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

Project Name:

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

# Deliverable D3.3
# 5G NORMA network architecture – Final report

| | | | |
|---|---|---|---|
| Date of delivery: | 16/10/2017 | Version: | 1.0 |
| Start date of Project: | 01/07/2015 | Duration: | 30 months |

## Document properties:

| | |
|---|---|
| Document Number: | H2020-ICT-2014-2 5G NORMA/D3.3 |
| Document Title: | 5G NORMA network architecture – Final report |
| Editor(s): | Christian Mannweiler (NOKIA) |
| Authors: | Ignacio Labrador Pavon (Atos), Alessandro Colazzo, Riccardo Ferrari (Azcom), Paul Arnold, Markus Breitbach, Heinz Droste, Dirk von Hugo (Deutsche Telekom), Stan Wong, Vasilis Friderikos (King's College London), Zarrar Yousaf, Vincenzo Sciancalepore (NEC), Marie Line Alberi-Morel, Mark Doll, Borislava Gajic, Sylvaine Kerboeuf, Christian Mannweiler, Peter Rost, Bessem Sayadi, Peter Schneider (Nokia), Sina Khatibi (Nomor), Julie Bradford (Real Wireless), Rafael López da Silva, Rafael Cantó Palancar (Telefonica), Marcos Rates Crippa, Bin Han (TU Kaiserslautern), Marco Gramaglia, Albert Banchs (Universidad Carlos III de Madrid) |
| Contractual Date of Delivery: | 30/09/2017 |
| Dissemination level: | PU[1] |
| Status: | Draft |
| Version: | 1.0 |
| File Name: | 5G NORMA_D3.3.docx |

## Revision History

| Revision | Date | Issued by | Description |
|---|---|---|---|
| 1.0 | 16.10.2017 | 5G NORMA WP3 | Version delivered to project officer. |

## Abstract

The key achievement of 5G NORMA WP3 is the design of a conceptually novel, adaptive, and future-proof 5G mobile network architecture allowing to adapt the use of the mobile network (radio access, core, and transport) resources to the service requirements, the variations of the traffic demands over time and location, and the network topology (including the available front-/ back-haul capacity). The functional design is characterised by unprecedented levels of customisability, enabling stringent performance and security. This is achieved by specifying the two innovative functionality groups "multi-service and context-aware adaptation of network functions" and "mobile network multi-tenancy". The technical enablers for these functionalities, "adaptive (de)composition and allocation of mobile network functions", "software-defined mobile network control", as well as "joint optimization of mobile access and core network functions", are integrated into the 5G NORMA architecture design. Network functions from both the radio access and the core network domain have been decomposed and adaptively allocated to antenna site, edge cloud, or central cloud locations. This deliverable further presents the assessment and verification of the architecture based on realistic evaluation cases from the London study area defined in WP2. The security analysis in radio and core network domain has fuelled

[1] PU = Public

novel security concepts specifically addressing the challenges of multi-tenant and multi-service mobile networks with decomposed network functions. Finally, the deliverable describes an incremental migration path from 4G to 5G NORMA networks, focusing on functional and infrastructure evolution and allowing for co-existence of legacy technology components, particularly 4G (LTE-A and evolved packet core, EPC), and novel 5G NORMA building blocks.

## Keywords

5G mobile network architecture; multi-tenant and multi-service networks; network slicing; software-defined mobile network control (SDMC); control and data layer architecture; centralised control; MANO architecture, MANO as a service; deployment architecture; flexible function (de-)composition and allocation; QoE-aware network control; 5G NORMA stakeholders and offer types; virtualised AAA, RAN security concepts; Trust Zone; architecture design and requirements verification, verification methodology; migration paths towards 5G NORMA networks, migration of infrastructure and functionality; co-existence of 4G and 5G NORMA networks

# Table of Contents

# List of Figures

# List of Tables

# List of Acronyms and Abbreviations

| | |
|---|---|
| 3GPP | Third Generation Partnership Project |
| 5G NORMA | 5G Novel Radio Multiservice Network Adaptive Architecture |
| AAA | Authentication, Authorization and Accounting |
| ABS | Almost Blank Subframe |
| A-CPI | Application-Controller Plane Interface |
| AI | Air Interface |
| AKA | Authentication and Key Agreement |
| ANN | Artificial Neural Network |
| AP | Access Point |
| API | Application Programming Interface |
| AS | Access Stratum |
| ASIC | Application-specific integrated circuit |
| BB | Building Block |
| BNG | Broadband Network Gateway |
| BRAS | Broadband Remote Access Servers |
| BS | Base Station |
| BSS | Business Support System |
| CAPEX | Capital Expenditure |
| CC | Central Cloud |
| CC | Charging Control |
| CCCM | Central Cloud Connection Monitoring |
| CIO | Cell Individual Offset |
| CME | Central Management Entity |
| cMTC | Critical Machine Type Communication |
| CMTS (HFC) | Cable Modem Termination System (Hybrid Fiber Coax) |
| CNE | Core Network Element |
| CNF | Core Network Function |
| CoMP | Cooperative Multipoint |
| CORD | Central Office Rearchitected |
| CP | Connection Point |
| CPU | Central Processing Unit |
| CQI | Channel Quality Indicator |
| C-RAN | Centralised RAN |
| CU | Central Unit |
| CUG | Closed User Group |
| CWDM | Coarse Wavelength Division Multiplex |
| D2D | Device-to-Device |
| DC-GW | Data Centre Gateway |
| D-CPI | Data-Controller Plane Interface |
| DECOR | Dedicated Core Network |
| DoS | Denial of Service |
| DPI | Deep Packet Inspection |
| DPI | Deep Packet Inspection |

| | |
|---|---|
| DSL | Digital Subscriber Line |
| DSP | Digital Signal Processor |
| DU | Distributed Unit |
| DWDM | Dense Wavelength Division Multiplex |
| E2E | End to End |
| EC | Edge Cloud |
| EC | European Commission |
| eCG | E-UTRAN Cell Global Identification |
| EDA | Event-Driven software Architecture |
| eDECOR | Enhancements of Dedicated Core Network |
| eICIC | Enhanced Inter-Cell Interference Coordination |
| EJB | Enterprise JavaBeans |
| EM | Element Management (Element Manager) |
| eMBB | Evolved Mobile Broadband |
| eMBMS | Evolved Multimedia Broadcast Multicast Services |
| eNB | Enhanced Node B |
| EPC | Evolved Packet Core |
| EPS | Evolved Packet System |
| ES | Emergency Services |
| ETSI | European Telecommunications Standards Institute |
| FBMC | Filter-Bank MultiCarrier |
| FCAPS | Fault, Configuration, Accounting, Performance, Security |
| FE | Functional Element |
| FFT | Fast Fourier Transform |
| FG | Forwarding Graph |
| FPGA | Field Programmable Gate Array |
| G.FAST | Fast Access to Subscriber Terminals |
| GDB | Geo-location data base |
| GPP | General Purpose Processing |
| GPU | Graphical Processing Unit |
| GUI | Graphical User Interface |
| GWCN | Gateway Core Network |
| H2020 | Horizon 2020 |
| H2020 | Horizon 2020 |
| HSS | Home Subscriber Server |
| HAT | Horizontal Topic |
| HW | Hardware |
| IaaS | Infrastructure as a Service |
| ICT | Information and Communication Technologies |
| IE | Information Element |
| IFFT | Inverse Fast Fourier Transform |
| IMS | IP-based Multimedia Services |
| iNC | iJOIN Network Controller |
| InP | Infrastructure Provider |
| IoE | Internet of Everything |
| IoT | Internet of Things |

| | |
|---|---|
| iRPU | iJOIN Radio Processing Unit |
| iSC | iJOIN Small Cell |
| KPI | Key Performance Indicator |
| LAA | Local Access Assistant |
| LCM | Lifecycle Management |
| LDAP | Lightweight Directory Access Protocol |
| LEC | Local Exchange Carrier |
| LI | Lawful Interception |
| LSS | Local Subscriber Server |
| LTE | Long-Term Evolution |
| LTE-A | Long-Term Evolution-Advanced |
| M2M | Machine-to-Machine |
| MAC | Medium Access Control |
| MANO | Management and Orchestration |
| MANO-F | Management and Orchestration Function |
| MBB | Massive Broadband |
| MCS | Modulation and Coding Scheme |
| MDD | Multi-Dimensional Descriptor |
| MEC | Mobile Edge Computing |
| MIC | Multiple Integrated Core |
| MM | Mobility Management |
| MME | Mobility Management Entity |
| mMTC | Massive Machine-Type Communication |
| mmW | mm Wave |
| MNO | Mobile Network Operator |
| MOCN | Multi Operator Core Network |
| MOS | Mean Opinion Score |
| MPLS-VPN | Multi-Protocol Label Switching - Virtual Private Network |
| MSAN | Multi-Service Access Nodes |
| MSC | Message Sequence Chart |
| MTC | Machine Type Communication |
| NaaS | Network as a Service |
| NAS | Non-Access Stratum |
| NEM | Network Element Manager |
| NF | Network Function |
| NFD | Network Function Descriptor |
| NF-FG | Network Functions Forwarding Graph |
| NFV | Network Function Virtualization |
| NFVI | Network Function Virtualization Infrastructure |
| NFVO | Network Function Virtualization Orchestrator |
| NGMN | Next Generation Mobile Networks Alliance |
| N-PoP | Network Point of Presence |
| NR | 5G New Radio |
| NRT | Neighbour Relation Table |
| NSSF | Network Slice Selection Function |
| OAM | Operation, Administration, And Maintenance |

| | |
|---|---|
| OCS | Online Charging System |
| OLT | Optical Line Termination |
| ONF | Open Network Foundation |
| OPEX | Operational Expenditure |
| OS | Operating System |
| OSS | Operation Support System |
| OTT | Over The Top |
| P Router | Provider Router |
| PA | Power Amplifier |
| PaaS | Platform as a Service |
| PCRF | Policy and Charging Rules Function |
| PDCP | Packet Data Convergence Protocol |
| PDN | Packet Data Network |
| PDU | Packet Data Unit |
| PE | Provider Edge |
| PE Router | Provider Edge Router |
| P-GW | Packet Gateway |
| PHY | Physical Layer |
| PLMN | Public Land Mobile Network |
| PNF | Physical Network Function |
| PoC | Proof of Concept |
| PON | Passive Optical Network |
| PoP | Point-of-Presence |
| QoE | Quality of Experience |
| QoS | Quality of Service |
| RAM | Random Access Memory |
| RAN | Radio Access Network |
| RANaaS | RAN as a Service |
| RAT | Radio Access Technology |
| RBAC | Role-based Access Control |
| RF | radio frequency |
| RG | Requirement Group |
| RISC | Reduced Instruction Set Computer |
| RLC | Radio Link Control |
| RLC-AM | Radio Link Control, Acknowledged Mode |
| RLC-UM | Radio Link Control, Unacknowledged Mode |
| RNE | Radio Network Element |
| RNF | RAN Network Function |
| RNM | Radio Node Management |
| ROADM | Reconfigurable Optical Add-Drop Multiplexer |
| RRC | Radio Resource Control |
| RRH | Remote Radio Head |
| RTT | Round Trip Time |
| SA | Security Auditing |
| SaaS | Software as a Service |
| SBI | Southbound Interface |

| | |
|---|---|
| SCTP | Stream-Control Transport Protocol |
| SDMC | Software Defined Mobile Network Control |
| SDM-C | Software-Defined Mobile Network Controller |
| SDMC+O | Software Defined Mobile Network Control and Orchestration |
| SDM-O | SDM Orchestrator |
| SDM-X | SDM Coordinator |
| SDN | Software Defined Networking |
| SF | Service Flow |
| SFC | Service-Function Chain / Service Function Chaining |
| S-GW | Serving Gateway |
| SLA | Service-Level Agreement |
| SoC | System on Chip |
| SON | Self-Organising Networks |
| SPTP | Small Packet Transmit Procedure |
| sSF | Sub-Service Flow |
| SW | Software |
| TeC | Technology Components |
| ToR | Top-of-Rack |
| TP | Transmission Point |
| TR | Traffic Reporting |
| TZ | Trust Zone |
| UCA | User-centric Connection Area |
| UND | Ultra-Dense Network |
| UE | User Equipment |
| UEID | User Equipment Identifier |
| UICC | Universal Integrated Circuit Card |
| uMTC | Ultra-reliable Machine-Type Communication |
| URLLC | Ultra-Reliable, Low-Latency Communication |
| V2I | Vehicle to Infrastructure |
| V2X | Vehicle-to-Anything |
| V-AAA | Virtualised-Authentication Authorization Accounting |
| VDSL2 | VDSL – Very High Speed Digital Subscriber Line 2 |
| veNB | Virtual eNB |
| vEPC | virtualized Evolved Packet Core |
| VIM | Virtualized Infrastructure Manager |
| VIM | Virtual Infrastructure Manager |
| VM | Virtual Machine |
| VMNO | Virtual Mobile Network Operator |
| VN | Virtual Network |
| VNF | Virtual Network Function |
| VNF-FG | VNF Forwarding Graph |
| VNFM | Virtual Network Function Manager |
| VNPaaS | Virtual Network Platform as a Service |
| WAN | Wide Area Network |
| ZM | Zone Management |

# 1 Introduction

The key objective of 5G NORMA has been to develop a conceptually novel, adaptive, and future-proof 5G mobile network architecture with a clear roadmap towards adoption of important components by standards developing organisations (SDOs). The designed architecture is characterised by unprecedented levels of customisability, enabling stringent performance, security, and cost requirements to be met; as well as an API-driven architectural openness, fuelling economic growth through over-the-top innovation.

These objectives have been achieved by specifying the two innovative functionality groups "multi-service and context-aware adaptation of network functions" and "mobile network multi-tenancy". The technical enablers for these functionalities, "adaptive (de)composition and allocation of mobile network functions", "software-defined mobile network control", as well as "joint optimization of mobile access and core network functions", have been integrated into the 5G NORMA architecture design.

The 5G NORMA architecture design process has evolved in three *design iterations*. This deliverable concludes the architecture design phase of the third iteration, which will be completed by the final socio-economic evaluation in WP2, specifically D2.3 due in M30.

The main technical achievement of work package (WP) 3 "Multi-service Network Architecture" is the design of a mobile network architecture allowing to adapt the use of the mobile network (radio access, core, and transport) resources to the service requirements, the variations of the traffic demands over time and location, and the network topology (including the available front-/back-haul capacity). Mobile network functions from both the radio access and the core network domain have been decomposed and adaptively allocated to antenna site, edge cloud, or central cloud locations, depending on (i) the specific service and its requirements, e.g., bandwidth and latency; and (ii) the transport network capabilities (e.g., available front/back-haul capacity).

WP3 has integrated the novel technologies developed in WP4 and WP5 into an overall architecture and protocol design that meet the 5G-PPP and industry expectations to be cost-efficient and adaptable to cope with current and next generation services and applications. Further, novel security mechanisms have natively been integrated into the overall 5G mobile network architecture. Specific achievements of WP3 include

- Natively incorporate multi-tenancy support into the architecture by specifying tenant-controlled MANO layer stacks ("t-MANO") and software-defined mobile network controllers (SDM-C)
- Flexible architecture allowing for different instantiations in terms of functional perspective and deployment perspective, matching the service requirements,
- Integration of WP4 and WP5 "sub"-architectures into a harmonised control and data layer architecture,
- Definition of important interfaces between functional network elements, including controller northbound and southbound interfaces, inter-controller interfaces, and management & orchestration (MANO) layer interfaces,
- Assessment and verification of the proposed architecture based on the use cases and KPIs defined in WP2,
- Security analysis in radio and core network domain,
- Novel security concepts specifically addressing the challenges of multi-tenant and multi-service mobile networks with decomposed network functions,
- Development of a migration path that allows for integration of legacy technology, particularly 4G (LTE-A and evolved packet core, EPC),
- Contribution to 5G-PPP program activities related to the coordination of architecture design, in particular, major contributions to both version one and two of the Whitepaper of the Architecture Working Group.

**Structure of the deliverable**

This deliverable contains eight chapters and one annex.

Chapter 1 motivates the scope of the deliverable, summarises the most important achievements, and outlines the structure of the document.

Chapter 2 describes the 5G NORMA architecture design. It iterates the important architecture design principles and objectives, elaborates on the high-level functional architecture, and introduces the three non-functional perspectives on the architecture (deployment, topological, and resource perspective). Further, it provides a detailed description of the design of the management & orchestration layer, the control layer, and the data layer and explains how the concepts of Software-Defined Mobile network Control and Orchestration (SDMC and SDMO) are realised by the 5G NORMA architecture.

Chapter 3 provides a selection of so-called 5G NORMA procedures that exemplarily illustrate the interaction (in terms of involved interfaces and utilised protocols) between network functions from different layers. The described procedures include network slice lifecycle management, mobility of VNFs, and SDM-C related procedures, as well as charging and lawful interception. The chapter concludes with a description of how multiple technology components from several WPs interact to realise two important 5G NORMA innovations: service-specific network slice composition and customisation as well as multi-tenant network control and resource allocation.

Chapter 4 analyses the 5G NORMA ecosystem in practical scenarios. A brief update of the 5G NORMA stakeholder roles and the network slice offer types is followed by a thorough analysis of two network slice deployments for industrial communications ("Industry 4.0"). Besides elaborating on the interaction between private networks and public mobile network functionality, the chapter discusses critical interfaces between different administrative domains and provides a security analysis for varying requirement levels regarding confidentiality, integrity, and availability.

Chapter 5 depicts the novel 5G NORMA security concepts that are particularly important for virtualised multi-tenant and multi-service networks. It includes the description of concepts on virtualised authentication, authorization, and accounting, on tokenisation for resource provisioning and deployment, on RAN (access stratum) security, and on local Trust Zones and details how these concepts are embedded into the 5G NORMA architecture.

Chapter 6 performs the architecture design verification by defining three evaluation cases set in the London sample area as defined in WP2: eMBB baseline case, multi-tenant case, and multi-service case. Each requirement group (as determined [5GN-D21]) is checked against at least one of the evaluation cases, thus rendering an overall verification of the architecture design.

Chapter 7 outlines possible migration paths from 4G networks towards 5G NORMA networks. It takes into account current network infrastructure assets of network operators and compares it to the infrastructure requirements of the 5G NORMA functional architecture. The identified gaps are used to sketch a gradual transition from 4G to 5G NORMA, in terms of both functional migration and infrastructure (hardware) migration.

Chapter 8 concludes the deliverable by providing a summary of the 5G NORMA architecture design, identifying open issues, and proposing areas of future research.

# 2 5G NORMA Architecture

This chapter describes the results of the final 5G NORMA architecture design, which has undergone three design iterations with continuous feedback from work packages (WPs) 2, 4, 5, and 6. In this approach, the early-defined high-level architecture has been refined on the level of architectural layers (management & orchestration, control, and data layer) and network domains (radio access and core network), particularly focusing on the novel cross-layer 5G NORMA concepts of Software-Defined Mobile network Control and Orchestration (SDMC and SDMO).

The 5G NORMA design principles and objectives are summarised in Section 2.1, briefly motivating the need for multi-service and multi-tenant networks and explaining how 5G NORMA realises the network slicing concept. Section 2.2 depicts the high-level architecture of the 5G NORMA system and elaborates on the four perspectives (functional, deployment, topological, resource) defined in WP3 and their respective purpose. Section 2.3 integrates the architectures of WP 4 and WP 5 into a common control and data layer architecture. The chapter concludes with Section 2.4 on the MANO layer. It specifically describes solutions for the integration of physical and virtualised network functions (NFs) into a common 5G NORMA lifecycle management framework, the deployment of tenant-controlled NFV MANO (t-MANO) stacks, and management & orchestration across multiple administrative domains.

## 2.1 Architecture design principles and objectives

In the past, mobile network architectures were designed to provide a limited set of services (mostly voice and Internet access), and they employed a single specific kind of deployment. For example, in 4G, base stations provided users with a single broadband packet data connection. Today, the introduction of new enabling technologies such as software-defined networking (SDN) and network function virtualisation (NFV) has opened the way to implement new concepts and explore myriads of possibilities. 5G NORMA has defined a versatile network architecture concept in the context of end-to-end (E2E) network slicing to provide and manage customized logical mobile network instances tailored to the individual vertical's service requirements while being cost- and energy-efficient at the same time. This section will highlight the key objectives and explore the architecture design principles in the context of those objectives.

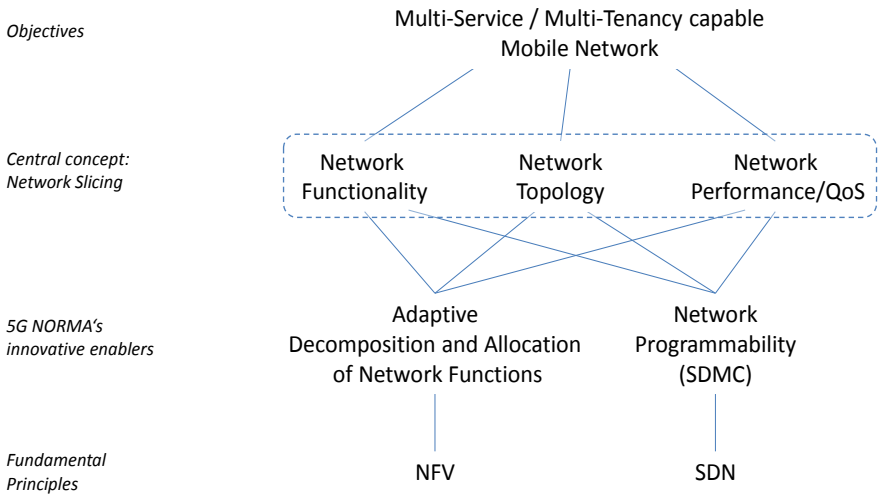



**Figure 2-1: 5G NORMA architecture design concept**

### 2.1.1 The need for multi-service / multi-tenancy capable networks

Mobile networks are becoming an important enabler for the ongoing digitization of business and daily life and the transition towards a connected society. Services in high priority areas for our society such as education, health, government or in vertical business areas such as smart grids, transportation, automation increasingly rely on a mobile infrastructure. In these areas, machine-type devices will increasingly contribute to mobile traffic, but with characteristics and requirements significantly different from today's dominant human-centric traffic. Hence, future 5G networks must combine high performance with the support for service-specific functionality.

Furthermore, 5G networks are expected to offer operators unique opportunities to address and offer new business models to consumers, enterprises, verticals and third-party partners. A large set of different customer types, each demanding specific communication solutions, is intrinsically difficult to handle due to the widely varying requirements that need to be addressed at any point in time of deployment.

Dedicated physical networks would be an obvious solution, as they could be perfectly tailored and adapted to service- or business-specific needs and deliver maximum performance. However, in many cases, the resources of such networks would remain underutilized most of the time as well as in large geographical areas. For economic reasons, a common infrastructure platform that is shared among multiple businesses is needed. It is in view of this objective that 5G NORMA proposes a multi-service and multi-tenant capable system architecture based on network slicing.

### 2.1.2 Network slicing

The idea behind network slicing is to substitute a single physical network by multiple logical networks running on a shared infrastructure. These virtual networks are called network slices in the following. Then each service or tenant can be provided its own network slice, i.e., its own dedicated logical network instance, to meet the specific requirements of the tenant's service.

Concepts to carry multiple traffic flows with heterogeneous quality requirements on a common network infrastructure have already been proposed in the past; e.g. in LTE, traffic flows / EPS Bearers can be assigned to QoS classes with specific QoS attributes [23.401]. Also for network sharing, several methods have been standardized, e.g., Multi-Operator Core Network (MOCN), orGateway Core Network (GWCN) [23.251].

5G NORMA goes beyond a pure QoS differentiation between traffic flows, as network slices can differ in (at least) three aspects, cf. Figure 2-1:

- ***Network slice functionality:*** the network functions included in a specific network slice instance can vary from slice to slice. In fact, each slice instance can have a dedicated functional architecture. E.g. slices could apply different service-specific mobility management schemes while running on the same infrastructure.
- ***Network slice topology:*** Topology refers to the set of infrastructure nodes (base stations / radio access points, compute nodes, edge and central data centres) utilised by an individual slice instance. This includes the location where Virtualised Network Functions (VNFs) are instantiated. Depending on service demands like latency or regional network coverage, slices may need to apply different topologies. E.g. a slice for automotive services will primarily cover roads, whereas a slice for factory services covers manufacturing plants. Nevertheless, both slices could wholly or partly share the same infrastructure.
- ***Network slice performance and QoS:*** Similar to EPS Bearers in LTE, slices may differ with respect to the QoS they provide. Resources have to be allocated accordingly, e.g. a slice for MBB services will be equipped with a greater amount of radio resources than a slice for Smart Metering services.

### 2.1.3 From monolithic network elements to network slices: the role of 5G NORMA's innovative enablers

Legacy mobile networks are characterized by monolithic network elements that have tightly coupled hardware, software, and functionality. In contrast, 5G NORMA decouples software-based network functions (Core Network Functions / CNF; RAN Network Functions / RNF) from the underlying infrastructure, cf. Figure 2-2. With monolithic network elements, a logical/functional view and a physical view are sufficient to describe a legacy system. A 5G NORMA system is much more flexible in the mapping of network functions to infrastructure resources in various locations. Therefore, 5G NORMA has defined one functional and three non-functional views on a 5G NORMA system. These views are explained in detail in Section 2.2.

The transition from a legacy system architecture to the 5G NORMA system architecture builds on two enablers, namely

(1) Adaptive Decomposition and Allocation of Network Functions and
(2) Network Programmability.

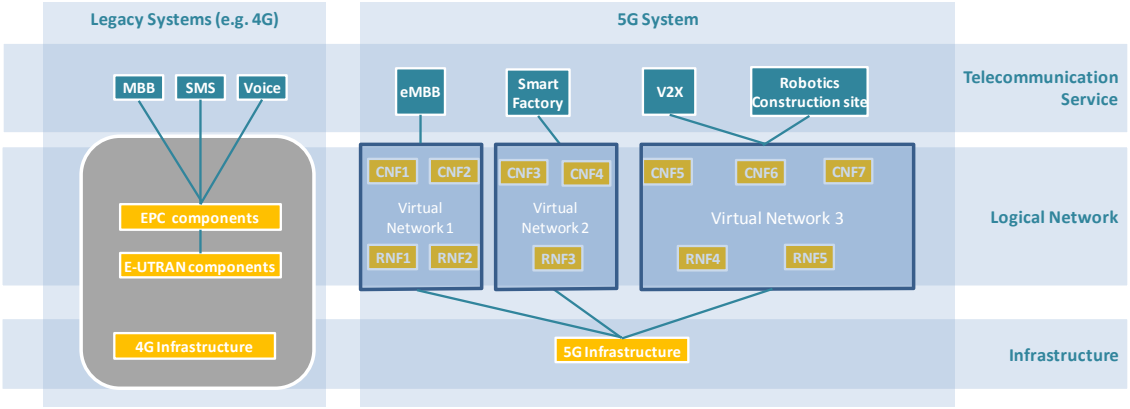



**Figure 2-2: 5G NORMA compared to legacy systems**

To enable the *Adaptive Decomposition and Allocation of Network Functions*, Network Function Virtualisation (NFV) methods for sharing compute and storage HW are advantageously complemented with well-known resource-sharing techniques like multiplexing and multitasking, e.g. WDM or radio scheduling, for physical network functions that cannot be virtualized. Benefits are threefold:

- All kinds of infrastructure resources can be shared between multiple users – obviously this is a key prerequisite for the envisaged end-to-end (E2E) multi-service / multi-tenancy network architecture.
- The way that E2E functionality of a network service is split into multiple modules resp. Network Functions can be oriented primarily on the demands of the network service, instead of the capabilities of the underlying HW. This yields functional building blocks of E2E network function chains that can be shaped individually per slice to provide a service-specific network functionality. E.g., multiple slices running simultaneously on the same HW component could apply different methods for the mobility management network function, like a 3GPP-compliant mobility management vs. SDN-based mobility management.
- Network Functions can be allocated and instantiated in different locations (5G NORMA refers to this as Software-Defined Mobile network Orchestration, SDMO), as they are no longer bound to a specific execution HW. (Note that the implementation of a NF may still be HW-specific, whereas the functionality as such is HW-independent.) This enables the realisation of service-specific network topologies.

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

The former two categories are a rather natural extension of the application of SDN principles, while the latter captures the key aspect of 5G NORMA's SDMC concept: the implementation of selected wireless control functions will no longer be bound to specialised hardware (e. g., LTE eNB), but rather become independent software entities that can be managed using a software-defined approach. These functions are executed and performed by a programmable and logically centralised controller that abstracts and thus homogenises different network technologies and implementations. Such a controller will make network slices programmable by controlling the topology and functionality of the service chains as well as the resources inside the network slices.

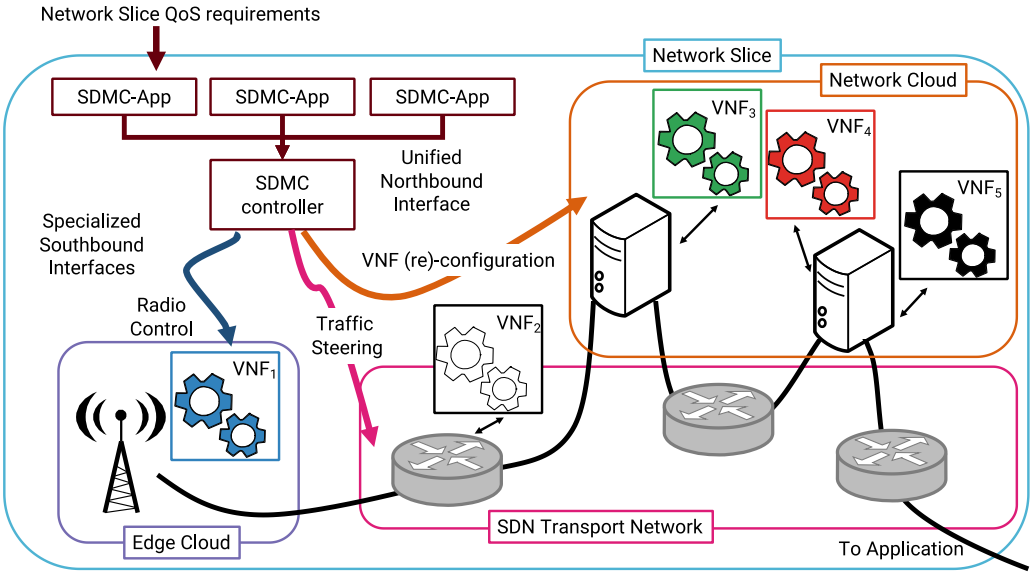



**Figure 2-3: Example of the network programmability concept**

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

### 2.1.4 Design principles of the 5G NORMA architecture

The considerations elaborated above are reflected in the design principles for the 5G NORMA architecture embodying the following capabilities and characteristics:

- Decoupling of stakeholders/players:
  The NGMN 5G White Paper [NGMN] distinguishes Infrastructure resources layer, Business enablement layer and Business application layer. Correspondingly, there are different stakeholder roles in 5G NORMA, namely Infrastructure Provider (InP), Mobile Service Provider (MSP) and Tenant. 5G NORMA allows many-to-many relationships between these stakeholders: a MSP will typically serve multiple tenants, and of course a tenant is free to use services from multiple MSPs. Furthermore, an MSP can utilize infrastructures from multiple (e.g. locally operating) InPs. In turn, an InP can offer its infrastructure to multiple MSPs. Stakeholder roles and different types of services offered to the tenants are discussed in detail in Section 4.1.
- Multi-service network management by separate management and orchestration (MANO) entities for MSPs and Tenants:
  For some business models, it may be desirable that a tenant can manage and orchestrate parts or even all of his network slice on his own. In that case the MSP's MANO stack is responsible for inter-slice resource management and allocates resources to a tenant. Then, the resources allocated to the tenant can be managed by the tenant himself using his own MANO stack in parallel to that of the MSP. Details on this concept can be found in Section 2.4.2.
- Separation between control layer, data layer, management & orchestration layer:
  The separation of network functions and the Operation Support System (OSS) has been state of the art for a long time. Furthermore, the separation of data forwarding and processing functions from functions that are controlling them follows from the SDMC concept, splitting between logic and agent for any network function in the network.
- Flexible allocation of modularised NFs in both spatial and temporal dimension:
  Network Functions can be allocated to the central cloud or to the Edge cloud, wherever it is best suited to the demands of the requested service and the available physical network infrastructure. Network functions may even be re-located dynamically, e.g. to follow a moving UE, adapt to fluctuations of traffic load in the network or minimize the energy consumption of the data centres.
- Network Programmability:
  Service function chains, including their topology (i.e. forwarding graphs) and their NF components, can be managed using SDN concepts. For further details, cf. Section 2.2.
- Efficient resource sharing between slices:
  In 5G NORMA, Network Functions can be either dedicated to a particular slice or common to multiple slices. Inter-slice coordination mechanisms on both control layer (Software-Defined Mobile network Coordinator, SDM-X) and on the MANO layer (Inter-slice Resource Broker, ISRB) ensure that in both cases resources are utilized most efficiently: For dedicated NFs, they adjust dynamically the allocation of resources to slices, while for common NFs, SDM-X controls the access and utilisation of the common NFs by the slices and ISRB manages the resource allocation to running slices.
- Integration of VNFs and PNFs in a common architecture framework:
  VNFs achieve high flexibility and scalability, while PNFs are advantageous in terms of performance and energy efficiency. Accelerator techniques are trying to combine the flexibility of SW and the performance of HW implementations, thus filling the gap.
  For efficient slicing and lifecycle management of both the RAN and core network, 5G NORMA integrates PNFs and VNFs in a common architecture framework, thereby extending the ETSI NFV architecture. This impacts not only the data layer, where the PNFs and VNFs are located, but also their management and orchestration in the MANO layer. This topic is addressed in more detail in Section 2.4.1.

## 2.2 High-level architecture

The system architecture for 5G networks shall incorporate the performance and flexibility to support multiple telecommunications services, with heterogeneous KPIs and sharing the same infrastructure. 5G NORMA will give vendors, mobile service providers, and infrastructure providers the opportunities for fine-grained offerings in order to target new customers on the B2B level, such as, private companies, non-profit organisations, or public authorities. To support their requirements, the 5G NORMA functional architecture is designed in a modular manner and incorporates four layers. For each of these layers, it defines the architectural elements that deliver the system's functionality. It includes the key functional elements, their responsibilities, the interfaces exposed, and the interactions between them.

### 2.2.1 Functional perspective

The high-level functional perspective of the 5G NORMA system architecture is depicted in Figure 2-4. It shows the separation into four layers as well as the differentiation into intra-slice and inter-slice functions.

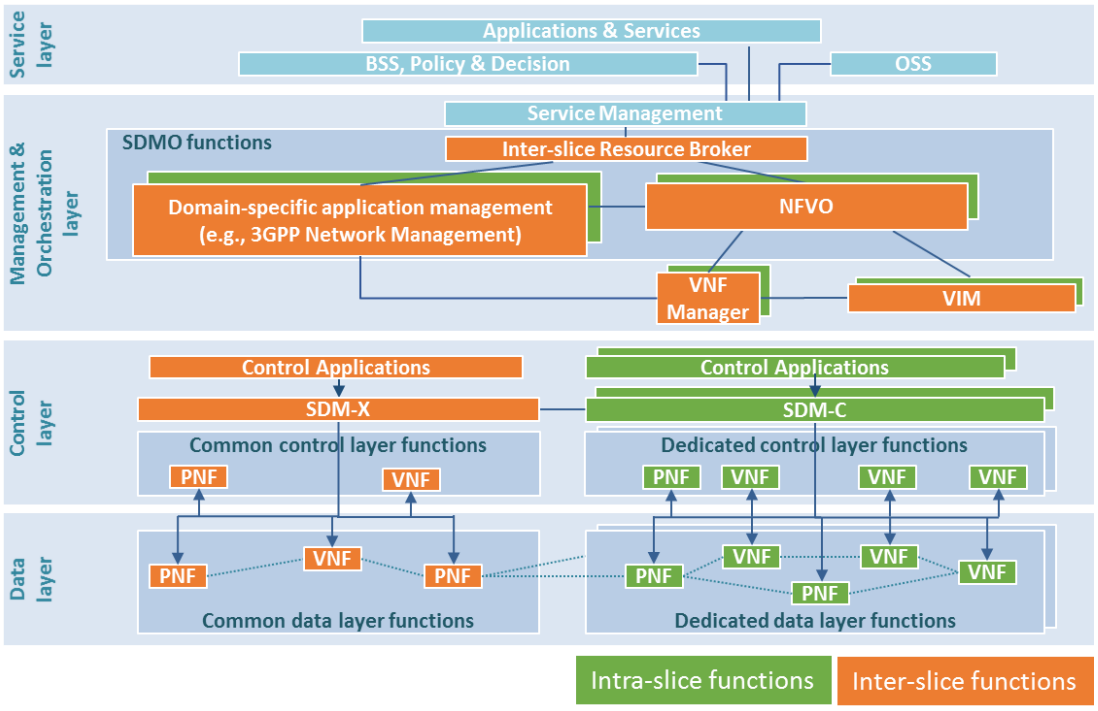



**Figure 2-4: Functional perspective of the overall 5G NORMA architecture**

The **Service Layer** comprises Business Support Systems and business-level Policy and Decision functions as well as applications and services operated by the tenant. Except for the reference point between Service Management function and functions in the Service Layer, this layer is not in the scope of the project. Rather, reader is referred to, e.g., [5GEx].

The **Management & Orchestration (MANO) Layer** realizes 5G NORMA's Software-defined Mobile network Orchestration (SDMO) concept by extending the ETSI NFV management and orchestration (NFV MANO) architecture towards multi-tenant and multi-service networks. The layer therefore comprises the Virtual Infrastructure Manager (VIM), the VNF Manager (VNFM) and the NFV Orchestrator. Further, the layer accommodates application management functions from various domains, e.g., the 3GPP domain, IEEE domain, or private enterprise network domains. For example, in the case of 3GPP domain, this can comprise Element Managers (EM), Network Management (NM), and selected OSS functions. Such functions shall also implement

ETSI NFV MANO-defined reference points between VNFM and EM as well as between the NFVO and OSS/NM. The Inter-slice Resource Broker (ISRB) determines and enforces policies for cross-slice resource allocation, particularly in the case of shared network functions. Finally, the Service Management is an intermediary function between the Service Layer and the ISRB. It transforms consumer-facing service descriptions into resource-facing service descriptions and reports selected key performance indicators to the Service Plane functions. Further details of this layer and the interaction with control layer are described in Section 2.3.5.

The **Control Layer** accommodates the two main controllers: (1) the Software-Defined Mobile Network Coordinator (SDM-X) for the control of common (shared) NFs (depicted in orange) and (2) Software-Defined Mobile Network Controller (SDM-C) for dedicated NFs (depicted in green). Following the SDN principles, SDM-X and SDM-C abstract form the technological and implementation-related details of controlled network functions. They translate decisions of the control applications into commands towards Virtualized NFs (VNFs) and Physical NFs (PNFs) in both Data and Control Layer.

Finally, the **Data Layer** comprises the VNFs and PNFs needed to carry and process the user data traffic. Further details of both Control and Data Layer are described in Section 2.2.

Figure 2-4 further depicts the split into common or so-called inter-slice functions and dedicated (intra-slice) functions. This split is maintained from the MANO Layer down to the data layer, i.e., dedicated NFs are controlled (and managed, respectively) by the tenant's own instances of SDM-C and MANO Layer functions (i.e., ETSI NFV functions as well as domain-specific application management functions). Shared functions are controlled (and managed, respectively) by the SDM-X and the respective MANO Layer functions, which are usually in the domain of the Mobile Network Operator (MNO) or the Mobile Service Provider (MSP), cf. Section 4.1. The policies regarding the utilization of shared functions, particularly the resource allocation to active slices) are determined by the ISRB and communicated towards the respective control, management and orchestration functions for further enforcement. More details on multi-domain orchestration are given in Section 2.4.2.

## 2.2.2 Complementary system perspectives

5G NORMA has defined three further, non-functional perspectives of the overall system architecture, each of them depicting specific architectural particularities. The **deployment perspective** utilises four location categories, representing the level of topological aggregation towards central nodes typically to be found in network infrastructures, for mapping the architecture functions to these four location categories. The **topological perspective** depicts the geographical distribution of infrastructure resources, particularly the topology and characteristics of the nodes and of the links interconnecting them. Finally, the **resource perspective** depicts which hardware and software resources are available at each of the four location categories.

### 2.2.2.1 Deployment perspective

The deployment perspective illustrates one of 5G NORMA's key innovations: The adaptive allocation of network functions depending on the service requirements and deployment needs. It depicts the different possible locations of functional blocks and the mapping of functional blocks to the location in which a block is intended to be instantiated and to operate. This also includes the possibility that a functional block may be deployed in different locations. The four location categories comprise:

- PNFs at antenna site, representing functions that are characterized by a tight coupling of software and hardware due to performance reasons (while PNFs are most commonly located at the antenna site, they can also reside in other locations of the network infrastructure),
- Edge Cloud at the antenna site, representing general purpose compute, storage, and networking hardware (e.g., NFVI resources) at or close to the antenna site,

- Edge Cloud at an aggregation site, representing medium-size data centres at aggregation sites of the network, and
- Central Cloud, referring to large-scale data centres, e.g., at an operator's central offices.

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

Figure 2-5 shows an illustrative deployment perspective example. It features four location categories, PNFs at antenna site, general purpose resources ("edge cloud") at antenna sites, edge cloud at aggregation sites, and central cloud data centres. It furtherdepicts one MNO that utilizes its network infrastructure to host own network slices and associated NFs (depicted in orange colour) as well as network slices for two tenants. These tenants operate own MANO layer functions that manage and orchestrate dedicated NFs in the control and data layer (depicted in green colour). Components of compound functions, e.g., VIM, can be distributed across several locations (depicted with dashed outline).

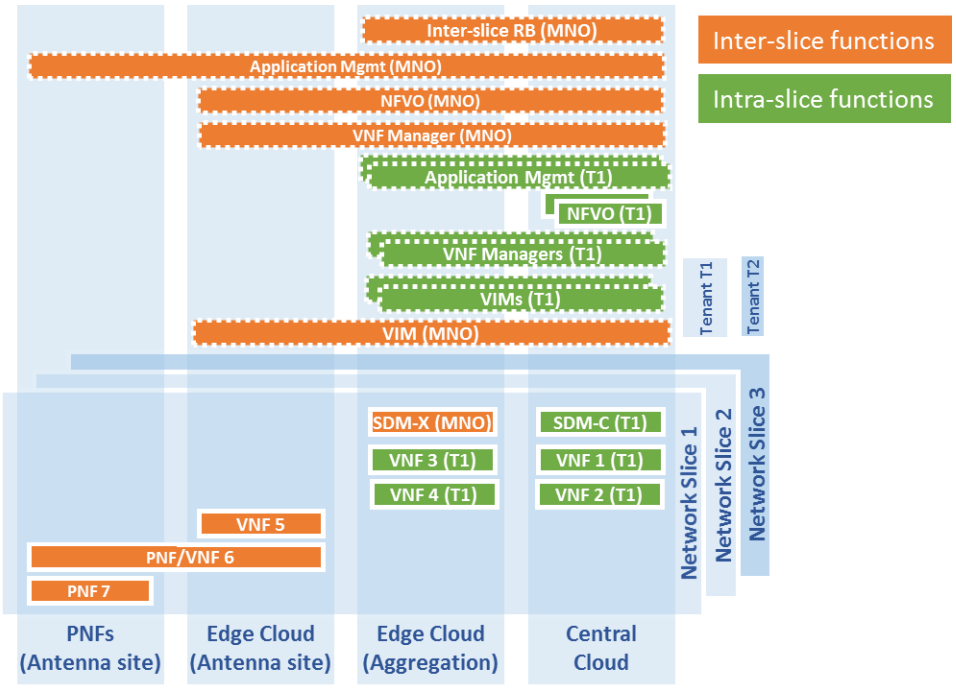



**Figure 2-5: Deployment perspective of the 5G NORMA architecture (illustrative example)**

In the example depicted in Figure 2-5, both tenants (T1 and T2) operate their own ETSI NFV MANO stack instance (NFVO, VNF Manager, VIM). The functions are distributed across edge and central cloud locations in order to manage the dedicated NFs of the respective network slice. In addition, both tenants run some dedicated Application Management functions. The MNO operates the full set of MANO Layer functions, which are distributed across all location categories. Particularly, some management functions can even be integrated with the managed PNF, e.g., with an eNB. Further, the figure shows the set of NFs comprising network slice 1 that serves T1's customers. This slice contains shared NFs (depicted in orange), which are controlled by the SDM-X which is run by the MNO, as well as dedicated VNFs running on cloud infrastructure and controlled by T1's SDM-C (depicted in green). Since a subset of functions is typically shared across slices, particularly PNFs in the RAN, it is a preferred solution to locate SDM-X closer to the antenna site than SDM-C. The interworking between dedicated functions and common functions is further elaborated in Section 2.3 as well as 5G NORMA Deliverable

[5GN-D42], including recommendations and constraints regarding the deployment location of the involved NFs.

#### 2.2.2.2 Topological perspective and resource perspective

Two further views on the architecture, the topological and the resource perspective, respectively, are depicted in Figure 2-6.

The topological perspective depicts the geographical distribution of the network infrastructure and thereby includes the notion of distance and associated latency characteristics, which in 5G NORMA determines the main difference between edge and central cloud. The central cloud typically comprises only a few data centres, which may be several hundred kilometres apart and connected through a wide area network (WAN). The WAN also connects the data centres of the central cloud to the data centres of the edge cloud. Compared to central DCs, edge DCs have less capacity but are more numerous. Furthermore, the topological perspective also depicts bandwidth and latency of transport media between distinct sets (nodes) of resources, therefore providing an additional dimension compared to the resource view.

