ppp.eu/5g-ppp-security-work-group-outcomes/
[5-17] https://www.owasp.org
[5-18] RFC 7519 - JSON Web Token (JWT) - IETF Tools, https://tools.ietf.org/html/rfc7519

6 Impact to standardization

To be added by the 5GPPP Pre-standard WG

7 Conclusions and Outlook

To be added at the end

8 List of Contributors

To be added