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

Hardware and physical resources include both general purpose (“cloud”) and application-specific hardware (i.e., PNFs). It thus comprises memory, compute, storage, networking, and other fundamental capabilities as well as radio spectrum. These HW resources either can be made available for virtualised network functions (VNF) or are part of physical network function (PNF), i.e., bound to a specific function.

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

The library of network slices comprises descriptions of all executable network slice templates including the necessary metadata such as QoS parameters. A network slice template refers to the set of VNFs and PNFs that implement the telecommunication service, as well as the NF-FG (network functions forwarding graph) that specifies how these NFs must be interconnected. A network slice template contains network service, link, and connectivity descriptors.

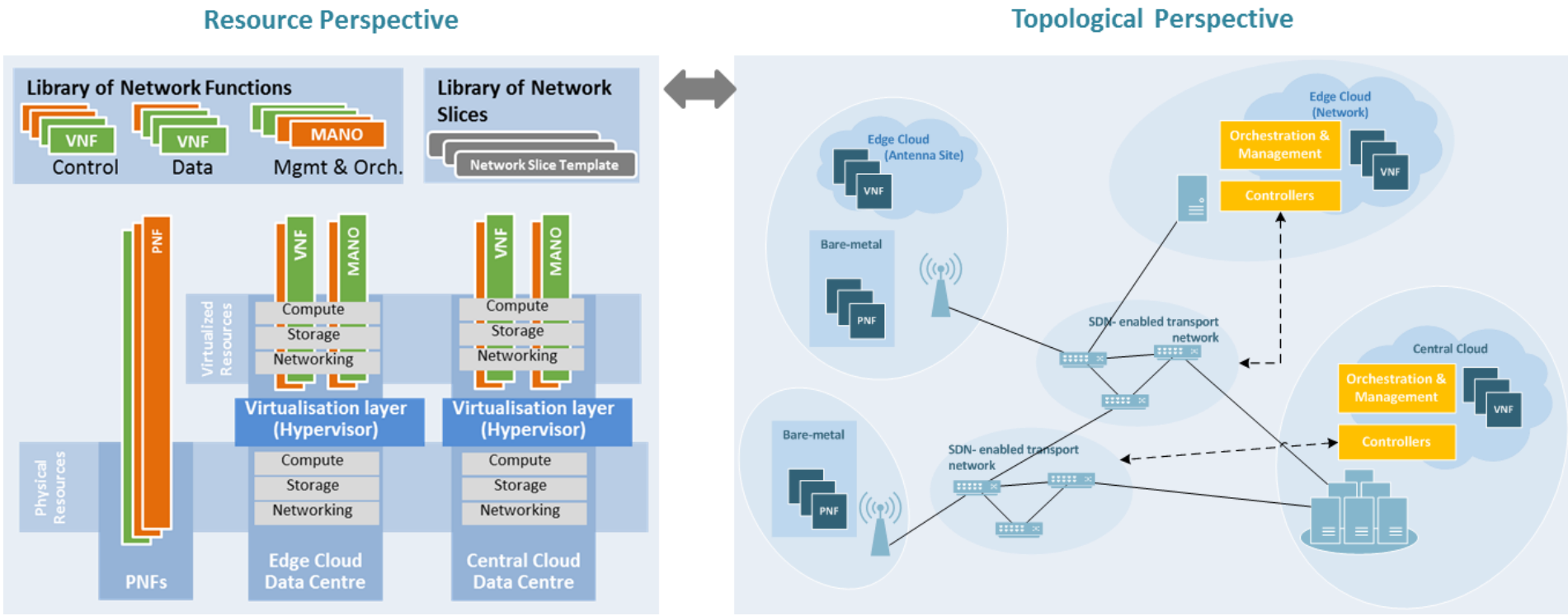



**Figure 2-6: Resource and topological perspectives of the 5G NORMA architecture**

## 2.3 Integrated control and data layer architecture

The 5G NORMA control and data layer architecture is depicted in Figure 2-7.

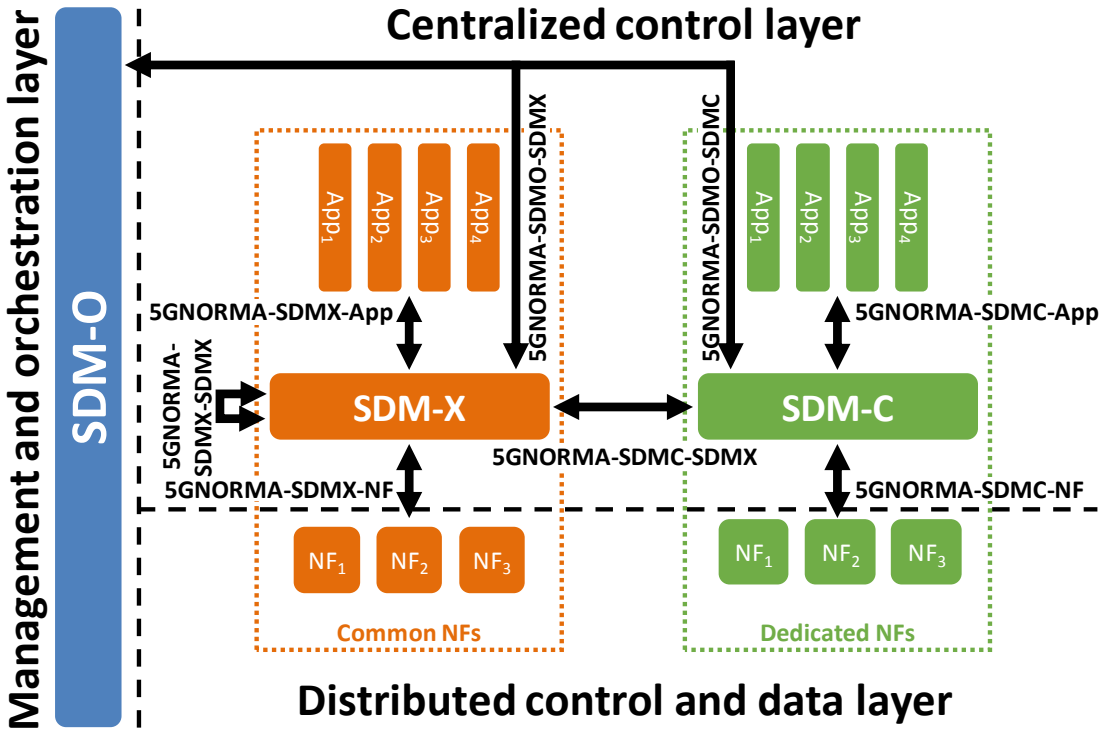



**Figure 2-7: The 5G NORMA SDMC interfaces**

The main components of the architecture are heavily linked with the network slicing concept, as all the modules are categorized according to their role with respect to a network slice: following the colour code in Figure 2-3, green elements deal with control within a network slice, orange ones take care of controlling resources (i.e., NFs) shared across slices, while the SDM-O orchestrates those resources.

### 2.3.1 Centralized control

Due to the extensive application of the Software Defined Networking concept to all the NF in a mobile network, the control and data layer will necessarily be split by the controllers. As controllers split the functionality between the application *logic* (i.e., the intelligence, that runs in the applications, cf. [5GN-D52] for a thorough discussion on this novel concept) and the *agents* running in the NFs. Therefore, the 5G NORMA control and data layer architecture pivots on the controllers.

SDM-C and SDM-X and the applications running on top of both represent the centralized 5G NORMA control layer that comprises an ecosystem of applications controlling the underlying NFs (dedicated or shared), exploiting the advantages of the SDN approach:

- Joint access and core optimization: the transition towards a VNF based architecture is blurring the current concepts of mobile access and core networks. The possibility of running VNFs almost everywhere in a cloudified network enables the possibility of jointly optimizing functions that have always been in separated boxes. This fact has important implications for the network control and orchestration. Therefore, the

5G NORMA control and data layer architecture allows for such optimized algorithm: the Video Aware Pre-Scheduler (cf. [5GN-D52]) is an example that integrates core and RAN to optimize the scheduling of video flows.

- One common interface to all network functions: having one reference point 5GNORMA-SDMX-App respectively 5GNORMA-SDMC-App on the northbound interface of SDM-C and SDM-X allows for diversity in both NF and application developers. SDM-C and SDM-X offer the necessary abstraction to provide efficient centralized control (cf. [5GN-D52] for several examples of SDM-C/X applications)
- Efficient resource sharing: The controllers offer a common view of the underlying network function through the NBI. In this way, the resources assigned to one (or a set) of NF can be efficiently controlled by a single application that enforces the resource sharing according to the quota each tenant is entitled to. This is a cleaner approach compared to different peer-to-peer interfaces across network functions.

Between controllers, 5G NORMA foresees the 5GNORMA-SDMC-SDMX and the 5GNORMA-SDMX-SDMX interface. Through the 5GNORMA-SDMC-SDMX interface, each SDM-C is able to control, up to the extent exposed by the SDM-X, those common NFs that are part of its slice but that are shared with other slices and which are therefore controlled by the SDM-X. The 5GNORMA-SDMX-SDMX interfaces enables the SDM-X of different stakeholders to coordinate each other, e.g., for seamless mobility and improved RRM, as exemplified later in Section 4.3.

### 2.3.2 Distributed control

The centralized control approach provided by the SDM-C and SDM-X applications is however, just one of the layers of the 5G NORMA control and data layer architecture. Not all the functionalities can be split into a logic running on top of an SDM-C (or -X) and the agents running in the underlying NFs. Hence, for different reasons, some of the control functionality should be managed in a legacy distributed way:

- Legacy PNFs: legacy PNFs that shall be integrated into a network slice may not support the interworking with SDM-C or SDM-X applications. Therefore, their behaviour should be integrated through the suitable 5GNORMA-SDMX-NF respectively 5GNORMA-SDMC-NF protocol plugins on the southbound interface of the controllers. For example, a legacy eNB can be integrated into the 5G NORMA controller-based architecture by managing the former control plane functions (e.g., the S1-AP) through centralized applications, implementing a mild network slicing architecture such as e-DECOR.
- Data Locality: some control NFs build on information available locally that should be processed with very low timing constraints. In this case, the performance gains obtained by a centralized approach may not be enough for certain NFs. Examples therefore are link adaptation or HARQ (Hybrid Automatic Repeat Request).
- Scalability: for some NFs, the overhead introduced by a fully softwarised and centralized control may be too much, especially when configured for extreme situations. As an example, a centralized MAC scheduler that controls several base stations may hardly be implementable as an SDM-C (or -X) application (not precluding hybrid approaches that combine distributed MAC schedulers with a centralized coordinator application). Therefore, some of the functionality is necessarily offloaded to distributed schedulers that can operate at wire speed. An example of this approach is provided by the Demo 1 (cf. [5GN-D61] for more details).

Accordingly, 5G NORMA introduced three distributed control functions: *MAC Scheduling*, *RRC User* and *RRC Cell*. These control functions do not run as SDM-C/X applications but as (legacy) distributed NFs. For RAN slicing Option 2 and Option 3 (cf. Section 2.3.5), these distributed control NFs communicate with the SDM-X via the 5GNORMA-SDMX-NF interface (for RAN slicing Option 1, the distributed control NFs are tenant-specific and instead interface with the SDM-C via 5GNORMA-SDMC-NF). This enables the MSP operating all common NFs to

customize the inter-tenant resource sharing through suitable SDM-X applications. Furthermore, via the 5GNORMA-SDMC-SDMX interface, the MSP offers tenants means for influencing how common NFs should treat their own traffic.

### 2.3.3 Data layer

Data layer NFs are inherently distributed due to their purpose of providing data forwarding end-to-end. There are two options to control data layer NFs: first, they may be controlled directly by (distributed) control layer NFs as it is done in current mobile networks; second, they may be controlled by SDM-C or SDM-X, which allows more degrees of freedom through programmability instead of today's mere parametrization of data layer NFs through OSS. Shared common NFs are controlled by the SDM-X via 5GNORMA-SDMX-NF while dedicated (per tenant customized) NFs are controlled by the SDM-C via 5GNORMA-SDMC-NF.

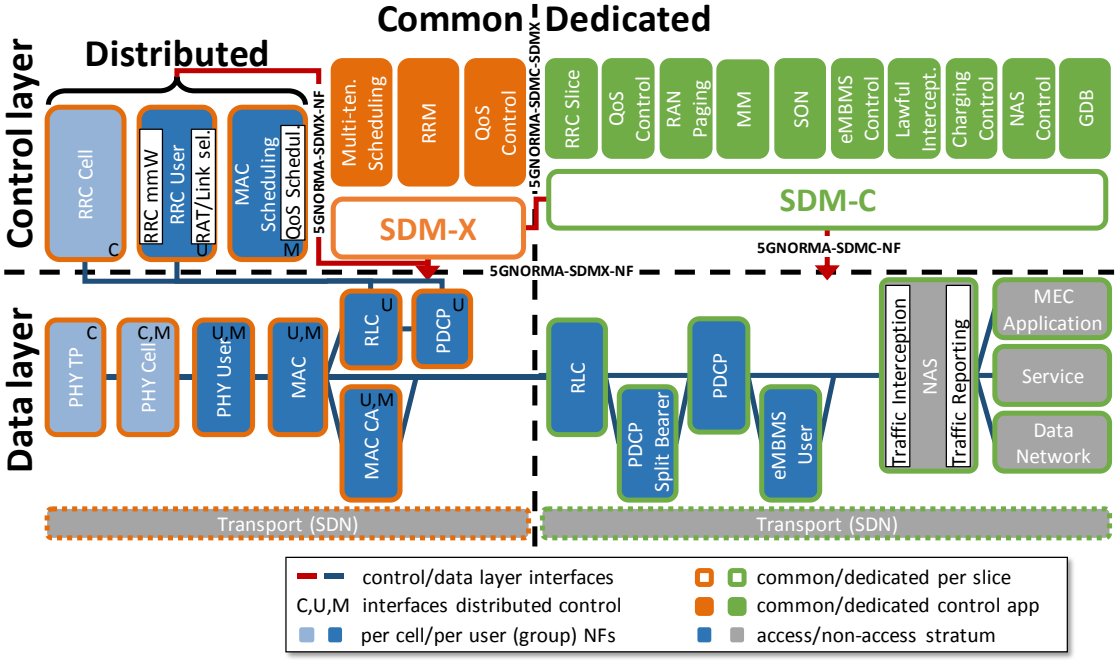



**Figure 2-8: Control and data layer functional architecture of a 3GPP LTE/NR telecommunication service (RAN slicing Option 2)**

5G NORMA defined the following (distributed) data layer functions and categorization (cf. Chapter 4 in [5GN-D42]):

- Access stratum physical layer (typically PNFs for energy/cost efficiency reasons)
  - *PHY TP* (all functions that inherently cannot be virtualized)
  - *PHY Cell* (functions mapping one-to-one to a single radio carrier)
  - *PHY User* (functions mapping one-to-one to a specific (group of) users)
- Access stratum link layer (typically VNFs)
  - *MAC*, *RLC*, *PDCP* (mandatory, 3GPP LTE/NR)
  - *MAC CA*, *PDCP Split Bearer* and *eMBMS* (optional, present if specific feature is used, namely carrier aggregation, multi-connectivity incl. multi-RAT or multimedia broadcast and multicast service)
- Non-access stratum (typically VNFs)
  - *NAS* (subsumes various VNFs like mobility anchor, policy enforcement, firewall, deep packet inspection, gateway). More details about mobility management SDM-C applications are provided Section 2.3.4.1 below.

- *MEC Application*, (end user) *Service* and (Packet) *Data Network* (hosting end user services within the mobile network resp. providing access to them, e.g. to the Internet)

Putting all together, Figure 2-8 shows the control and data layer functional architecture of a network slice running a 3GPP LTE or NR (broadband) telecommunication service, here for RAN slicing Option 2. The optional data layer VNFs MAC CA, PDCP Split Bearer and eMBMS User lie on an alternative branch of the data layer network function chain. These NFs can be left out in case their respective functionality is not employed by the slice. The shown set of SDM-X and SDM-C applications relates to specific WP4 and WP5 innovations, e.g. Multi-tenancy Scheduling and RAN Paging [5GN-D42], complemented by general "catch-all" applications like RRM and NAS Control that subsume mandatory control functionality. Like in the data layer, not all applications must be present, e.g. eMBMS Control is only orchestrated if eMBMS User has been orchestrated, too.

## 2.3.4 SDM-C and SDM-X applications

As introduced above, the control layer in 5G NORMA follows the *Software Defined Networking* concept: SDM-C and SDM-X applications control the underlying distributed control and data layer network functions. Deliverables [5GN-D41] and [5GN-D51] provide details of how this concept is applied. The next subsections describe two examples, namely, software-defined mobility management and radio resource control.

### 2.3.4.1 Mobility management application

Within a software-defined mobile network the capability of flexibly supporting mobility of users and their terminals, as well as sessions and flows and even (virtual and physical) network entities is predominantly seen as a challenge in terms of changing performance (e.g. throughput) and user perceived QoS/QoE. On the other hand, the continuing need to adapt the network performance to the service-specific demands can also be a chance to flexibly assign scarce resources to a connection in terms of shifting them within in-homogeneously loaded pool of cells, data centres, transport network entities, etc. Problems and solutions of this approach are introduced next. The reader can read further details in [YGF+17] and [RAH+17].

*Binding Mobility Management to a Network Slice*: In order to support a service tailored Mobility Management (MM), a network slice is designed which includes specific network functions enabling a specific MM scheme. A way to realize this is to maintain specific, mobility related flavours of network functions and/or specific configurations of network functions and instantiate them according to the mobility related context of the network slice. The selection of appropriate mobility management scheme needs to be provided through a dedicated functionality, i.e. *binding functionality*. The binding functionality provides the mapping between the mobility related context of the slice, i.e. MM requirements and the MM scheme that supports the mobility requirements in the most suitable way. Furthermore, it translates this mapping into a concrete configuration of the network slice. The binding functionality takes into account not only the network slice context but also the predetermined policies. MM schemes can differ in many ways, e.g., requiring special handover policies and settings in the RAN, flexible mobility anchoring, adaptive gateway relocation rules, or customized network elements (e.g., local gateways or gateways with specific mobility support).

*Flexible SDM-C steering of flows*: In the current mobile network architecture, the Mobility Management (MM) functionality is a process that involves different physical entities in the network (i.e., eNB, MME, gateways, etc.). The current *softwarisation* trend will trigger the transition from a *network of entities* to *a network of functions*. To that end, an essential functionality as MM must also be transformed and made aware of the novel QoS/QoE, Orchestration and control mechanisms. Following the design and architectural principles defined in 5G NORMA, the mobility management as a whole can be managed as a SDM-C application. Taking advantage of a unified QoS/QoE framework [5GN-D5.2], the management of user

mobility is a thorough process that involves network function control and orchestration to achieve an optimized functionality on a per slice basis. By exploiting these characteristics, the network flexibility is increased: the adaptation of the network slice capacity according to the instantaneous traffic demands and required KPIs entails the re-configuration and re-orchestration of the network at many levels. Therefore, besides the selection of the most appropriate MM algorithm or the parameters that may influence the MM algorithm behaviour, the MM shall be able to control different network configurations seamlessly. One of the key technologies for the enhancement of the flexibility is RAN as a Service (RANaaS) [iJOIN-D53]. This capability, envisioned as one of the future pillars of 5G networks, allows to split the current monolithic RAN protocol stack into atomic functions that may be orchestrated in different ways, exploiting either the multiplexing gain of baseband processing centralization or the flexible resource utilization of decentralization of edge computing. In this very heterogeneous context, an enhanced MM shall i) jointly optimize RAN and Core network functions by leveraging on the centralized network control capabilities of SDM-C and, ii) steer user flows across different network functions according to the RANaaS functional split implemented in the network. The former functionality is implemented within an SDM-C application, while the latter is provided by a set of *plugins* installed on the southbound interface of the controller. The overall ideas of a software defined MM algorithm that can cope with the changing environment of a RANaaS-enabled network are sketched next.

Inter-slice mobility: Design considerations on the amount and type of parameters to be configured within a slice-specific MM application (MM-App) include the issue whether the MM is invoked on a dedicated per-slice function or across multiple slices: E.g. a simple only on-demand MM would be associated only to a low/no mobility slice (e.g., for IoT, fixed/home network slice etc.) whereas an MM-App supporting a range of terminal speeds and seamless session continuity might apply for multiple sessions/UEs within same regional context (within train/bus, on highway, etc.) but belonging to different slices (e.g. verticals' automotive slices and an enhanced mobile broadband slice). Such an MM-App would then be reused across different slices. Further design criterion could be the availability and usage of layer-sensitive information across layers (e.g. to proactively invoke handover on MAC or IP layer based on signal strength at the physical layer). Such a feature as well as the capability to adjust to variable service demands (e.g. in terms of QoS/QoE) allowing for higher granularity in terms of mobility support of course would depend on the specifics of the involved access technologies. Another parameter guiding the MM-App design is the potential differentiation for a hierarchical mobility treatment (e.g. local and global mobility). Finally, the degree of flexibility which a chosen MM approach supports (e.g. whether a feature may be changed on the move according to changing environments) is restricted to certain variables (e.g. no multi-link or access heterogeneity is possible in case of single-interface devices). Based on such parameters considered in a mobility support design, the correspondingly required effort in terms of process complexity and amount as well as frequency of necessary signalling messages can be estimated and used to decide on the most appropriate MM module to be applied to a specific network slice. Also, the number of network entities included in the required MM processes and the related messages to be exchanged may determine the effort spent by a specific MM application. Entities involved cover UE and RAP nodes and beside MM-App also core network entities as a GW-App (i.e. an SDM-C application for configuring the gateway between 5G NORMA architecture and the outside world) and customer/subscription data base (CDB/HSS) for storage of subscription policy information.

#### 2.3.4.2 Radio resource management applications

As described above, the SDM-X (SDM-C in case of RAN slicing Option 1, cf. Section 2.3.5) is in charge of controlling shared radio resources. To meet slice-specific service requirements in cases when BSs need to serve multiple slices with more than one service data flow each, it is crucial to react dynamically on critical interference situations in the network. The SDM-X needs the opportunity to optimize temporary appearing critical interference constellations in local parts of the mobile network. In case of RAN slicing Option 2 and Option 3, it should influence synchronous inter-cell interference coordination (ICIC) schemes by dynamically adapting BS

clusters (e.g. for Joint Transmission (JT), Coordinated Multipoint (CoMP) or coordinated beamforming) or it can adapt and de-/active asynchronous ICIC schemes (e.g. frequency reuse schemes, Carrier Aggregation (CA) based ICIC, enhanced (e)ICIC). It might be even possible to de-/activate the Medium Access Control (MAC) scheduler, if alternative schemes were orchestrated by the SDM-O.

Figure 2-9 shows the principle idea. A database with RRM applications, such as ICIC and scheduling schemes is defined and some of them will be placed to the physical nodes of the mobile network during the orchestration process and executed as NBI application through SDM-X. The SDM-X takes care of the control of the RRM schemes during life cycle management.

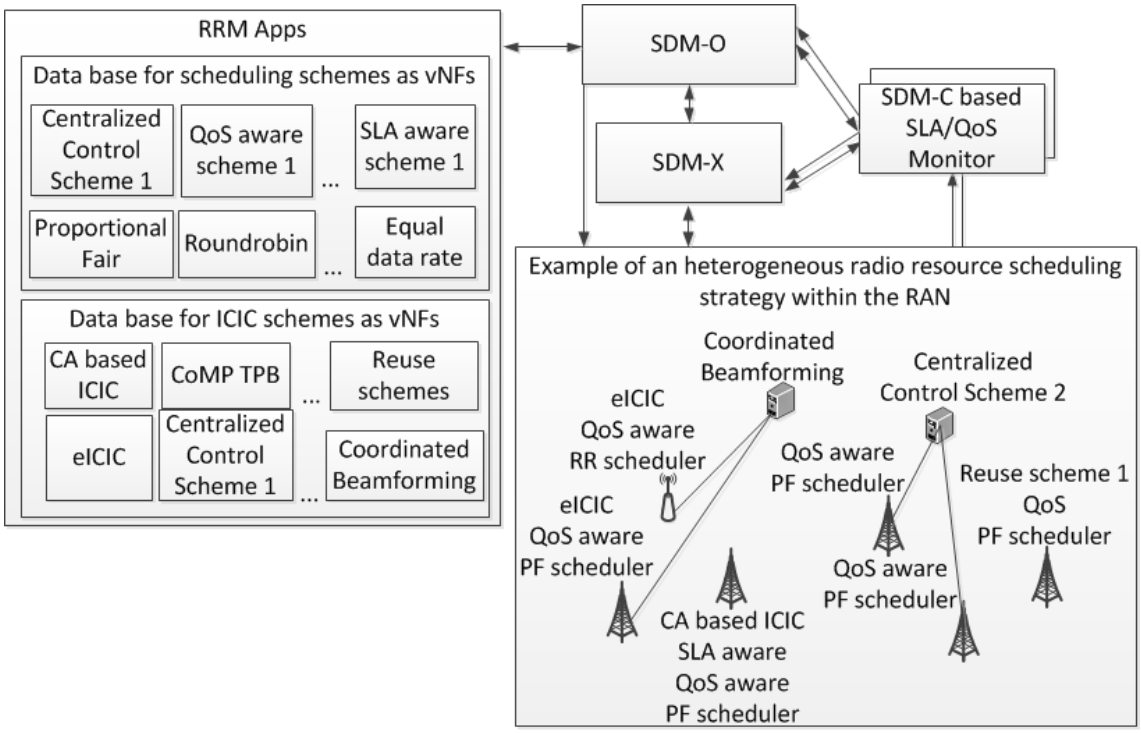



**Figure 2-9: SDM-X control of ICIC and scheduling schemes**

The controller architecture that builds the control layer of 5G NORMA fully supports the stakeholder models detailed in Section 4.1. For example, the Radio Resource Management application will usually work on shared radio resources, through the SDM-X. In case of a single MSP scenario as described in Section 4.3.1.1, shared network function will be under control of the same SDM-X that manages network slices belonging to the same tenant. In contrast, in scenarios with multiple MSP domains (e.g., scenario described in Section 4.3.2.1), the best possible operation of shared resources is guaranteed by SDM-X to SDM-X interfaces belonging to different MSPs that will provide non-mandatory synchronisation parameters such as the ones defined by 3GPP.

## 2.3.5 RAN slicing

5G NORMA introduces three RAN slicing options, shown in Figure 2-10 [5GN-D42]:

- Option 1: Slice-specific RAN,
- Option 2: Slice-specific radio bearer,
- Option 3: Slice-aware shared RAN.

In Option 1, slices just share antenna sites and some basic physical (PHY) layer processing (analogue and mixed signal processing), but are otherwise fully independent. This includes dedicated spectrum per slice, which is assumed to be assigned semi-statically. Although dynamic

spectrum reassignments as of as every tens of milliseconds are in principal not precluded, the following two slicing options are more suitable for dynamic radio resource sharing. The complete RAN is under control of SDM-C, except the spectrum assignments among multiple tenants, which is controlled by SDM-X.

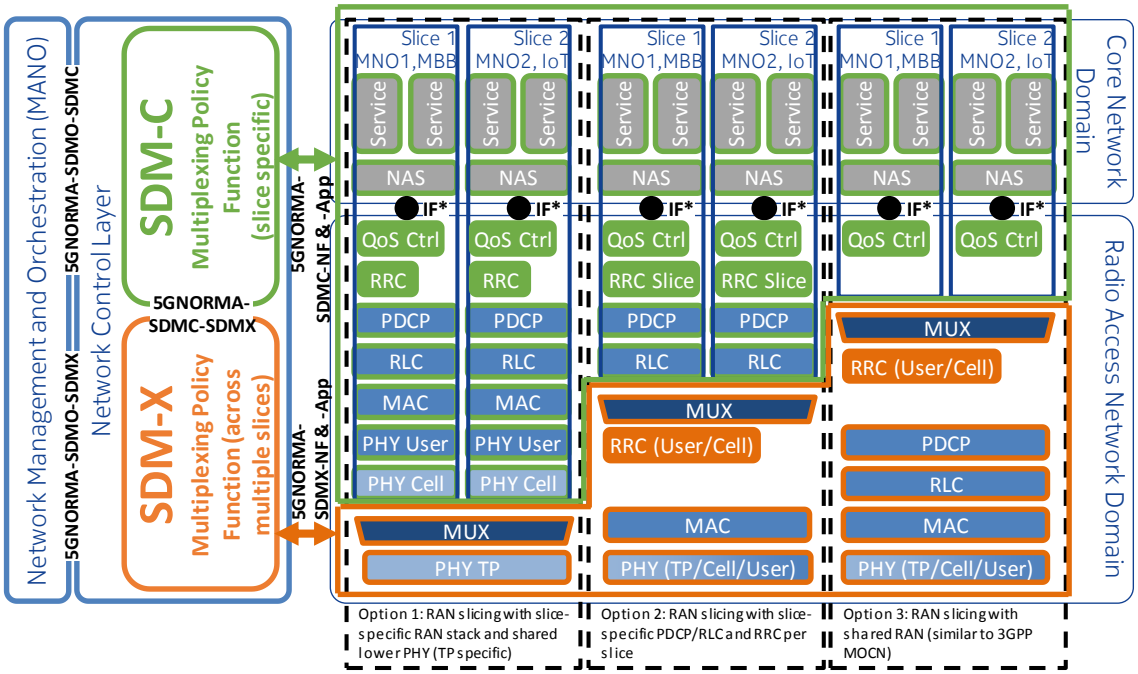



**Figure 2-10: Three RAN slicing options of 5G NORMA**

In Option 2, physical (PHY) and medium access control (MAC) layer in both data and control layer, i.e., including the MAC scheduler (RRM) and access stratum configuration (RRC), are operated by the MSP. This enables a fully flexible dynamic sharing of spectrum/radio resources among tenants, controlled by SDM-X. Multiplexing of slices' traffic takes place on a per radio bearer basis. Tenants may therefore fully customize through dedicated VNFs the protocol layers RLC, PDCP and the whole non-access stratum (NAS). Otherwise, the tenant's SDM-C application *RRC Slice* has only indirect control over common NFs in PHY and MAC up to the extent exposed by the MSP's SDM-X at the 5GNORMA-SDMC-SDMX interface.

In Option 3, the MSP operates the complete RAN or access stratum. Only (core) network slicing is used, while the RAN is not sliced, i.e. the same NF instances handle all traffic of all slices that map to the same (RAN) service. In other words, transport over the air interface is abstracted into QoS classes with latency and reliability targets, of which a certain quota of resources (data bandwidth or radio resources) is assigned to each slice. Within its own quota, a tenant may control the individual user flows via the SDM-C application *QoS Control* (in the figure abbreviated as QoS Ctrl). QoS Control influences how the MAC Scheduler respectively its *QoS Scheduling* sub-block (cf. Figure 2-8) should treat the tenant's flows. Like in Option 2, full radio resource sharing among slices allows to maximise the utilisation of radio resources. On top of that, the MSP has further potential to optimize performance through its full control of mobility and (RAN-based) multi-connectivity including multi-RAT support, i.e. in Option 3 the SDM-X respectively its SDM-X applications become most powerful, while differentiation between tenants primarily takes place in the non-access stratum.

# 2.4 MANO Layer Architecture

## 2.4.1 Integration of PNFs into 5G NORMA Lifecycle Management framework

5G NORMA pursues the objective to develop a network architecture that makes mobile radio networks multi-service and multi-tenant capable from end to end, i.e. to enable E2E network slicing. Such an E2E network architecture is only meaningful if PNFs are fully integrated in a way that is comparable to VNFs. In the foreseeable future, PNFs as well as NFs using special accelerator technologies will be indispensable, whenever performance and low energy consumption are more important than superior flexibility. Furthermore, PNFs, e.g. radio base stations, are deployed widely and in large quantities in today's networks, and they will stay in operation during the transition phase to a future 5G NORMA-based network. Hence the integration of PNFs into 5G NORMA's lifecycle management framework is a must.

However, PNFs differ from VNFs in some important respects:

- Typically, PNFs cannot scale as flexible as VNFs. Often, PNFs can only be switched on or off, i.e. they scale in steps of 0% or 100% of total capacity. Scaling of LTE base stations is a bit more flexible: Their transmission capacity can be changed by variation of their radio bandwidth in steps of 1.4, 3, 5, 10, 20 MHz.
- Consequently, from the limited scaling possibilities, PNF resources must be assigned based on assumed peak capacity demands. Accordingly, they are assigned over longer periods of time and used in a rather static manner either as a dedicated NF exclusively by a single slice or as a common NF shared between multiple slices. A highly dynamic reassignment of PNFs across slices currently seems inappropriate.
- PNFs can, in contrast to VNFs, not be migrated dynamically from one data centre to another, and hence are not portable.
- As long as PNFs comprise hardware, firmware, and software components, these components are fully integrated and cannot be separated during normal operation. FW or SW updates are usually not done during regular operation, but only as part of maintenance activities.

For these reasons, the subsection elaborates on some typical characteristics of PNFs, using the example of accelerator technologies, and consequently depicts how PNFs can be integrated into the LCM framework of 5G NORMA.

### 2.4.1.1 Accelerator technologies in 5G NORMA

PNFs and accelerators, both hardware and software, are used to offload and speed up the cloud computing applications as already defined in [5GN-D32] and they differ by the following aspects:

- flexibility,
- computational cost, already analysed in [5GN-D32],
- latency overhead,
- power consumption.

The accelerators are introduced in ETSI architecture [NFV_IFA004]. Since they affect the performance, but also the migration and the placement of the NFs in the clouds, it is necessary to consider them also in the 5G NORMA architecture.

Accelerators could be both dedicated hardware and specific software running on generic programmable units. The life cycle management of the accelerators is already introduced in ETSI, [NFV_IVA001], and it can be easily integrated in the 5G NORMA architecture. Accelerators are part of the NFVI and are controlled by the VIM, which shall be able to establish connectivity in the NFVI. The communication with the accelerators is done through a network controller, and the same considerations made by ETSI in [NFV_IFA002] can be extended to 5G NORMA. The

NFVO and the VNF Manager shall be able to request to the VIM the allocation and release of necessary acceleration resources to support the acceleration capabilities requirements of the VNF, and in addition to obtain acceleration capabilities requirements of the VNF from the deployment template of the VNF.

In [5GN-D32] the analysed accelerators were the GPUs, which were compared to the pure PNFs. The accelerators could be also other dedicated hardware components, such as FPGAs and ASICs. The silicon industry produced, in fact, a variety of chips which can be generally categorized, as reported in Fig. X, in programmable processors (CPUs, GPUs, DSPs), reconfigurable architectures (FPGAs) and Application Specific Integrated /circuits (ASICs).

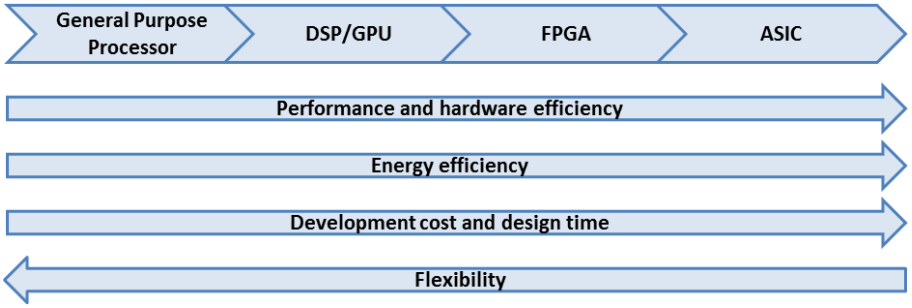



**Figure 2-11: Design features for different hardware platform (reworked from [Prab17])**

The use of application-specific accelerator can greatly increase the performance of data centres, given a fixed power budget as reported in [KS16], where different frameworks of hardware acceleration in data centre have been analysed for the most widely used cloud computing applications. According to [KS16] the use of a mix of general purpose processors and hardware accelerators (e.g. DSPs, ASICs, FPGAs) comes at the cost of an increased programming complexity, which can be tackled leveraging on programming languages such us OpenCL [KGI] or C. Moreover, other issues regard the acceleration resource abstraction and its sharing, orchestration and management.

The latency overhead is also a key factor in accelerators' virtualization, among others, for the mobile network architecture. Part of the processing is offloaded on accelerators at the cost of an increased signalling overhead due to the virtualization and queue management. In fact, the accelerators could be shared between different NFs and every offloaded packet has to wait the previous one has been processed. Moreover, PNF are optimized since all hardware components are on the same silicon chip, virtualized accelerators could be not located in the same silicon chip.

In the scenario of LTE eNB virtualization, some layers of the protocol stack have stringent timing synchronization requirements. For example, considering the LTE physical and the MAC layers, FDD LTE HARQ needs a round trip time of 8 milliseconds that imposes an upper-bound of less than 3 milliseconds for the entire eNB processing, as reported in [NKG]. This runtime complexity can limit the performance of a virtual eNB implemented on VMs leveraging only on general purpose processors. The usage of virtualized accelerators is attractive in this scenario since they allow the achievement of predictable latency. Consequently, a number of attempts to allow the usage of accelerators as general-purpose computation resources have been done. [CSZ+14] introduces the Acceleration Pool (AP) abstraction where FPGA chips are considered as a pool of accelerators with various functions and performance. In this approach, each FPGA chip has a number of pre-defined accelerator slots. Each slot can be considered as a virtual FPGA chip with virtual standardized resource types, capacity and interfaces design. Accelerators can be mapped on each slot if they follow the standardized design without caring about the hardware details of FPGA chip. Both pre-defined hardware accelerators which are stored in central repository and specific design can be submitted by the cloud tenant (in the second case the cloud owner compiles the bitstream for the FPGA slot). The AP approach enables the virtualized accelerators resource management by the cloud system, since standard slots allow regular resource allocation. The authors of [CSZ+14] implemented a prototype on an x86-based Linux-KVM environment with

attached FPGA trough PCIe interface, and deployed in a modified OpenStack cloud environment. The proposed framework shows only less than 4 microseconds of latency overhead introduced by virtualization which seems of interest also for the above-mentioned eNB virtualization case.

However, the best performance is achieved by the ASIC technology [CSZ+14]. ASICs show indeed large reductions in energy consumption versus CPUs, GPUs, DSPs and FPGAs because they can be highly customized for the required computation, furthermore improving the overall TCO. On the other hand, the flexibility and re-programmability of ASICs is limited. For that reason, a number of ASIC Cloud frameworks have been already studied, and several proprietary solutions are available today (e.g., Google's second-generation Tensor Processing Unit available on Google Cloud Platform). Also in this case the objective is to allow the usage of ASICs as general-purpose computation [MKG+16].

In summary, acceleration technologies are an important complement to virtualisation. Studies are still ongoing as to how more flexible architectures, such as the one proposed by 5G NORMA, can integrate accelerator technologies with a very minimal impact and in a future-proof manner. In any case, VNF lifecycle management procedures and object models need to be extended because they will have to cope with PNFs also in the future.

#### 2.4.1.2 Lifecycle management procedures

As pointed out in Section 2.2.1, management and orchestration in 5G NORMA are built on the concepts developed in ETSI's NFV industry specification group (ISG). However, the NFV-MANO architectural framework has to be extended to integrate PNFs and NFs using accelerators into the E2E architecture. Figure 2-12 shows the NFV-MANO architectural framework.

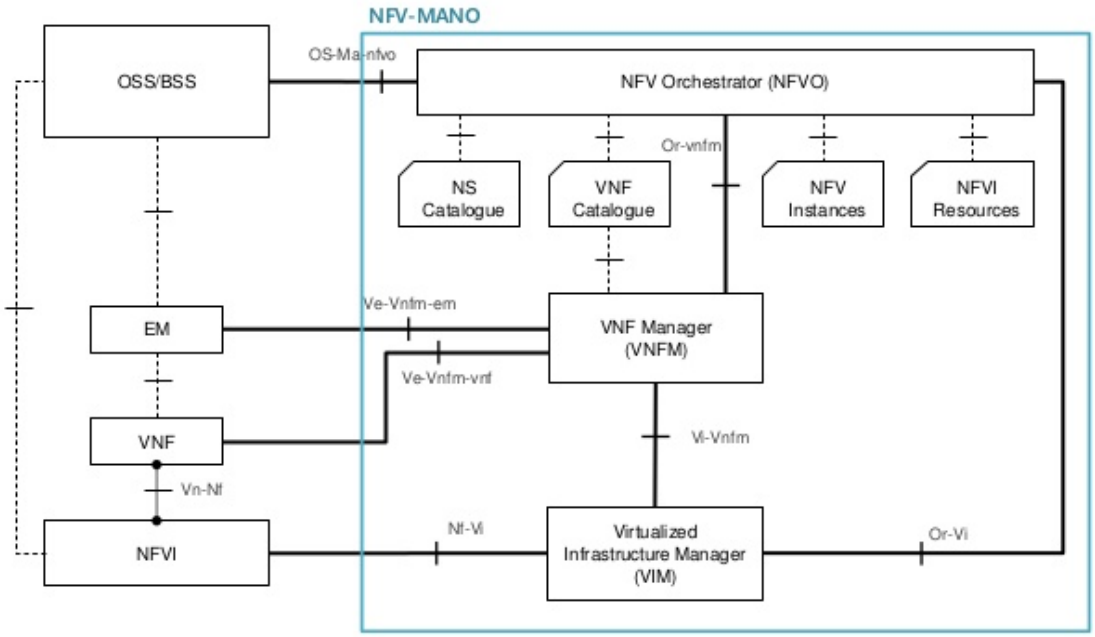



**Figure 2-12: NFV-MANO architectural framework [MANO]**

The NFV-MANO framework distinguishes three levels:

- The infrastructure and its corresponding management functions reside on the bottom level. There, HW resources (i.e. compute, storage and networking resources) are virtualised and provided to the middle level in the form of VMs. The VIM gathers the VMs under its responsibility in one or more NFVI-PoPs and informs the NFVO on the resource availability in these NFVI-PoPs via the Or-Vi interface.
- The middle level is responsible for individual Network Functions (NFs) and their management.

- On the top level, the NFVO assembles NFs to function chains that perform the network service as requested by the tenant.

5G NORMA aims at integrating PNFs, in particular conventional network elements with a very limited or even no decoupling of HW and SW, as well as network elements needing accelerators into this NFV MANO framework. Furthermore, 5G NORMA wants to automate the corresponding Lifecycle Management (LCM) processes. In the sequel, the LCM processes on each level are analysed with respect to their applicability for PNFs:

- Orchestration flows on the lowest layer are described in [MANO], Annex D, and basically comprise the creation of NFVI-PoPs, adding / removing resources and informing the NFVO about the availability of the NFVI-PoP and its resources.
- Typically, the VIM functionality is addressing generic HW and therefore is not well suited to the special-purpose HW of PNFs or to NFs utilizing accelerators. This makes it difficult to transfer these flows one-to-one from VNFs to PNFs. On the other hand, also legacy networks possess the necessary functionality to manage and configure PNFs. This functionality is part of OSS and EM, which are captured by the block "domain specific application management" in Figure 2-4 showing 5G NORMA's architecture.
  To enable the NFVO for its task to manage network services, it needs to know which resources it has available. This information is provided to it via the *Or-Vi* interface. A possible and simple solution for management and orchestration of PNFs in 5G NORMA would be to leave the infrastructure management functionality for PNFs in OSS and EM, but serve the NFVO with similar information as provided by the VIM. For this, the already existing *Os-Ma-Nfvo* interface could be extended by corresponding information elements, or an additional new, Or-Vi-like interface could be introduced between OSS and NFVO. In this way, a function block for PNFs with comparable functionality as VIM for VNFs would not need to be introduced.
- [MANO], Annex B, treats the LCM of VNFs. This includes the onboarding of VNF packages, the instantiation of VNFs, scaling them in or out, their termination and the fault management. With the exception of fault management, these tasks are not applicable to PNFs, as SW, FW and HW of PNFs are managed only jointly. The fault management of PNFs in legacy networks, in turn, is performed in the OSS. Thus, there are no tasks in the management of PNFs on an NF level that need to be taken by a similar function block as the VNF-M for VNFs. Like the VIM on infrastructure level, the VNF-M on a NF level does not need to be complemented by a new, additional function block for the management of PNFs.
- The LCM of network services as well as the management of the underlying VNF Forwarding Graphs can be found in Annex C of [MANO]. This comprises the onboarding of the network service description, the instantiation, scaling, update, and termination of the network service as well as the creation, update, query, and deletion of the network service's VNF Forwarding Graph. Basically, the NFVO is assembling function chains from NFs, which could likewise be VNFs or PNFs. Thus, far-reaching modifications of the NFVO functionality are not necessary. However, it has to be kept in mind that PNFs have slightly different properties than VNFs when it comes to scaling and portability. To describe these properties of PNFs appropriately, the VNF Packages of ETSI-NFV MANO have to be extended by some information elements. These extensions are described in the next Section 2.4.1.3.

5G NORMA distinguishes between logical models and implementation models for network slices and network functions. Logical models can be used to instantiate network slices and VNFs on different NFV Infrastructure having different controllers. They are defined in a generic manner and thus have no dependency on the underlying infrastructure. In contrast, implementation models are specific to the underlying NFV Infrastructure, controllers, and utilized technologies. The following subsection therefore focuses on logical models for the managed objects and associated LCM procedures.

#### 2.4.1.3 Managed objects

Lifecycle management needs to be supported by a catalogue and templating system as well as a resource inventory with an associated modelling language for object representation. In ETSI NFV ISG, a derivate of the OASIS standard "TOSCA" (Topology and Orchestration Specification for Cloud Applications) is used to model both VNFs and network services ("TOSCA Simple Profile for Network Functions Virtualization (NFV) Version 1.0", [TOSCA-NFV]).

The object models from ETSI NFV relevant in the context of lifecycle management include VNFs, network services (NS), and VNF forwarding graphs (VNF-FG). The latter describes the topology of the network service, i.e., it denotes the physical NFs (PNFs) and VNFs forming the network service as well as the (virtual) links connecting the NFs. ETSI NFV SOL working group has started to specify solutions for so called "descriptors" (VNF and NS) as well as "packages" (for VNFs only). 5G NORMA extends these concepts to make them suitable for integrated LCM of both PNFs and VNFs as well as LCM of network slices comprising both VNFs and PNFs.

Figure 2-13 depicts the network resource model for managed elements that 3GPP has defined for inventory management of generic NEs (referred to as PNFs in the context of 5G NORMA) via Itf-N, i.e., the interface between Element Manager and Network Manager. It shows the association diagram consisting of the NE structure, hardware, software and license data inventory. The tight coupling between hardware and software is modelled by the *SWHWRelation* association. Figure 2-14 depicts a further breakdown of the managed object, taking the example of an E-UTRAN eNB function.

In contrast, Figure 2-15 shows the high-level association diagram of a VNF descriptor as defined by ETSI NFV. It focuses on the breakdown of a VNF into so-called Virtualisation Deployment Unit(s) (VDU) as well as associated connection points and virtual links. A VDU supports the description of the deployment and operational behaviour of a VNF component (VNFC). A VNF is composed of one or multiple VNFC(s), and each VNFC maps to a single virtualisation container. The so-called "deployment flavour" information element carries deployment-specific information of the VNF, such as, available LCM operations, configuration parameters for the VNF LCM operations, or the various levels (i.e., amount) of resources that can be used to instantiate the VNF using this particular deployment flavour [NFV-IFA011]. Clearly, the model is quite independent of the application logic of the VNF, but rather focuses on NFV orchestration and LCM operations for the virtualisation container or VM.

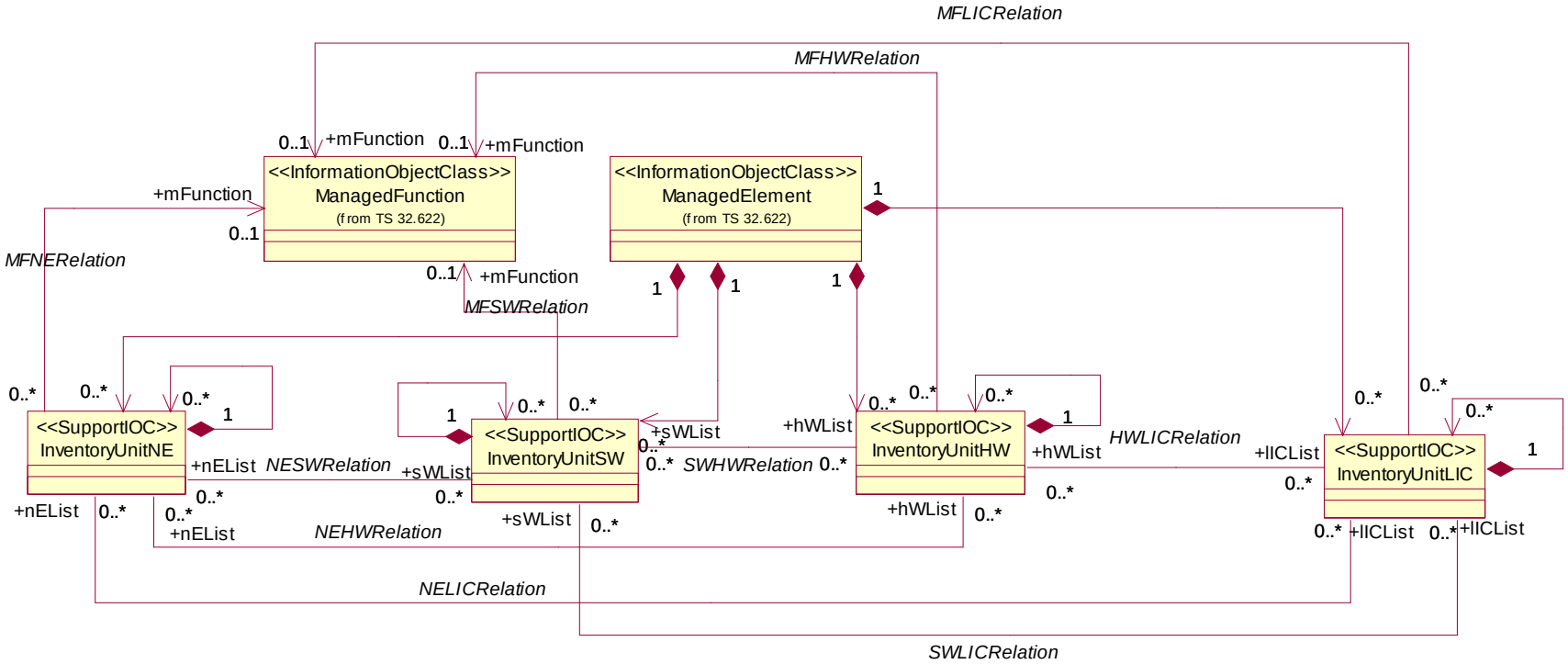



**Figure 2-13: PNF-centric 3GPP network element association diagram [32.692]**

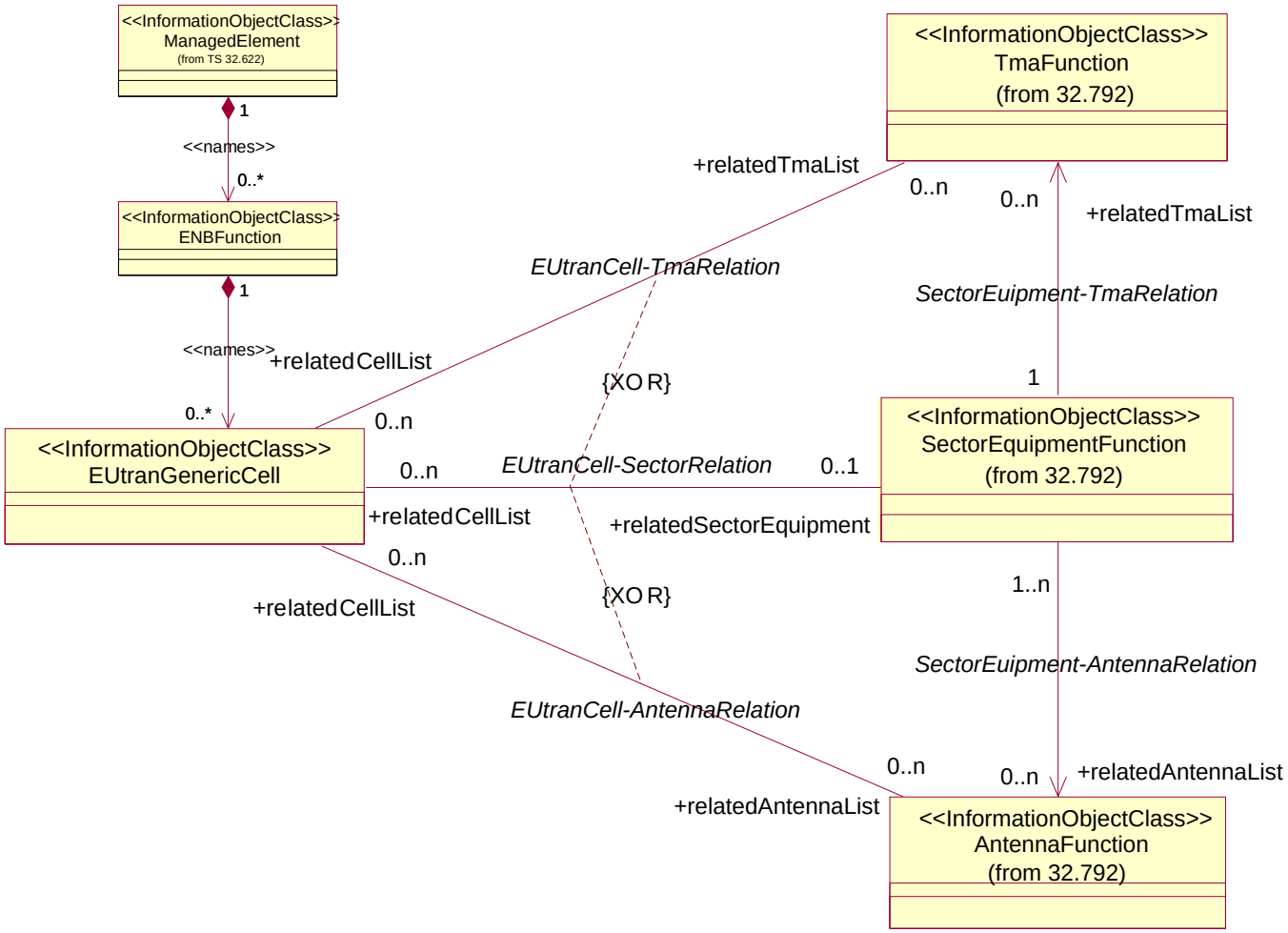



**Figure 2-14: Association diagram of eNB function, adapted from [32.762]**

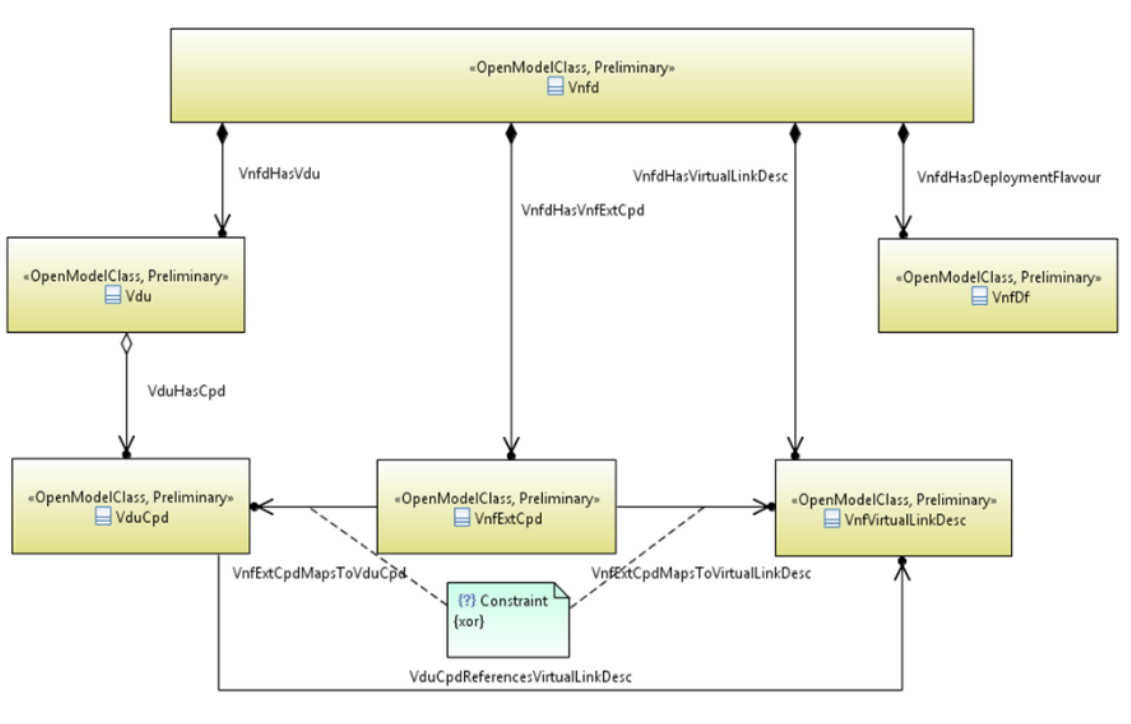



**Figure 2-15: NFV-centric high-level structure of a network function [NFV-IFA011]**

In order to bring together the two modelling approaches, the following touching points can be exploited:

- *InventoryUnitSW* (cf.Figure 2-13) <-> s*wImageDesc* information element (attribute of the *vdu*, which in turn is an attribute of the VNFD, cf. Figure 2-15)
- *InventoryUnitHW* (cf. Figure 2-13) <-> (virtualComputeDesc,virtualStorageDesc), attribute tuple of the VNFD (cf. Figure 2-15)
- *lifeCycleManagementScript*, attribute of the VNFD. In 5G NORMA, this is extended to allow for the definition of LCM operations for PNFs, such as, scaling of available resources (e.g., scalable carrier bandwidth in LTE, cell (de-)activation, adaptation of RAN sharing configurations, etc.)

### 5G NORMA NF Package

Adapted from the ETSI VNF package [NFV-SOL004] and using the 3GPP network resource models for inventory management, 5G NORMA defines an NF package to include the NF descriptor (NFD) in the form of a YAML file containing the TOSCA model of the NF as well as additional files. These additional files comprise change history files, testing files for package validation, licensing information files, certificate files, and TOSCA metadata for providing an entry information how to process the NF package. Moreover, a manifest file shall comprise the basic NF package metadata (name-value pairs). Table 2-1 depicts the adapted ETSI version that covers both PNFs and VNFs.

Further, the manifest file shall contain all references to required internal (within the NF package) as well as external files.

**Table 2-1: Conventions for names and values for NF package metadata**

| Name | Value |
|---|---|
| nf_provider_id | A sequence of UTF-8 characters |
| nf_name | A sequence of UTF-8 characters |

| | |
|---|---|
| nf_release_data_time | String formatted according to IETF RFC 3339 |
| nf_package_version | A sequence of groups of one or more digits separated by dots |
| Nf_package _description | Human-readable short description of the NF |

The NF descriptor (as the most important part of the NF package) is a deployment template which describes an NF in terms of deployment and operational behaviour requirements, including connectivity, interface and resource requirements. Thus, the main parts of the NFD are information on the NF's topology, deployment aspects, and LCM operations. Table 2-2 depicts the attributes of the NFD element.

**Table 2-2: Attributes of the NFD element (adapted from [NFV-IFA011])**

| Attribute | Cardinality | Content | Description |
|---|---|---|---|
| nfdId | 1 | Identifier | Identifier of this NFD information element. This attribute shall be globally unique. |
| nfProvider | 1 | String | Provider / vendor of the NF and of the NFD. |
| nfProductName | 1 | String | Name to identify the NF product. Invariant for the NF product lifetime. |
| nfSoftwareVersion | 1 | Version | Software version of the NF. This is changed when there is any change to the software that is included in the NF Package. |
| nfdVersion | 1 | Version | Identifies the version of the NFD. |
| nfProductInfoName | 0..1 | String | Human readable name for the NF product. Can change during the NF product lifetime. |
| nfProductInfoDescription | 0..1 | String | Human readable description of the NF product. Can change during the NF product lifetime. |
| vnfmInfo* | 1..N | String | Identifies VNFM(s) compatible with the VNF described in this version of the NFD. |
| applMgmtInfo | 1..N | String | Identifies application management functions (e.g., EMs) compatible with the NF described in this version of the NFD. |
| isrbInfo | 1..N | String | Identifies Inter-slice Resource Brokers compatible with the NF described in this version of the NFD. |
| localizationLanguage | 0..N | String | Information about localization languages of the NF |
| defaultLocalizationLanguage | 0..1 | String | Default localization language that is instantiated if no information about selected localization language is available. |
| swImageDesc | 0..1 | Separate information element | Describes the software components (or SW images) which are either<br>• installed on the PNF hardware component(s) or<br>• directly loaded on the virtualisation container (*vdu* points here if it uses the same software image) |
| pnfHw+ | 0..N | Separate information object | Defines descriptors of hardware components (physical resources) of the PNF, incl. type(s) and capabilities, and (static) deployment location |

| | | | |
|---|---|---|---|
| vdu* | 0..N | Separate information object | Virtualisation Deployment Unit |
| virtualComputeDesc* | 0..N | Separate information element | Defines descriptors of virtual compute resources to be used by the VNF |
| virtualStorageDesc* | 0..N | Separate information element | Defines descriptors of virtual storage resources to be used by the VNF |
| intVirtLinkDesc* | 0..N | Separate information element | Represents the type of network connectivity mandated by the NF provider between two or more CPs which includes at least one internal CP |
| nfExtCpd | 1..N | Separate information element | Describes external interface(s) exposed by this NF enabling connection with a (virtual) link |
| deploymentFlavour* | 1..N | Separate information element | Describes specific deployment flavours (DFs) of a NF with specific requirements for capacity and performance |
| configurableProperties | 0..1 | Separate information element | Describes the configurable properties of the NF (e.g. related to auto scaling and auto healing). |
| modifiableAttributes | 1 | Separate information element | Describes the modifiable attributes of the NF |
| lifeCycleManagementScript | 0..N | Separate information element | Includes a list of events and corresponding lifecycle management scripts performed for the NF.[2] |
| elementGroup* | 0..N | Separate information element | mechanism to associate elements of an NFD for a certain purpose, e.g., for lifecycle management. |
| nfIndicator | 0..N | Separate information element | Declares the NF performance indicators that are supported by this NF and that other functions (e.g., VNFM, EM) can subscribe to. |
| autoScale* | 0..N | Rule | Rule that determines when a scaling action needs to be triggered on a VNF instance e.g. based on certain VNF indicator values or VNF indicator value changes or a combination of VNF indicator value(s) and monitoring parameter(s). |

* only applicable to VNFs (not to PNFs)
+ only applicable to PNFs (not to VNFs)

**5G NORMA Network Slice Package**

In ETSI MANO, the concept of the Network Service Descriptor defines a deployment template which consists of information used by the NFVO for life cycle management of a network service [NFV-IFA014]. 5G NORMA extends this concepts for the purpose network slice lifecycle

[2] The *lifeCycleManagementScript* attribute contains a list that maps events to respective LCM actions. This is very homogeneous for VNFs but can considerably vary for PNFs, depending on the possibilities to scale and re-allocate the resources as described in the *pnfHw* attribute.

management. The description of a network slice ("network slice descriptor", NSD) is the key component of the Network Slice Package and includes or references the descriptors of its constituent objects: (1) zero, one or more NFDs (as defined above) and (2) zero, one or more nested NSD.

In addition, the network slice package contains NF forwarding graph (NF-FG) that describes how the NFs listed in the NSD are connected to one another, regardless of the location and placement of the underlying physical network elements and transport links. As a component of the logical model, the NF-FG defines a logical network topology and can thus map to several implementation models, which might, for example, deviate from each other in terms of placement of instantiated VNFs. The required properties, relationships, and other metadata of the connections are specified in link abstractions. To model how links connect to network functions, 5G NORMA uses the ETSI NFV concept of so-called Connection Points (CPs) that represent the virtual and/or physical interfaces of the NFs and their associated properties and other metadata.

Finally, similar to the NF package, the slice package contains a manifest file listing important metadata of the network slice, cf. Table 2-3.

**Table 2-3: Conventions for names and values for NS package metadata**

| Name | Value |
|---|---|
| ns_id | A sequence of UTF-8 characters |
| ns_provider_id | A sequence of UTF-8 characters |
| ns_name | A sequence of UTF-8 characters |
| ns_release_data_time | String formatted according to IETF RFC 3339 |
| ns_package_version | A sequence of groups of one or more digits separated by dots |
| ns_package_description | Human-readable short description of the NS |

The depicted extensions of the ETSI NFV network service descriptor now allow to perform model a network slice comprised of VNFs and PNFs and to apply LCM operations on the e2e slice by breaking them down to the constituting NFs. As a simple example, a capacity reduction (e.g., for night time) of a network slice composed of a completely VNF-based core network and a PNF-based RAN, can now be performed by reducing the number of instances for all core VNFs, re-configuration of the eNB to a reduced carrier bandwidth as defined in the PNF *lifeCycleManagementScript* (cf. Table 2-2) and adapting the reserved capacity in the backhaul transport network using SDM-X or SDM-C controllers, respectively.

## 2.4.2 NFV MANO as a Service (MANOaaS)

One of the key features of the 5G NORMA architecture is the support of multi-tenancy. This section will highlight the details on the provisioning of such an NFV MANO stack that forms the basis of MANOaaS. Subsequently, MANOaaS will be extended orchestration across multiple infrastructure domains.

### 2.4.2.1 General MANOaaS concept

The advantage of providing each tenant with its own NFV MANO stack (i.e., MANOaaS), referred to as tenant-MANO (t-MANO), is to give each tenant management and orchestration control over its own set of allotted resources and network services. This has the advantage of:

- Reduced processing load/delay on the central MANO (c-MANO) system.
- Increased effectiveness and flexibility when slices span more than one InP.

- Enabling a distributed MANO framework architecture that is more reliable and fault tolerant as compared to the management under a single MANO framework.
- Providing greater autonomy for the tenants to manage their own resources, slice(s) and policies.
- Higher level of scalability by providing added flexibility to the tenants to act as virtual infrastructure providers to other tenants and thereby managing and orchestrating its own respective tenants (i.e., nested tenants or sub-tenants).
- Controlled access to the MANO features and capabilities for the respective tenants.

Another fundamental feature of the MANOaaS concept is the capability of spanning multiple infrastructure domain. As detailed below, this feature is enabled by the 5G NORMA overall architecture described in Section 2.2 and allows dynamic stakeholders models as detailed in Section 4.1.

Figure 2-16 provides the conceptual overview of the MANOaaS, where each tenant (Tenant-1 and Tenant 2), is provided its own t-MANO stack. The c-MANO, which is assumed to be owned by the mobile network operator (MNO), is responsible for the provisioning of the t-MANO stacks to the respective tenants. As illustrated in Figure 2-16, each t-MANO instance consists of own NFVO, VNFM and VIM instances that the tenant can use to manage and orchestrate its own respective network slices with minimum dependence on the c-MANO. The t-MANO stack manages slices on the service and functional level (service orchestration for "service slices", sSlice) and on the resource level (resource orchestration for "resource slices", rSlice). Each service slice on the functional level is supported by a corresponding resource slice on the infrastructure level. Both sSlice and rSlice descriptions are contained in the annotated slice template.

Each t-MANO stack will provide either a full or partial set of the features, capabilities, and services of the c-MANO system depending on the SLA negotiated between the tenant and the MNO as the operator of the c-MANO system. However, the c-MANO system will have full administrative rights over the respective tenant's t-MANO stack. It monitors the t-MANO stack for SLA compliance and, if necessary, provides features, capabilities, and services outside the SLA bounds to the tenants. Under specific situations, it overrides or modifies decision or actions of the t-MANO stack. In other words, the c-MANO system has the added responsibility of validating and controlling the t-MANO stacks so that they do not interfere or impact the performance of network slices belonging to other tenants.

Prior to commissioning a t-MANO stack in its respective domain, it is the responsibility of the c-MANO system to allocate the required rSlice(s) for the respective tenants based on tenants' demands and requirements. The SLA negotiation shall mainly determine which functional elements and which features, capabilities, and services the tenant's t-MANO system will have, including specification of access rights and levels. The tenants can request from the MNO or MSP for resource blocks and t-MANO stacks via the Service Management function of the MNO or MSP, which provides a GUI and API for making requests and negotiating the relevant SLAs.

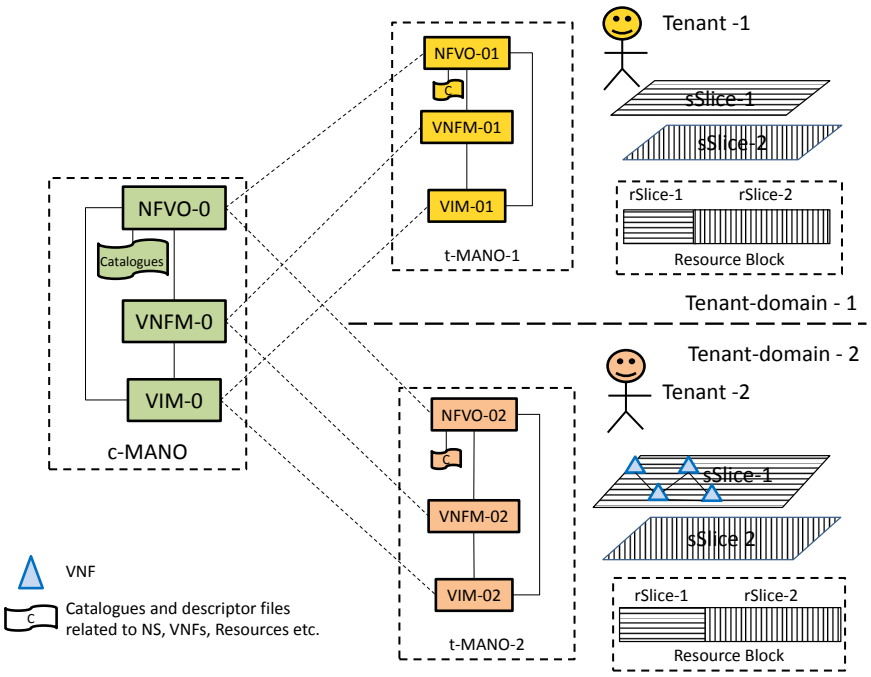



**Figure 2-16: MANOaaS conceptual overview**

Once a t-MANO stack is deployed and instantiated for the respective tenant, it will be able to partition their respective allocated resource block into rSlice(s) and then create, deploy, configure and instantiate their own sSlice(s) without involving the c-MANO. In other words, Tenant-1 and Tenant-2 will have autonomy in performing fault, configuration, accounting, performance, security (FCAPS) and lifecycle management operations and orchestration actions over its respective resource and service slice(s). Implementing own policies without involving the c-MANO is facilitated by allowing each t-MANO stack to maintain its own catalogues local to the tenant domain. However, the degree of autonomy of the respective t-MANO stack instances will depend on the agreed SLA with the c-MANO. The MANOaaS paradigm will also allow the tenants to recursively lease out resources.

### 2.4.2.2 MANOaaS implementation

This section provides an overview of the MANOaaS implementation. For the sake of simplicity, it focuses on a single domain scenario and will be extended to a multi-domain scenario subsequently.

#### 2.4.2.2.1 Deployment of t-MANO stack

A t-MANO stack can be realized as either a virtualized function or a container. This section will assume the t-MANO stack being deployed as a Virtualized Management Function (VMF) (similar to a VNF), where the NFVO, VNFM and VIM are the VMF components (VMFC) interlinked by virtual links (VL) that implement the necessary interfaces between them. VMFCs of the t-MANO stack maintain peer relationship with the c-MANO stack via the ISRB.

Figure 2-17 shows an example of the deployment and provisioning of a t-MANO system stack for Tenant-1 by the c-MANO that is owned by the MNO. Tenant-1 is allocated a quota of NFVI resource block by the c-MANO system. Upon specific request by Tenant-1 for provisioning of a t-MANO stack, the c-MANO system deploys, instantiates, connects, and configures the different VMF Components (i.e., VIM-1, VNFM-1, NFVO-1) to create a t-MANO stack instance in almost a similar manner as it would deploy, instantiate, connect, and configure VNFs and/or its components forming a network slice. In case the NFVO-1, VNFM-1 and the VIM-1 components of the t-MANO stack are on-boarded as separate VMFCs, a VMF Forwarding Graph (VMFFG)

and VMF descriptor file (VMFD) need to be present in order to describe the interconnectivity between the three essential t-MANO functional blocks. VLs are used to interconnect the VMFCs whose characteristics are described in the VL Descriptor (VLD) file. Thus, with the help of these descriptor files, the c-MANO system is able to deploy and instantiate the t-MANO stack in a manner that is similar to the deployment and instantiation of the VNF and the network slice.

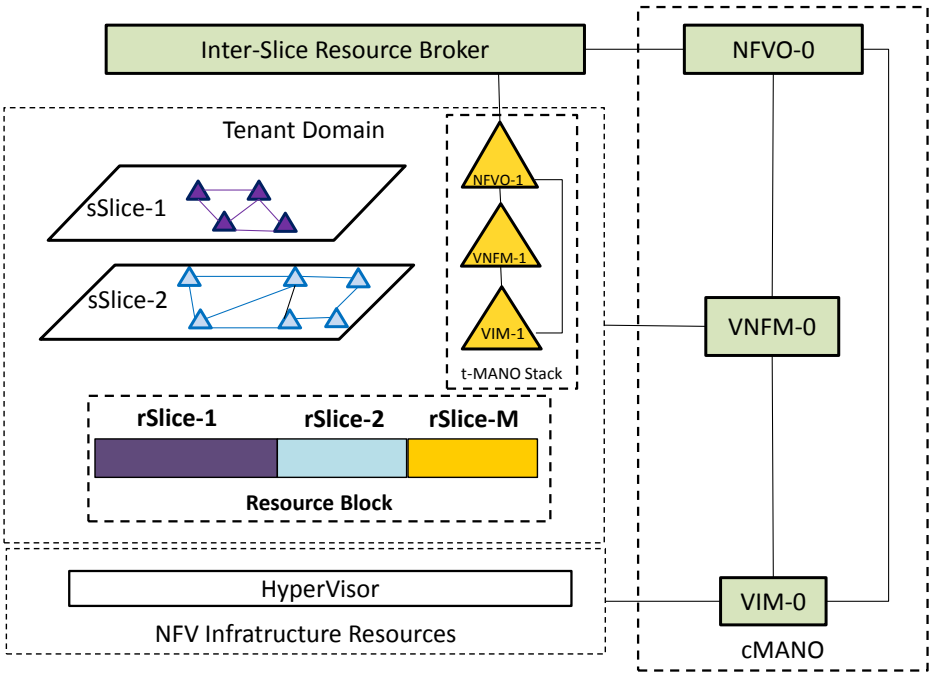



**Figure 2-17: Deployment and provisioning of t-MANO as a VNF**

#### 2.4.2.2.2 t-MANO deployment process

Figure 2-18 provides the process overview for the instantiation of the t-MANO stack. The process involves the following main steps:

(1) The tenant requests the c-MANO system of the MSP or MNO for the allocation of NFVI resource block by specifying the type (e.g., compute, network, storage, memory etc) and capacity of each resource type. Such a request is made through the customer portal and it is handled by the Service Management. This procedure is a part of slice setup and deployment phase described in [5GN-D52]. The flavour of the resource block is based on the tenant's service requirements.

(2) The tenant then requests for the provisioning of the t-MANO stack indicating the required VMF components (VMFC) such as NFVO, VNFM, VIM and specifying required privileges to access specific MANO features, capabilities, and services that the tenant requires from the t-MANO stack.

(3) The c-MANO will create a t-MANO Catalogue based on the parameters indicated in the request.

(4) An SLA negotiation process will take place through Service Management during the slice setup and deployment phase [5GN-D52]. The agreed SLA determines the scope of tenant's autonomy (with respect to c-MANO) to access MANO features, capabilities, and services that the tenant is allowed to execute independently using its t-MANO stack within its domain.

(5) A VMFD instance is created, which is updated with the SLA parameters agreed between the tenant and the MNO (MSP).

(6) The c-MANO system instantiates the t-MANO stack and configures its relevant management components (i.e., NFVO-1, VIM-1, VNFM-1) as per the agreed SLA indicated inside the VMFD.

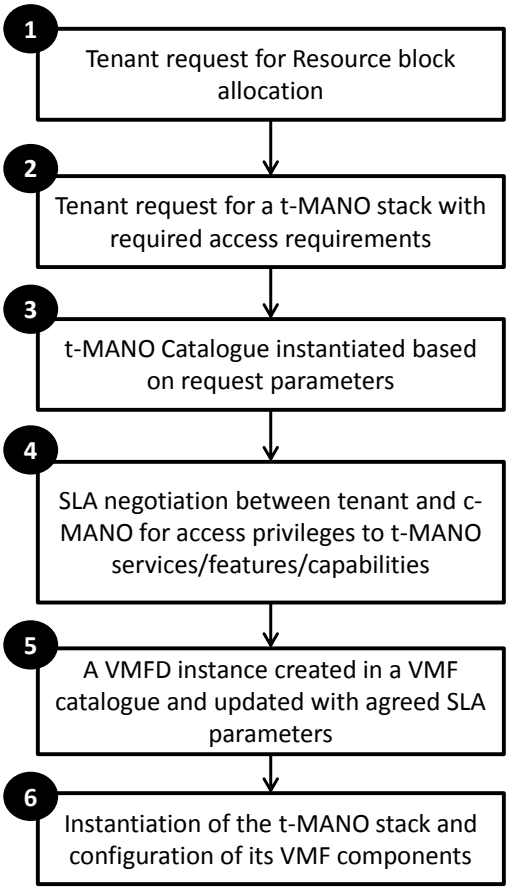



**Figure 2-18: Process overview for t-MANO instantiation**

The t-MANO stack is now ready to be used by the tenant for creating own rSlices. According sSlices can be instantiated, managed and orchestrated by the t-MANO stack within the negotiated SLA.

## 2.4.3 MANOaaS: Extension to multi-domain orchestration

### 2.4.3.1 Motivation and requirements

Multi-domain orchestration is a required extension of the default single domain orchestration that allows MSPs to deploy network slices in multiple domains, e.g., due to performance requirements or specific tenant requests

Multi-domain orchestration is a concept that has repeatedly appeared in literature [KNE16] [Intel][CSC14][SCS15]. ETSI NFV MANO [NFV-MAN001] defines "administrative domain" as a "collection of systems and networks operated by a single organization or administrative authority". In the 5G NORMA context, this concept is aligned with the notion of an Infrastructure Provider (InP, cf. Section 4.1). In other words, multi-domain refers to orchestrating and managing E2E network slices across heterogeneous infrastructures of different InPs (or MNOs) and deployed in different geographical locations.

As an example, some specific cases requiring multi-domain orchestration could be industrial customers needing multi-national (or even worldwide) services. This usually involves multiple InPs and mobile service providers MSPs since they are usually organized in national/regional boundaries. Another use case comprises tenants requesting to integrate parts of their services already running in one InP domain into a different InP domain. For example, in order to implement a new functionality in an already deployed service, the tenant could request to extend its forwarding graph (FG) to include VNFs that must be executed in a different InP domain. This

can be used also to execute selected parts of services in specialized domains in order to increase productivity.

The main requirement for 5G NORMA multi-domain orchestration is the consolidation of management and control into a single virtual domain in order to avoid having one network operator per country or regions. A second important requirement is the support of the three offer types stated in Section 4. For this purpose, E2E service orchestration across multiple domains extends the network slice concept beyond the boundaries of a single InP, i.e., a network slice consumes infrastructure resources from multiple administrative domains.

The main leverage to provide multi-domain orchestration comes from the ETSI NFV MANO framework [NFV-MAN001] and the MANOaaS concept.

ETSI NFV MANO assumes that in a typical scenario for NFV there will not be a single organization controlling and maintaining a whole NFV system [NFV_003]. Multiple organizations, either departments of the same "root" organization (e.g. a network and an IT datacentre department) or different companies, provide different functional blocks of the NFV Architectural Framework. The ETSI framework identifies two types of domains:

(1) Infrastructure Domain and
(2) Tenant Domain.

An Infrastructure Domain may be defined by different criteria (e.g., by organization, by type of resource such as networking, compute, and storage, by geographical location, etc.), and multiple Infrastructure Domains may co-exist. Furthermore, an Infrastructure Domain may provide infrastructure to a single or multiple Tenant Domains.

Similarly, a Tenant Domain may be defined by different criteria (e.g. organization, by type of Network Service, etc.), and multiple Tenant Domains may co-exist. The 5G NORMA multi-domain orchestration corresponds to a scenario where a Tenant Domain uses infrastructure from multiple Infrastructure Domains as defined in [NFV-MAN001]. VNFs and Network Services residing in the Tenant Domain can consume resources from multiple Infrastructure Domains by using the Infrastructure Domain orchestration functionality. In this way, a Tenant Domain and an Infrastructure Domain can be mapped against different organizations (particularly in the case when the Infrastructure Domain is offering infrastructure to a Tenant Domain). ETSI NFV states that the existence of different administrative domains leads to the requirement to support different management and orchestration functionality in those domains. However, the specific mapping of administrative domains to NFV-MANO functions is deliberately left out of scope in the ETSI NFV MANO model.

#### 2.4.3.2 Multi-domain orchestration in 5G NORMA

According to the ETSI NFV model, the NFVI is the totality of all hardware and software components which build up the environment in which VNFs are deployed. The ETSI specification specifically declares that NFVI can spans across several locations (i.e., multiple NFVI-PoPs). The network providing connectivity between these locations is regarded to be part of the NFVI [NFV_003]. Also, NFVI resource management across operator's Infrastructure Domains can be performed using one or more VIMs as needed.

These VIMs are responsible for controlling and managing the NFVI resources. However, they are intended to be used for controlling and managing the NFVI resources usually within one operator's Infrastructure Domain. 5G NORMA proposes to extend the concept of network slice towards a "multi-domain slice", thus allowing tenants to integrate resources from different Infrastructure Domains. Necessary resources from these different domains are integrated into a single NFVI block that is orchestrated from just one administrative domain. From the business perspective, specific agreement among the different administrative domains should be reached to make this possible (which is realized by a separate "offline" procedure). From the technical

perspective, the domain designated for orchestration is granted access to the resources in the "external" domain(s).

Figure 2-19 shows an example of the described scenario considering two infrastructure domains: NFVI-0 (depicted in green) and NFVI-1 (depicted in blue). These could be two InPs (e.g., two MNOs), each providing a specific set of resources. Moreover, each InP could have instantiated different t-MANO stacks for different tenants (in the green domain two t-MANO stacks are represented (orange and pink) beyond the c-MANO system depicted in green. For the sake of simplicity, no c-/t-MANO stacks are depicted in the NFVI-1 domain.

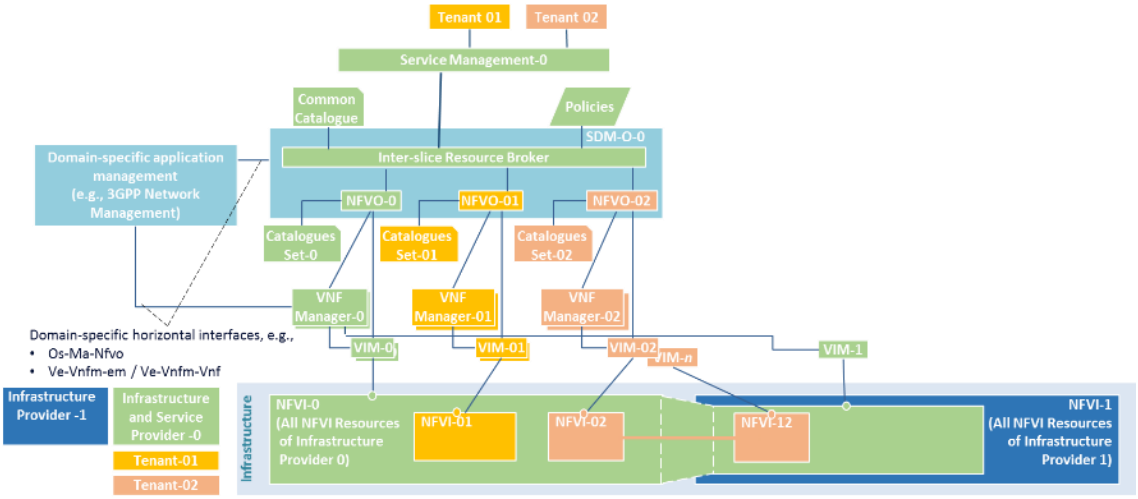



**Figure 2-19: Multi-domain orchestration on infrastructure level**

In this situation, the tenant operating the pink t-MANO stack could request to utilize resources from the blue domain to operate his network slice, which is already being executed in the green NFVI-0 domain. The technical solution is to integrate these "blue" resources into NFVI-0 block and using specialized VIMs to manage these resources. More specifically, VIM-1 enables the green InP to integrate these "foreign" resources into its own domain. From the point of view of the VIMs of the pink t-MANO stack, all resources are consolidated in single domain again. In other words, the complexity of multi-domain orchestration is hidden from the tenant[3]. In the ideal case, this would limit the need for business agreements, i.e., only InP-0 (green) and InP-1 (blue) would have to close an agreement.

5G NORMA considers this business agreement as an "offline" procedure (i.e., outside of the usual NFV MANO procedures) that should be reached before the technical implementation, cf. Section 2.4.2.2.2.

InP-0 (depicted in green) and InP-1 (depicted in blue) have to agree on:

- The amount and type of InP-1 NFVI resources to be allocated into the InP-0 Infrastructure Domain; this includes the typical NFVI resources (compute, storage and networking).
- The commitment model to allocate them (static quota, dynamically scalable quota, allowed/supported protocols, etc.).
- Exceptionally, the agreement could also include the VIM(s) in case the rented resources would require a specific VIM. From the technical point of view, the only condition for integration is that such "external" VIMs shall provide the ETSI NFV *Or-Vi* and *Vi-Vnfm* reference points in order to connect with the VNFM(s) and the NFVO.

[3] Technically, it would be necessary to also register the allocation of resource subsets to administrative domains in the NFVI Resources Repository of the involved c-MANO and t-MANO stacks.

- Additional operational policies (including, but not limited to, scaling rules, security requirements, redundancy and overprovisioning levels).

Regarding the fundamental realization of multi-domain orchestration, 5G NORMA considers two scenarios:

(1) InP-0 extends the NFVI-0 infrastructure in order to provide a better service to the hosted tenants. E.g., Tenant-02 could request more infrastructure resources from InP-0 than NFVI-0 can satisfy. Then, the InP-0 signs a business contract with another InP (InP-1) to include resources from that provider into his infrastructure domain. The request of Tenant-02 can now be satisfied and the new resources can be made available to the pink t-MANO stack, no additional business contract between InP-1 and Tenant-02 is required. From a technical point of view, the fact that some NFVI resources used by Tenant-02's t-MANO stack are actually located in the InP-1 domain is transparent from the tenant perspective. In other words, for Tenant-02 it looks like resources from the InP-0 domain are used and only a single business relationship is maintained.

(2) The tenant explicitly wants to extend its slice using specific infrastructure from another InP (InP-1). This could happen when a tenant already has certain infrastructure up and running on a different InP and does not want to take the risk/effort of migrating that infrastructure into the InP-0 domain. I.e., the tenant already has a business contract with both, InP-0 and InP-1. In this case, besides the corresponding update of these both contracts, a new contract involving both InPs needs to be agreed.

Even though these considerations above only cover a simple setup (one tenant, two InPs), the situation is conceptually similar if more parties are involved. Besides the business contracts and agreements between the different parties, relevant technical considerations include:

- InP-1 shall isolate and assign the requested NFVI resources and provide the necessary interfaces/reference points to InP-0. In particular, this applies to the *Nf-Vi* reference point when the InP-0 deploys its own VIMs, or the *Or-Vi* and *Vi-Vnfm* reference points when the VIMs are supplied by the InP-1, as described above.
- InP-0 shall take the newly integrated NFVI resources and associate those to one (or multiple) t-MANO stacks, thus forming a "merged", single-domain set of resources for these stacks.
- While Figure 2-19 does not assume any deployment constraints of VIMs or the entire c-/t-MANO stack instances, for performance reasons (e.g., latency) it might be necessary to add such constraints. However, this does not change the functional perspective of the architecture.
- Tenant's t-MANO stacks should have the full information about logical node topology and resources, address space, etc. within its specific Tenant Domain (NFVI-02 and NFVI-12 in Figure 2-19), independent of whether these resources originally come from the blue domain (InP-1) or the green one (InP-0). Hence, NFVI-02 and NFVI-12 can be used, together with the associated VIMs, to set up the interconnections between VMs in the Tenant Domain accordingly.
- Security issues: As detailed in Section 4.4, users of infrastructure cannot enforce well-behaviour of the InP but need to trust the InP. Similarly, tenants need to trust their MSP. A tenant may not want to rely on trust in arbitrary InPs – hence an SLA between tenant and MSP may restrict the choice of InPs the MSP can use to host tenant functions. It can be further noted that in a multidomain scenario, inherently the number of involved parties, data centres, interfaces and software routines will be higher, thus increasing the attack surface. There is no specific remedy against this – the general security rules apply, in particular designing a sound security architecture (cf. Chapter 5) and implementing it carefully in order to minimize the threat of exploitable vulnerabilities.
- Although the adjustment of resources assigned to the slices depends on an offline procedure, the policies to assign resources could be the same as for the single-domain use case. I.e., t-MANO stacks could be assigned with a specific quota of resources (regardless

of whether those resources come from a single domain or multiple domains) that should be properly dimensioned to dynamically scale during the slice operation. In both cases (single-domain or multi-domain), an offline re-negotiation is necessary if the assigned quota of resources were not enough to meet the tenant's necessities for its slices.

Resource allocation in a single-domain NFVI is a potentially complex task because a lot of requirements and constraints need to be met at the same time. Adding multi-domain capabilities adds a supplementary degree of complexity. However, as previously described, this can be addressed by the design of the 5G NORMA architecture.

### 2.4.4 Function selection and placement

Function selection and placement constitute key decisions taken by the MANO layer entities. The network of functions as proposed by 5G NORMA allows for adapting the NF graph to best provide the requested telecommunication service. Two basic levels of flexible function adaptation can be distinguished, namely *function selection* and *function placement*: first, the most appropriate NFs are selected and, second, the selected NFs are then deploying at the most appropriate physical location. In the following, this approach is exemplified with three typical services, focusing on the data layer, cf. Figure 2-20.

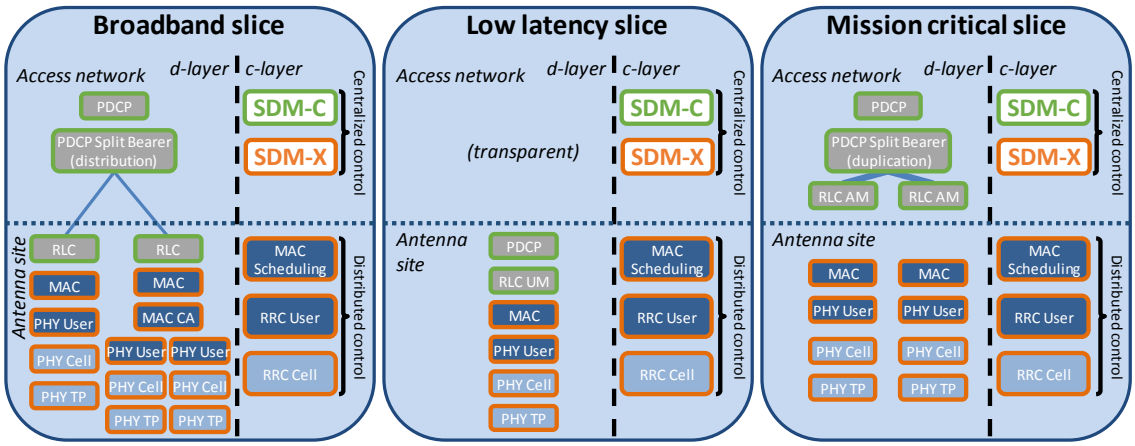



**Figure 2-20: Function selection and placement (RAN slicing Option 2).**

The broadband service targets maximum data rate and efficiency for high performance UEs. To increase throughput, the traffic is split both at PDCP and MAC layer through selecting VNFs *PDCP Split Bearer* in distribution mode for multi-connectivity and *MAC CA* for carrier aggregation for increased system bandwidth. *PDCP Split Bearer* is deployed at the edge cloud located within the access network to facilitate multi-connectivity across antenna sites.

The low latency service targets minimum latency. The edge cloud in the access network is bypassed and the complete AS protocol stack is deployed at the (edge cloud at the) antenna site. *RLC UM* (unacknowledged mode) is selected to avoid additional delay through ARQ, relying solely on the much faster HARQ on MAC/PHY layer. For the same reason, no multi-connectivity is used to avoid any reordering delays.

The mission critical service target maximum service availability. Data is duplicated across all available connections by selecting *PDCP Split Bearer* in duplication mode and *RLC AM* (acknowledged mode) provides additional retransmissions. All VNF instances are deployed in an edge cloud in the access network to facilitate coordination among RLC instances. This helps avoiding unnecessary further transmissions of one *RLC AM* instance, if another *RLC AM* instance already successfully conveyed that data, freeing radio resources for additional retransmissions where really needed.

# 3 Procedures, Interfaces, and Protocols

After the 5G NORMA system architecture design has been presented in Chapter 2, Chapter 3 describes interfaces between network functions and associated protocols. Due to the modular architecture and the resulting high number of interfaces, the chapter makes a selection of the most relevant and innovative procedures and elaborates on the involved functions and utilized protocols. It thus complements the deliverables [5GN-D32], [5GN-D42] and [5GN-D52], but also occasionally refers to them for further details.

Section 3.1 builds on Section 2.4.1 and describes the 5G NORMA procedures for end-to-end network slice lifecycle management. Section 3.2 covers a further MANO layer aspect: migration and context transfer of VNFs, including the design and application of VNF placement and re-location rules. Two control-layer-related (SDM-C/-X) solutions are depicted and evaluated in Section 3.3: QoE-aware enhanced Inter-Cell Interference Coordination (eICIC) and QoE-aware video packet pre-scheduling. Section 3.4 shows how charging and lawful interception are realised in a customised (service- or tenant-specific) manner within the 5G NORMA architecture. The chapter concludes with two major enabling procedures that combine several 5G NORMA novelties from multiple WPs (3, 4, and 5): *service-specific network slice composition and customization* and *multi-tenant network control and resource allocation*.

## 3.1 Network slice lifecycle management

According to [28.801] and depicted in Figure 3-1, lifecycle management is composed of four distinct phases: (i) preparation phase, (ii) instantiation, configuration and activation phase, (iii) run-time phase, and (iv) decommissioning phase.

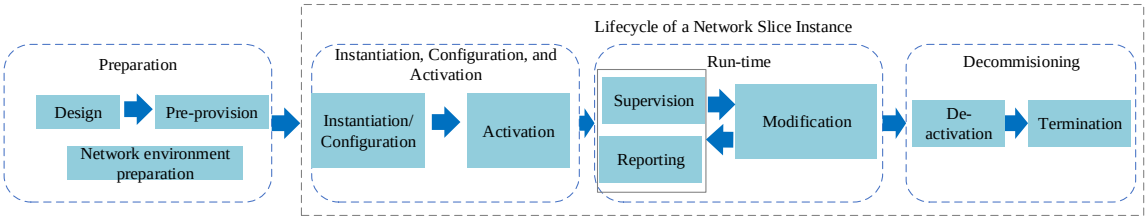



**Figure 3-1: Network slice life-cycle phases from [3GPP TS 28.801]**

**Preparation phase (i), design of network slice template**

The preparation phase includes the creation and verification of network slice description template(s), the onboarding of these, preparing the necessary network environment which are used to support the lifecycle of network slice and any other preparations that are needed in the network.

For defining a network slice properly, it is essential to capture the use case from the tenant, the related service requirements and to map or translate these into a network slice description template.

The 5G NORMA service layer handles the SLA from the tenant or from the mobile service provider. It holds process for defining the "network slice as a service" shorten to service template according to the use case from the tenant. Each possible use case is decomposed into modular, human understandable service specifications. Each service is described by a unique set of non-overlapping service attributes and their value range. The non-overlapping service attribute sets and their range are used to define the service description template. Those cover set of service performance attributes with range value such as the service throughput, the number of devices, the geographical coverage, the latency, the availability, the set-up time, the security, the reliability, the out-of-order tolerance, etc. They also contain service operation attributes such as

the duty cycle, the transmission characteristics (sporadic to continuous), the service priority (pre-emption), the mobility type, the user density, the traffic model, etc.

The mobile service provider offers an API to the tenant over the service layer ‘s interface to provide exposure of the service template(s), selection of the service template and customization with specific attributes from the tenant. The tenant gets the monitored SLA and associated service KPIs over this same interface from the service layer. The specific service requirements and SLA from the tenant are captured into the annotated service description template.

The annotated service template is the output of the 5G NORMA service layer and is provided to the service management entity in 5G NORMA Management and Orchestration Layer over the SI-Sm interface.

The service management entity from 5G NORMA Management and Orchestration Layer holds the catalogues for network slice, network functions, infrastructure resources as well as the repository for slice policies. The service management entity uses the annotated service template to select an adapted network slice template. In case there is no available corresponding template already in the catalogue for the requested service, it goes into a design phase for defining a network slice template fulfilling the service needs. The service management entity controls the process for designing the network slice along with the definition of the various policies for the slice management and control in run-time, i.e. to carry out phase (ii)-(iv) of the slice life-cycle. The output of the service management entity is an annotated network slice template with tenant’s SLAs. The annotated slice template is further delivered to the 5G NORMA SDM-O via Os-Ma-Nfvo interface.

The service management entity offers over the Sl-Sm interface some APIs onto the slice template(s) to other MSPs who want to use a sub network slice with specific attribute for building their own E2E network slices. The APIs over the Sl-Sm interface also support the tenant to design/compose a network slice and/or to define set of service policy rules for the network slice operation. Details are provided in next section. The service management holds procedures for checking and incorporating tenant’s certified functions in the catalogue of VNF.

The service management realizes the following operations for the network slice design and network slice template definition. The service description information is used for the selection of specific network functions, network services and their forwarding graphs. The KPI value ranges are used to define the parametrization of the network functions as well as for the forwarding graph. In 5G NORMA, the network slice is end-to-end with wide coverage area. In case the E2E network domain is split into multiple sub administrative domains, the network slice template is split into sub-network slice templates associated to each sub administrative domain. The requirement on E2E service attributes are divided onto the sub network slices. 5G NORMA addresses end-to-end network slices spanning across multiple vendors, multiple technologies domains and multiple operator/administrator domains. Therefore, it is necessary that the description language of an E2E network slice is generic enough that it is written in a universal description (extended from e.g. TOSCA) understandable from the multiple domains in charge of executing the activation of the slice based on its description. The network slice template contains the following parts:

(1) It comprises a specification of the NFV MANO stack instance (NFVO, VNFM, VIM, catalogues for network services and functions, etc.) that is dedicated to the lifecycle management of the network slice. It also includes the set of policies for the MANO stack to carry out the phase (ii) – (iv) for the slice lifecycle,
(2) It comprises a specification of the OSS dedicated to the slice domain specific operations (e.g. 3GPP domain) and the SDM-C/X that control the slice,
(3) The specific network functions, PNFs and VNFs from RAN, EPC, transport, service, application domain. It covers functions from data layer, control layer, and management & orchestration layer including the control applications to run on top of the SDM-C controller. It is notified when functions are potentially shared with other slices and specifies the reference to the relevant SDM-X entity controlling those shared functions,

(4) The topology of the network made of the above network functions and the links for interconnecting those network functions,
(5) The set of network function forwarding graphs expressing the sequence of network functions a data or a control layer traffic flow has to traverse,
(6) The physical and virtualized resources required by each function and each link,
(7) The set of service KPIs,
(8) Indications on function location with respect to the infrastructure domains (e.g. in one of the edge datacentre(s) or in the central datacentre),
(9) The configuration of network functions,
(10) The set of domain specific policies for managing and controlling the slice (e.g. for QoS/QoE management, for SDM-C control, for FCAPS).

The design phase includes a verification of the slice design through an iterative process including slice network modelling using the infrastructure resource description to check the feasibility for the slice creation and its performance. Designed slice can also be tested, refined and verified by real deployment onto platforms in sand box. The iteration stops once the designed slice performance meets the requirements from the use case. Figure 3-2 depicts the process flow for network slice design.

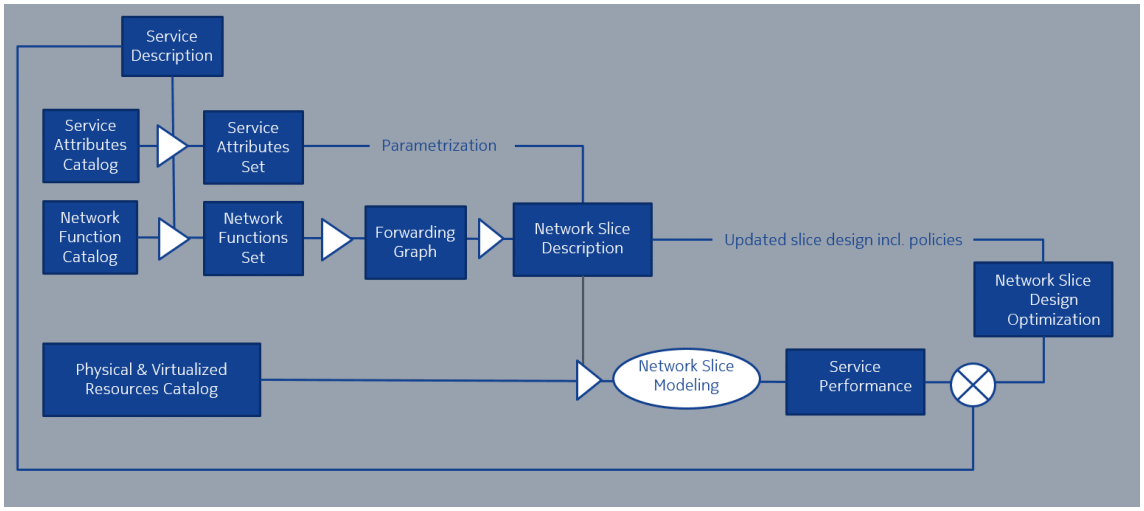



**Figure 3-2: Operations for network slice design**

### Operations in the Activation phase (ii)

Upon a new service/slice order from a tenant, the service management enters in a slice admission control. This step is done through an overall verification of the slice design done at the activation phase. It checks for the real amount of resource not used by already deployed network slices for deploying an instance of the network slice. Possibly, it results in adapting the final design of the slice or in adapting the infrastructure resource allocation to already deployed slices. In case re-allocation of resource from other deployed slices is needed, the service management trigger the SDM-O for requesting a new resource brokering decision among the slices. If successful the slice management accepts the service request from the tenant, then onboards the new network slice to the SDM-O and orders the instantiation of the slice via the *Os-Ma-Nfvo* reference point.

The Inter-Slice Resource Broker, as part of the SDM-O, uses part (a) of the network slice descriptor, i.e., the NFV MANO descriptor, to commission a new NFV MANO stack. In second step, the rest of the network slice descriptor is utilised to generate the necessary objects and models that the NFV MANO instance operates on, i.e., NFV service catalogue, VNF/PNF catalogues, NFV instances, and NFVI resources. For the allocation of the NFVI resources that are under control of this MANO stack instance, the Inter-Slice Resource Broker uses a combination of the resource commitment models as outlined in [5GN-D3.2]. Commissioning of the network slice control and data layer functions is triggered by the Inter-Slice Resource Broker via the Os-

Ma-Nfvo reference point of the NFVO by providing or referring to the set of network service descriptors to be instantiated. When activation phase (ii) is terminated, end-devices can attach to the new slice.

**Operations in phase (iii) and decommissioning (iv)**

In next steps, the network slice lifecycle management is now delegated to the NFV MANO instance and the according domain-specific application management functions, for the slice life cycle phase (iii) – (iv). The run-time operations between MANO and control layer are further detailed in the sub-section on the description of the interface between MANO and control layer.

# 3.2 Mobility of VNFs

## 3.2.1 VNF migration

The 5G NORMA architecture provides support for the adaptive (de)composition and allocation of network functions - they can run either on the central cloud or on the edge cloud. Their location is based on service requirements and deployment needs. To meet all those requirements and needs, network functions (specifically, virtual network functions) might have to be moved within the network. In this section, the focus is why and how a VNF could be migrated or replicated. Some strategies for performing this replication/migration are also discussed.

Replication/migration is a re-orchestration process, triggered by the SDM-C or the SDM-X, and implemented by the NFVO using the VNFMs and VIMs, whenever the network cannot fulfil the service requirements of a given service, cf. Figure 3-3. There are two types of events that will trigger a replication/migration:

- Inability of meeting defined QoE/QoS targets, for example
  - Inability of meeting QoS goals for a specific parameter (the most common example are time-critical functions being moved to the edge cloud),
  - Sudden changes leading to service interruption:
    A large group of users attached to a base station manage by an edge cloud moves to another edge cloud, for example, a large group of soccer fans takes the train from the suburbs to the stadium area,
- Resource shortages, such as an overloaded edge cloud, requiring functions to be moved to the central cloud

When the decision for replication/migration has been made, two potential strategies for replication/migration exist:

- Context migration and replication:
  Transfer of state information and reassignment of flows from a function on one cloud to the same function in another cloud. The target function could be already instantiated or be instantiated on demand. Function deployment will follow the same blueprint used by the SDM-O for the original function. Functions with negligible amounts of state information and tolerant of long interruptions will use this strategy.
- Live Migration for stateful functions:
  Live migration is a procedure that entails the instantiation of the same VNF followed by the live transfer of memory, storage and network connectivity from the original VNF to the new VNF (e.g., similar to what is currently implemented in OpenStack Nova Live Migration). This strategy applies to functions like Video Caching. If a video is being frequently requested in an edge cloud, it could be moved closer to the users (together with the video server VNF) using live migration.

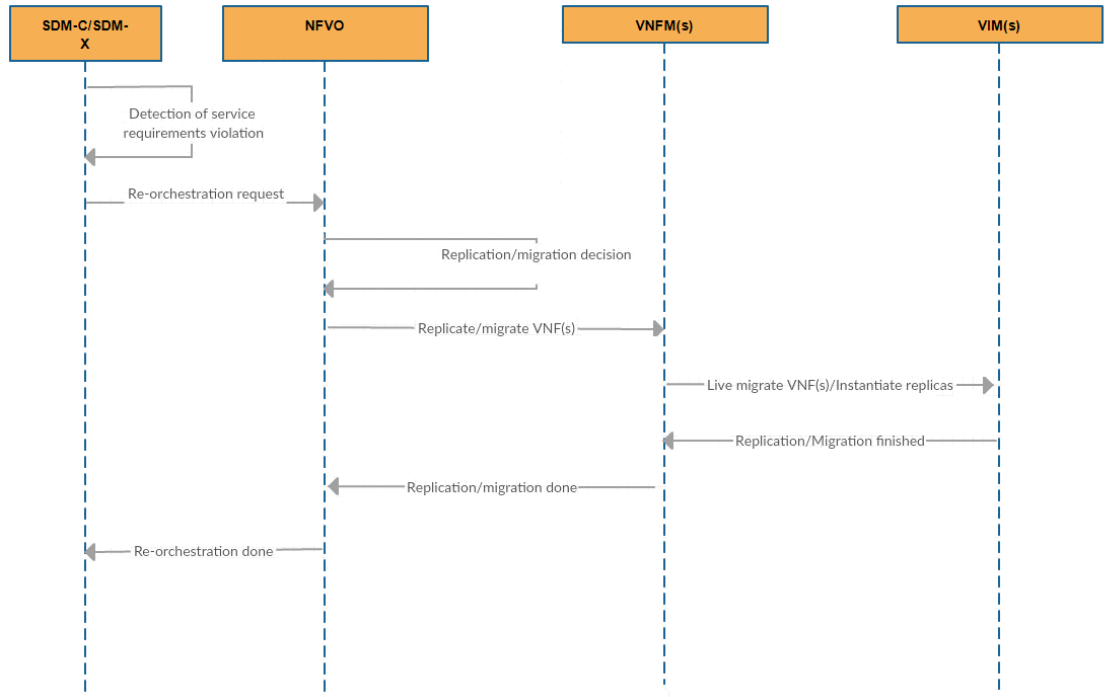



**Figure 3-3: Message Sequence Chart describing the general process of replication/migration in 5G NORMA**

### 3.2.2 VNF placement rules

This section provides a comparison between the proposed set of optimization problems developed within the 5G NORMA project with a close relevant work presented in [HKL17]. The work in [HKL17] aims to provide a flexible allocation of virtual machines (VMs) on distributed clouds. In a similar manner, flexible allocation of VNFs has also been one of the main design characteristics of the proposed 5G NORMA architecture. This has been in line with other 5G NORMA proposed techniques and solutions aiming to allow greater flexibility in the network operation and management which can be considered as the cornerstone for introducing efficient network slicing. An important element in that architectural framework is to allow for flexible VNF placement in order to allow for more efficient use of the limited resources from the physical (substrate) network infrastructure. This is also extremely important when considering the highly diverse requirements that might arise from different tenants/slices, applications and services which broadly speaking can range from enhanced mobile broadband, to massive machine-type communications and ultra-reliable low-latency communications to support haptic-type applications.

The work in [HKL17] can be considered as a generalized VM placement methodology that allows VM allocation across spatially distributed clouds. The proposed techniques can provide optimal and/or near-optimal allocations by taking into account various optimization criteria and request characteristics (such as delay for example) and assume a number of different VM types and in terms of resources the emphasis is placed on CPU, memory and storage space. The proposed optimization problems consider solving the allocation across a time horizon by assuming that each request requires resources for a specific time period. In terms of the main objectives has been the maximization of the revenue assuming a known function when hosting different VM types at different clouds. However, the framework is general and does not take into account the particularities of NFV framework. One of the main characteristics of the NFV paradigm is that it can steer traffic through sequences of VNFs deployed in edge and/or core clouds to form service chains, in the so-called VNF forwarding graph. One key aspect is that the specific ordering (i.e., sequence) of the visited VNFs per service need to be preserved. This requires special constraints

construction in the mathematical programming framework which has not been considered in [HKL17], however since 5G NORMA has been focusing explicitly on VNF chains this issue has been considered. Moreover, the issue of mobility has also been considered in order to allow optimal VNF placement by taking into account user mobility. The work in [HKL17] does not consider such mobility issues which explicitly affect the routing performance and implicitly the quality of service at the end user. In addition to mobility, the issue of routing including multi path inter-VNF chain routing has explicitly been considered and it was reflected as one of the main optimization objectives, I.e., to minimize overall routing cost of the VNF chains. However, it has to be mentioned that even though such multi path routing support has not been explicitly considered in the framework developed within [HKL17] it is possible to include such aspects in the what they call "multiple Group Case" (but this need to include VNF ordering in the chain which is not considered).

Another key difference is that in addition to the offline versions of the proposed optimization problems, there are also on-line allocation algorithms with provable performance. Such an on-line operation where requests are treated in a one-by-one fashion is highly desirable since it can be deemed as more realistic since requests inherently arrive (stochastically) in such a manner. On the other hand, the salient assumption behind the optimization problems defined within 5G NORMA project assume that all requests are known in advance or an operator can provide batch processing within a pre-defined controlled time interval (in this case there might be some delay on constructing the VNF graph but depending on the use case scenario this might be acceptable). Such an assumption means that generally speaking such optimization problems can be defined as offline methods since they require that requests are known in advance. Another strength of the work in [HKL17] is that it considered a time-domain based optimization where VM allocation has a pre-defined life span. The work within 5G NORMA did not consider (by incorporating it within the modelling framework) such time domain allocation but assumed a specific set of VNF requests that need to be allocated; this can include both average cases or instantaneous request dealt as a bunch.

In summary, the work in [HKL17] provides a solid foundation for allowing optimal and/or near-optimal general VM allocation in multiple clouds but does not take explicitly into account particularities of the VNF framework as mentioned above. By taking into account the work that has been conducted within 5G NORMA this framework can be adapted to be more 5G VNF specific. And, vice versa, some of the mathematical foundations found in [HKL17] can be utilized for the VNF allocation, routing and chaining problem that has been studied within the project, especially the issue of on-line operation which can be deemed as an issue worth studying since it relates to real-world operation.

## 3.3 SDMC-related procedures and interfaces

### 3.3.1 Summary on SDM-C/-X interfaces

Figure 2-7 depicts SDM-C, SDM-X and SDM-O together with the interfaces in between. There are four different types of interfaces:

- Interfaces between the SDM-O and both controllers (5G NORMA-SDMO-SDMC resp. 5G NORMA-SDMO-SDMX)
- Interface between both controllers (5G NORMA-SDMC-SDMX)
- Interfaces between both controllers and their applications (5G NORMA-SDMC-App resp. 5G NORMA SDMX-App)
- Interfaces between both controllers and NFs under their control (5G NORMA-SDMC-NF resp. 5G NORMA SDMX-NF)

**Interfaces between SDM-O and both controllers**

These two interfaces are used to inform the controllers about the orchestration decisions of the SDM-O and to convey the status of slices and the pool of common resources, resp. to the SDM-O. This status comprises the configuration of network slices, the resources allocated to them and the configuration of the NFs in the slices. This exchange is necessary when a slice is newly instantiated, reconfigured or deleted.

Furthermore, the controllers use this interface to indicate the need of a re-orchestration to the SDM-O. Usually, the SCM-C applications control NFs (e.g. by changing their parameters) such that the requested SLA targets are met with the available resources. However, when the SLA targets for a slice cannot be met due to a lack of transmission or processing resources, the SDM-C or SDM-X can send a re-orchestration request on the SDM-O via these interfaces.

**Interface between both controllers**

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

**Interfaces between Controllers and Applications**

These interfaces are used to enforce the conditions defined by the Control Applications that have to be realized for a given traffic identifier on dedicated (SDM-C) or common (SDM-X) functions and resources, in order to fulfil the targeted SLA. Via this interface an Application can convey to SDM-C/X the slice configuration, and in the reverse direction, information regarding the current slice performance can be reported to the corresponding Control Application.

**Interfaces between both controllers and the NFs under their control**

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

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

The properties of these interfaces could not yet be investigated in detail in 5G NORMA. In particular, it is open if there is a single kind of interface for all NFs, or if multiple kinds of interfaces for different types of NF are needed. The following examples show that the properties of interfaces between logic and agent can differ significantly:

- Routers in the transport network are the "classical SDN devices". Hence, for those the OpenFlow protocol would be suitable as well.
- Mobility Management: Aside the mobility management schemes standardised for LTE, other schemes like PMIP, VertFor, OFNC and LIME have been discussed in the literature. For some of them e.g. the OpenFlow protocol would be suitable for communication between the router logic and the forwarding agent that redirects data packets when mobile terminals move.
- In UMTS, the Radio Network Controller (RNC) decides on the configuration of logical channels and transport channels, and its decisions are communicated via the NBAP protocol [25.433] to the NodeB for execution. In 5G NORMA, a similar interface protocol could be used to convey SDM-X / SDM-C decisions on the QoS of traffic flows to the NFs processing these flows.

It remains to be seen if a single kind of interface is sufficient for all kinds of NFs, of if multiple kinds of interfaces have to be defined for different types of NFs.

## 3.3.2 SDMC-driven QoS/QoE Procedures

### 3.3.2.1 QoE-aware eICIC

This part focuses on the specifications of the implementation of the QoE-aware eICIC mechanism, that is presented in [5GN-D51] and studied in [5GN-D52], in 5G NORMA architecture.

The motivation of including QoE in the eICIC is to drive inter-cell interference coordination algorithms with user satisfaction indicator for benefiting eICIC from QoE intelligence and to meet users QoE requirements, cf. [5GN-D52]. However, driving inter-cell interference coordination functions with QoE implies three challenges for the mobile network:

- A deeply insertion of the QoE intelligence in the mobile network, in the radio layers of mobile edge cloud, to allow eICIC function a direct access to QoE awareness,
- A centralized control of QoE/QoS on the inter-cell interference management,
- A metadata sent by the QoE/QoS mapping component to QoE-aware eICIC mechanism, and a bilateral exchange between QoE-aware eICIC mechanism and the controlled network infrastructure.

5G NORMA can take advantages with SDN-based architecture to introduce the QoE intelligence in the radio layers, in a simple software-based way. The controllers SDM-C/X introduced makes a direct bridge between the QoE/QoS mapping component, the QoE-aware eICIC mechanism and the mobile radio Infrastructure. Furthermore, the SDN controllers enable to flexible acquire the QoE-related information from the network infrastructure, for Over-The-Top (OTT) on-demand multi-services [5GN-D32].

**QoE-aware eICIC framework**

The QoE-aware eICIC function is in charge to derive the optimal radio settings ABS and CIO that are used then by all the local schedulers of all eNodeB of HetNets for transmissions in pico or macro cells and for user attachment to pico cells. ABS fix silent periods of BSs deployed and CIO triggers the user switch in pico cells. eICIC connected to SDM-C/SDM-X operates in three main steps:

- Data collection step: SDM-C/SDM-X has the function of a central controller over the physical and virtual infrastructure. It takes in charge the collection of the required metrics to resend them to the optimizer. It collects also:
    - the QoE abstraction reports of new services requested by HetNet users that are derived in/sent by the QoE/QoS mapping module of QoS/QoE Assessment system (cf. Section 4.6.2.1 5 in [5GN-D32].
    - the new repartition of services (per zone, per cluster or per cell) once a modification occurs among the services requested by users (service repartition change or new service(s) request),
    - the updated radio capabilities or user distribution in the zone supervised by the controller which are transmitted by all eNBs of Physical or Virtual Mobile Edge Infrastructure (e.g., via X2 interface for physical structure), periodically or when a change occurs.
- Optimization step: eICIC derives the optimal values of (ABS, CIO) pairs using Game theory-based iterative algorithm, referred as Best Response, that maximizes a QoE-based utility. The QoE abstractions, the service/radio states and the deployment configuration of the mobile edge infrastructure of the controlled zone are involved in the maximal utility derivation.

- Distribution and Execution step: SDM-C/SDM-X sends periodically to the eNBs of P/V Mobile Edge Infrastructure the updated values of optimal radio settings (ABS and CIO) that are derived by eICIC. They are then used by the local schedulers for cell transmissions or muting and pico cell switch.

**QoE-aware eICIC application integration in 5G NORMA architecture**

Figure 3-4 and Figure 3-5 depict the integration of QoE-aware eICIC into 5G NORMA architecture at the NBI of SDM-C. Two integration schemes are proposed depending on the deployed RAN slicing option (or scenario of resource sharing between the slices) in the whole E2E network.

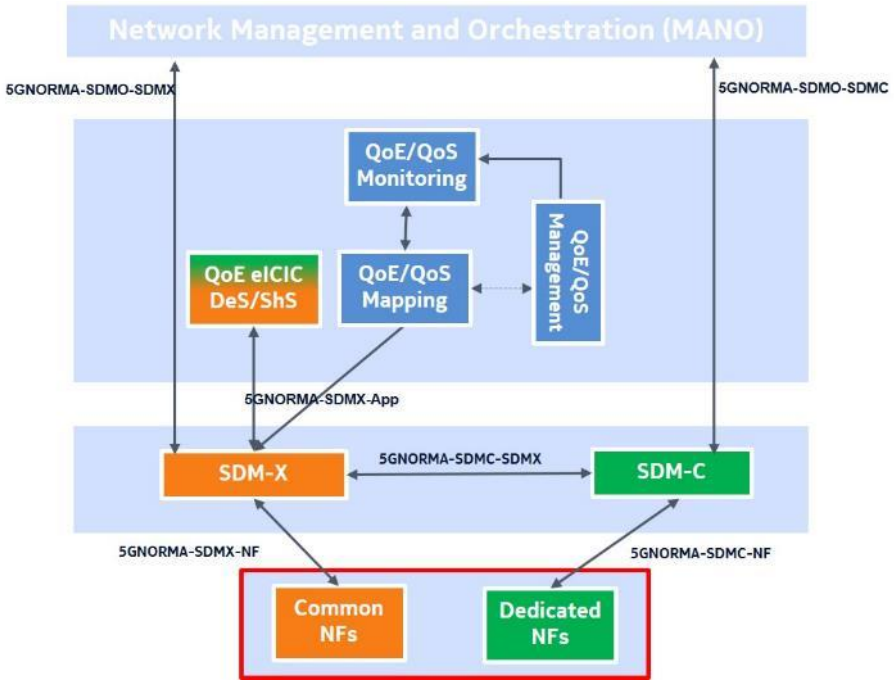



**Figure 3-4: QoE eICIC integration into 5G NORMA architecture with a unified application**

Different RAN slicing or slice multiplexing scenarios that correspond to different ways of resource sharing (NF and spectrum) between the slices are considered. The resource can be either completely isolated or shared. As pointed out in [5GN-D41], two main options "related" to the OSI protocol stack (cf. Figure 4-4 in [5GN-D32]) can be applied: standalone slice (own/dedicated NFs and spectrum), slice with shared resource (shared NFs and spectrum).

Depending on the multiplexing scenario of the deployed slice, ABS and CIO are used by the local MAC schedulers of macro and pico-cells hosting a standalone slice or a slice with shared radio resource. Consequently, the QoE-aware eICIC function shall operate in:

- An optimization mode compliant with Intra Slice control corresponding to the RAN slice scenario, standalone slice and slice with own spectrum, where the dedicated *MAC* NF are SDM-C-controlled (referred to QoE eICIC DeS algorithm) or,
- An optimization mode compliant with Inter Slice control corresponding to the RAN slice scenario, slice with shared resource where the common *MAC* NF are SDM-X-controlled (denoted to as QoE eICIC ShS algorithm).

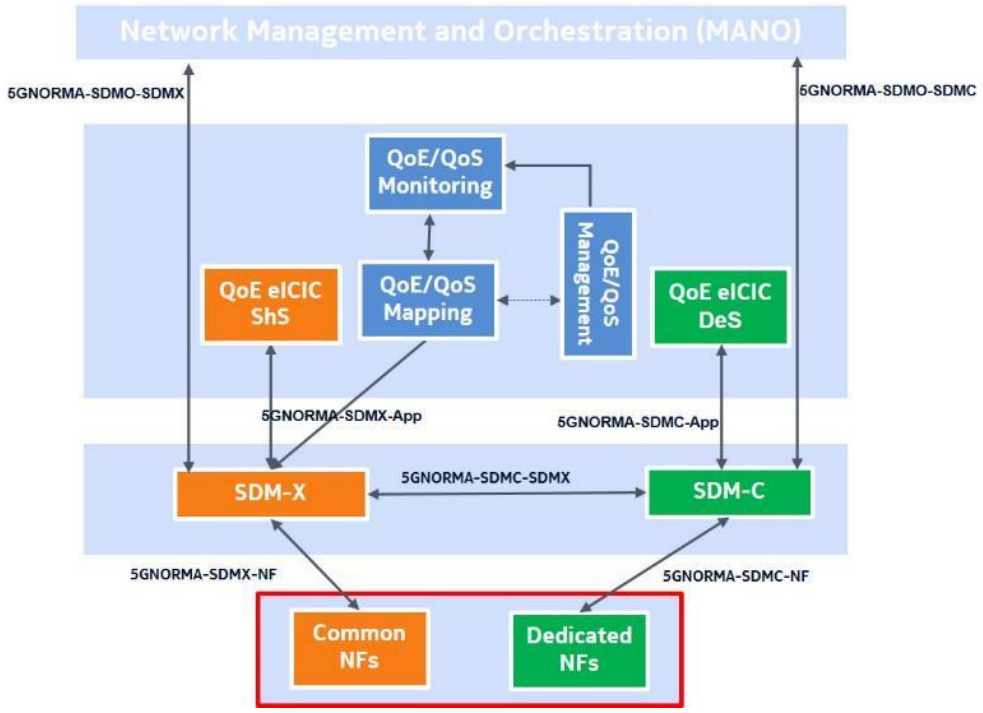



**Figure 3-5: QoE eICIC integration into 5G NORMA with two stand-alone applications**

The first architecture scheme integrates in one optimizer module the two optimization modes running at the top of SDM-C. It presents the advantage to achieve a direct bridge between the QoS/QoE components and the optimizer operating in one or other mode but a supplementary delay is added in the phases of collection and execution since SDM-X manages the shared network functions.

The second one presents two optimizer modules one running at the top of SDM-C and the other at the top of SDM-X. In this scheme, the additional delay disappears in the step of execution but remains in the step of collection of QoE abstractions.

**QoE abstraction generation process in QoS/QoE Execution Environment**

The QoE requirements are derived by the QoS/QoE Execution Environment (alias QoS/QoE Mapping) to the QoE-aware NBI applications. Before their transmission to the QoE-aware eICIC optimizer, they are converted into abstractions that are modelled by time-varying and content-dependent vectors of $\mathbf{q} = [\, q_1, q_2, q_3 \,]$. Then, the dependency model between the encoding rate $R$ and QoE depends upon the vector $\mathbf{q}$. Based on its accuracy, objectivity and low complexity, the popular video metric referred as SSIM is adopted to assess online QoE (namely user satisfaction) for video-based services of eMBB slices. The QoE-rate curve model then is written as follows:

$$\mathrm{SSIM} = f(R, q_1, q_2, q_3) = q_1 \cdot \log(q_2 \cdot R + q_3) \qquad \text{(Eq. 3-1)}$$

with $\mathbf{q} = [q_1, q_2, q_3]$ and the encoding rate $R$ is defined in the interval of interest $[R_{\min}; R_{\max}]$.

The generation of SSIM measurement points for different encoding rate values is done in the side of the content provider that owns the source videos. The derivation of vectors q is done in the side of the operator mobile network.

In case of HTTP Adaptive Streaming (HAS), the mobile users request directly the video services to HAS servers that are owned/managed by the content/service provider. They are assumed to be located at entry of the operator network. The HAS servers have the roles to store and to deliver both the streaming videos and associated manifest files, also named Media Presentation Descriptor (MPD). The derivation is processed in two steps.

**Step 1: Video encoding and *SSIM* computation in the content provider side**

Video encoding into chunks: The videos are encoded into different formats with multiple bit-rates and resolutions (named representation) and split into subsequent pieces of chunks (each of which of duration of 2-10 seconds). The HAS servers encode the video sequences at multiple bit-rates and, after video segmentation, generate a manifest file, also named media presentation descriptor (MPD). The latter contain the different available video representations to offer the possibility to the HAS client of users to select the most appropriate one according to its channel rate.

Pairs of (*R*, *SSIM*) computation and storage: A plurality of (*R*, *SSIM*) pairs are computed off-line. *SSIM* values are computed for a discrete set of encoding rate values *R* per each chunk and for overall the video content. Then, they are inserted in MPD files after video segmentation. The MPD files contain also the set of all actual discrete empirical pairs of (*R*, *SSIM*) per chunk (or segment) for all the video stored in HAS servers. The generation process is described in Figure 3-6.

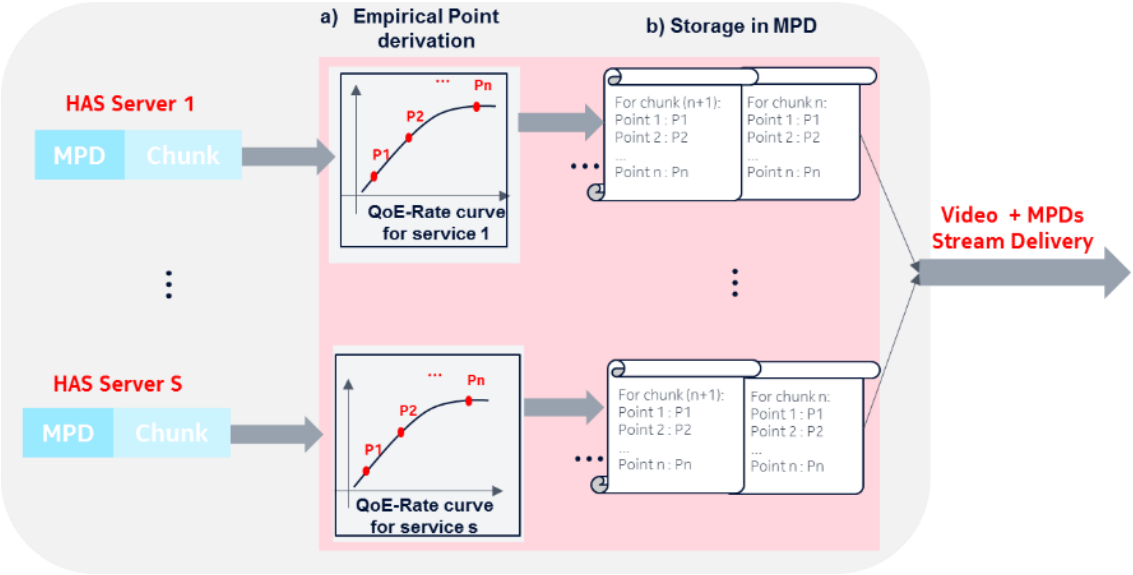



**Figure 3-6: Generation process of empirical pairs of encoding rate and *SSIM*[4]**

**Step 2: Computation of QoE abstractions (q vectors) in the network side**

When a user requests a new video content, namely a new segment encoded at the rate *R* selected by the HAS client of the user, the HAS server sends to the operator mobile network both the wished video data and the corresponding MPD file containing among others M empirical pairs of (*R*, *SSIM*) of video segment transmit.

A Media-Aware Network Element (MANE) close SDM-C/X is able to intercept and to process the MPD file in order to extract the synthetic quality information of the video segment. A quality-rate profile of M points is then extracted from the MDP file for each chunk. Applying the curve-fitting methods over the M actual discrete empirical pairs of (*R*, *SSIM*), the set of all admissible values of the vector $\mathbf{q} = [q_1, q_2, q_3]$ of the SSIM-Rate model are derived on-line in the module of QoS/QoE Execution Environment.

The QoS/QoE Execution Environment (alias QoS/QoE Mapping) designed in [5GN-D52] hosts the following components:

[4] MPD: The Media Presentation Description provides sufficient information for a DASH/HAS client for adaptive streaming of the content by downloading the media segments from a HTTP server.

- Input Adapter (IA): collects the pairs of (*R*,*SSIM*) extracted by the module MANE that represents Input Parameters,
- Mapping Function Block (MFB): implements the mapping function that represents a curve fitting method,
- Output Adapter: delivers the vector **q** composed of three real-valued numbers ($q_1$, $q_2$, $q_3$) for all chunks of videos requested by users,

with two main interfaces defined (Northbound and Southbound Interfaces). The architecture of the overall system is described in Figure 3-7.

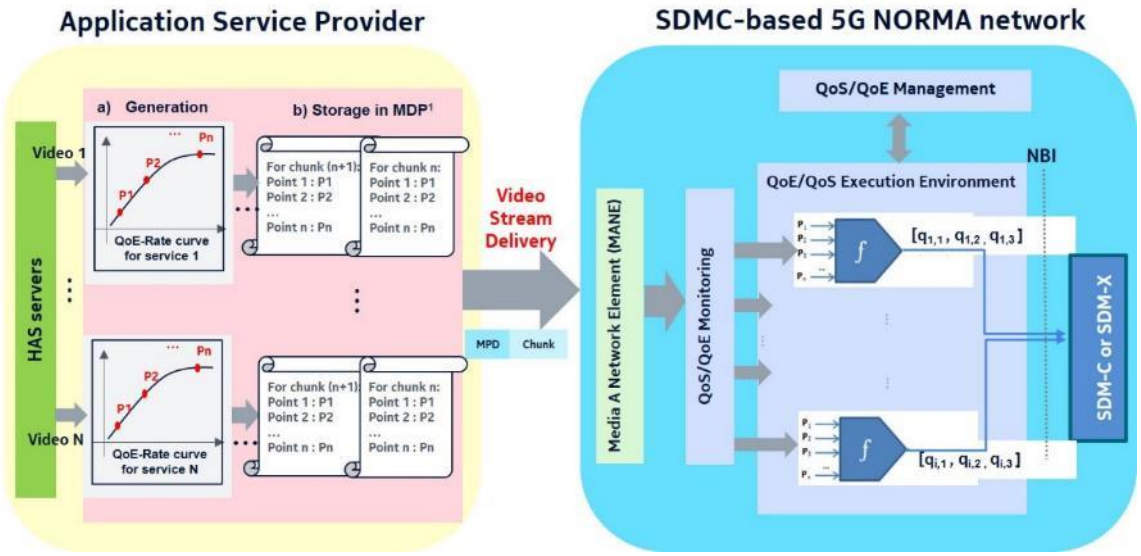



**Figure 3-7: QoS/QoE Execution Environment specification for QoE-eICIC**

In conclusion, knowing the set of all admissible values of q, the QoE-eICIC optimizer is then able to reproduce online the curve of rate-distortion of each video using (Eq. 3-1) and then to quantify and to track the evolution of the *SSIM*-based utility during the session. This way, the operator can exploit quality information for online and real-time inter-cell interference management during the data delivery.

**Required parameters and Interfaces**

To derive the optimal values of (ABS, CIO), the QoE-aware eICIC application requires the following features: radio capabilities, deployment scheme characteristics, service information and QoE abstractions.

Information exchanged between SDM-C/X and dedicated/shared infrastructure:

- Attached cells: eCG Id (E-UTRAN Cell Global Identification = Global eNB ID + E-UTRAN Cell Identity)
- Per cell: transmit power, path loss model, number of users (User IDs connected to cell or cluster), ABS and CIO (current and optimized values), ABS and CIO pattern, neighbour MC and SC: neighbour Id
- Per cell: no. of video flows active, video flow types

The controllers interact namely with QoE/QoS mapping application and the QoE eICIC application:

(1) Information exchanged between QoE/QoS module and SDM-C/X:
- Flow-based QoE abstractions for new service(s),
- Video flow type (s)

(2) Information exchanged between QoE eICIC and SDM-C/X:
- Attached cells: eCG Id (E-UTRAN Cell Global Identification = Global eNB ID + E-UTRAN Cell Identity)
- Per cell: transmit power, path loss model, number of users (User IDs connected to cell or cell cluster), ABS and CIO (current and optimized values), ABS and CIO pattern, neighbour MC and SC: neighbour Id
- Per cell: no. of video flows active, Video flow types
- Flow-based QoE abstractions for new service(s), video flow type (s)
- Dedicated or shared mode (Intra-slice or inter-slice mode)

Three service modules and two data bases are considered for the structure of SDM-C/X: a network topology manager (radio and service), a network statistics manager and a network configuration manager, a RAN inventory[5] and a Service inventory[6].

**Message sequence charts specification**

The message sequence chart (MSC) in Figure 3-8 describes an example of the message exchanges between the QoE eICIC component and the other involved components. A sequence of messages is triggered when a modification occurs in the service or radio (mobile edge cloud) state:

- Users requesting/deleting new service or,
- Modification of service repartition in a cell or,
- Change in number of users.

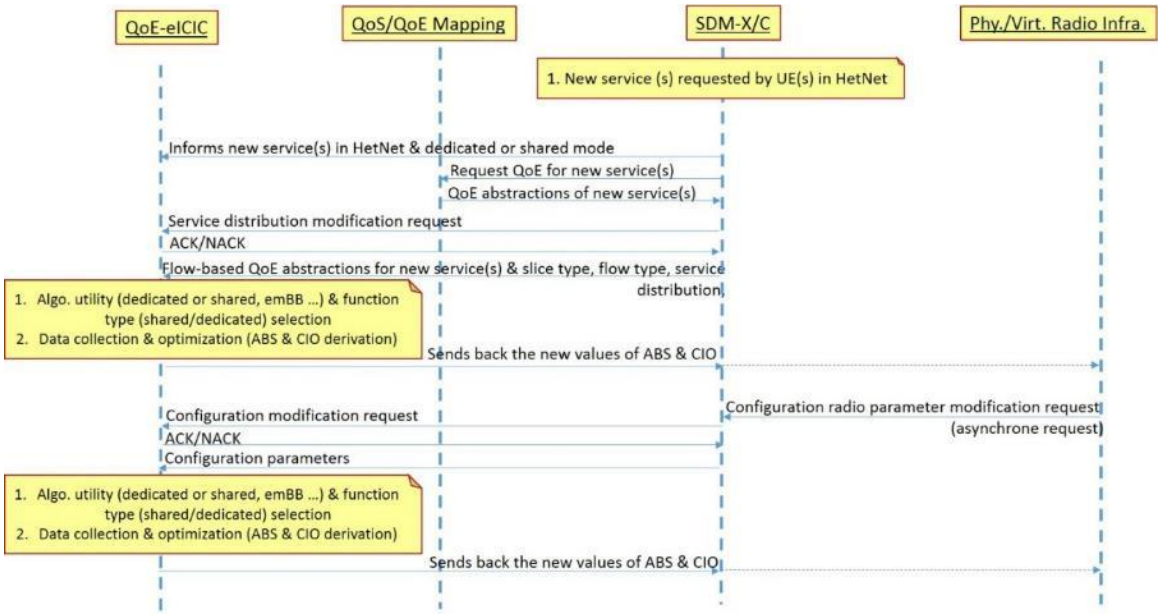



**Figure 3-8: Message sequence chart for QoE-aware eICIC**

[5] RAN inventory: network topology, neighbour information, an interference/user attachment information, configuration parameter (in a SDMC/X or external database) implementing an abstraction of the wireless network (neighbouring cell list, active User ID and radio features: transmit power, ABS and CIO current and updated, etc.)

[6] Service inventory: QoE abstractions, video repartition per cell (video IDs), user reparation per cell and user IDs.

#### 3.3.2.2 5G NORMA QoE-aware video pre-scheduling

In a generic view, SDM-C is composed by three layers: (i) the application and services; (ii) the controller functions and the network intelligence, and (iii) the elements for southbound communications.

The connection at the upper-level layers is based on northbound interfaces such as REST APIs, the most deployed one. On the lower-level part of SDM-C, protocol plugins interface the forwarding elements or the controlled VNFs. They provide a common interface for the upper layers, while allowing to use different protocol plugins (e.g., OpenFlow, OVSDB, SNMP, BGP, etc) to manage PNFs or VNFs. This is essential both for backward compatibility and heterogeneity, i.e., to allow multiple protocols and device management connectors.

The controller part of the SDM-C is characterized by a combination of the abstraction part and the network intelligence. The abstraction is used to implement the communications between the different plug-ins and the different network function modules (topology manager, switch manager, etc). The network intelligence hosts all the information like measurements, reporting, etc targeting to maintain a global view of the network. As a result, the network appears to the applications and policy engines as a single, logical switch.

In case of Video-prescheduling application? the SDM-C is interfaced to an OpenVirtualSwitch to be able to prioritize some video flows among others in one cell. The architecture and the algorithm are described in detail in [5GN-D52]. For clarity, this section reminda the main principle of the algorithm: it controls the amount of video data received by each client by giving more data to the client suffering from a bad channel. A client having a full playback buffer will not be impacted by the reception or not of a video segment. Under that perspective, designing the video-pre-scheduling application becomes a problem of managing the queues in an OVS between the Core and the RAN. For that, the interface between SDM-C and OVS should support:

- Create/Read/Update/Delete queues command from the applications running in the SDM-C northbound. The different queues will allow us to control the video bitrate received by each video client.
- Mapping video flows to queues using OpenFlow protocol.

Figure 3-9 depicts a synoptic scheme of the relationship between the application/SDM-C/ OVS.

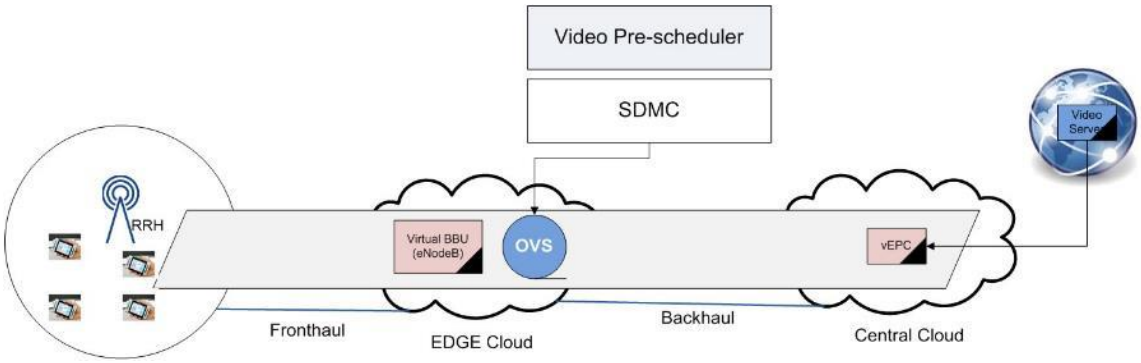



**Figure 3-9: Video pre-scheduling and SDM-C *interfaces***

Following Figure 3-9, the main components of the system are:

- **SDM-C:** The architecture of our solution is based on the SDM-C paradigm. The Open-Flow protocol supports the centralized monitoring of switches for different statistics that can be useful in the video prescheduling application. Secondly, SDM-C can add/modify forwarding rules; therefore, it provides a centralized point to decide routing of flows which is received as output from the pre-scheduler application to prevent queue buildups.
- **Ovswitch queues:** Using the northbound API exposed by the SDM-C, queues in the ovswitch are created and managed from the application layer.

OpenVswitch has three modules: OVS-DB server, OVS-Vswitch.d and OVS Kernel Module, cf. Figure 3-10:

- OpenVswitch is typically used inside hypervisors for packet switching between virtual machines.
- OVS-Kernel Module and OVS-Switch.d communicate via NetLink Protocol, OVS-DB Server and OVS-Vswitch.d communicate via JSON/RPC.
- OVSDB module in the SDM-C is responsible for backing up the status of OVS-DB Server time to time and also it can configure the OVSDB-Server from ODL. The configurations are related to creating virtual ports, bridges, tunnels, etc.
- OVS-Vswitch.d is managed by OpenFlow protocol plugin in the ODL via OF protocol. Main configurations are related to mapping virtual ports with openflow ports and packet forwarding logics. The status of OVS-Vswitch.d is backed up from time to time in the OVS-DB server module in the user space via JSON/RPC protocol.
- OVS-Kernel module is responsible for packet forwarding (Layer 2 look up) based on the simple cached table (simple tables to get rid of performance issues).
- If entry is not found in the kernel cache table it sends the packets to OVS-Vswitch.d module in the user space for decision and stores the decision in the kernel cache table for further forwarding of the packet related to the same flow. Also, if the entry is not configured in OVS-Vswitch.d it will be further forwarded to SDM-C.

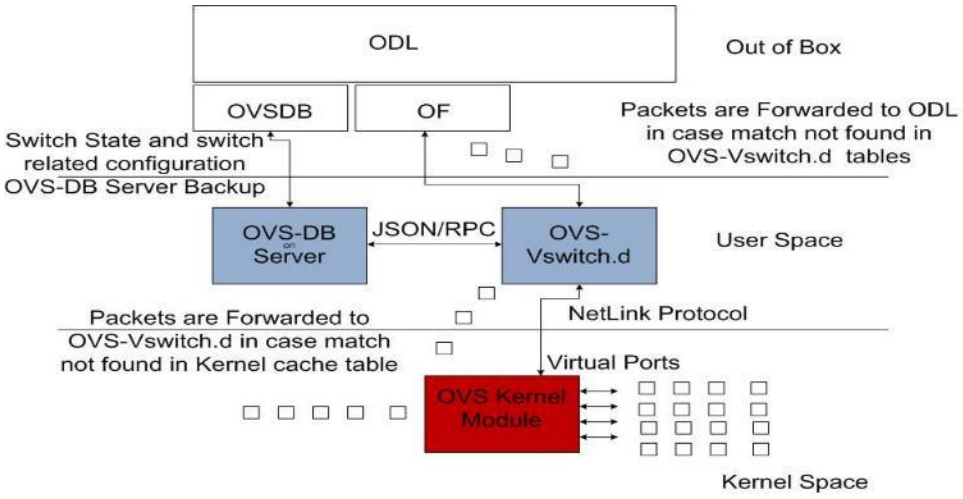



**Figure 3-10: OVS and SDM-C interfaces**

# 3.4 Charging control and lawful interception

The following two subsections elaborate how charging control and lawful interception are realized within the 5G NORMA system architecture.

## 3.4.1 Charging control in the 5G NORMA architecture

Charging control refers to the "process of associating packets, belonging to an aggregated set of packet flows, to a charging key and applying online charging and/or offline charging, as appropriate." A charging key contains "information used by the online and offline charging system for rating purposes" [23.203]. Roughly, online charging systems are usually applied for pre-paid subscription plans where a real-time balance of a subscriber's remaining credit (volume, time) is required, while offline charging systems apply for post-paid subscription plans. In mobile networks, charging is typically performed on the service data flow level or the application level (e.g., IMS). In either case, charging identifiers are required to unanimously identify bearers and packets as to apply the correct charging key.

Generally, the 5G NORMA system architecture shall support at least the following charging models: (1) volume-based charging, (2) time-based charging, (3) volume- and time-based charging, and (4) event-based charging. Further, both tenant-specific and service-specific charging must be enabled. Figure 3-11 depicts the charging control system within the 5G NORMA architecture. It consists of the following functions:

- Charging Control application
  This is an SDMC application located on top of a tenant-specific SDM-C that communicates the set of active charging control rules to the Traffic Reporting function.
- Traffic Reporting function
  Tenant-specific Data Layer function for monitoring and reporting traffic as well as enforcing charging control rules as selected by the Charging Control application.
- Charging Policies Management
  Tenant-specific Application Management function that configures and updates the charging control rule sets according to the service requirement, the subscription data, and the constraints introduced by the MSP or MNO. Updates on currently applicable rule sets are communicated to the Charging Control application (via 5GNORMA-SDMO-SDMC reference point).

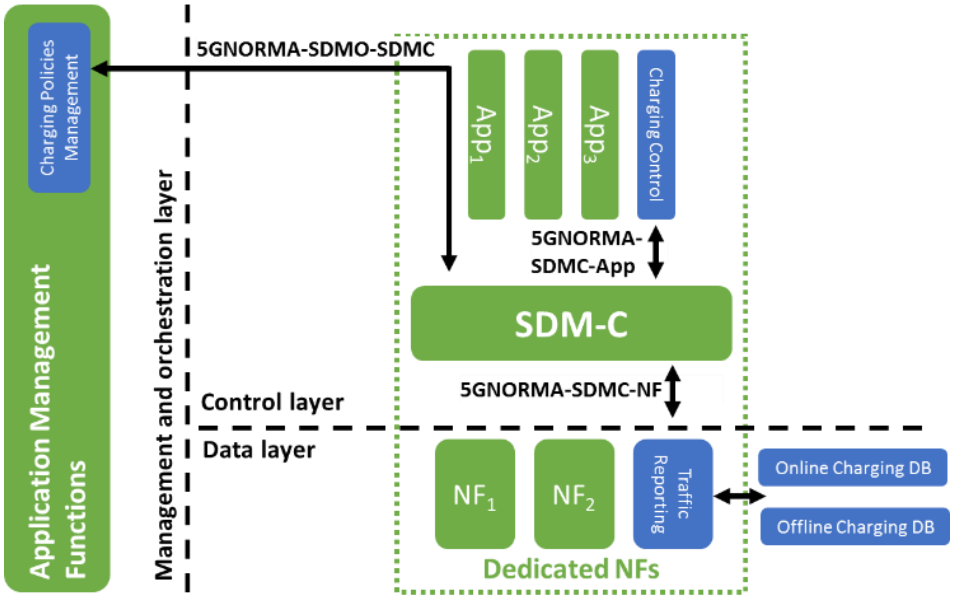



**Figure 3-11: Charging control system within the 5G NORMA architecture**

The Charging Control (CC) application provides network control for service data flow-based or application-based charging towards SDM-C via 5GNORMA-SDMC-App reference point. The SDM-C transforms this signalling to the protocol used by the associated Traffic Reporting (TR) function (via 5GNORMA-SDMC-NF reference point) By deciding on active charging control rules, the CC application decides how any identified service/application traffic is treated by TR, i.e., which charging control rules have to be applied per traffic identifier. Further, CC application updates CC rules for prepaid subscribers by getting according threshold event notifications (e.g., exhausted data volume) from the Online Charging System (OCS).

TR performs service data flow and/or application detection, which is based on identifying underlying network bearer, i.e., QoS class. TR can be located in the IP CAN (IP connectivity access network), e.g., within the SAE gateway function. Per network slice, there can be one or several TR functions. For a service data flow that is subject to charging control, the TR shall allow the service data flow to pass through data layer of the network slice if and only if there is a corresponding active CC rule. For online charging, the OCS has to authorize credit for the

requested charging key. TR reports the all measured service data flow- or application-based charging information to the according Online or Offline Charging Systems.

### 3.4.2 Lawful interception in the 5G NORMA architecture

Lawful Interception (LI) consists of a legally sanctioned interception process in which a service provider or network operator collects and provides law enforcement officials with the private communications of a certain target (either individuals or organizations). [ES 201 158]. Each country or territory will have specific regulations or laws dictating how this process should be done. Globally, these regulations usually follow the specifications laid out by the ETSI Technical Committee Lawful Interception [ETSI TC LI]. This committee works together with 3GPP, especially on the handover interface between IMS and the ETSI lawful interception framework. [33.107].

Lawful interception applies to two types of information: Intercept Related Information (IRI) and Content of Communication (CC). IRI means metadata associated with the target, including timestamps, location, and potentially signalling messages. CC means data exchanged between users, like a TCP/IP flow datagrams or emails.

For the 3GPP Evolved Packet System (EPS), the architecture for LI contains one main function (the Administration Function) that is responsible for managing the whole process. It has two main tasks:

- it mediates all communications between a law enforcement agency (LEA) and the network's internal components (MME, HSS, gateways), in a way that hides the LEA's identity from the components, and vice-versa;
- it manages all the intercepting elements within the architecture, informing them of the target and indicating them where and in which format should the data be sent

In 5G NORMA, each service can define its own LI rules, i.e., service -specific interception requests can be commissioned and according data can be provided to a LEA.

The 5G NORMA architecture will be extended to have following extra functions, cf. Figure 3-12:

- Lawful Interception Policies Management
  LEA-specific Application Management function responsible for configuring and updating interception rules based on the current laws and regulations, the intercepting request and any internal rules set by the MSP or MNO. This function uses the 5GNORMA-SDMO-SDMC reference point to interface with all the necessary Lawful Interception applications, since the request might cover multiple slices.
- Lawful Interception Control application
  An SDMC application that runs on top of a SDM-C which the interception covers. Its job is to manage all the Intercepting Element functions instantiated for this interception request, collect all data matching the request from these elements, and transmitting the data to the service layer.
  The LI application interacts with the SDM-C via 5GNORMA-SDMC-App reference point. The SDM-C then translates all given commands into signalling messages send to all relevant IE functions via the 5GNORMA-SDMC-NF reference point.
- Intercepting Element function
  A function responsible for monitoring and reporting traffic matching the interception request. There will be one for each other function relevant for the request. This means one IE function could monitor either another Control or Data layer function. Each network slice can have multiple IE functions. Data matching the request is sent to the SDM-C, which accordingly sends to the LI application.

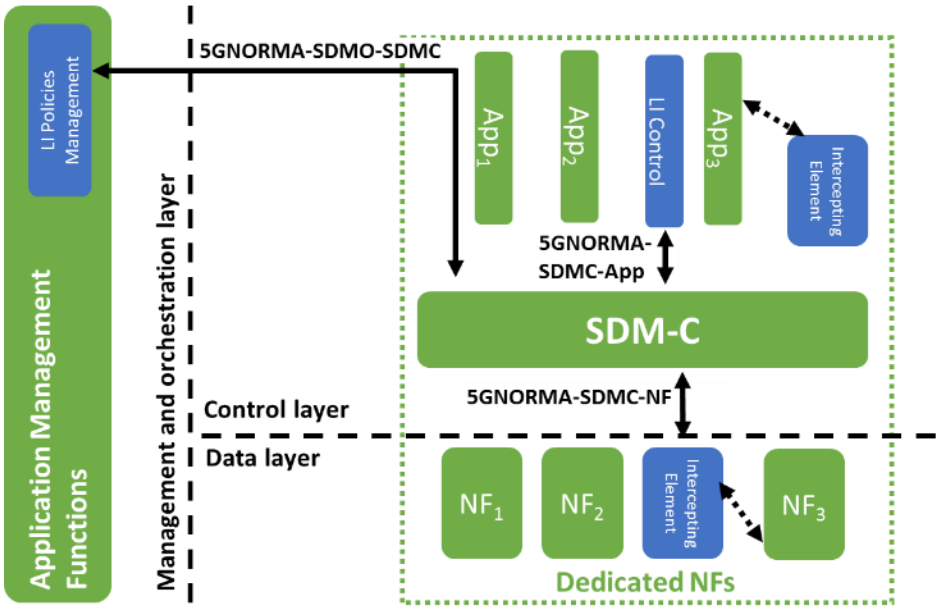



**Figure 3-12: Lawful interception control system within the 5G NORMA architecture**

# 3.5 Realization of 5G NORMA enablers

## 3.5.1 Service-specific network slice composition and customization

Service-specific network slice composition and function customization form two objectives that are realised by several functions and procedures throughout all layers of the 5G NORMA architecture, covering both slice preparation & instantiation as well as slice run time phases. Taking the example of a network slice instance for V2I services, the following subsections depict how several features from several project work packages interact to commission and operate such a specific slice.

**Phase 1: Pre-runtime network slice preparation & instantiation**

The V2I service provider submits most important V2I service parameters, in particular number of supported UEs, traffic profiles, coverage area, are provided to the Service Management function of the mobile network operator (MNO). Based on this input, Service Management selects the best-matching network slice template from the "Common Catalogues" set and extends it with request-specific deployment and configuration parameters, including

(1) Customization of the MANO layer stack: V2I-specific configuration of NFVO for conservative, failure-proof lifecycle management policies and operation of V2I application management functions in parallel to MNO 3GPP network management functions (cf. Figure 3-13),
(2) V2I-specific placement information for VNFs, in particular redundant placement of critical functionalities at the edge of the network,
(3) Selection of V2I-specific SDM-C instance(s) with customized SDMC applications, e.g., very reliable mobility management and QoE/QoS control with V2I prioritisation,
(4) Modifications of common RAN functions and SDM-X applications, changing the applicable policies in a way to reflect for increased reliability demands of V2I UEs (e.g., for scheduling and eICIC configurations),
(5) Selection and customization of dedicated RAN functions, and
(6) configuration of further V2I-specific features, such as, multi-connectivity support.

The resulting annotated V2I network slice template is provided to the Inter-slice Resource Broker (ISRB). Taking the current ("live") load situation in the overall network into account, the ISRB uses the VNF placement rules (cf. Section 3.2.2) further enhance the network slice template regarding the placement of NFs and the allocation of associated resources, including radio resources at different cell layers. At this stage, the network slice template is ready for instantiation and activation and is thus handed over to the NFVO and application management functions, which also coordinate and execute the operational tasks during slice runtime.

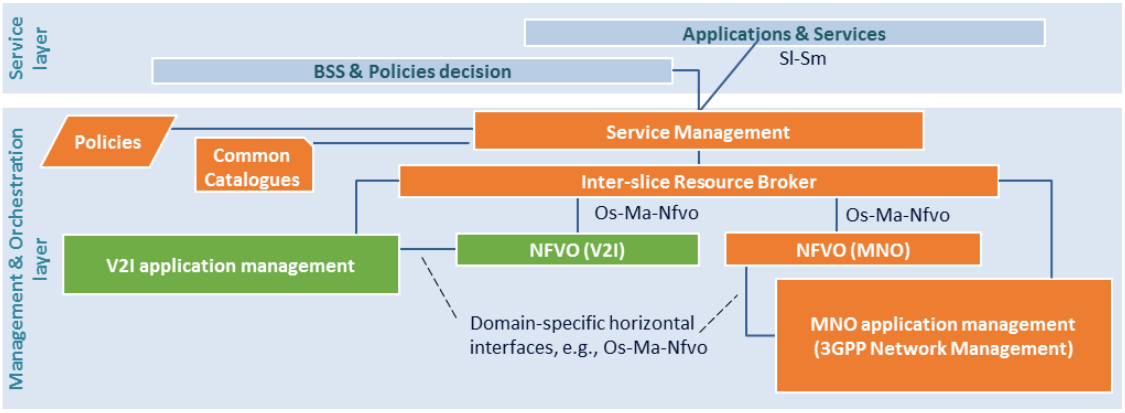



**Figure 3-13: Interaction between Service and MANO Layer**

Regarding the placement of functions (cf. item (2) of list above), Figure 7-4 describes the customized functional architecture and the deployment view of a V2I network slice, including multi-connectivity architecture (cf. item (6)). Specific adaptations for V2I include activation of *data duplication* mode as a special case of multi-connectivity, explicit operation of RLC in acknowledged mode (AM) and placement of RLC and PDCP functions in the edge cloud. The reader is referred to Section 7.2 for further details on function placement and configuration for a V2I slice.

**Phase 2: Runtime network slice adaptation and optimization**

For the RAN configuration (cf. item (5) of list above), Figure 3-14 shows a possible multi-service capable decomposed c/d-layer as proposed by 5G NORMA WP4. In this example, the access stratum supports two services, eMBB and V2I, where the latter can be extended with a V2V service. All services share a common carrier, i.e. share the mixed signal and analogue processing part (PHY TP). In a simplified implementation, a subset or even all services may employ the same subcarrier spacing and symbol length (assuming an OFDM-based system). This allows the lowest part of PHY Cell, namely the (i)FFT and CP insertion/removal, respectively the sub-band filtering for filtered waveforms like UF-OFDM, to be shared in addition to PHY TP. The upper part of PHY Cell and PHY User differs for each service, creating an optimized implementation for V2I. V2I-specific PHY User implementations are optimised for lowest latency at the cost of spectral efficiency, utilising very short TTIs and accordingly higher DM-RS overhead and an RRM able to pre-empt other services. Moreover, Figure 3-14 depicts joint RRC User for eMBB and V2I assuming V2I UEs are always also eMBB-capable and therefore reception/transmission of RRC messages is done via eMBB instead of via V2I. Further, it assumes that V2V UEs are always downlink (DL)-capable as well and can therefore receive RRC messages via DL. For RAN slicing Option 2 as depicted here, the data layer implementation of the access stratum becomes slice-specific at RLC and above. NF instances differ but NF types may be common among services (for further details, the reader is referred to Section 3.2 in [5GN-D42]).

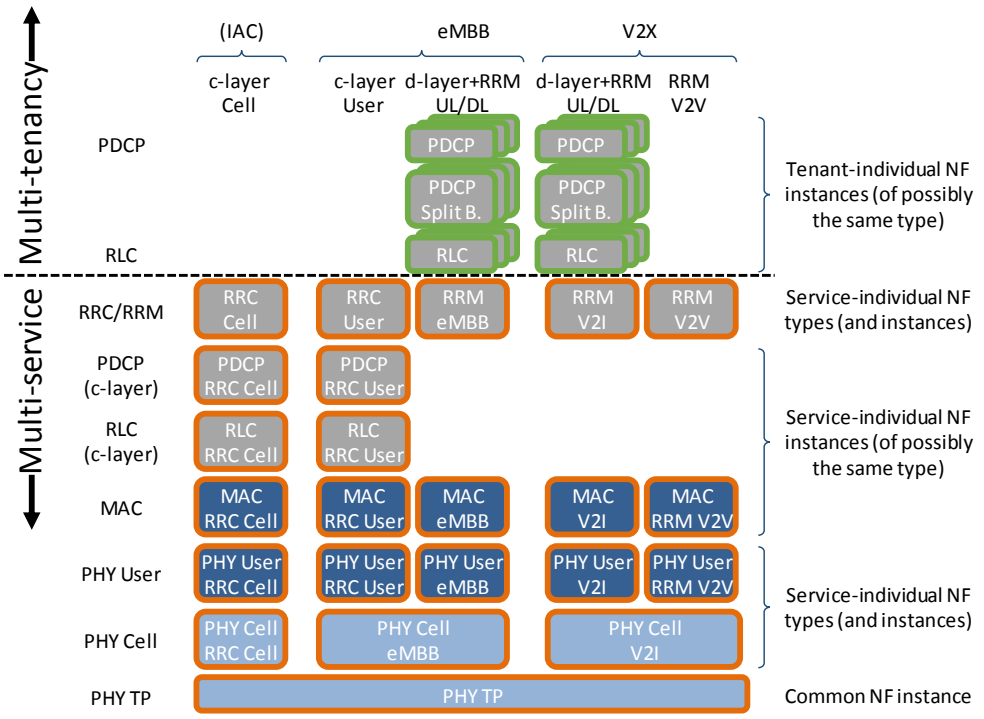



**Figure 3-14: Example 5G New Radio multi-service RAT for RAN slicing Option 2, adapted from [5GN-D4.2]**

The common part of the access stratum needs to maintain the one-to-one mapping of traffic to individual slices. Via the 5GNORMA-SDMC-SDMX interface, SDM-X provides each slice the means to influence how traffic of their own slice should be processed, cf. item (4) of list above.

For this purpose, 5G service flows (SF) and sub-service flows (sSF) were introduced in [5GN-D41] , which provides a more fine-grained QoS control, i.e., in-bearer/in-flow differentiation of QoE/QoS handling. Since the 5G NORMA architecture enables flexible placement of NFs, QoE/QoS enforcement functions (points), implemented as VNF, may have different deployment options as illustrated in Figure 3-15. A deployment at both edge and central cloud is preferred since the edge cloud enforcement points may mainly enforce the radio resource-related QoE/QoS and central cloud enforcement points focus on transport resources-related QoE/QoS.

Reusing the 5G SF and sSF concept, the sSF establishment and update logic, running as an application on top of the SDM-X and SDM-C, repsectively, is executed by the QoE/QoS enforcement point. Based on monitored data layer traffic and the applicable policies, it works as follows:

- Unknown/unidentified traffic: no sSF establishment is needed (default service).
- Identified but deliberately unmanaged V2I traffic (e.g., traffic that falls under the best effort category): in case unknown and unmanaged traffic needs to be differentiated, a separate sSF that handles all unmanaged V2I traffic can be established. If an V2I sSF already exists, new V2I traffic is mapped into the existing V2I sSF.
- Managed V2I traffic: establish a dedicated V2I sSF for V2I traffic according to the SDM-X policies updated by the Service Management function before activation of the V2I slice, i.e., the service parameters of the sSF are derived from the QoS/QoE requirements of the corresponding traffic. The QoE/QoS enforcement point also detects when the sSF service parameters need to be changed (e.g., due to change in the QoS/QoE requirements of the application sessions) and thus initiates an sSF modification.
- The QoE/QoS enforcement point terminates an V2I sSF in case V2I traffic is terminated.

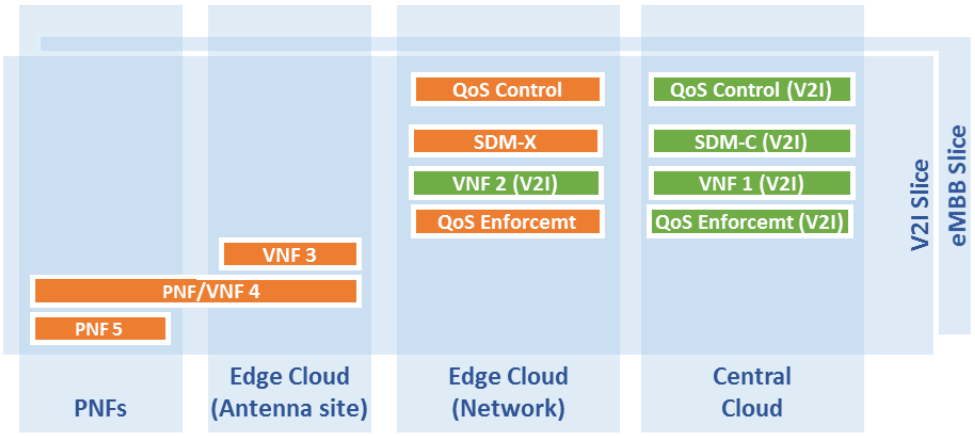



**Figure 3-15: V2I-specific QoE/QoS control and enforcement (adapted from [5GN-D5.2])**

SDM-X-controlled utilization of shared radio resources is a further 5G NORMA feature to realize V2I-specific behaviour in the RAN, e.g., in order to rectify temporary appearing critical interference constellations in local parts of the mobile network. Based on the feedback of the V2I-specific QoS/QoE monitoring (e.g., threshold violations), SDM-X applications influence synchronous inter-cell interference coordination (ICIC) schemes by dynamically adapting BS clusters (e.g. for Joint Transmission (JT), Coordinated Multipoint (CoMP) or coordinated beamforming to further increase transmission reliability) or adapt and de/-activate asynchronous ICIC schemes (e.g. frequency reuse schemes, Carrier Aggregation (CA) based ICIC, enhanced (e)ICIC), or change the applied scheduling scheme. These options can be repeated several times to further improve the SLA of a "threatened" V2I slice. For example, for eICIC, 5G NORMA WP5 introduced a novel optimization method by integrating a direct QoE measurement into the computation of the V2I-aware utility. Controlling inter-cell interference via QoE allows limiting the interference experienced by the mobile users in cell edge while improving the QoE over the network. More specifically, the QoE-driven eICIC function has been designed as a controller application running on the northbound interface of SDM-X. It derives the optimal radio settings for Almost Blank Subframe (ABS) and Cell Individual Offset (CIO). ABS and CIO are used by the local MAC schedulers of macro and pico cells hosting V2I slice(s) that share available radio resources with other slices. In this setup, the V2I-aware eICIC function operates in an inter-slice control mode corresponding to the RAN slicing scenario depicted in Figure 3-14.

As a further V2I-specific customization feature, the UCA (user centric connection area) concept [5GN-D4.2] as developed in WP4 has been amended with service-specific enhancements in WP5. While the initial design target included signalling reduction during small and sporadic data transmission, WP5 utilized UE mobility characteristics such as speed, trajectory along with anticipatory information on user movement to flexibly adjust the size of a UCA. Furthermore, the number of users in the UCA as well as current load on anchor nodes and backhaul influence the decision on UCA size. In the case of V2I services, the rather strict requirements in terms of e.g. latency and reliability result in a rather small UCA size when compared to, for example, mMTC traffic.

The Mobility Management App on top of the V2I-SDM-C (cf. item (3) of list above) has insight on the V2I traffic and UE characteristics and supports the definition of a suitable UCA size based on, among others, QoS requirements, acceptable level of service degradation, maximum delay, number of users served by anchor node, current load of a backhaul, UE mobility pattern and any available data on mobility prediction. The UCA size recommendation for V2I terminals is then signalled towards the SON function which finally implements the UCA cluster in the RAN as described in [5GN-D41].

## 3.5.2 Multi-tenant network control and resource allocation

Enabling new business models through the availability of fully customizable network slices that are bound to a specific service provided to one tenant is one of the fundamental requirements for future 5G networks. This capability, also known as multi tenancy, entails several research challenges that need i) a specific architecture that allows for the multi-slicing network control and orchestration and ii) specific solutions that change the current (single sliced) NF behaviour to a multi-sliced one.

This section addresses those research challenges and describes how the 5G NORMA project solves the issue of multi-tenancy through its end to end architecture. They are grouped according to the problems tackled: network orchestration or control.

### 3.5.2.1 Network orchestration and resource allocation

The multi-tenant network orchestration and management of several network slices running on the same infrastructure require three fundamental operations: the network slice admission control, their blueprinting/onboarding and, finally, the dynamic resource sharing across slices.

- As further described in Section 4.1, network slicing opens the mobile network ecosystem to new players:
- Infrastructure Provider (InP), which operates the infrastructure;
- Mobile Service Provider (MSP), which offers the (mobile) telecommunication service (realized by a network slice); a Mobile Network Operator (MNO) can be considered to combine the roles of InP and MSP; and
- tenants, which acquire a network slice from the MSP to deliver a specific application-level service to own subscribers.

In this new ecosystem, MSPs issue to the InP requests for spectrum and computational resources (IaaS) in order to set up their slices, which are finally used by subscribers of the tenant. Since spectrum is a scarce resource, for which overprovisioning is not possible and its availability heavily depends on SLAs and users' mobility, the InP cannot apply an "always accept" strategy for all the incoming requests from MSPs. In the same way, MSPs cannot serve all incoming requests from tenants. Thus, the new 5G ecosystem calls for novel algorithms and solutions for the allocation of network resources among different tenants.

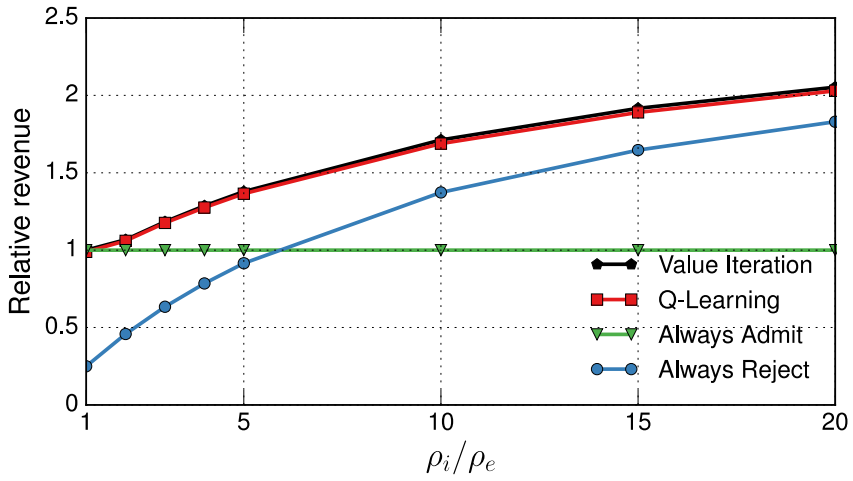



**Figure 3-16: Relative revenue according to the network slice price**

The 5G NORMA partners designed a network capacity brokering algorithm executed by the MSP in order to decide whether to accept/reject a request from a tenant with the goal of maximizing the MSP revenue, satisfying the service guarantees required. This module, that runs in the ISRB

block, adopts a reinforcement learning based technique to perform admission control. The current approach focuses on spectrum, but also other kind of resources may be managed in the same way.

The full algorithm is specified in [5GN-D42] and briefly summarized here for the sake of completeness. The approach, based on Q-learning can select the decision that maximizes the revenues associated to accepting or not a new network slice (modelled as elastic/inelastic) requests.

Figure 3-16 shows the relative trends (compared with the "always admit" policy) of the proposed algorithm, emphasizing its advantages with respect to static policies. The figure shows how, depending on the current price of slices (in this case, elastic or inelastic traffic slices $\rho_e$ and $\rho_i$), these solutions can significantly be suboptimal.

Another important challenge in multi-tenancy is the definition of a sharing criterion and the design of an algorithm that follows it to enable statistical multiplexing of spatio-temporal traffic loads. Again, a naïve solution like Static Slicing (SS) may still be a baseline solution (as "always admit" in the admission control). The idea is to design a criterion that maximises the network utility while it

- allocates resources fairly among operators;
- takes into account the number and the location of the active users of each operator

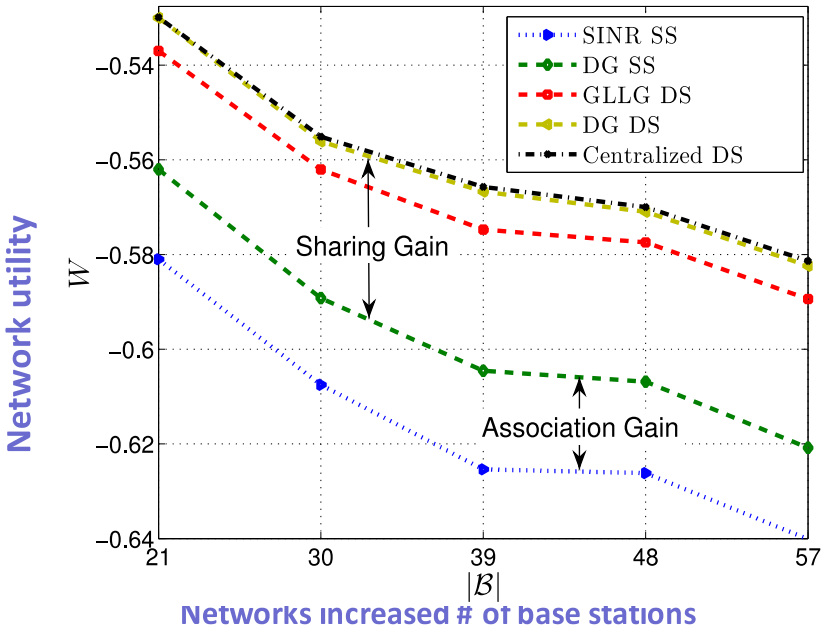



**Figure 3-17: Utility gains for different approaches as a function of network size**

The information involved include channel capacity, the number of users in the network, their spatial distributions and thus their mobility. The reader is referred to [5GN-D4.2] for more details about the detailed algorithm and how this component runs as part of the ISRB module. A teaser of the results obtained in Figure 3-17 is included here. There, it can be seen how a centralized approach, such as the one that will be running in the ISRB, can obtain a higher network utility due to the fact that a more efficient resource allocation is performed. This is even more flagrant when compared with Static Sharing (SS) approaches; Dynamic Sharing approaches (DS), indeed perform better up to the centralised approach, which provides the best results.

As a joint effort with WP5, WP3 designed a multi network sliced MANO stack that implements the Virtualised Network Platform as a Service (VNPaaS) concept, that is the capability for each tenant of running its own MANO stack. This stack, described in Section 2.4.2, includes several modules that have been designed for especially managing multiple slices running on the same infrastructure. For the sake of simplicity, this full ecosystem is named software-defined mobile network orchestration (SDM-O), the module that is in charge of the resource orchestration of a

network slice. One of the responsibility of SDM-O is to map the slice templates representing the slice requirements along with the corresponding tenants' SLAs to 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 SDM-O.

According to the slice requirements, different network slices are placed in different locations of the network: those with more stringent latency requirements are likely to be placed at the network edge, while the less demanding ones are in the central cloud, so as to maximize the multiplexing gain. Also, the decision of which functions will be shared across slices is taken by the ISRB, that finally instructs the SDM-C and SDM-X about this decision. A full specification of the Network Slice onboarding procedure is available in [5GN-D52].

As mentioned in Section 2.4.2, the SDM-O deploys a specific NFVO component for each slice in order to implement the specific intra-slice orchestration tasks. This NFVO block is the same as the one defined in the ETSI MANO framework [MANO], and communicates with the external OSS/BSS systems using the *Os-Ma-Nfvo* reference point through the Inter-Slice Resource Broker and the Service Management function (cf. Figure 2-19).

Hence, 5G NORMA delegates the intra-slice network orchestration tasks to the ETSI MANO NFVO component. As the ETSI NFV MANO framework defines, the NFVO is in charge of the network orchestration and management of NFV resources (infrastructure and software) [ETSI_GS_NFV], but in this case, applied to each single slice. Basically, the NFVO operates, manages and automates the NFVI associated to its own slice, helping to deploy, manage and configure the different NFV service topologies on that NFV Infrastructure, having control and visibility of all VNFs running inside the NFVI quota assigned to that slice.

Since services are defined per slice, the service orchestration is resolved at this intra-slice level; i.e., the intra-slice orchestration mechanism is the responsible to transform the slice specific SLA parameters (received from the IS-RB for each slice) into the corresponding network service. But in this case, the orchestration process is not as static as the inter-slice orchestration process described in Section 2.4.2; on the contrary, quick and dynamic adaptation of the service to different load situations or to fulfil the agreed QoE/QoS requirements is needed. For this, besides the regular life-cycle management and control mechanisms, the application of actions such as scaling the different VNFs, re-route traffic among them or updating the (V)NF-FG in a rapid and adaptive manner is expected to be performed. So, main functions performed at this orchestration level are focused to adapt the service to the current system and load situations; this can be performed in the following ways:

- Scaling up/down/in/out individual NFs;
- Restructuring an NS graph;
- Modifying the placement of functions
- Rerouting traffic to or between different instances of such functions,

The trigger to perform these re-orchestration actions comes from the 5G NORMA control blocks (SDM-C and SDM-X), and is forwarded to the associated NFVO in the corresponding slice or, in case of SDM-X, to the ISRB.

Low level resource orchestration (i.e., connectivity, compute, and storage resources) across multiple network functions is performed at this level also. For this case, the NFVO delegates on the VNFM and VIM as it is defined in the ETSI NFV MANO framework.

The NFVO is accessed from the Domain Specific Application Management for each slice through the *Os-Ma-Nfvo* reference point. This specific Application Management is available for each tenant through the ISRB in the SDMO, so each tenant gets access to selected management functions to manage their slices from their corresponding OSS/BSS.

Operations available to such OSS/BSS systems are those already defined for the *Os-Ma-Nfvo* reference point [NFV-MAN001], that is:

- NS Descriptor and VNF packages management.
- NS instances lifecycle management (instantiation, termination, scaling, query and update).
- VNF lifecycle management.
- Policy management and/or enforcement for Network Service instances, VNF instances and NFVI resources.
- Querying relevant Network Service instance and VNF instance information.
- Forwarding of relevant events, accounting and usage records and performance measurement results.
- Shared network control and orchestration for multi-tenant purposes are intertwined operations that build the 5G NORMA multi-tenant environment. The results presented above just hint the advantages of how such algorithms may improve the network performance by optimizing them. For example, the slice admission control described in [5GN-D41] just consider two kinds of slices (elastic and inelastic traffic). A fine-grained solution, may improve even more the performance, by taking into account, e.g., the number of users that may be served using the virtual cell strategy described next.

#### 3.5.2.2 Virtual cell and multi-tenancy

The proposed new architecture for 5G network in the framework of this project introduces the concept of software defined networking (i.e., SDMC) in mobile network in addition to many other architectural innovations to offer flexible service-aware architecture. As the results, new set of innovations such as centralised radio resource management can be considered. On-demand formation of the virtual cells is one of the centralised radio resource management techniques studied in [5GN-D4.2]. In this section, after briefly describing the concept of virtual cells, the procedure of forming virtual cells for a multi-tenant environment is addressed.

Formation of virtual cells originally presented in [SSP+14] and extended in [CSS+16] for TD-LTE. Using bearer split, the terminals in the coverage of two neighbour cells are allowed to use the available radio resources in both cells. The terminals serving from multiple cells form a logical cell, which has different TDD-pattern from the primary cells. This logical cell is referred as a virtual cell. The key architectural enablers, as it is shown in Figure 3-18, are the SDM-C for slice-specific functions and SDM-X for common functions, which enables controlling and coordinating multiple base stations.

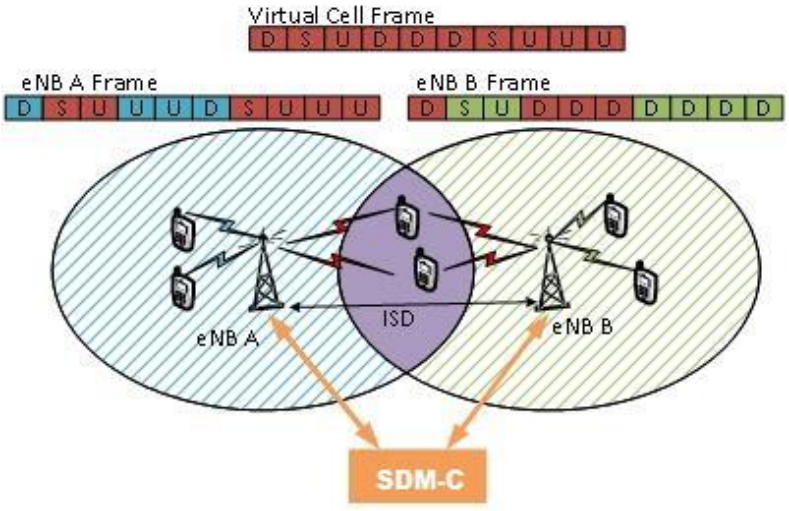



**Figure 3-18: Simple illustration of the virtual cell concept (extracted from [5GN-D42]).**

The procedure of formation of virtual cells has been extended in the scope of WP4 by means of common MAC (as well as common PDCP) multi-connectivity, i.e., each cell has its own PHY

layer coordinated for radio resource management by a common MAC scheduler [5GN-D41]. In addition, the new algorithm improves the performance of formation of virtual cells by introducing the edge cell threshold. This threshold is the difference between the RSRP (Reference Signal Received Power) and it has been used to define the size of virtual cells. Finally, a Q-learning approach to achieve near optimal edge cell threshold and form the virtual cell has been described in [5GN-D42].

In the following, the procedure for formation of virtual cells using the aforementioned approach in a multi-tenant environment with different network slicing options is described. The primary challenge for resource management in a multi-tenant environment is meeting the different QoS requirements and SLAs of the tenants. SDM-C and SDM-X configure the radio resource pool and allocate the proper amount of resources to each of the tenants/slices. However, the centralised radio resource management techniques can improve the resource usage efficiency in addition to the network flexibility. For instance, the TDD-pattern selection can be done based on the latency requirements and by forming the virtual cells improve the total network throughput. The effect of dynamic TDD patterns on the improvement of throughput and latency has been studied in [PLS15].

The three slicing options are presented in Figure 2-10. The control and coordination of multiple base stations required for formation of virtual cells are enabled by SDM-C (for RAN slicing Option 1) and SDM-X (for option 2 and 3). For TDD mode, the selection and reconfiguration of the pattern for the cells or cluster should be done based on the requirement of all the slices sharing the PHY layer.

**Option 1:**

In this case, each slice has its own frequency band, on which the TDD-patterns can be configured independent of the other slices as it has been addressed in [CSS+16]. The formation of virtual cell for each of the slice can be done simply based on the requirement of the relative slice. In this case, balancing the radio resource usage and traffic demands among the cells are the key achievements. The comprehensive description and the numeric results are presented in [5GN-D4.2].

**Options 2 and 3:**

In these two slicing options, multiple slices have to share the same radio resource pools. The reconfiguration of TDD patterns cannot be done as easy as the option 1. The slice sharing the radio resource pool may not have the same requirement. Hence, the (re)-configuration of the radio resource pool (i.e., the TDD-patterns) parameters has to be done considering the requirement of all the slices sharing it. However, the formation of virtual cells can increase the flexibility of the radio resource pool and reduce the signalling messages for reconfiguring the TDD-patterns of cells.

In the first step, the SLA requirements (the throughput, latency, and priority [KHC14]) are translated to a set of policies for SDM-C in charge of forming the virtual cells. SDM-C may change the patterns of the cells to meet the throughput as well as latency requirements [LGL+13] of the related slices. SDM-C chooses the pattern for the cells to meet the latency requirements (i.e., selects pattern with a more frequent uplink subframe to reduce the delay in uplink) and compensate the throughput requirements by means of forming virtual cells using the proposed approach. In better words, the selection of the patterns with more frequent uplink sub-frame improves the latency performance but it may reduce the throughput in downlink. Hence, the formation of virtual cell can improve the throughput by using the unallocated resources in neighbour cells.

The objective function (Q-learning the reward function) should be changed to the weighted throughput of the network given as:

$$R^{T}_{b[\text{Mbps}]} = \sum_{i=0}^{N_s} w_i R_{b\,i[\text{Mbps}]} \qquad \text{(Eq. 3-2)}$$

where:

- $R_b^T$: The weighted total network throughput,
- $w_i$: The slice serving weight selected based on the SLAs and $w_i \in [0,1]$,
- $R_{b_i}$: The throughput of the slice.

The weights in Eq. 3-2 enable the SDM-X/C to prioritise the network slices sharing the same radio resource pool.

# 4 The 5G NORMA Ecosystem

The purpose of this chapter is to analyse the 5G NORMA ecosystem in practical scenarios. Section 4.1 first revisits how the relationships between the various stakeholders sustain the realization of "Network Slice-As-A-service". The different offer types provided by the mobile service provider to tenants are summarized from [5GN-D32]. These offer types have been examined in [5GN-D32] in terms of potential impact on components and interfaces and security considerations. That analysis is now extended by applying offer types to two practical scenarios. In the first scenario, Section 4.2 investigates dynamic on-demand slices used by the MSP to monetise OTT traffic, where the MSP's offer deals with guarantying every slice resource in accordance with each tenant's SLA. In a second scenario, Section 4.3 considers the case of industrial communication slices where security is a high concern and where tenants want to be less dependent on the trustworthiness of an MSP or InP. Finally, Section 4.4 investigates the security implications to provide full slice isolation for Industry 4.0 verticals.

## 4.1 Stakeholders and offer types

The 5G NORMA framework allows "Network Slice-As-A-service" to become a new business-to-business opportunity between the ecosystem stakeholders. The following relationship between the various stakeholders sustain the realization of "Network Slice-As-A-service":

The 5G NORMA **infrastructure providers** (**InPs**) own and manage parts of or all infrastructure of the network. It can be further distinguished between RAN infrastructure provider and datacentre or cloud infrastructure provider. The former owns the physical infrastructure such as the antenna sites and the HW equipment for the antenna. The latter owns and manages local and central datacentres. It provides virtual resources such as virtual computing, storage and networking by deploying a virtualization environment to logically abstract the physical infrastructure.

The **Mobile Service Provider (MSP)** provides various telecommunications services to end users (subscribers). Furthermore, the MSP sells dedicated mobile network instances ("network slices as a service"), each realizing a specified telecommunication service (e.g. mMTC), to tenants. The MSP leases all needed physical and virtual resources from one or multiple InPs to deploy the end-to-end mobile network. The **Mobile Network Operator (MNO)** owns and operates the physical and virtual network functions and the communications links to realize a mobile network and provides mobile connectivity towards mobile end-users.

The **tenant** is usually a business entity that rents and leverages a 5G NORMA network slice provided by the MNO. It can be a Mobile Virtual Network Operator (MVNO), an enterprise (e.g. from a vertical industry) or other organisation that requires a telecommunications service for their business operations. The tenant can rent a 5G NORMA slice with different options for managing and controlling it. Different options of offer provided by MSPs to tenants have been described into D3.2 and are summarized in the following:

- **Offer type 1 (No control):** MSP operates slice and provides communication services on behalf of the tenant
  Here, a tenant requests the commissioning of a network slice by providing the high-level requirements of the telecommunication service to be provided. Operation of the network slice is completely handled by the MSP (or MNO), the tenant only receives coarse-grained performance reports. This allows for a flexible Network Slice market that is ultimately enabled by the algorithms running in the Inter-Slice Resource Broker (ISRB).
- **Offer type 2 (Limited control):** MSP allows for limited slice configuration and control options for the tenant
  Here, in addition to offer type 1, a tenant can specify more fine-grained configuration options for the requested network slice. Moreover, selected network operations (e.g.,

subscriber data management, QoS control) are performed by the tenant. Still, the major part of network operation is handled by the MSP or MNO. The network slice can integrate a set of tenant-owned functions customised/certified for its needs. Nevertheless, the interaction of onboarded functions with the MSP's systems is strictly monitored and controlled by MSP entities.

- **Offer Type 3 (Extended to full control**): MSP allows extended slice configuration and control options for the tenant
  In addition to offer type 2, the tenant has a rather wide control over deployed network functions. This can go as far as the tenant onboarding own network functions for selected areas, e.g., mobility or session management, contributing own infrastructure, and operating a part of the network slice independent of the MSP (or MNO).

These offer types have been examined in [5GN-D32] in terms of potential impact on components and interfaces and security considerations. However, it is worthwhile to examine what offer types could apply into two practical scenarios.

## 4.2 On-demand network slices

Nowadays, a common business model for the telecommunication sector is represented by the OTT paradigm. Different enterprises (mostly content providers, such as Youtube or Netflix) use the underlying mobile network as a "data pipe" to the end user, relying on the standard "best effort" configuration. This has several drawbacks, being the most notable i) the content provider cannot tune the network or ask for enhanced KPIs and ii) the MSP (or the MNO) cannot monetize this traffic, as it is not offering any added value to the OTT. The possibility of dynamically instantiating Network Slices can solve these problems, allowing thus for a flexible Network Slice Market.

The 5G NORMA architecture supports this setup, while [5GN-D42] and [5GN-D52] define mechanisms for this use case (e.g., Sec 5.4.1 in [5GN-D42]).

This scenario considers that the MSP is a MNO, owning the RAN infrastructure along with data centres at edge while he may lease cloud resource to an InP for the central cloud. The MSP needs to deploy various customized slices to sustain dynamic requests from various tenants. Slice requests are occurring at different times, for various slice lifetime durations and covering specific areas. Examples of customized slice and tenants are presented below:

- "Reliable video quality" slice:
  The tenant is an OTT video player and requires a "congestion free" network HD video slice for wide (national) coverage. To fulfil such demand, the MSP needs to set up a slice which is customized with respect to a "common eMBB" slice type with some specific congestion mitigation mechanics. The deployed network must provide high throughput DL – medium latency – medium mobility- high reliability- nationwide coverage – optimized for specific video using transmission protocols like DASH, improved with congestion mitigation technique.
- "Event-video" slice:
  The tenant is an event organizer who requires a "video upload slice" for a localized area (e.g. in a stadium arena) for a given time duration corresponding to the event (e.g. a football match). The requested network needs a high throughput in uplink, an average latency, with limited mobility and medium reliability in a localized coverage.

In the next sections, network slicing and stakeholder relationships are examined under two situations, depending on the slice offer SLA between the MSP and the tenants.

**SLA between tenant and MSP covers guaranteed service KPIs (and data isolation)**

In this situation, each tenant is mainly interested into the slice Offer Type 1 in which the MSP provisions and operates a slice for customized communication services for the tenant. The tenant

has no expertise in networks and doesn't want to manage slice mechanics. He is only interested in getting some specific service KPIs, including requirements such as the desired coverage area, the number of devices, and possibly the device types (sensor, smartphones, etc).

The MSP can quite easily optimize the allocation of resource between the slices as he has full control on those resources. For example, it can rely on 5G NORMA dynamic allocation offered by the ISRB - beyond static resource quotas - while maintaining each tenant's SLA (service level only) and tenant's data isolation. Leaving the full control of the network slice mechanics to the MSP can make easier the prediction of traffic and users variations, allowing thus for a more efficient resource usage.

The MSP cannot rely on an "always accept" strategy for slice requests from the tenants. The ISRB (Inter Slice Resource Broker), part of SDM-O entity and hold by the MSP, is also entitled for the Dynamic Admission control mechanism (cf. the specific algorithms in Section 5.1.2 and Section 5.2.1 in [5GN-D42]). Enhanced admission control algorithms that leverage multiplexing gains of traffic among slices are key to the optimisation of network utilisation and monetisation. Finally, the SDM-X (that is entirely controlled by a single stakeholder, the MNO) makes sure that radio resources are correctly shared across slices.

**SLA between tenant and MSP covers guaranteed resource for slice.**

This case mainly considers slice offers from Offer Type 2 (limited control for the tenant) and Offer Type 3 (extended control). The SLA between the MSP and the tenant specifies the set of infrastructure resource assigned to the tenant's slice. The tenant is paying for the resource and he will not accept any resource reduction in case another tenant is experiencing some lack of resource. The MSP needs to carefully verify the availability of resource before accepting a new slice request from a tenant. The tenant can monitor and even (in Offer Type 3) control the set of resource assigned to him (cf. Section 2.4.2).

The MSP needs a dynamic slice admission control mechanism in the ISRB for the on-demand slices. The admission control can take decision based on prediction of load/traffic, always considering conservative quotas.

The MSP owns the ISRB, responsible for allocating set of resource quotas to each tenant slice. Then, the procedure follows a similar path as described in Section 3.5.2, but the possibility of resizing the NFVI domain associated to each slice is limited, so the multiplexing gains are reduced. When compared with the previous alternative, flexibility is exchanged for increased resource separation (an operational mode similar to the "Industry 4.0" scenario described in the next section). For this reason, it is expectable that this second operational mode will be more expensive (for the OTT) than the previous one.

## 4.3 Industrial communications network slices

Tis section considers the special case where the tenant is a vertical enterprise that owns some Industry 4.0[7] campus factory sites. The vertical wants to deploy a network to address several use-cases of services selected to illustrate the various requirements along Industry 4.0 networks (or Industry 4.0 slices). Those use cases are described as follow:

- Critical IoT: the tenant wants a secure network for highly sensitive traffic from monitoring sensors in the product line of the factory floor with low latency, an average throughput and high reliability. This is for indoor only coverage.

[7] Industry 4.0 refers to a fourth industrial revolution combining production methods with state-of-the-art information and communication technology, cf. https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Factories-of-the-Future-Vertical-Sector.pdf

- Non-critical IoT: The tenant wants a network related to tracking of forklifts and trucks equipped with sensors. This is less sensitive traffic which requires average latency, medium/high mobility and LTE-like throughput. This is for both indoor and outdoor coverage, e.g. in wide areas around the factory buildings.
- eMBB: the vertical wants a network for mobile broadband communications to its corporate data network which is also connected to the Internet. The network is possibly accessible from inside the factory building as well as by commercial vehicles fleet dispersed outside the factory over wide area coverage: it requires network with average latency, high mobility, high throughput, medium/high reliability, less sensitive traffic - this is for both indoor and wide area outdoor coverage.

In next sections, network slicing and stakeholder relationships are examined under two situations, within the factory building when the vertical wants to deploy several services as described above and outside the factory buildings where the tenant will lease a subnetwork slice from an MSP or MNO for reaching wider outdoor coverage.

## 4.3.1 Industry 4.0 slices deployment into private infrastructure

The vertical relies on the resource provided by his own private network infrastructure to deploy several slices needed to realize the afore mentioned services (critical IoT, non-critical IoT and eMBB) inside the factory campus. In that case, the vertical combines the role of tenant, MSP and infrastructure provider. More precisely, it is the various organisations of the vertical (like production line, delivery line entities) that are the inner tenants of the vertical. In this ecosystem, the MSP becomes a possible business partner, selling to the vertical its expertise into designing and rolling out IoT and eMBB networks onto the vertical's private infrastructure. The MSP is further involved for providing wide area coverage for the eMBB slice, which is detailed in Section 4.3.2.

The driver for this scenario comes from the vertical's requirement for a secure private network within the factories fully isolated for its critical monitoring sensors network. Full isolation of the vertical's traffic against any network provider or network user is only possible by running a private network on infrastructure owned by the vertical. The vertical owns and deploys all the infrastructure including its own base stations. Base stations can be physical equipment (hosting PNFs) or 5G gNBs deployed onto distributed units (on antenna sites) and central units at datacentres (or edge clouds). The vertical owns all infrastructure for the core domain and the corporate data network. It includes any X-haul and core transport elements (switches, routers and physical links) to interconnect the base stations, the core network and the data network. The vertical manages its customer subscriber database. The tenant doesn't want to rely on any third party to administrate the network. For example, the vertical wants to be free to switch on/off some base stations or to re-deploy or scale upon his own decisions. It owns all needed software assets for management, orchestration and control functions and applications (i.e. it owns BSS/OSS layers, ETSI-MANO layers and SDM-O, SDM-X/C entities). Alternatively, one could envisage that some software assets are bought from an MSP/MNO (for realizing the non-critical slices).

When end- devices connect through the base stations located inside the factory floor they are attached to one dedicated slice (e.g. for critical sensors devices) or possibly to multiple dedicated slices (e.g. for smartphone devices). The authentication and security procedure are handled by the vertical.

### 4.3.1.1 Slicing into the industry 4.0 private infrastructure

In the factory campus scenario, the vertical is using network slicing for optimizing its own network for its various organization entities (product line, delivery line, commercial service, etc.). An example of three slices (critical IoT, non-critical IoT, and corporate eMBB) deployed by the Industry 4.0 vertical on its own private infrastructure is depicted in Figure 4-1.

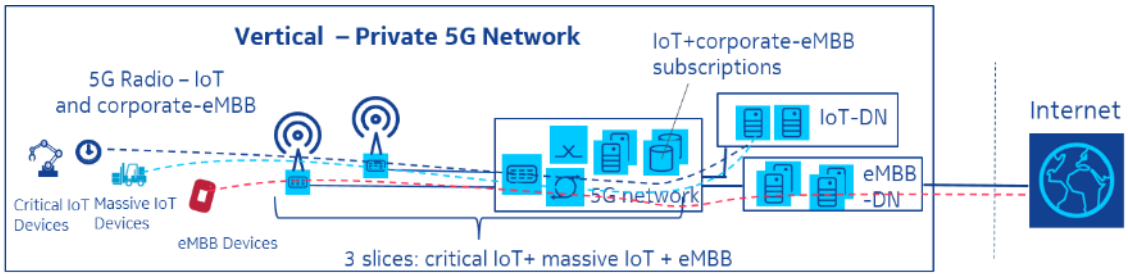



**Figure 4-1: Architecture for deploying industry 4.0 slices onto a vertical private network**

The vertical has three slices within its own private network. Those slices are deployed as follows:

(1) Private slice for critical IoT, indoor coverage:
This IoT use case of critical monitoring sensors inside the production line may require customization down to the PHY layer which would correspond to the RAN slicing Option 1 as described in [5GN-D42]. Only transmission (and reception) point specific functionality is shared across other network slices while all other functionality is implemented specifically by the critical IoT slice. Such an IoT slice will implement its own specific radio scheduler and this will translate to high complexity for the vertical's SDM-X when managing radio resource for multiples slices. However, this may be alleviated by reserving fixed spectrum resources for this slice. Compared to other slices, the slice for critical IoT may require reduced NAS signalling. Especially, critical sensors inside the factory do not need any mobility management when they are immobile.

(2) Private slice for non-critical IoT, factory campus coverage:
The vertical deploys a second IoT slice for covering the forklift sensors inside the factory campus. The non-critical IoT slice may not need customized PHY and MAC layer which also increases the deployment flexibility of sensor nodes as well as reduces their costs because non-proprietary technologies are used as argued in [5GN-D42]. In such a situation, Option 2 for RAN slicing is possible, which allows for slice specific radio bearer. The sensors connect to the slice through the vertical's base stations.

(3) Private slice for eMBB, factory campus coverage:
The vertical deploys on its own infrastructure inside the factory an eMBB slice for providing to employees an access to its private corporate data network (eMBB-DN). The eMBB-DN may be connected to the Internet and thus allow also Internet access for the vertical's eMBB subscribers. The smartphone devices and the subscriptions for corporate access are solely managed by the vertical. (i.e. the vertical's employees have mobile devices and subscriptions to access the corporate network). Again, Option 2 for RAN slicing would be preferred as it allows for slice specific radio bearer. When the smartphones are inside the factory, they connect to the private eMBB slice through the vertical's base stations up to the eMBB-DN.

Figure 4-2 summarizes the multiplexing of the three slices onto the private network and provides an indication on the technical domain ownership. The vertical acts as a private MNO and manages all 5G NORMA NFV Management and Orchestration (MANO) layer functionality through adequately designed APIs of according entities (e.g. SDMO functions, SDM-C, Service Management). Depending on the slice, the SDM-X handles various RAN functions depicted in orange colour in Figure 4-2, whereas one SDM-C entity deployed per slice handles the RAN functions customized per slice and depicted in green colour in Figure 4-2. The interworking of the different functional entities interwork is detailed in [5GN-D42].

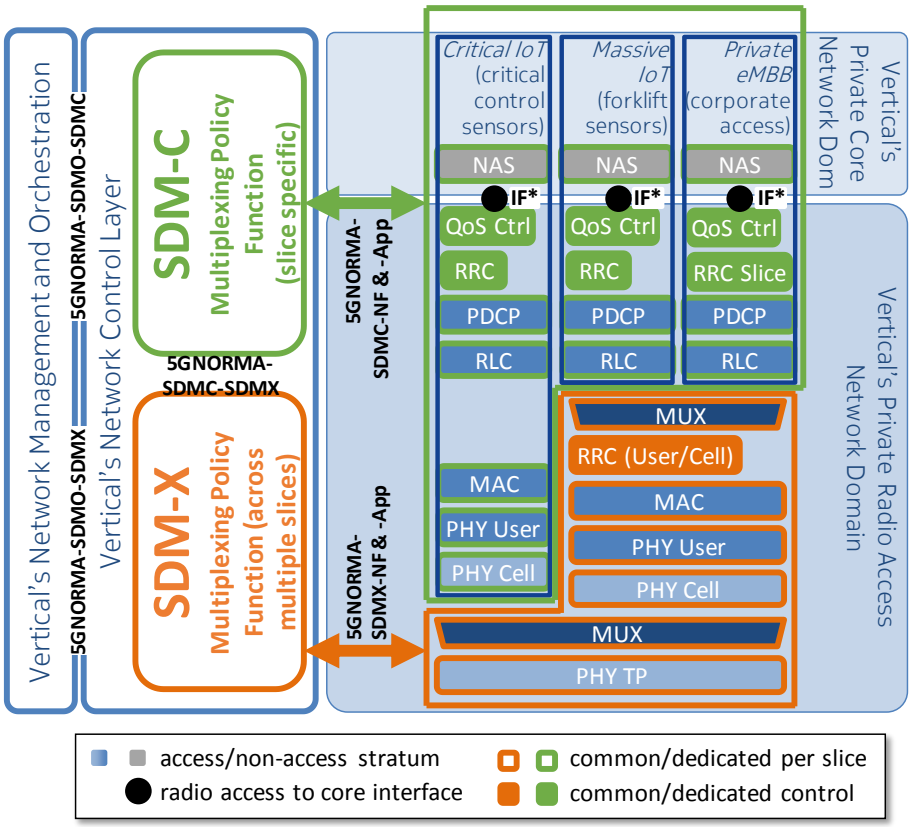



**Figure 4-2: Exemplary deployment using slicing inside industry 4.0 factory floor on a private infrastructure owned by the industry 4.0 vertical**

#### 4.3.1.2 Business relationship with an MNO and impacted 5G NORMA cross-domain interfaces

Possibly, the MSP can sell its expertise into designing and rolling out IoT and eMBB networks to the vertical. The vertical will avoid investment for designing the network but can concentrate investment on controlling the network operation. The vertical buys the template of such an IoT or eMBB network (slice) including network functions software along with the full set of management and control layer functionalities from the MSP. The vertical further activates and operates both IoT and eMBB network templates onto its own infrastructure.

With respect to the 5G NORMA architecture, a single cross domain interface is involved for this special case and corresponds to the interface produced by the MSP service layer functions and consumed by the vertical's service layer functions. Authorization policies determine the access rights for the tenant. The interface is used for provision of credentials to access to the MSP platforms for the vertical for further retrieving slice template and downloading network functions images from the catalogue (e.g. providing a URL to the MSP platform and a tenant account/password).

**Considerations on spectrum**

The spectrum for operating the private IoT slices would mostly belong to unlicensed spectrum as the vertical control every equipment inside its building using the given spectrum (no other interfering devices). Alternatively, the vertical can be a "private" MNO, who purchased some "private" licensed spectrum for his IoT slice. Alternatively, the vertical can have an agreement for using a licensed spectrum from an incumbent MNO who authorizes the vertical to use the spectrum with a low transmission power in a limited geographical area (factory). For private non-critical IoT and private eMBB, when such slices are to be extended for wide outdoor coverage (cf. Section 4.3.2) with the support of a public MNO network, the spectrum used for those slices

would mostly belong to some licensed spectrum shared with the MNO. Therefore, the ecosystem will not only offer the ability to purchase network slices as a service but has also to consider mechanisms for offering spectrum sharing where a vertical may lease some shared part of licensed spectrum from the MNO for use in its own private network. The MNO may have interest in monetizing the sharing of his spectrum with the vertical rather than supporting high cost expense for deploying and operating by himself the network in particular for indoor coverage.

## 4.3.2 Slice extension into the mobile network operator infrastructure for industry 4.0 outdoor coverage

In this scenario, the vertical has an agreement with a public MNO for enabling his corporate devices to connect to the corporate data network from outside the building factory, for both non-critical IoT (forklift & truck sensors) and eMMB services. Within this agreement, the tenant leases some dedicated sub-network slices over the MNO access network for conveying traffic from its devices to its own core network. Such a situation will be covered by Offer Type 3, through which the vertical will be provided with some means for controlling a sub network slice deployed within the mobile operator network.

### 4.3.2.1 Industry 4.0 slices into the MNO network

Figure 4-3 highlights an example of deployment of those various slices deployed by the Industry 4.0 vertical on its own private infrastructure and on the MNO infrastructure.

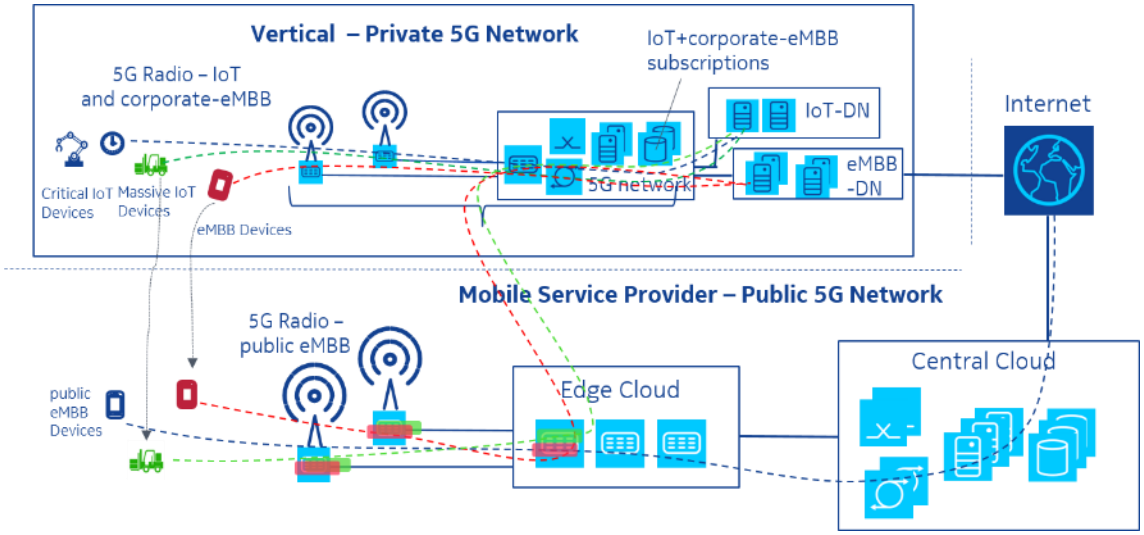



**Figure 4-3: Example of three Industry 4.0 slices deployment into vertical private network and MNO infrastructure: private critical IoT, private massive IoT and private eMBB**

**Private eMBB slice for Industry 4.0 vertical via the MNO network**

For providing indoor connection to its corporate data network, the tenant also operates a corporate eMBB slice on its own private network inside the factory with its own base stations deployed for covering the inside factory's floor. Only mobile devices from the vertical's subscribers can connect to this eMBB slice. The tenant has an agreement with a public MNO for enabling his corporate devices to connect to the corporate data network from outside the factory building. Involved security mechanisms are further described in Section 4.4.2. Within this agreement, the tenant will have requested a dedicated slice over the MNO's access network for conveying traffic from the tenant's devices up to the tenant owned core network.

**Private non-critical IoT slice for industry 4.0 vertical via the MNO network**

In the scenario for forklift and truck tracking, the tenant considers that it is less sensitive data. Moreover, it requires outdoor coverage over an area not covered by the vertical base stations. The

vertical sets an agreement with an MNO who owns base stations covering the factory outdoor for leasing a mobile access slice covering the outdoor wide areas. The vertical uses this access slice for connecting his forklift and truck sensors to his own IoT core and data network.

Figure 4-4 highlights an example of deployment onto the MNO infrastructure of those various slices deployed by the Industry 4.0 vertical crossing its own private infrastructure and the MNO network. It is assumed that RAN slicing Option 2 as described in Section 2.3.5 is used for those two slices as well as for the public eMBB slice provided by the MNO. Within this option, both the transmission point and user specific part is shared across network slices, and the service (or bearer) specific part is implemented in each network slice. Hence, in this option, the individual network slices rely on the same radio access technology but customize their operation at lower layers through parameterization and at higher layers through customized implementation.

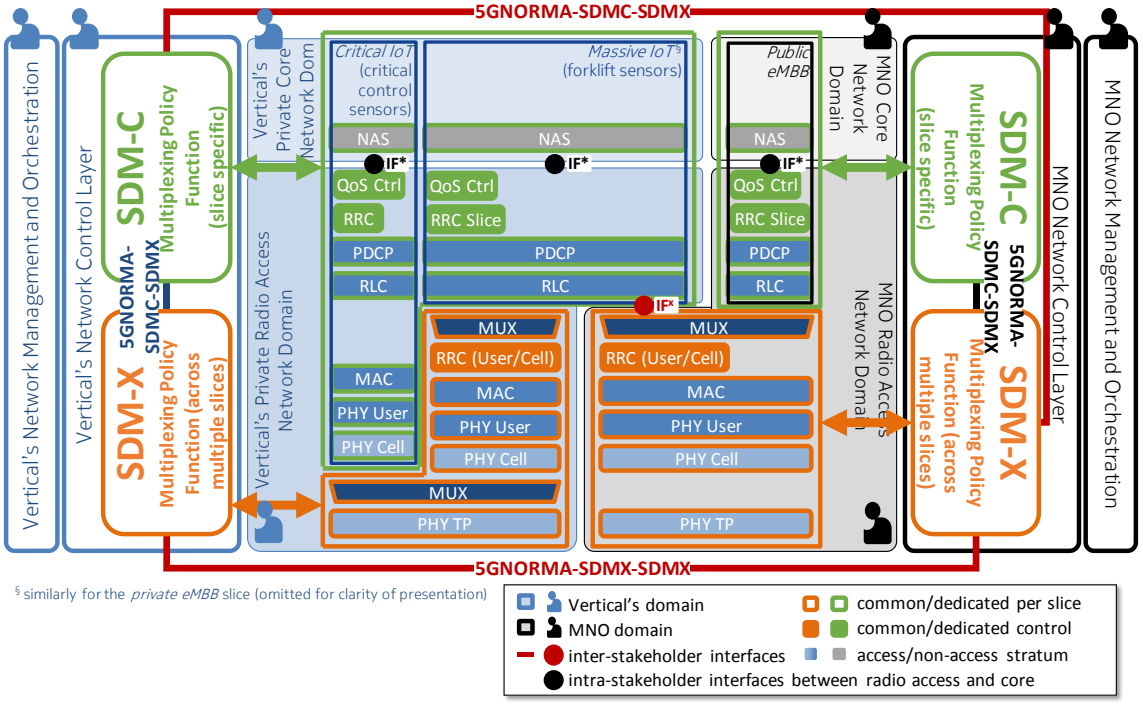



**Figure 4-4: Exemplary deployment of RAN slicing for outdoor Industry 4.0 communications in a public 5G MNO network**

### 4.3.2.2 Business relationship with an MNO and impacted 5G NORMA cross-domain interfaces

In addition to considerations in Section 4.3.1.2, the ownership of the various 5G NORMA entities for managing, orchestrating and controlling the various slices onto the operator infrastructure is identified. The vertical manages its own customer subscription and subscriber database. The vertical owns and operates the OSS/BSS layers, ETSI-MANO and SDM-O and SDM-X/-C controllers in its data network and core network domain. The massive IoT (private eMBB) slice comprises the core functions and the corporate data network and two branches for access, one branch adapted to local base stations owned by the vertical and another branch using a private IoT (private eMBB) RAN slice provided by the MNO for radio access network covering the outdoor area. The slices are provided by the MNO to the vertical under the Offer Type 3 which involves the following cross-domain interfaces between the vertical and the MNO.

(1) Interface between the service layer functions from the MSP and service layer functions from the tenant:
These interfaces are produced by the MSP service layer functions and consumed by the tenant's service layer functions. Authorization policies determine the access rights for the

tenant. The interface is used for defining the required service KPIs for the leasing the RAN slice, in accordance with the service level agreement. It requires provision of credentials to access to the MNO platforms for sub-network slice template sharing and for orchestration/control operation (e.g. providing a URL to the MSP platform and a tenant account/password).

(2) Interfaces between Service Management & Orchestration layer from the MSP and Service Management & Orchestration layer from the tenant:
These interfaces are exposed by the Service Management & Orchestration layer from MSP and consumed by Service Management & Orchestration layer functions from the tenant. Authorization policies determine the access rights for the tenant. The interfaces are used for exposure and sharing of the RAN slice templates and for allowing the tenant to order a limited set of network slice operations such as RAN sub slice activation/stop/scaling (granted by the MSP Service Management).

(3) Interfaces between the SDM controllers of MSP and tenant (control layer) and between the MSP-controlled multiplexing function and the tenant-controlled higher protocol layers (data layer):
These interfaces are exposed by the SDM control layer from MSP and consumed by SDM control layer functions from the tenant. Authorization policies determine the access rights for the tenant. The interfaces cover the following requirements:

- Allowing to interconnect VNFs across domains: in the considered example, as detailed in [5GN-D4.2], each slice specific PDCP/RLC function instance belonging to the vertical domain should be interconnected (through $IF^{+}$ interfaces as represented in Figure 4-4) with the multiplexing entity instance facing the shared lower level RAN functionalities and owned by the MNO. The interconnection is realized by the establishment of a link between the two instances. SDM-C from the vertical and SDM-X from the MNO should exchange the information on their respective 5GNORMA-SDMC-SDMX interconnection point and a dedicated VPN can be established between the vertical's datacentre hosting the vertical's core and the MNO edge cloud hosting the central units of the operator's RAN. Moreover, standardization of 5G interfaces between the above functions are required for interoperability.
- Allowing for example the tenant to enforce QoS/QoE change under the network slice operation (granted by the MSP Service Management): The transmission point (cell), PHY and MAC layer in the data plane and the RRC in the control plane are shared across all network slices and then operated by the MNO. The MNO operates the SDM-X. The vertical can customize its slice through configuration and parametrization based on the service. It owns the SDM-C and can implement his own QoS control for its slices. On the interface between the vertical SDM-C and the MNO SDM-X, information about the individual services and their QoS requirements is provided to the MAC scheduler as part of the SDM-X, which then enforces those QoS constraints as part of its multi-service framework. In such scenario, the interface between the shared and dedicated controllers would coincide with the interface between the distributed units and the edge/central cloud. The PDCP layer sits in the edge cloud of the vertical which can offer guarantee for slice security as further detailed in Section 4.4.3.2.
- The 5GNORMA-SDMX-SDMX interfaces between tenant's and MSP's SDM-X controller facilitates advanced RRM functions in the coverage overlap of the tenant's base stations with the MSP radio access network. In the example, the tenant's forklift sensors benefit in the outskirts of the tenant's premises from coverage provided by the MSP's base stations. The inter-SDM-X interfaces enables seamless mobility between MSP and tenant (primarily important for the private eMBB slice, which is not depicted in Figure 4-4 for clarity of presentation) and consistent performance through mutual interference control and coordination even for sensor devices with

their typically low power low end transceivers, incapable of advanced (beamforming and interference rejecting) transmit/receive processing.

## 4.4 5G NORMA ecosystem security considerations

[5GN-D32] already provided security considerations for the basic stakeholder relationships, such as the necessity of trust in InPs and in software-vendors. Moreover, for all three offer-types of [5GN-D32], the specific security and trust implications were analysed. In all cases, the MSP/MNO can secure its resources and APIs against erroneously or maliciously acting tenants and need not trust tenants. Vice versa, tenants need to trust the MSP/MNO in many respects, including correct resource assignment, keeping tenant data secret, and maintaining integrity of tenant data. While such a trust relationship is a reasonable assumption in many cases, there may be other cases where a tenant requires enforceable isolation even against the MSP/MNO.

### 4.4.1 Over-the-top security

A vertical that rents a slice to connect mobile devices to a vertical-owned data network (DN) has the option to apply over-the-top (OTT) security between mobiles and an entity in the DN. For this, an entity within the DN, e.g. a VPN gateway, performs access authentication to the DN, combined with key agreement. A secure tunnel is setup between mobile device and DN. (To avoid double encryption, radio interface security mechanisms could be disabled for the data layer.) The authentication may be performed in the data layer, e.g. by doing an IKE or TLS handshake. Alternatively, the mobile network may support transport of authentication messages in the control layer, for example by support of the EAP framework, allowing authentication before data layer connectivity is established. (Such concepts are under consideration in 3GPP standardization groups.) OTT security requires that the vertical operates an own user[8] database and provisions credentials to its users. Therefore, a mobile device in this scenario needs two sets of subscription credentials, one for attaching to the mobile network and the other for accessing the DN.

OTT security can provide confidentiality and integrity protection of the tenant traffic even against the MSP/MNO. However, the MSP/MNO will still be able to retrieve a lot of metadata on the traffic, like which MSP subscribers get access to the DN, their location when they access the DN, their communication times and volumes, communication relationships and so on.

### 4.4.2 Security by using a private network

The scenarios described in Section 4.3 focus on verticals with highly sensitive traffic that is handled on tenant-owned infrastructure. In Section4.3.1, the vertical operates a private network, thus achieving the best possible isolation. (This kind of isolation may still not be perfect, because the radio interface, even when used indoors, may still allow external attackers to capture and analyse radio traffic and gain some (limited) insight into the private network by doing this.)

It should be noted that the vertical in this setup still needs to trust other parties, for example the vendor of the network equipment and software, and the service personnel that sets up and operates the private network.

[8] The term “user” does not necessarily refer to human users. “user” refers here to any entity that has an id and possible credentials to access the DN.

### 4.4.3 Security for networks built on both private and public infrastructure

Section 4.3.2 describes cases where a vertical with its own private (indoor) network rents slices of a public network in order to get extended coverage for the subscribers of the private network. At least the following two security aspects require attention in these setups:

(1) How are the private network subscribers authenticated and authorized to use the respective slices in the public network?
(2) How is the sensitive traffic protected in the public network?

#### 4.4.3.1 Roaming Model

One solution could be a roaming model. Here, the verticals have roaming agreements with the MNO that allows the verticals' mobile devices to use the public network. When a mobile belonging to a private network attaches to the public network, the MNO recognizes this via information provided by the mobile during attachment. The MNO then authenticates the mobile based on subscription information that it retrieves from the private network and authorizes it to use the respective network slice in the public network. Traffic can subsequently be protected by a key derived during authentication, but according to the EPS-AKA authentication procedure as used in today's LTE networks, this is a key shared between mobile and serving network, i.e. the public network in this scenario. Hence, this key cannot isolate the traffic against the MNO. The vertical operating the private network would need to apply OTT security to achieve traffic protection against the MNO.

However, 3GPP will introduce in 5G a new, EAP based authentication procedure. Figure 4-5 visualizes this approach. It uses terminology from [23.501]. As the first step, EAP-based mutual authentication (cf. [33.501]) is done between the mobile IoT device and the AUSF (Authentication Server Function) in the private network. Subsequently, the AUSF passes a key agreed with the mobile to the AMF (Core Access and Mobility Management Function) in the public network, which is used to establish security between the mobile and the public network. But in addition, the private network may retain another key agreed with the mobile during the EAP run, which can be used to establish data layer security.

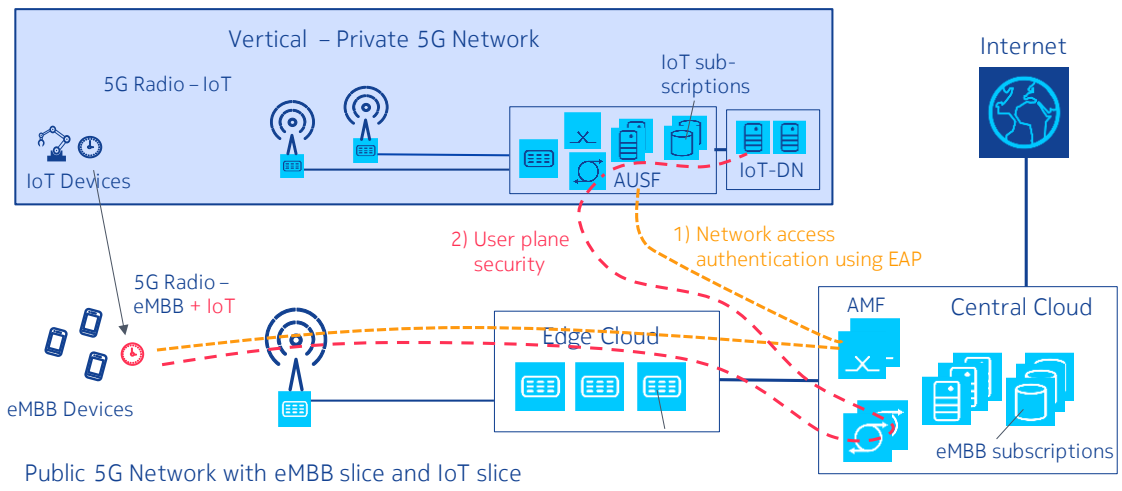



**Figure 4-5: Roaming approach with OTT security for networks built on both private and public infrastructure**

This approach has the advantage that mobiles do not need an additional ID and subscription for the public network, and that a single authentication run is sufficient to establish a key for security between mobile and public network and a key for OTT security.

Concerning network slicing in the public network, this model fits to a slicing setup relying on common control functions in the core, which perform authentication when a mobile attaches to the network. After authentication, the mobile is authorized to use a specific slice – in this case a slice rented by the vertical for connecting the vertical's mobiles to the vertical's private network.

The above example shows only a single service (IoT) in the private network. However, the same approach works also for a private network with several services supported by several slices. As described in Section 4.3.2, one slice in the public network can be rented per service, and a mobile attaching to the network may be authenticated to use one or more of the available slices.

It can be noted that this model does not necessarily require tenant-specific slices in the public network that can be rented by verticals, but could also work with a network that only provides one slice per service type (e.g. eMBB and IoT). Clearly, the network must be suitably engineered by the MNO to be able to provide the service required by all the verticals, as agreed upon between MNO and verticals in service level agreements (SLAs). However, a slicing mechanism including proper resource management may facilitate providing such services with high SLA-compliance and at a large scale.

### 4.4.3.2 Pure Slicing Model

Another model is suggested by Figure 4-4. In this model, the slice rented by the vertical only includes the lower layer RAN stack and connects to the private network that provides higher RAN functions and core network functions. There may be several options for the split between the slice in the public network and the private network. From an isolation point of view, the entities terminating radio interface security should be in the private network – this ensures isolation even against the MNO. In LTE as well as in 5G phase 1 (according to current considerations in 3GPP), radio interface security is terminated in the PDCP, which is located in a RAN entity, but this RAN entity may be implemented in a strongly centralized way in 5G. Making the split below the PDCP could thus allow to keep all security related functions within the private network. The setup is visualized in Figure 4-6.

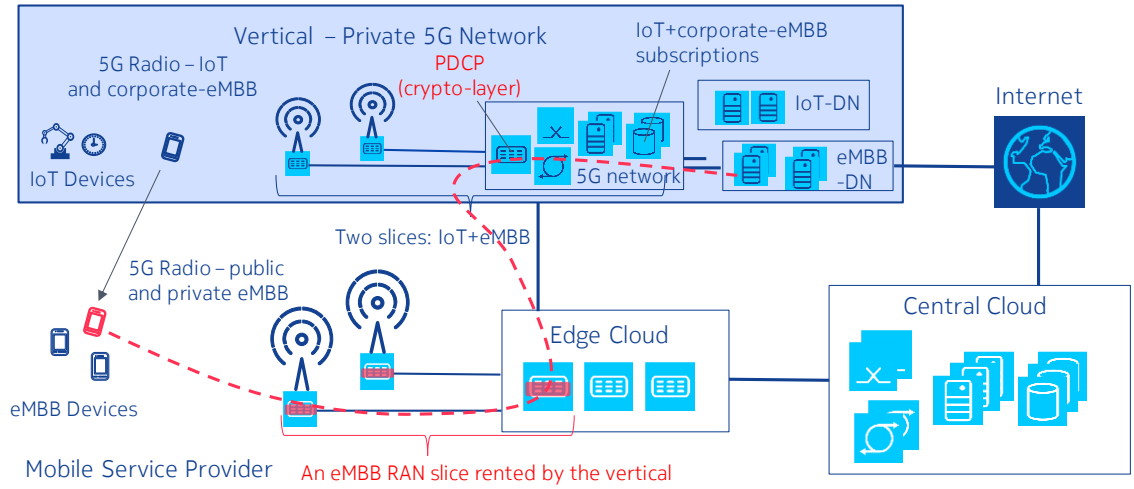



**Figure 4-6: Slicing approach for networks built on both private and public infrastructure**

In this approach, it is required that a mobile attaching via the public network provides sufficient information to allow the public network to recognize to which slice the mobile needs to be connected.

With this approach, the vertical can achieve high isolation against the MNO without the need to apply OTT security.

**Note**: In [5GN-D42], the split below the PDCP layer is described as RAN slicing Option 2, and in this option, everything below PDCP is considered as belonging to the "common part". However, this does not mean that this common part is common for all possible slices.

Rather, as shown in [5GN-D42], Section 2.1.4, there can be other slices which follow RAN slicing Option 1 and implement their own MAC schedulers, running in parallel to the MAC scheduler in the common part of Option 2. In this sense, the common part of slicing Option 2 can be considered as another slice according to Option 1, and this is the slice rented by the vertical in the scenario described above.

Radio interface security for the data layer need not necessarily be terminated by PDCP. In 3GPP, proposals to terminate it in the UPF (User Plane Function, equivalent of a 5G NORMA data layer NF) in the core have been discussed, and this option may become part of 3GPP 5G phase 2. With this option, the vertical's slice in the public network could comprise also RAN functions such as the PDCP and the RRC, but still the radio interface security would provide isolation of the data layer traffic against the MNO. Section 5.4.1 introduces an access stratum security function and argues that this function can be centralized. Applied to the present scenario, this means that the access stratum security function can run as part of the private network and thus enjoy full isolation against the MNO. Only keys for the protection of the signalling bearers would be passed to the slice in the public network, while the keys for the protection of the data radio bearers would be passed to the UPF(s) and thus not leave the private network.

# 5 Security in 5G NORMA Networks

Security is of paramount importance for future 5G networks. 5G NORMA has substantiated the need for security by analysing important 5G use cases and setting up (black-box) security requirements in [5GN-D21]. In a next step, taking into account the envisaged architectural principles of 5G NORMA, dedicated security requirements have been specified in [5GN-D31]. Subsequently, the threats to a network adopting the 5G NORMA architecture have been analysed, and innovative security concepts to mitigate these threats and to fulfil the security requirements have been developed. This is documented in [5GN-D32]. During the last phase of the project, the security concepts have been refined and adapted to the final 5G NORMA architecture as documented in [5GN-D42], [5GN-D52] and the present document.

This chapter refrains from reproducing the results that are already documented in [5GN-D32], but focuses on the refined security concepts and their mapping to the final architecture. Hence, in the following, only a very brief summary of the earlier results (cf. Section 5.1) is given, and then the detailed, final description of the 5G NORMA security concepts (Section 5.2 through Section5.5) are provided. Finally, Section 5.6 concludes the chapter.

## 5.1 Summary of previous results

A **study on the impact of security breaches** was carried out. The results provide additional motivation for the security work by showing the high socio-economic importance of providing a supreme level of security in 5G communication networks, cf. [5GN-D32] Section 5.1.

**Potential security risks associated to new concepts and procedures defined by 5G NORMA** have been analysed, and guidelines have been given in order to make sure that these risks are suitably mitigated, cf. [5GN-D32] Section 5.2.

The **applicability of LTE security concepts** to 5G NORMA network was investigated, with the result that substantially new security concepts are required to cover network function virtualization, multi-tenancy and software defined mobile network control, cf. [5GN-D32] Section 5.3.

**Innovative security concepts to secure 5G NORMA networks** have been investigated. A first description is given in [5GN-D32]. The final description is provided in the following sections.

## 5.2 Virtualised authentication, authorization, and accounting

The objectives and design requirements of Virtualised-Authentication, Authorization, Accounting (V-AAA) are illustrated in [5GN-D32] [VAAA]. Based on the principle of 5G NORMA flexible RAN, V-AAA is devised. The V-AAA takes a two-level design approach to secure the 5G NORMA flexible RAN. This two-level design adopts the traditional central governance approaches for authentication and authorization. It also adds the tenant and subscriber identification features to the core network (central cloud) and the access network (edge cloud). Moreover, it secures the flexibility and elasticity of resources demand from tenants and their network slices. The flexible RAN network slicing definition is classified into three categories in [5GN-D42]. These network slicing categories require a secure isolation and protection in between network slices and network entities. Therefore, a secure communication mechanism must be applied when the network entities initiate an exchange of control signalling within the network slice or across network slices. Particularly, when 5G NORMA logical network entities demand an access or manipulation of resources across different security domains, a secure communication mechanism (e.g., TLS, DTLS and IPSec etc.) in protecting the session's confidentiality and integrity of the control signalling instructions is essential. For instance, a MSP provides multi-

network slicing services to its tenants. Sometimes, these tenants/network slices require to access or manipulate the common resource blocks via SDM-X, and more frequently, these tenant/network slices require to access or manipulate the dedicated resources blocks within the security domain. These inter- and intra-security domain communication considerations have been discussed in [5GN-D42].

In this section, we aim to provide an aspect of V-AAA extension and security considerations when control signalling exchange within (intra) security domain or across (inter) security domains. Particularly, when 5G NORMA network entities request to access network resources within (intra) security domain or across (inter) security domains, the communication session must be protected by a security association. Typically, network slices or network resources provisioning and deployment are required for a secured environment to be executed. In order to produce a secured environment, the V-AAA is responsible to ensure the proof of identity, grant the level of access, and delegate the access to the specific resources. After network slices have been deployed, tenants might request an extension of their network slice or network self-optimisation to trigger the manipulation of network resources. The V-AAA extension is designated to tailor such requests more securely. A security association should be used to protect the entities of communication session. For example, in an inter-security domain network resources request, the V-AAA acts as a software agent to establish a security association with V-AAA Manager and via the V-AAA Manager to request for extra network resources. These network resources could be owned by the MSP or other MSP tenants. On the other hand, the V-AAA agent can also initiate a security association with other MSP tenant's V-AAA agent to request for extra network resources. The tenants might have the same types of service operating in their network slices and they might also have private service level agreements between them. Moreover, in an intra-security domain network resources request, the dedicated network slice V-AAA agent establishes a security association to protect the session with the V-AAA Manager for obtaining extra network resources. These V-AAA agent communication sequences are given in [5GN-D61]. The direct and indirect V-AAA agent requests are illustrated in Figure 5-1 and Figure 5-2.

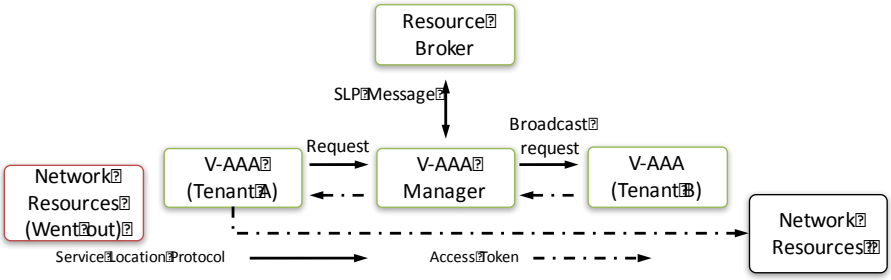



**Figure 5-1: Indirect agent request via V-AAA Manager**

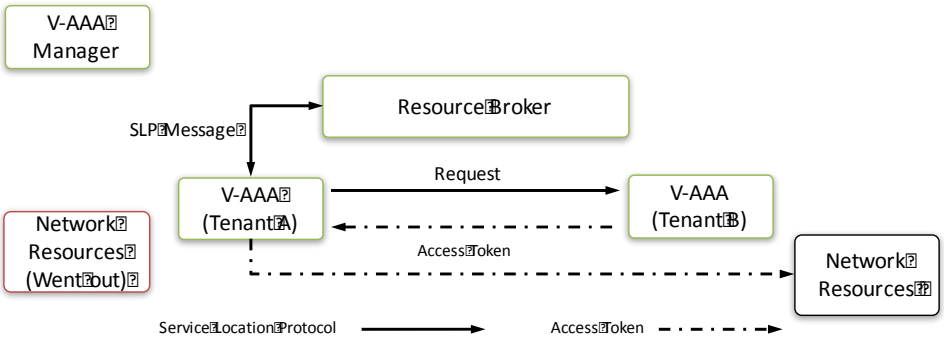



**Figure 5-2: Direct request to V-AAA agent network**

We consider the tenant needs extra network resources after their network slice has already been deployed. V-AAA agent extends to support two options to deliver a secure approach for allocating extra network resources for the tenant. The first option is a direct control signalling that is sent to

another tenant network slice's V-AAA agent. This signalling session would remain in the edge cloud and possibly across different security domains. In this case, we assume that there is a private SLAs between the tenants. However, in this type of requests, the MSP should be informed. Also, the trust procedure should be established and enforced by the MSP in this kind of cross-security domain transactions. The second option is another direct control signalling that is sent from the tenant's network slice V-AAA agent to the V-AAA Manager. This signalling session is established from the edge cloud to the central cloud but it remains in the same security domain. Hence, the MSP would have all the transactions and it remains directly informed with full control of tenant's network resources. Furthermore, of course, the V-AAA Manager would contact the network resource broker for obtaining the network resources. On the other hand, in this option, the V-AAA Manager has a full visibility of tenant's network slice. The V-AAA Manager can also redirect the request for obtaining extra network resources. We assume the V-AAA Manager can be acted as a software agent and securely obtains the availability of network resource information when it needed. These network resources availability information could come from the network resource broker or OSS/BSS. However, in runtime operation, the network resources information could be available from SDM-O which depends on the network configuration.

In fact, this direct and redirected approach in V-AAA adapts the 5G NORMA flexible architecture and provide a security to the 5G NROMA architecture. This V-AAA extension resolves some of the insider attack issues, reduces the loss of login credential risks and removes the exposure of network resource from the OSS/BSS network provisioning platform directly. Basically, we add an identity layer to the network provisioning platform, tokenize the login credential and give a specific objective and duration to the token. The tokenization technique with identification and delegations can be referred to OpenID connect open authentication protocol [RFC6749] and json web token [RFC7519].

## 5.3 Tokenization on provisioning and deployment

Tokenization technique is a typical approach for reducing the access of service or resource risks. It isolates and segments from data processing. There are many types of token which are used to encrypt curial information into a standard format with a specific expiry time of accessing the resources [RFC7519].

We tokenize the tenant's identity to reduce the insider attack and minimize the direct access of sensitive data or network resource errors. More importantly, the proof of identity should be carried out in the multi-factor authentication and with time limit while using the identity to login the network provisioning platform. For instance, a tenant could have many different administrators. Each of the administrator could belong to different access groups with different levels of service access authorities. When a tenant would like to extend their network slices, a network infrastructure administrator logs in with his/her credential to the service identity server and obtains the identity and access token. This token has an expiry time for accessing a specific network resource or entity. In 5G NORMA, this identity and access token can be applied when a common network resource to be accessed. We send this identity and access token to SDM-X and then the SDM-X decrypts the token. The SDM-X would check the expiry time of token and establish a security association with the authorisation server to get a confirmation of identity and accessing right. However, there might have some draw back in applying the tokenisation technique to access or manipulate the common resource such as SDM-X due to low latency and fast reaction of changing resource on demand.

In order to have a full protection of 5G NORMA flexible RAN and network entities, after solving the potential insider attack, we also need to protect the communication session confidentiality and integrity. For example, SDM-O receives an instruction to manipulate a particular network slice resource. Without any protection of the communication session, this instruction could be tampered or eavesdropped. Therefore, the SDM-O or other 5G NORMA communication session confidentiality and integrity must apply a security association (IPSec, TSL or DTSL) to establish

a session for protecting the control instructions and data. Another example, without any protection of the SDM-X communication session confidentiality and integrity, the damage could be affected the overall flexible RAN common network resources.

## 5.4 5G NORMA RAN security concepts

The 5G NORMA flexible RAN architecture is described in [5GN-D42], and this description includes a discussion of the security aspects relating closely to this architecture. In particular, security considerations are given for

- the flexible selection and dynamic allocation of network functions ([5GN-D42], Section 3.3);
- RAN multi-tenancy ([5GN-D42], Section 5.3);
- and the interfaces in the RAN control and data layer, with a special focus on how to secure the interfaces in the data layer ([5GN-D42], Section 4.4).

Earlier, [5GN-D32] introduced

(1) a novel access stratum (AS) security architecture to secure the 5G NORMA radio interface;
(2) a discussion of how the AS security concept can be used together with RAN slicing, considering several different slicing approaches;
(3) multi-service support by tailored radio interface security algorithms;
(4) and concepts for securing 5G NORMA RAN entities and the front- or back-haul interfaces.

The subsequent two sections now provide the final description of the 5G NORMA access stratum security concept, and its mapping to the 5G NORMA RAN architecture, in particular to the different RAN slicing options. This replaces items (1) and (2) from the list above, while items (3) and (4) hold as described in [5GN-D32], Section 5.4.4.3 and Section 5.4.4.4.

### 5.4.1 5G NORMA access stratum security concept

As described in [5GN-D32], in LTE the radio interface is terminated at a single eNB (or at most at two, in LTE Dual Connectivity). In contrast, 5G NORMA features a much more flexible RAN comprising edge clouds as well as bare metal equipment, and RAN functions can flexibly and dynamically be allocated. Also, multi-connectivity is a native 5G NORMA feature. This requires a far more flexible Access Stratum (AS) security concept, which has been introduced in [5GN-D32] in an initial version and has been extended to cover the final 5G NORMA RAN architecture described in [5GN-D42], in particular the various slicing options that have been introduced there. This enhancement mainly comprises carving out an "AS security function" from the former control layer security termination function.

Figure 5-3 visualises the proposed new AS security concept. Here, based on a key $K_{AS}$ derived in the core, an entity called "AS security function" derives a key pair comprising an encryption key and an integrity key for the control layer as well as multiple key pairs for multiple possible data layer termination functions (that may be allocated at different physical entities). Note that both encryption and integrity may be optional in the data layer and could be replaced by application layer security, if network policies allow this.

Figure 5-3 further shows a single control layer security function, as we assume a single control connection, as a rule. However, the approach clearly could also support multiple control connections terminated by multiple control layer security functions, should this become a valid RAN architecture option.

For a mapping of the security functions to the identified 5G NORMA RAN functions according to [5GN-D42], cf. Section 5.4.2.

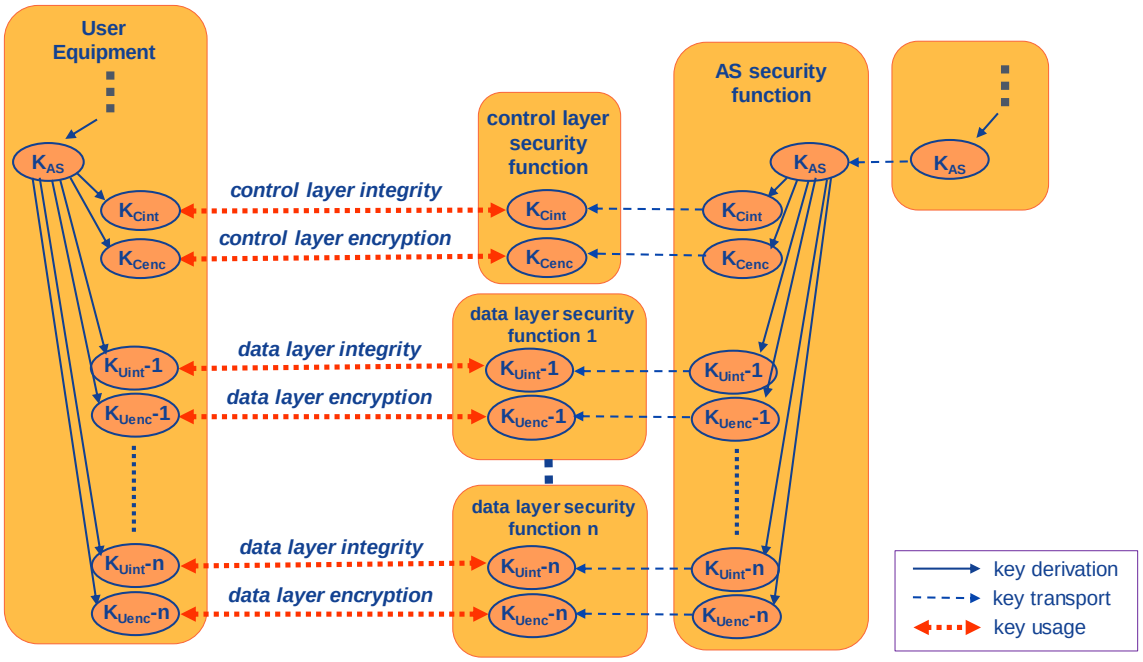



**Figure 5-3: Flexible 5G NORMA Access Stratum security approach**

Notably, in this approach, a compromised data layer security function instance has no means to decipher or fake control layer messages or data layer messages terminated at other data layer function instances. A data layer security function may thus be allocated also on physically exposed entities, very close to the antenna, without endangering the security of the control layer and of data layer traffic handled by other data layer security functions. In contrast, the AS security function is supposed to be located typically in an edge cloud, thus less exposed to attacks exploiting physical access.

The proposed setup allows to refresh a key pair for one data layer radio leg while keeping all other keys unchanged. A data layer security function that needs to refresh a key pair (e.g. to prevent a repetition of the key stream) can trigger the AS security function to perform the refreshing. Likewise, the keys of the control layer security function can be refreshed independently. Figure 5-4 visualizes the key refresh procedure.

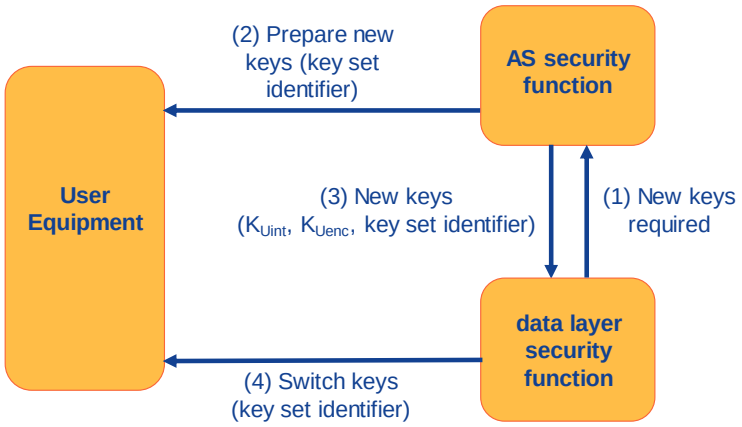



**Figure 5-4: Exemplary Key Refresh Procedure**

A new key may also be needed when a data or control layer security function is relocated, in order to prevent that consequences of a possible security breach at one location are propagated to other locations. Relocation may be necessary due to mobility of the user, but also due to network-side reconfiguration. Relocation may also require a change of the security algorithms, if the platform

to which the termination function is relocated does not support the current algorithms or has other preferences concerning algorithms. All this can be executed individually per data or control layer radio leg.

A relocation may also be required for the AS security function. In this case, like in an LTE handover, it may be reasonable to refresh the complete AS key hierarchy. A new $K_{AS}$ may either be received from the core, to avoid any dependency on the previous one. However, to optimize speed, a new $K_{AS}$ may also be derived from the old $K_{AS}$. In this case, if the old $K_{AS}$ was compromised, also the new $K_{AS}$ must be considered as compromised. Hence, care must be taken that an independent new $K_{AS}$ is generated in the core sufficiently often in order to limit the impact of a potential compromise of an AS security function.

### 5.4.2 Supporting RAN slicing

[5GN-D4.2] describes three basic RAN slicing options, cf. Figure 2-10.

**RAN slicing Option 1** is called "slice specific RAN", where only transmission point specific functionality is shared among network slices, while all other functionality is instantiated specifically for each network slice. This option is illustrated in Figure 2-2 of [5GN-D42].

In this model, all relevant security functions are instantiated per slice – there is no common security function in the RAN. Security termination happens in the "PDCP" blocks, which exist once per radio bearer, in the control layer as well as in the data layer. According to the description above, a key pair per termination point is required, but not necessarily a key pair per radio bearer. However, rather than using a single key pair for multiple collocated PDCP blocks, an implementation may choose to allocate a key pair per PDCP block, i.e. per bearer. This leads to potentially some more key pairs, but has the advantage that each bearer's endpoint at the network side can be relocated without any impact on other bearers.

The AS security function introduced above is a central function that need not be invoked on a per packet base. Therefore, it can be implemented as one of the blocks that in the figure sit on top of the SDM-C, either as a dedicated block, or as part of one of the blocks shown in the figure.

**RAN slicing Option 2** is called "slice specific radio bearer", and visualized in [5GN-D42], Figure 2-3, reproduced as Figure 2-8 in the present document.

In this option, the control layer functions and the signalling radio bearers are shared between slices, while data radio bearers are slice specific. This raises the question, where the AS security function should reside. The answer depends on how slicing is done in the core and where the key $K_{AS}$ is derived – in a common part or in a slice specific part. This in turn depends on what authentications are performed when the UE attaches to the network and establishes a PDU session using a specific slice. It is beyond the scope of this document to discuss the different possible approaches. Assuming a scenario, where $K_{AS}$ is derived within the slice, also the AS security function should be within the slice, i.e. in the picture above, it would sit on top of the SDM-C. Keys for signalling bearers would then be passed through the SDM-C and SDM-X from the AS security function to the signalling bearers' PDCP entities. It is an obvious advantage of this scenario, that the common parts have no access to slice specific keys. This is an additional protection against an attacker that may have access to common functions (e.g. the attacker is another tenant) and may successfully compromise a common function.

**RAN slicing Option 3** according to [5GN-D42] is called "slice-aware shared RAN". Here, all security termination points reside in the common part. This may best fit to a core slicing scenario where the key $K_{AS}$ is derived in a common core part. In this case, the AS security function should also reside in the common RAN part, as one of the function blocks sitting on top of the SDM-X (cf. Figure 2-5 in [5GN-D4.2]).

**Slicing from the UE perspective**

From the UE point of view, the presence of several slices does not affect the AS security procedures and the UE internal key derivations. However, depending on core network slicing scenarios, a UE connecting to several slices simultaneously may need to perform several slice-specific authentication procedures, resulting in several instances of $K_{AS}$ and several sets of key pairs derived from them, that must be maintained by the UE simultaneously in slice specific security contexts. This should not be a major burden for a UE that is capable and resourceful enough to communicate with several slices simultaneously.

# 5.5 Trust zone

## 5.5.1 Trust zone in the 5G NORMA architecture

A Trust Zone (TZ), as defined in Section 5.4.5 of [5GN-D32], is a geographical area served by a local base station i.e. an edge cloud, where different policies are autonomously implemented to ensure data security, while as many services as possible can be provided, regardless of the connection status between this edge cloud and the central cloud. Generally, among the reference use cases defined in [5GN-D21], TZ is highly related to the following ones with concerns of data security in emergency situations, where edge clouds can be disconnected from central clouds:

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

Generally, the TZ is aware of the edge-cloud-to-central-cloud connection (EC4) state, in order cognitively invoke the security functions either in the central cloud or in the local edge cloud, and activate variant essential emergency services to users of different trust groups. To achieve this, the TZ function must be tenant-dependent and tightly integrated with the V-AAA framework. A five-state behavioural model has been proposed in [5GN-D32], defining three steady states upon the EC4 availability, two transient states, and available transitions, as shown in Figure 5-5. The functional entities needed to implement a TZ were also listed, including:

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

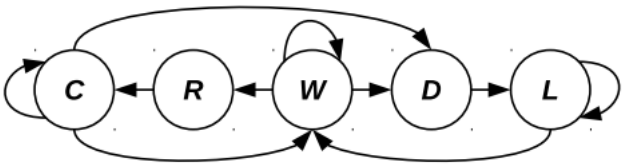



**Figure 5-5: State model of Trust Zone. State abbreviations C, R, W, D and L stand for Connected, Reconnecting, Weak Connection, Disconnecting and Lost Connection, resp.**

Details of every entity are available in [5GN-D32]. In [Han17], interfaces between these entities and other 5G NORMA modules have been defined to build an architectural entity model and to integrate it with the 3GPP architecture for the 5G system, cf. Figure 5-6.

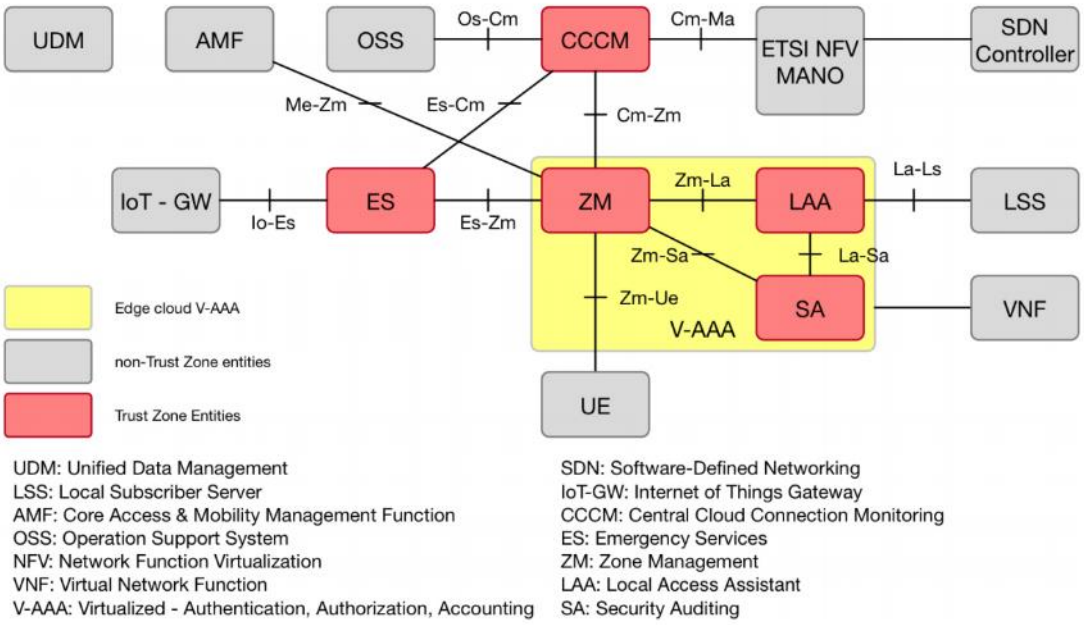



**Figure 5-6: Trust Zone integrated with edge cloud V-AAA server and the 3GPP architecture for the 5G system [23.501]**

### 5.5.2 Secured access management transferring

As TZ is supposed to cognitively and flexibly invoke security functions in different domains with respect to the EC4 state, the management of UE access will be transferred between the central cloud and the edge cloud when the TZ state switches. However, due to the incomplete set of security functionalities in the edge cloud (local V-AAA servers) in comparison to the central cloud (V-AAA managers), the edge cloud security functions are usually less secured than the central cloud ones, and therefore makes the local TZ an easier target to attack than the central cloud. This brings a risk that attackers may initiate attacks to disconnect the central cloud and the edge cloud, hack the local TZ, and eventually try to obtain access to the central cloud during the reconnection, when the edge cloud hands its security functions back over to the central cloud.

To mitigate this risk while providing a high availability of network services, an asymmetric approach of transferring the access management between the central cloud and the TZ has been designed and reported in [Han17], as follows:

- When a disconnection takes place (state *D*), the ZM considers all UEs that have already been authenticated as trusted devices. These devices are able to retain maximal access to the TZ according to the respective policy under the current situation, until they lose their connections to the edge cloud.
- When a UE tries to access the TZ and the central cloud is unavailable (state *L*), the ZM invocates the LAA and the LSS to gain an access for the UE. If the subscriber data of the UE is available in LSS and the security check is passed, the UE can be considered as a trusted device. Otherwise, it remains untrusted and is only granted the basic emergency services.
- When the edge cloud is reconnected to the central cloud (state *R*), the ZM disconnects all locally authenticated and authorized UEs in a prescheduled order, so that the UEs have to be re-authenticated and reauthorized by the central security server, in order to regain full access. The UEs that have been authenticated and authorized by the central security server before the disconnection do not have to reconnect.

With this mechanism, emergency services are ensured to remain available for all users, edge cloud services are as much attainable as possible for legal users, while fake devices are prevented from accessing the central cloud.

### 5.5.3 Impact on the network performance requirements

As the local TZs can only derive security keys generated in the central cloud, but never locally generate new keys, user information must be prepared in the edge cloud before a possible disconnection or limited connection occurs. This user profile synchronization generates extra data traffic and hence raises the requirement of backhaul networks.

Consider a reference scenarios of the V2X communication use case: a TZ is implemented at an intersection for V2X services, which rely little on servers out of the local edge cloud. However, if the EC4 is disconnected or seriously limited, devices entering the local edge cloud may fail to be authenticated or authorized, and thus unable to exploit the V2X services. Therefore, the user profiles of devices in neighbour edge clouds must be earlier synchronized to the local TZ, depending on the mobility model and expectation of EC4 state in the future. Depending on the amount of prepared user profiles and the frequency of updating, a trade-off is taken between the generated extra backhaul traffic and the availability of V2X services. To optimize the QoS and QoE, a statistical model of the EC4 quality, a mobility model of the UEs and the topology of edge clouds will be needed to adapt the synchronization preferences.

## 5.6 Summary and conclusion

By building on new networking paradigms such as NFV and SDN, and by introducing multi-tenant, multi-service and multi-connectivity concepts, the 5G NORMA architecture doubtlessly introduces new risks into mobile networks. We have analysed these risks carefully and proposed ways how to mitigate them. We have further investigated and specified a number of innovative security approaches that can be integrated into the new architecture (although they are mostly applicable also in more general contexts): Virtualised AAA, tokenization technique for provisioning and deployment, a new AS security approach supporting flexible allocation of RAN security functions and the Trust Zone approach. We are confident that the proposed measures, applied carefully, together with other relevant security measures that are not in the focus of the project (as a complete coverage would require an effort at much larger scale), will result in highly secure networks that comply with the challenging security requirements raised by the expected 5G use cases.

# 6 Architecture Design Verification

In order to make sure that 5G NORMA architecture design meets requirements of use cases and stakeholders, the iterative architecture design process has been accompanied by quantitative and qualitative evaluations conducted in close cooperation between WP2 and WP3. For this purpose, WP2 defined use cases and KPIs [5GN-D21] in an early project phase. Based on this, WP3 developed an evaluation concept that covers a broad range of evaluation criteria checking performance, functional, operational, security, and economic requirements [5GN-D31]. As the original use cases turned out to not be sufficient for testing requirements on flexibility and adaptiveness, evaluation cases for a baseline, a multi-tenant and a multi-service network have been introduced [5GN-D32]. Intermediate verification results have been compiled in [5GN-D22] (economic perspective) and [5GN-D32] (technical perspective). This chapter compiles final results of architecture design verification from a technical point of view (WP3 perspective). Final results of economic evaluations will be published in [5GN-D23].

Section 6.1 sets the scene by providing a brush up of 5G NORMA verification methodology. In order to be more precise, use case and KPI definitions from [5GN-D21] have been updated and complemented by traffic demand descriptions in Section 6.2. Basic objectives and assumptions taken during elaboration of the three evaluation cases are described in Section 6.3. Technical verification results structured along evaluation cases and evaluation criteria are compiled in Section 6.4. Finally, Section 6.5 concludes on the most important verification results. We refrain from reproducing results from earlier deliverables instead references to passages discussing different evaluation topics including former deliverables are given in Annex A.1.2.

## 6.1 Brush up on methodology

The basic objective of architecture design verification is, in conjunction with the techno- and socio-economic evaluations in WP2 as well as demonstrator development in WP6, to contribute to a proof of concept of the 5G NORMA architecture design and its key innovations:

- Multi-service- and context-aware adaptation of network functions,
- Mobile network multi-tenancy,
- Adaptive (de)composition and allocation of mobile network functions,
- Software-Defined Mobile network Control (SDMC),
- Joint optimization of mobile access and core network functions.

For this purpose, different deployments of the 5G NORMA system in a London study area have been emulated and denoted in evaluation cases [5GN-D32].

Use cases and traffic models providing performance requirements and also functional requirements defined at the beginning of the project in [5GN-D21] have been updated by WP2. Descriptions covering all components of the generic 5G services, namely

- extended Mobile Broadband (eMBB),
- massive Machine Type Communication (mMTC) and
- Vehicular-to-Infrastructure (V2I), as an example for ultra-reliable communications (uMTC)

are summarized in Section 6.2.1.

The service selection reflects results of socio-economic evaluations making sure that most valuable services from revenue or socio-economic benefit perspective are taken prioritised. Besides quantitative and qualitative requirements, service definitions also include traffic models so that the network behaviour can be studied under realistic usage assumptions in the study area.

Evaluation criteria depicted in Figure 6-1 have been derived from an overall KPI list and grouped into sub-categories of similar features that can be discussed jointly.

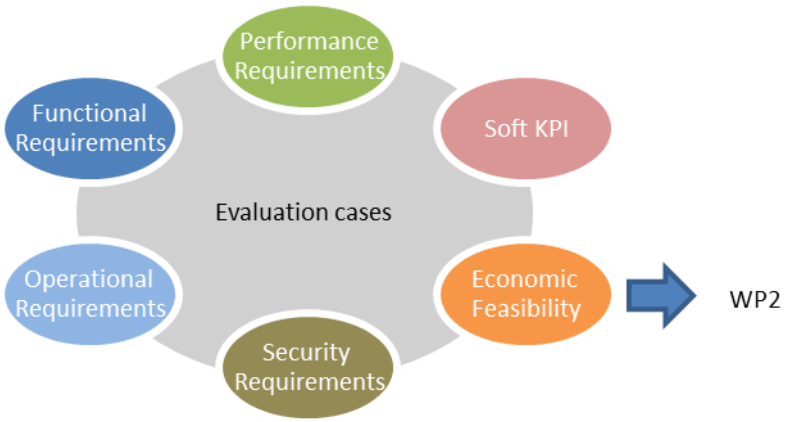



**Figure 6-1: 5G NORMA evaluation criteria**

In the following, the sources of requirements and KPIs are listed:

- Performance requirements are driven by the selected services and compiled in Section 6.2.1.
- Functional network requirements have been taken from the use cases eMBB, mMTC and V2I in [5GN-D21]. They have been adapted to selected services as well (Section 6.2.2).
- Operational requirements summarize the operational needs from a stakeholder point of view and have been formulated in [5GN-D31]. Some additional topics have been added (cf. Figure 6-3).
- Security requirement definition on top of LTE shall make sure that the network design allows for protection against cyber-attacks. The requirements identified in [5GN-D31] have already been addressed [5GN-D32].
- Soft-KPIs, in a qualitative way, measure the feasibility of envisioned network flexibility, complexity and standardisation effort. More generally, by checking of this soft-KPIs, it shall be ensured that the architecture design provides mature results that can be handed over to next step realisation activities. A couple of topics have already been addressed in [5GN-D32]. The discussion is continued in here (cf. Figure 6-3).
- Investigation of economic feasibility will be covered by WP2 in the upcoming deliverable [5GN-D23]. The definition of overarching evaluation assumptions has assured that technical and economic evaluations fit to common assumptions.

The investigation of some service-related performance requirements (e.g. user throughput, cell edge throughput, mobility, peak data rates and feasible device density) has been out of scope of 5G NORMA. Results may be taken from other R&D projects, e.g., from the EU H2020 programme.

An emulated network roll-out of base station sites and edge clouds in a London study area is depicted in Figure 6-2. By this tangible application, project results shall be presented in an overall context, technical challenges shall be identified and the capability of the 5G NORMA architecture to cope with them shall be checked.

In order to provide test scenarios for 5G NORMA key innovations, three so-called evaluation cases have been defined

(1) Baseline case
(2) Multi-tenant case and
(3) Multi-service case

Evaluation case descriptions have already been provided by former deliverables (e.g. [5GN-D32]). They basically include investigation objectives, roll-out assumption, and verification topics to be studied. They build a common basis for technical evaluations in WP3 and socio-economic evaluations in WP2.

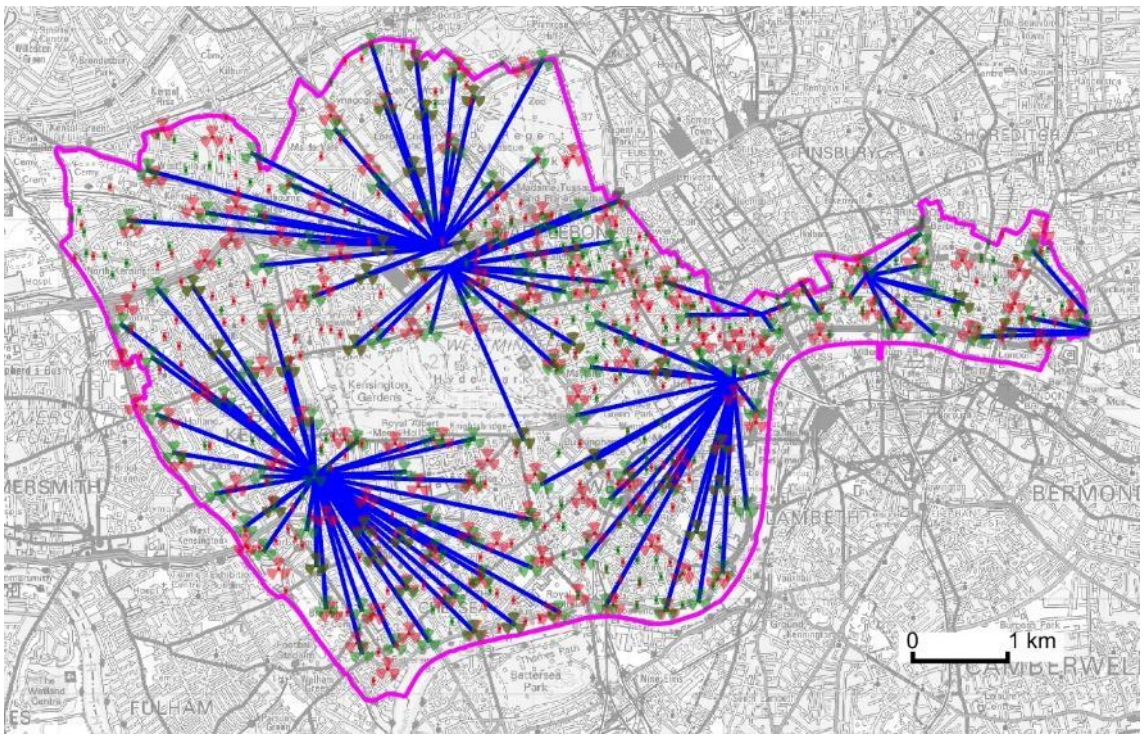



**Figure 6-2: Location of edge clouds and antenna sites in the London study area**

An overview of verification topics and mapping to respective evaluation cases and criteria is depicted in Figure 6-3. Most of the verification topics discuss fulfilment of requirements in a qualitative way.

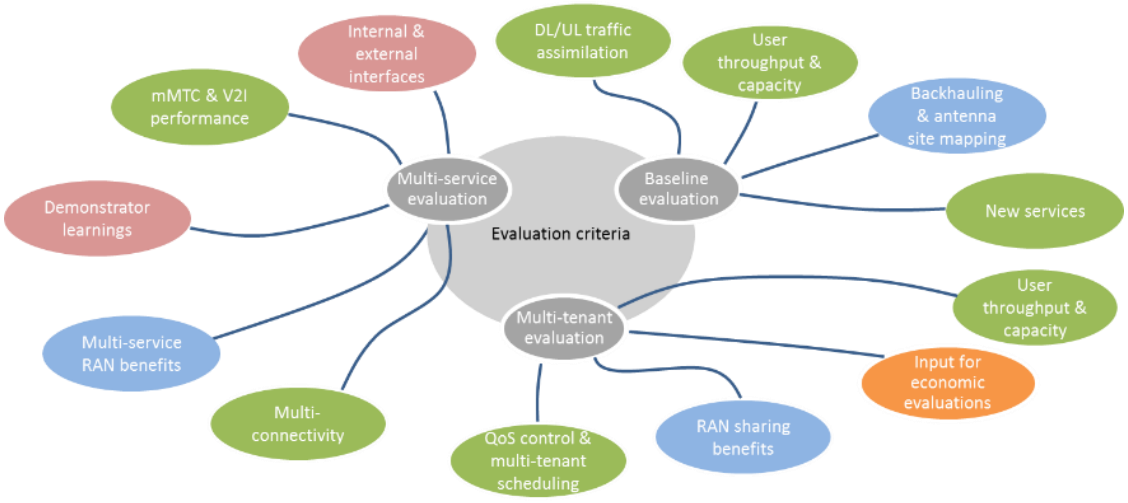



**Figure 6-3: Verification topics and mapping to evaluation cases and evaluation criteria**

Architecture design verification done by WP3 highlights deployment topics from a technical point of view. To provide the whole picture interrelations, with techno- and socio-economic analysis done by WP2 have to be considered.

- The service definition described below build the common verification basis for both views. All discussions in this chapter are based on service-related requirements defined herein.
- Roll-out assumptions like spectrum deployment, radio node density, transport network availability, functional split options, and HetNet deployment are carefully aligned with WP2.
- The discussion of architectural integration of service enablers like massive MIMO (M-MIMO), 5G NORMA multi-tenant MANO Layer, RAN sharing, multi-connectivity, etc. may impact cost calculations significantly. Results from technical verification compiled in Section 6.4 may, as much as possible, be considered in the final WP2 outcome on socio-economic evaluation of architecture design [5GN-D23].

The results of architecture design verification are presented with three layers of detail:

- The main part, covering all verification topics of Figure 6-3, can be found in Section 6.4.
- To improve readability, detailed assumptions, descriptions of verification tools, and detailed analyses have been shifted to Annex A.
- A high-level summary is provided in Section 6.5.

## 6.2 Services and related requirements

### 6.2.1 Service definition

Evaluation case 3 focuses on the multi- service benefits of the 5G NORMA architecture. This evaluation case examines the full range of services, thus requiring high throughput (eMBB), high device densities but small packet sizes (mMTC) and high reliability (uMTC). A number of specific example services covering these three broader categories of service classes have been selected to assess KPIs in the context of the central London study area. The example services have been selected based on:

- Being applicable to the central London study area
- Showing promise for generating significant incremental revenues and or social benefits from the on-going techno-economic assessment in WP2.

Considered example services include:

- eMBB for consumer portable devices (such as smartphones, tablets and laptops) consisting of:
  - Enhanced MBB for up to 4k live streaming of video
  - Extreme MBB for AR applications
- V2I
  - Infotainment and advertising to passengers
  - Information services on road and driving conditions and navigation
  - Assisted and automated driving services
- Smart cities
  - Environmental monitoring, road traffic congestion control, and waste management sensors
  - Smart metering and smart grids
- Logistics
  - Sensor data for tracking goods in transit

**eMBB for consumer portable devices**

eMBB for consumer portable devices considers enhanced mobile broadband to today's typical consumer portable devices such as smartphones, tablets and laptops. These services are included as the baseline "traditional" consumer demand which is most likely to drive the dimensioning of mobile networks going forward, with other services being added on top of this baseline. While

these devices consume a mix of voice, messaging, and data, the key applications driving demand tend to be video-based with video forecast to account for 75% of traffic by 2022 [ERI2016]. We assume that 4k video streaming will be a sufficient resolution for these small form factor devices in outdoor environments and also allow for challenging uplink throughput requirements to facilitate emerging apps such as live streaming [ERI2016]. Augmented and Virtual Reality (AR/VR) are frequently identified as extreme broadband applications requiring 5G for a consistent user experience. It is assumed that VR will be largely applicable to indoor environments[9]. However, there may be some situations where AR services would be required from an outdoor mobile network such as in the central London study area. Hence, these services are included in the technical evaluations as well[10].

**Vehicle to Infrastructure communications**

The following services are part of the vehicle-to-infrastructure services:

- Infotainment and advertising to passengers,
- Non-critical driver information services on road and driving conditions and navigation,
- Assisted and automated driving services.

Infotainment includes providing a cellular connection to the vehicle which then may be distributed locally in the vehicle to passengers via Wi-Fi. Presumably, requirements for this service will be driven by demand for high resolution video as in the case of eMBB services. The service definition is dimensioned for providing 4k video streaming to up to three passengers per vehicle (not applicable to drivers).

Driver information services on road and driving conditions are mMTC applications where the information passed between the vehicle and roadside infrastructure does not have a critical component, but is to inform the driver.

Assisted and automated driving services reflect connectivity requirements for different levels of autonomous driving and a move to more automated services over time. Fully autonomous driving is not expected to be realized before 2030. Still, it is anticipated that messages associated with assisted or semi-automated driving will constitute high priority and high reliability services. Hence, they are classified as uMTC.

The potential new revenue streams from vehicular services and the forecasted significant social benefits make these interesting example services to consider in the 5G NORMA assessment.

**Smart cities**

Since *Smart city* is a concept in a rather early stage, the set of required/associated communications services is still evolving. Services cover both business to business and business to consumer services. For example, the GSMA smart cities report includes: smart meters, electric vehicle charging, microgeneration of electricity, intelligent transport systems (ITS), smart parking, smart waste and water management (some of which could come under smart grid), street lighting, environmental monitoring, congestion charging, and road tolls. Typically, smart city applications are considered as examples of mMTC connection requirements although a limited amount of smart grid signalling for controlling the main electricity distribution networks could fall under uMTC.

The on-going WP2 economic assessment has identified significant social value from smart city services such as smart grids and metering. WP2 anticipates that direct revenues from smart city

[9] In contrast to Section 4.2 where industrial communications have been described, the verification analysis focuses on outdoor services.

[10] AR is not directly considered in the cost modelling within WP2 since it will be served by localised mmWave cells which are beyond the scope of the current techno-economic model.

services may be limited particularly when considering the incremental revenues over what could be delivered with existing LTE systems. However, the ability of 5G NORMA to support network slicing and enable the rapid roll-out and trial of these evolving services without the costly roll-out of proprietary networks could bring their deployment and any potential revenues from these services earlier than expected.

**Logistics**

Finally, sensor data for tracking goods in transit is selected as another case of mMTC traffic requiring an outdoor mobile network. We consider the message sizes from goods in transit would be small (like the smart city applications), i.e., in the order of 200 bytes but transmitted reasonably frequently and throughout the day. Further, their number scales with the volume of commercial vehicles in an area.

Potential incremental revenues and social benefits are still to be investigated for this case but WP2 anticipates they may be significant due to more efficient delivery of goods.

Potentially, *Logistics* could be extended to monitoring of driver performance for compliance with safety regulations.

The defined services map to the three services classes of eMBB, mMTC and uMTC as shown in Table 6-1.

**Table 6-1: Categorisation of example services to service classes**

| | **eMBB for consumer portable devices** | | **Vehicle to infrastructure** | | | **Smart cities** | | **Logistics** |
|---|---|---|---|---|---|---|---|---|
| | *Enhanced MBB - Up to 4k streaming* | *Extreme MBB - AR/VR* | *Infotainment* | *Information services* | *Assisted driving* | *Environ-mental monitor-ing, ITS, and waste manage-ment* | *Smart energy* | *Tracking goods* |
| **eMBB** | X | X | X | | | | | |
| **mMTC** | | | | X | | X | X | X |
| **uMTC** | | | | | X | | | |

Performance, capacity, and coverage requirements are summarised for the example services under each of the three service categories in A.1.2.

## 6.2.2 Qualitative evaluation criteria

Based on the service descriptions in the previous section where mainly performance requirements from a user perspective have been introduced, this section continues with a selection of quantitative and qualitative requirements regarding the other evaluation criteria. Links to sections where fulfilment of these requirements are discussed are given in Annex A.1.2.

### 6.2.2.1 Functional requirements

For the service components described above the following functional requirements have been identified:

**eMBB social media and V2I infotainment**

- *Application awareness:* The network will expose its capabilities to other parties through a set of open APIs, allowing different provider business models to be implementable (e.g. XaaS). The network will implement application awareness for OTT.
- *Multi-layer and multi-RAT connectivity*: The network should provide multi connectivity in order to improve user throughput (user plane aggregation) or coverage and reliability (user plane diversity)
- *Efficient backhaul:* The network should be able to adapt function placement and user data flows according to availability of transport technology performance.
- *User privacy and security:* User privacy and security is required at least at the level provided in LTE, and should be enhanced by options for even better protection (e.g. "IMSI-catching" protection). While security is important for mobile broadband, it is not in the main focus of this use case.
- *Capacity for uplink and downlink:* Capacity for uplink and downlink can be flexibly allocated and optimized on cell and sector level based on just-in-time user requirements and used applications.

**V2I – assisted driving (uMTC)**

- *Fast and targeted dissemination of safety messages*: safety messages shall be transmitted with high reliability (cf. Section 6.2.1)
- *Optimizations for control plane and data plane functions:* The system should enable optimizations for control plane and data plane functions such as optimal routing and handover minimization
- The system should guarantee the coexistence of safety and non-safety vehicular applications operating over the same scenario.
- Very high network availability and therefore superior robustness against attacks, in particular DoS attacks, is required. This includes strong authentication between devices and network in order to prevent unauthorized communication. Moreover, integrity protection and encryption is required for the signalling traffic and – unless the applications build on application layer security mechanisms – also for the user plane. Security mechanisms must be robust against loss of network nodes; security mechanisms must be available also in RAN parts that are isolated from central components. Security aspects are of high importance for the use case.

**V2I – driver information service (mMTC)**

- The system should guarantee the coexistence of safety and non-safety vehicular applications operating over the same scenario
- The mobility management should support stationary, nomadic, and highly mobile devices and should consider also roaming across network boundaries.

**Environmental monitoring, waste management, and congestion control (mMTC)**

- Depending on device type the network access should be applicable via dedicated RATs and frequency bands or in a flexible way
- The mobility management should support stationary and nomadic devices and should consider also roaming across network boundaries.
- The system should support both unidirectional as well as bidirectional communication between sensors and other radio nodes.
- The network should provide flexible security and authentication procedures for mMTC as well as means for easy security credential provisioning for massive number and high density of devices.

**Smart meters - sensor data, meter readings, individual device consumption (mMTC)**

- Depending on device type the network access should be applicable via dedicated RATs and frequency bands or in a flexible way

- The mobility management should support stationary devices
- The system should support both unidirectional as well as bidirectional communication between sensors and other radio nodes.
- The network should provide flexible security and authentication procedures for mMTC as well as means for easy security credential provisioning for massive number of devices.
- The network should provide coverage for smart meters in difficult indoor locations like basements

**Smart grid sensor data and actuator commands (mMTC)**

- The 5G system should support appropriate authentication for low power devices/sensors.
- The 5G system should be able to accept unsolicited information from large numbers of sensor devices without the need for bearer establishment or mobility signalling (mobility signalling is not required because it is not "connected" to any particular node)
- The system should support an infrequent uplink data transfer in a "non-connected" mode
- The system should be able to deactivate the service and sensors, possibly for future use.

**Service-overarching functional requirements are (defined below)**

- Network programmability: Third parties shall be enabled to acquire network resource on-demand satisfying their individual SLAs. In addition, programmability shall enhance the user perceived QoE by customizing the network resource accordingly.
- QoE based routing: The network shall introduce flexible routing path in order to provide an additional degree of freedom for QoE improvements.
- Edge function mobility: The network shall allow demand oriented re-orchestration of service chains.
- Slice and service specific mobility concepts: The network shall allow for service specific and context aware adaptation of mobility concepts.

#### 6.2.2.2 Operational requirements

Operational requirements are not that much service related. They have already been checked in [5GN-D32] but are listed below for the sake of completeness.

**Multi-tenant dynamic resource allocation [5GN-D31]:** Network resources such as communication, storage, processing, and function resources are provided in a sliced manner based on different service requirements. This will be performed using a pool of resources, which are reserved for a given network slice to achieve specific performance goals. Thus, a resource optimization mechanism is required to optimally allocate resources optimizing metrics such as spectral efficiency or network energy consumption. Distinct tenant requests can result in different profits.

**Saving of operational and capital expenditures [5GN-D31]:** An important requirement is represented by OPEX and CAPEX reduction. Shared utilisation of resources and network equipment, e.g., to accommodate and balance tenants' capacity requests, helps to realise multiplexing gains and reducing costs significantly.

**Service specific and context-aware derivation of service requirements, adaptation, and placement of VNFs [5GN-D31]:** Requirements on QoE and QoS, mobility, security, or others must be considered dynamically based on selected services as well as network context:

- Flexible vertical-specific and service-specific detection of traffic and dynamic network monitoring [5GN-D31]
- Adaptation and placement of VNF [5GN-D31]
- Capability of spectrum sharing or reuse (NGMN)

#### 6.2.2.3 Security requirements

Seven security requirements have been defined in [5GN-D31]. A concise recapitulation is given below.

**Tenant isolation:** Ensure that tenants are restricted to their assigned resources and cannot attack other tenants by stealing their resources, modifying their resources, or modifying or reading any content held by these resources.

**Secure Software Defined Mobile Network Control:** The SDN controllers must distinguish application-authentication and -authorisation mechanisms for different application roles and respective permission classes for the different control-operations.

**Physical VNF separation:** Means must be provided that allow physical separation of VNFs without sacrificing the principle of flexible and efficient resource allocation.

**Flexible security:** Security procedures must adapt to the specific needs of a service or network slice.

**Support of reactive security controls:** It is required that the architecture allows the dynamic use of reactive security controls, i.e. means to detect possible security breaches and to react on them accordingly in an automatic way.

**Security orchestration:** The protection mechanisms and the security controls need to be rather dynamic, flexible and autonomous. For this purpose, security orchestration functions are needed.

**Reliable fallback:** The system must allow for forced reset and it must be guaranteed that compromised hardware can be reset.

The fulfilment of requirements above has already been checked in [5GN-D32].

#### 6.2.2.4 Soft KPI's

Soft-KPI that measure in a qualitative way the feasibility of envisioned network flexibility, complexity and standardisation effort have been defined with view on the three evaluation cases. Most important questions with respect to feasibility have been addressed.

**Interfaces between Service Management and Management and Orchestration:** The automatic interfaces between service and management & orchestration layer as well as between stakeholders shall allow for all required communication in terms of

- Exposure and sharing of slice templates
- Allowing the tenant to order a limited set of network slice operation such as network slice activation/stop/scaling (granted by the MSP Service Management)
- Allowing to interconnect VNFs across domains
- Allowing the tenant to enforce QoS/QoE change under the network slice operation (granted by the MSP Service Management)
- Allowing the tenant to provide subscribers information to be registered in the MSP HSS
- Reporting Accounting/Charging data
- Allowing tenant to design/compose a network slice either from a list of available VNF/VNF sub-graphs or by adding its own VNFs (certified for tenant's use)
- Allowing tenant to define set of service policy rules for the network slice operation
- Validating/authorizing the network slice design as well as service policies customised by the tenant
- Providing means to check and incorporate tenant's certified functions in the catalogue of VNF
- Allowing the tenant to order a limited set of network slice operation such as network slice/service activation and stop (this does not cover any network slice management operations such as NS scaling nor FCAPS)

- Exposure of monitored KPIs to the tenant (for changing policies settings)
- Allowing the tenant to change of service policies under network slice operation

**Scalability of centrally arranged management and control functions:** Orchestration functions must ensure that scalability of management and control functions can be checked before re-instantiation at more central locations within the network topology.

**Feasibility of growing number of slices:** Architecture design must ensure that bottlenecks of management and control functions appearing with growing number of slices can be detected sufficiently in advance.

**Roles of external & internal interfaces:** Standardisation effort must be held feasible by identification of mandatory and optional standardisation topics.

**Feasibility of C&LI:** Requirements for charging and lawful interception are defined in a service specific and context aware manner.

Requirements for service charging are listed in Table 6-2.

**Table 6-2: Requirements for service charging**

| Service | Charging |
|---|---|
| eMBB – consumer portable devices | Subscription based charges either unlimited or up to various monthly data limits (in GB) |
| V2I – infotainment (eMBB) | Subscription based charges either unlimited or up to various monthly data limits (in GB) |
| V2I – assisted driving (uMTC) | Subscription charge (vehicle makers) for network slice with taking into account number of devices and data usage OR flat fee per embedded device registered to auto maker. |
| V2I – driver information service (mMTC) | End user subscription charge – unlimited or based on data usage up to specified monthly limits |
| Environmental monitoring, waste management, and congestion control (mMTC) | Fixed fee contract, potentially multi-year OR periodic subscription charge based on number of sensors. |
| Smart meters - sensor data, meter readings, individual device consumption (mMTC) | Fixed fee contract, potentially multi-year OR periodic subscription charge based on number of sensors. |
| Smart grid sensor data and actuator commands (mMTC) | Subscription (energy company) for network slice taking into account number of devices and data usage |
| Logistics sensor data for tracking goods (mMTC) | Subscription (logistics application provider) taking into account number of devices and potentially data usage. The customer may or may not be a tenant. |

## 6.3 Evaluation cases

High level descriptions of 5G NORMA evaluation cases are given in [5GN-D32]. In the following objectives of the different evaluation cases are briefly replicated. An overview of discussed topics is given in Figure 6-3.

### 6.3.1 Baseline evaluation

Objectives of baseline evaluation are

- to provide a baseline economic cost case evaluation for (e)MBB services deploying legacy LTE-A Pro technologies within the years 2020 to 2030 with final results in [5GN-D23],
- to compare this legacy cost case with the deployment of 5G technologies, identifying most important differences in case of single operator networks with final results in [5GN-D23],
- to check performance, functional, and operational conditions in the London study area against respective requirements for (e)MBB originating from [5GN-D21] and Section 6.2 (covered in Section 6.4.1)
- to establish a baseline C-RAN vs. D-RAN cost basis, i.e., to identify possible cost penalties when deploying and running a 5G NORMA network [5GN-D23].

A detailed description of baseline assumptions is already available in [5GN-D32]. For single operator 5G NORMA networks, the mobile network operator (MNO) is assumed to own all infrastructure including RAN and edge as well as central clouds. The focus of the baseline evaluation case is on fulfilment of eMBB performance, back- / X-haul as well as mapping of antenna sites to edge clouds. A baseline topological view depicted in Figure 6-4. Function placement for a baseline network can be described as follows:

- Core network functions are placed at central could.
- Required MANO functions and control functions (incl. SDM-C) are placed at the central cloud. SDM-O can be skipped for baseline.
- In order to improve multi-connectivity performance (PDCP split bearer), PDCP functionality is placed at the edge clouds (central units, CU) whereas the lower part of RAN protocol is placed at the antenna sites (distributed units, DU). This applies to macro as well as to small cells at low and med. bands. Mid-haul split is arranged according to option 2 of [38.801] (s. Figure A-1)
- For M-MIMO at macro sites (up to 64 virtual antennas) as well as mmW small cells the same mid-haul split option may be applied.
- In order to provide sufficient user mobility at mmW small cells initially at least 3 nodes per coverage area have to be deployed.

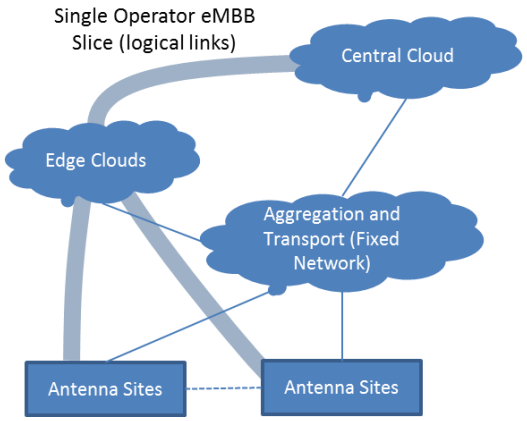



**Figure 6-4: Baseline topological view**

### 6.3.2 Multi-tenant evaluation

The multi-tenant evaluation case extends the base line network roll-out for use by multiple operators. Objectives of multi-tenant evaluations are

- building on the 5G NORMA network considered in the baseline evaluation, compare the cost and deployment of MBB networks for single and multi-operators 5G NORMA networks with final results in [5GN-D23],
- identifying key benefits of multi-operator deployments (performance, operational and functional view), mostly covered in Section 6.4.2.
- checking for suitability of service and management / orchestration layers for multi-tenancy applying the roles described in [5GN-D32] and Section 4.1.

Instead of one mobile network operator owning all physical and cloud infrastructure resources for evaluation, Offer Type 3 is selected as described in Section 4.1. The mobile service provider is offering mobile services to tenants (former MNOs) that own their RAN infrastructures including spectrum resources. Hence these former MNO change their roles into RAN InP, MSP and tenants. Tenants rely on an SLA with one or more MSP for provisioning of their end-to-end eMBB slices. The MSPs may have multiple SLAs with infrastructure providers (for both cloud and RAN infrastructure). Tenants would have to rent mobile services from the MSPs but own their software including element managers, service management, OSS, and other MANO functions (except VIMs). Tenants orchestrate the telecommunication services (commission, operate, decommission) for its own business. The evaluation assumes up to four RAN-InP as well as up to four tenants that share the services provided by the MSPs. Transport network is provided by SLA's with fixed network operators.

A topological view for a 5G NORMA multi-tenant network is depicted in Figure 6-5. The function placement within the topology can be described as follows:

- Dedicated core network functions for each tenant are placed at central cloud.
- Required functions of management and orchestration layer and as well as control layer (SDM-C, SDM-X, common and dedicated control applications) are placed at the central cloud.
- In order to improve multi-connectivity performance by PDCP split bearer similar to the baseline case, PDCP functionality is placed at the edge clouds whereas the lower part of RAN protocol is placed at the antenna sites. This applies to macro as well as to small cells at low and medium and high bands.
- In case of collocated sites hosting multiple former MNO due to selection of functional split option 2 (Figure A-1) sharing of transport infrastructure between edge cloud and antenna site as well as antenna panels is possible.

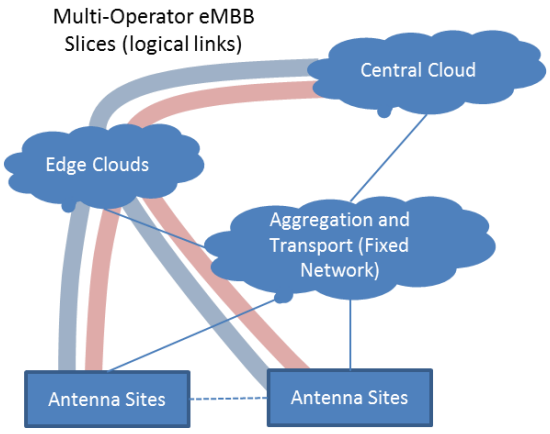



**Figure 6-5: Multi-tenant topological view.**

## 6.3.3 Multi-service evaluation

The multi-service evaluation case extends the network investigated in the baseline case by the two additional generic services mMTC and V2I. Hence, the evaluation spans the whole space of 5G services.

Objectives of the multi-service evaluation case are (results are compiled in Section 6.4.3)

- Feasibility check for mMTC and V2I performance requirements.
- Demonstration of multi-connectivity potentials.
- Investigation of architectural aspects including multi-service MANO layer, security, protocols, and interfaces.
- Identification of key elements for multi-service cost benefits.

Stakeholder roles are slightly different compared to the multi-tenant evaluation case:

- MSP(s) and RAN InP(s) are subsidiaries of the MNO(s)
- Tenants are either a subsidiary of MNOs (eMBB) or verticals (mMTC and V2I)
- Cloud and transport network InPs provide services based on SLAs.

The tenants (verticals) may prefer Offer Type 1 (cf. Section 4.1) where the MSP operates the slices on behalf of the tenants. For reason of coverage and reliability improvement, for multi-service networks, MSPs are assumed to have SLAs with multiple RAN InPs.

In order to achieve the required coverage and availability for

- environmental monitoring, waste management, and congestion control (mMTC),
- smart meters - sensor data, meter readings, individual device consumption (mMTC),
- smart grid,
- sensor data and actuator commands (mMTC), and
- Logistics - sensor data for tracking goods (mMTC)

the mMTC network slice is deployed at sub 1 GHz frequency bands. With RAN slicing Option 2, part of the spectrum of the different RAN InPs may be used jointly based on slice specific parameterisation. The same applies for part of the V2I slice namely

- V2I – assisted driving (uMTC)
- V2I – driver information service (mMTC).

V2I – infotainment (eMBB), which is belonging to the same slice as the services above, may be mixed to spectrum at low and medium frequency bands used even by the eMBB slice. Due to the fact that mobility is higher as in the usual eMBB cases small cells at mmWave frequency bands are excluded from the V2I slice. Spectrum might be split into RAN InP owned spectrum and spectrum given to a pool that jointly can be used (owned by the joint venture).

A topological view for a 5G NORMA multi-service network is depicted in Figure 6-6. Function placement in a multi-service network is more detailed in Section 7.2.1.

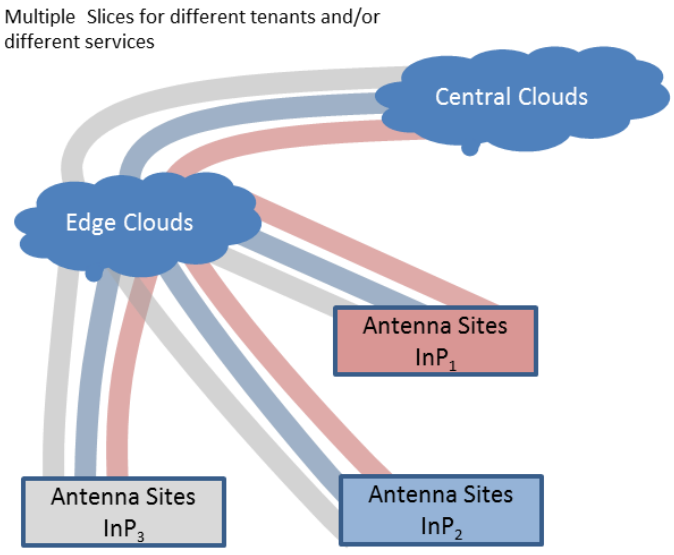



**Figure 6-6: Multi-service topological view**

# 6.4 Verification results

## 6.4.1 Baseline evaluation

### 6.4.1.1 Performance requirements

Fulfilment of eMBB performance requirements have already been checked in [5GN-D32]. The performance requirement check is updated and complemented by consideration of the latest service definition (cf. Section 6.2.1). Analysing these service definitions from a performance perspective the following topics have to be discussed:

- DL/UL traffic demand assimilation
- User Experienced Data Rates
- eMBB traffic demand
- New services

#### 6.4.1.1.1 DL/UL traffic demand assimilation

Challenging from today's perspective will be an increasing assimilation of downlink and uplink traffic. Besides the fact that more and more applications (i.e. video conferencing) will contribute to increased UL traffic demand this demand may be temporarily and locally very fluctuating.

Currently capacity considerations concentrate mainly on the downlink. Due to low device transmit power, link budgets for DL and UL are rather asymmetric hence UL capacity will be much more difficult to provide. In addition, FDD with fixed assignment of DL and UL spectrum allows only limited flexibility.

The London roll-out emulation applies 20 MHz unpaired spectrum at medium frequency bands for macro TDD operation. As M-MIMO is applied with TDD and for economic reasons with preference is to be deployed at macro sites it will not only be an important enabler for high spectrum efficiency (64x4[11] three time as high spectrum efficiency compared to 4x4) combined with low to medium mobility it also contributes to improved assimilation of UL/DL capacity.

Even small cells at unpaired medium (40 MHz spectrum available per operator) and high frequency bands (100 MHz spectrum available per operator) will enable more flexible adaptations of the network to changing DL and UL demand. But of course, provisioning of more UL capacity comes at cost of DL capacity.

In this context, the virtual cell concept described in [5GN-D42] not only improves cell edge performance by adapting virtual cell frames of neighbouring cells it also allows adaptation of DL/UL resources to changing traffic demand.

#### 6.4.1.1.2 User Experienced Data Rates

Data rate requirements are expressed in terms of user experienced data rate, measured in bit/s at the application layer. The required user experienced data rate should be available in at least 95% of the locations (including at the cell-edge) for at least 95% of the time within the considered environment.

Even if quite moderate realizing User Experienced Data Rates of 10 Mbps DL/UL (s. Table A-2) over big parts of the service area is not easy to achieve with today's LTE technology. Virtual cells including multi-cell coordination, as described in Section 6 of [5GN-D42], will enable steadier user experience as they allow for shifting radio resources from one cell to another if locally required. Complementing the virtual cell concept, the QoS framework introduced in Section 6 of

[11] First figure = number of transmit antenna ports, second figure = number of UE antennas

[5GN-D51] will provide the needed control mechanism in order to initiate resource shifting between cells. In addition, interaction between SDM-O, SDM-C and SDM-X provides a flexible means of shared resource management that allows for improved interference management as well as resource allocation and hence improves user experience.

Round trip times in current fixed national operator networks exhibit a big spread. One reason may be plenty of idle – active state transitions of current network protocols that introduce non-predictable additional latency. Another reason is that e2e network paths are variable and cannot be tailored with respect to steady latency conditions. By allocating core network functions (P-GW, S-GW, etc.) in edge clouds future 5G networks will enable more steady and lower round trip times and e2e latencies. Economic considerations have to balance performance benefit and additional costs arising from those service improvements. However, cost modelling in 5G NORMA will focus on user experienced data rates.

#### 6.4.1.1.3 eMBB (DL) traffic demand

eMBB daily traffic demand for 2030 has been estimated to be $t_D$ = 2.85 GB per device. Device density in London $d_D$ may reach at the same time 47,000 per km$^2$. Traffic density T is computed as follows:

$$T = t_D \, d_D \, f_S / t \qquad \text{(Eq. 6-1)}$$

Assuming that a (service-specific) fraction of $f_S$= 1/7 of this data volume is to be processed in the network busy hour (t = 3600 s) and applying Eq. 6-1, traffic density becomes T ≈ 42.5 Gbit/s per km$^2$. In order to get an economically viable solution, the capacity calculations performed in [5GN-D32] assumed (for high traffic area scenarios) an average macro site inter-side distance of 200 m, three macro sectors per site as well as five small cells < 6GHz and twelve small cells > 6GHz. The resulting capacity density has been in the range of 83 Gbit/s per km$^2$ which proofs the feasibility of the MBB service definitions documented in Table A-2.

#### 6.4.1.1.4 New services

Use case business models investigated by WP2 in [5GN-D22] reveal that besides improved performance new services like virtual or augmented reality (VR/AR) may enable increased ARPU. Those services however will require user experienced data rates in the range of 50 Mbps. Even with increased flexibility in 5G networks it will be challenging to offer such conditions in an area covering manner. We can assume that AR/VR will be feasible outdoors at selected points of interest (POI). At those POI operators may deploy nodes at high frequency bands. In order to enable low mobility in traffic hot spots at least three nodes should be clustered providing PDCP level multi-connectivity [5GN-D42] and providing extreme MBB services.

### 6.4.1.2 Operational requirements

#### 6.4.1.2.1 Backhaul aspects

To enable the most challenging services targeted by 5G, networks are being specified so that they will permit to increase performance over current mobile networks. Increased performance capabilities will lead to a similar increase in requirements for the transport networks. Capacity-wise, backhaul pipes will need to grow (tens of Gbps optical and wireless e.g. at mmWave frequencies). Latency-wise, the backhaul needs fast and resilient forwarding as well as intelligent leveraging of the edge networking and computing infrastructure. Cost-wise, there is a need for using less fibres where possible (e.g. DWDM and/or fibre-like wireless). Programmability of the backhaul control and its centralization becomes essential in order to add flexibility and agility for shorter service deployment times and traffic-aware backhaul network management so that one can achieve higher (energy) efficiency.

**Backhaul and fronthaul requirements for heterogeneous functional splits**

The EU H2020 project 5G XHAUL [XHAUL-D23] derives theoretical results for data rate in the transport for different RAN functional splits, corresponding to 5G scenarios defined by 3GPP [38.913]. 5G XHAUL focuses on the evaluation of certain splits (defined as A, B, C and D by the project), considered as the most promising ones by the project, as shown in Figure 6-7. Calculation of required transport capacity is described in Annex A.2.1.1 (cf. Table A-9).

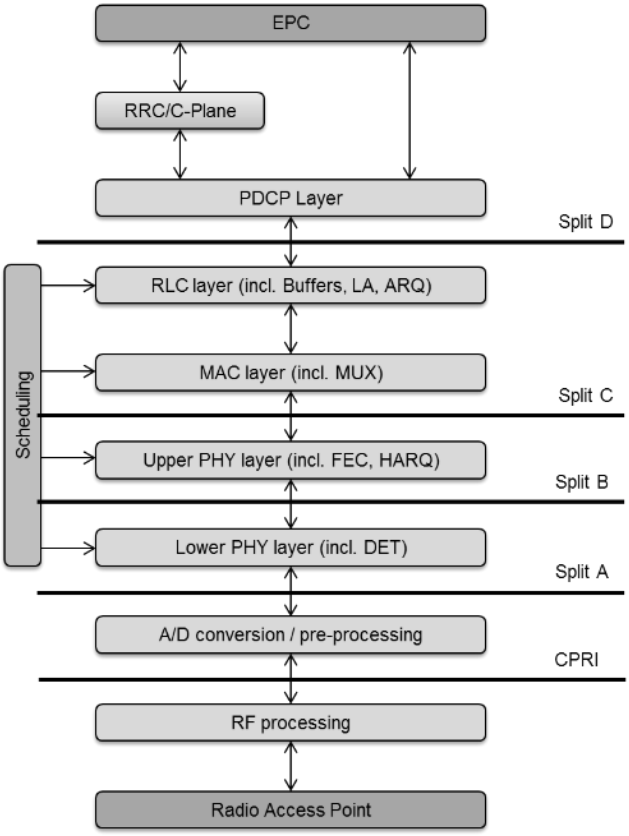



**Figure 6-7: 5G XHAUL Functional Splits**

As shown in Table 6-3 the most demanding 5G scenarios would require a fronthaul capability ranging from 5 Gbps to more than 5 Tbps in CPRI that not any current transport technology may currently provide cost-effectively.

The split to be eventually deployed will be conditioned by the characteristics of the transport network available (capacity, delay). 3GPP considers preferable that several options are supported to accommodate to the specific transport network characteristics.

It has been proposed that traditional CPRI is split [ZTE] into two separate entities CU (central unit, processing higher layer functions) and DU (distributed unit, processing lower layer functions), interconnected by midhaul, and one between DU and RRU (*fronthaul*).

**Table 6-3: 5G XHAUL backhaul data rates for several functional splits [XHAUL-D23]**

| Parameter | X-Haul Type | LTE | NR< 6GHz | NR>low mmW | NR>high mmW |
|---|---|---|---|---|---|
| Carrier freq. [GHz] | | 2 | 2 | 30 | 70 |
| Channel size [MHz] | | 20 | 100 | 250 | 500 |
| No. of Antennas | | 4 | 96 | 128 | 256 |

| | | | | | |
|---|---|---|---|---|---|
| Data rate CPRI [Gbps] | Fronthaul | 4.9 | 575 | 1533 | 5108 |
| Data rate Split A [Gbps] | Midhaul | 4.9 | 96 | 144 | 200 |
| Data rate Split B [Gbps] | Midhaul | 1.6 | 35 | 50 | 73 |
| Data rate Split C [Gbps] | Midhaul | 0.46 | 17 | 22 | 27 |
| Data rate Split D [Gbps] | Midhaul | 0.46 | 17 | 22 | 27 |
| Latency CPRI [ms] | Fronthaul | 0.07 | 0.07 | 0.07 | 0.07 |
| Latency Split A [ms] | Midhaul | 0.25 | 0.25 | 0.25 | 0.25 |
| Latency Split B [ms] | Midhaul | 0.25 | 0.25 | 0.25 | 0.25 |
| Latency Split C [ms] | Midhaul | 0.25 | 0.25 | 0.25 | 0.25 |
| Latency Split D [ms] | Midhaul | 1.5 | 1.5 | 1.5 | 1.5 |

**Opportunities for multiservice and mobile backhaul/fronthaul convergence**

Current C-RAN fronthaul deployments are based on microwave technologies (E-band) or fibre deployments. Current fibre deployments do not use any optical technology, but grey interfaces. As interface between baseband and remote radio units, *fronthaul* technology uses Common Public Radio Interface (CPRI) specification.

CPRI does not allow for any statistical multiplexing, aside including a limited number of different sectors of the same RRU, and is limited to point to point (dark-fibre) connections, synchronization and bandwidth being some of the technical reasons behind these limitations.

When the number of RRUs increase, the operator must deploy optical transmission to reach the BBU site. In 5G, targeted data rates imply the need for massive number of antenna elements (e.g. massive MIMO) and large spectrum e.g. mmWave in the access, leading to waveform samples for fronthaul that would lead to data rates exceeding 1 Tbps per macro-cell site (cf. Table 6-3).

The adoption of heterogeneous functional splits with reduced fronthaul bandwidth requirements as described before open the possibility of developing technologies allowing also for statistically multiplexing and coexistence with classical backhaul and the rest of the services in the network.

H2020 project Crosshaul [CROSS-D21] is developing such a 5G integrated fronthaul and backhaul transport network solution, enabling a flexible interconnection of the radio access with the core network by software-defined reconfiguration of all network elements.

In the standardization area, IEEE “Next Generation Fronthaul Interface” [NGFI] working group is defining a standard in which Ethernet, as a common switching protocol, is being used as the supporting technology for fronthaul, paving the way for a converged fronthaul and backhaul, i.e., a Crosshaul, featuring common data, control, and management planes. Also, CPRI is defining eCPRI, an open specification for 5G fronthaul with a new split point, announced to enable a ten-fold reduction of the required bandwidth, with Ethernet as underlying technology.

This possibility will be key for the success of 5G deployments, since it will enable the deployment of a single converged transport network, in which resources which are efficiently shared between 5G and other existing broadband services in the network with the consequent benefits in infrastructure investment and operational efficiency.

**Available technologies for fronthaul/backhaul**

Among the different currently existing technologies available (optical wireless, fibre, wireless, PON, xDSL, etc.), it can be assumed that fronthaul and backhaul will mainly consist of fibre deployments, complemented for specific scenarios (e.g. small cells deployments) with wireless technologies. An overview of available transport technologies including capacity and latency as well as maximum distance properties is given in Table 6-4.

**Table 6-4: Available transport technologies**

| Technology | Medium | Capacity | Latency | Max. Distance (1) |
|---|---|---|---|---|
| mmW E-Band | Wireless | 2.5-10 Gbps | ~ 50 µs/hop | 0.5 - 4 km |
| mmW V-Band | Wireless | 0.5-1 Gbps | ~ 50 µs/hop | <0.5 km |
| Sub 6 GHz | Wireless | ~ 0.2 Gbps | ~ 100 µs/hop | < 10 km |
| OTN | Optical fibre | N x 100 GbE<br>(N number of DWDM channels N<92) | ~ 40 µs/hop(2) | ~ 100 - 1000 km (4) |
| CWDM PON | Optical fibre P2MP (wavelength selective) | 16 x 25 GbE<br>16 x 50 GbE | ~ 0 µs (5) | 20 km |
| DWDM | Optical fibre | ~ 1Tbps | ~ 40 µs/hop(2) | 20 km |
| NG-PON2 TDM<br>NGPON2 PtPWDM | | 40G/40G<br>8x10G | < 1 ms<br>~ 0 µs (5) | 20 km |
| XGS-PON | Optical fibre P2MP (splitter) | 10G/10G | < 1 ms | 20 km |
| Air laser | Optical wireless | 1-10 Gbps | ~ 1 ms | ~ 1 km |

(1) Maximum distance depends on technology latency, and link budget. Physical medium propagation adds ~ 5µs/km. Jitter introduced by the technology also needs to be considered
(2) Due to G.709 wrapper. It is expected that 5G compatible solutions with lower latency are delivered in the short to medium term.
(3) 5G optimized solutions are expected to come up with 5G fronthaul compatible figures
(4) Depending on optical transmission technologies used for the optical link
(5) Just passive WDM. Upper layers latency (i.e. CPRI, eCPRI or FlexE) adds on top

**Opportunities for a convergent fixed-mobile deployment in the last mile**

To finish with the analysis, it is included a specific mention to the possibility of using already deployed broadband networks for the mobile backhaul, either based on copper or fibre

In general, copper networks will not satisfy the required bandwidth for mobile backhaul in 5G. Most of xDSL technologies do not satisfy the required bandwidth for mobile backhaul in LTE or 5G systems, although there might be an opportunity for the use of G.FAST [G.9701], able to

provide 150 – 300 Mbps DL in short links (50-100m) for the deployment of small cells backhaul. There could also be an opportunity of using cable networks [DOCSIS3.1] able to provide 10G/1G in a point to multipoint architecture for mobile backhaul or mid-haul of small cells, although there is no experience of its use in commercial networks for LTE backhaul [HFC].

In any case, latency, synchronization and jitter requirements for 5G mobile backhaul or mid-haul should be carefully checked in any case prior to the deployment of the solution. It should be also noted that the strong asymmetry DL/UL bandwidth of this solutions (typically 1:10) would greatly limit the applicability of the solution

Fibre technologies offer higher bandwidth and lower latency than copper, and are in general more suited to backhauling applications. However, for 5G scenarios, they are still somehow limited in bandwidth for the general 5G targeted scenarios, due its point to multipoint nature in which available bandwidth (2,5G/1,25G in GPON, and XGS-PON (10G/10G) would be shared between a number of fixed users and base stations.

Future NGPON2 [NGPON2] technology increases available bandwidth to 40Gbps symmetrical, which increases its applicability and includes an optional PtP WDM PON, able to map host OTU2 or CPRI option 7 (10G) line rates, which could represent a real alternative for small cells deployments of 5G backhaul in a convergent fixed-mobile architecture

#### 6.4.1.2.2 SDN & Network Slicing and antenna site to edge cloud mapping

The variety in service requirements for 5G and the necessity to create network slices on demand will also require an unprecedented flexibility in the transport networks, which will need to create dynamically connections between geographically distributed sites (likely across different network domains), network functions or even users, providing resource sharing and isolation.

The versatile consumption of resources and the distinct nature of the functions running on them can produce very variable traffic patterns on the networks, changing both the overlay service topology and the corresponding traffic demand. In order to adapt the network to the emergence of 5G services it is required the provision of capacity on demand through automatic elastic connectivity services in a scalable and cost-efficient way. Those requirements for flexibility and dynamicity across different network domains, along with the need for efficient consumption of resources, reinforces the demand for network programmability that transport networks already face.

SDN decouples network control and forwarding planes, and places control in a (logically) centralized controller. Northbound, SDN controller provides an API to higher layer control applications, abstracting network resources to them. Southbound, it controls connectivity of forwarding nodes, typically through their embedded SDN agents or NETCONF interfaces. Also, the centralized control plane capabilities provide e2e visibility of network resources for establishing and maintaining and optimized connectivity.

For many years, a combination of SDN and non-SDN enabled elements, physical and virtual elements will coexist. To facilitate an e2e view of the network, network resources need to be treated as generic resources, leaving the specifics of each technology to specific domain controllers during this coexistence period.

Figure 6-2 depicts the locations of edge clouds and bare metal nodes in the London study area. Exemplary the antenna sites of the green InP are connected to central office (CO) locations where roll-out of edge clouds might be possible. The chosen arrangement assumes that each of the antenna sites is mid-hauled (or backhauled) to the nearest edge cloud location. This of course is just an assumption that is near at hand but is not mandatory. In real 5G deployments, the controllers (SDM-C and SDM-X) might determine the service chaining dynamically in a context and load-aware manner. Hence, depending on the network slice or service, an antenna site may be connected to different edge cloud locations. However, due to their distributed nature, edge cloud locations have to be highly optimized and there is a need to have a certain minimum

compute power to be installed and an overall WAN orchestrator has to interact with the 5G NORMA SDM-O. In addition, it would be helpful if the InP of transport and edge clouds NFVI would be identical in order to avoid cumbersome gateways for demarcation [5GN-D52].

## 6.4.2 Multi-tenant evaluation

### 6.4.2.1 Performance requirements

#### 6.4.2.1.1 Network capacity and user experienced data rates in shared RANs

Network capacity by definition is measured in fully loaded network states. Hence the capacity of integrated RAN infrastructures would not be higher than the sum of the single RAN capacities that are limited by available spectrum, antenna site densities and spectrum efficiency of the air interfaces. However, if spectrum pools are commonly available to multiple tenants due to local traffic diversity, the maximum available spectrum can be virtually increased (Table 6-5). Elastic traffic (e.g. file download) in this case benefits from decreased air times. This lower air time consumption again results in less interference so that even non-elastic traffic benefits from virtually relaxed cell load. In total spectrum sharing enables to move the load operating point of the networks closer to the maximum load (capacity). Improved multi-connectivity due to integration of multiple RAN InP infrastructures will lead to increased user experienced data rates. In addition, increased multi-connectivity (compared to single tenant deployments) improves coverage probability and reliability.

#### 6.4.2.1.2 QoS control and multi-tenant scheduling

As described in Section 2.1, multi tenancy will allow to achieve a better network usage from two different perspectives: i) maximise the monetization of the network slice infrastructure market by accepting just those slices that are feasible with the available resources and ii) enforcing the QoE/QoS requirements set by a network slice with the available amount of resources. Figure 3-16 and Figure 3-17, respectively, show the advantages of multi tenancy algorithms for both perspectives, showing how naïve techniques can significantly deviate from the optimal behaviour.

### 6.4.2.2 Operational requirements for RAN sharing

In this section RAN sharing concepts enabled by 5G NORMA architectures are investigated. Starting with a description of state of the art sharing options and based on the assumptions defined by the multi-tenant evaluation case (Section 6.3.2) options for multi-operator full blown spectrum deployment are discussed. In that context, different stakeholder relations and resulting stakeholder requirements (e.g. spectrum, radio node management (SON), backhaul capacity for different RAN functional split options) are investigated.

RAN sharing side effects like improved multi-connectivity and virtual base station densification enabled by multi-operator side grid integration could be important enablers for mMTC and uMTC deployments.

#### 6.4.2.2.1 Current RAN sharing options

Today's network sharing is typified by different levels. Site sharing is the most common and is mostly motivated by lowering the cost. In urban areas in addition increasingly complex site acquisition can be avoided.

The main driver for RAN sharing – the next more complex option – is to reduce operational cost. Whereas site sharing allows certain degree of freedom to competing operators RAN sharing demands much more readiness for collaboration especially as it demands certain assimilation of operational mechanism and control. Operator may in addition share their core networks transmission rings and / or core network logical entities. Whereas European operator cost savings

due to RAN sharing are estimated in the range of 20%, core network sharing is not seen as providing substantial saving potentials [GSMA].

Current drivers for RAN sharing are quick roll-out for new entrants including cost savings even for incumbents. In addition, for urban areas, it is motivated by lack of suitable site locations.

#### 6.4.2.2.2 Potential future drivers for RAN sharing and related benefits

The drivers for RAN sharing described in the proceeding section of course will be still relevant in future networks. But due to rapidly *increasing mobile traffic demand* and *appearing new services* introducing additional requirements the list of future RAN sharing drivers might be extended.

New 5G services like mMTC and uMTC require improved network coverage, reliability and availability. Even these service requirements might be enabled by future RAN sharing concepts. Integration of multi-operator RAN infrastructures will enable decreased macro cell ranges and increase multi-connectivity options dramatically. Hence integration of RAN infrastructures will enable improved network coverage, reliability and availability and unlock the potential to deliver non-eMBB more niche services with potentially new revenue streams.

#### 6.4.2.2.3 5G NORMA enablers for RAN sharing

**RAN slicing**

5G NORMA introduces three *RAN slicing Options* described in Section 2.3.5. These options permit different degrees of RAN sharing. If tenants set great value on their unique selling points the slice-specific RAN (Option 1) would fit best (cf. Figure 2-10). The whole network slice down to the physical layer transmission point (PHY TP) can be customized individually. A critical point for macro antenna sites is the size and number of mounted antenna panels as well as the maximum radiated power that depending on the frequency band must not exceed a certain power flux density at locations where humans can be expected. Hence multiband antennas should be applied as much as possible and number of antenna elements should allow for reasonable form factors. An important advantage is that, in Option 1, the antennas can be operated jointly which alleviates fulfilment of EMF requirements and enlarges spectrum deployment limitations (cf. A.2.2.2). On the other hand, this option requires tight synchronisation of the common functions which lead to challenging requirements at the antenna sites and the achievable multiplexing gain is limited.

RAN slicing Option 2 provides sharing of PHY and MAC in the data layer and RRC in the control layer. This enables a reasonable compromise in terms of flexibility and complexity allowing for more multiplexing gains utilizing common resources (spectrum, compute, storage). As scheduling is performed under control of SDM-X, the spectrum of multiple RAN InP can be deployed in a flexible way by splitting it into InP specific and jointly used fractions to be done by slice specific parameterisation by an SDM-X multi-tenant scheduling app. By doing so, local multiplexing gains first of all with respect to the scarce spectrum resources can be achieved and user experience (user throughput) can be improved as well as headroom with respect to cell load can be reduced (Section 6.4.2.1.1).

Option 3 allows sharing of the complete RAN by multiple tenants. This case is similar to the ongoing discussion in 3GPP SA [23.799], which considers slicing only in core network and treats the RAN as a common resource. Compared to this 5G NORMA will allow for more flexibility, as with SDM-C and SDM-X RAN functionality may be more powerfully implemented as applications on top of the controllers. E.g., 3GPP LTE allows for connecting a UE to multiple PDNs while using the same Service GW. This is the main limitation, which is alleviated by 5G where each UE may be connected to more than one network slice (i.e., core network). Hence, this option would provide seamless migration and requires minimal changes to current standards.

**More flexible spectrum sharing**

In RAN slicing Options 2 and 3, spectrum sharing will be enabled by common MAC scheduling performed under control of the SDM-X. The common network functions realize sharing by multiplexing. Radio resources in this context can be assigned to different slices by slice specific parametrisation. Spectrum is assumed to be owned by the RAN InP (which might not be mandatory). Per RAN (InP), this yields (cf. [5GN-D32] and [5GN-D22]):

- 100 MHz macro spectrum, 80 MHz FDD @ low and medium frequencies, 20 MHz TDD @ medium frequencies
- 60 MHz TDD @ low and medium frequencies for SC
- 200 MHz TDD @ high bands for SC.

Upper limits for site specific spectrum deployments have been elaborated in Annex A.2.2.2.

- Typical macro sites 200 MHz
- Small cells < 6 GHz per node max. 60 MHz
- Small Cells >6 GHz per node max. 400 MHz

Already today antenna sites are jointly used by multiple operators. An assumption for the percentage of operator collocation in the London study area is given in Figure 6-8.

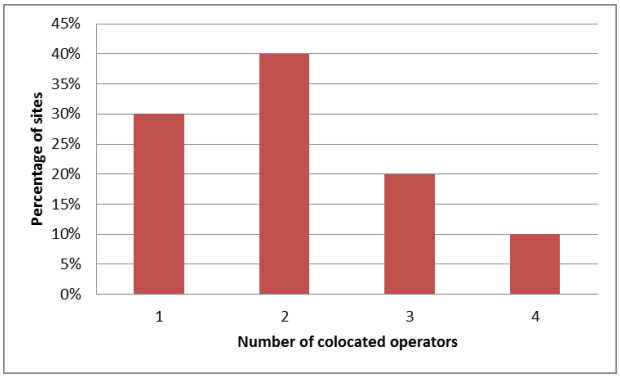



**Figure 6-8: Assumed distribution of macro site collocation in the London study area**

The fraction of the InP specific spectrum decreases with increasing number of joint site usage. If spectrum is fair distributed between InPs the resulting maximum spectrum deployment per macro site is given in Table 6-5.

**Table 6-5: Spectrum limitation at macro sites**

| Number of collocated RAN InP | Max. spectrum per InP [MHz]* |
|---|---|
| **1** | 100 (120) |
| **2** | 100 (120) |
| **3** | 70 (120) |
| **4** | 50 (120) |

*Number in brackets apply in case of spectrum pooling with 20 MHz joint spectrum.

As can be seen from the table above at sites shared by up to two RAN InPs each InP at typical sites may fully deploy his own spectrum (100 MHz). In case of spectrum sharing (20 MHz out of 400 MHz total spectrum applicable for macro sites) as long as the maximum usage of 200 MHz is not exceeded (EMF limit) the InP can provide up to 120 MHz spectrum offering to slices using his infrastructure. As in our London multi-tenant case RAN InP and tenant is assumed to have a one-to-one mapping this means that local traffic peaks of a single tenant may be intercepted by

use of additional 20 MHz from the spectrum pool. This of course will only be possible in case of multi-tenant traffic diversity[12].

Even sites that are shared by more than two RAN InP may benefit from local multi-tenant traffic diversity and the flexibility of multi-tenant dynamic spectrum assignment. If no spectrum sharing between InP is arranged single InP may offer their owned spectrum (100 MHz) as long as the maximum usage of 200 MHz per site is not exceeded. In case of spectrum sharing they may offer even up to 120 MHz. If traffic peaks of all InP correlate the maximum spectrum bandwidth drops below 100 MHz.

For economic reasons, the transmit power of small cell nodes must not exceed a certain limit (e.g., 5 W). Hence link budget restrictions in terms of minimum EIRP power density will restrict the amount of spectrum to be deployed per radio node (cf. Annex A.2.2.2.2). If spectrum between RAN InP is equally distributed there should be no need for spectrum sharing. This is because a node owned by an InP does not allow for deploying more than his owned spectrum. Anyhow, small cell spectrum sharing make sense when Introducing frequency reuse > 1 between densely arranged small cell sites. As spectrum efficiency with decreasing inter-site distance drops dramatically this frequency planning beyond the InP site grids would improve spectrum efficiency significantly. Particularly small cell usage is heavily fluctuating with time. Hence sharing of SC nodes between tenants (slices) again enables benefit from multi-tenant traffic diversity.

**Flexible functional split**

5G will allow more flexibility in terms of function placement. Details on functional split are described in Section 2.2 of [5GN-D42]. For matters of simplicity (variety of options can be found in [TR 38.801] (depicted in Figure A-1), the evaluation focuses on RAN split 2 (or split D according to Figure 6-7) for mid-haul and RAN split option 8 (or CPRI) for fronthaul. Requirements for traditional S1 backhaul are given for comparison.

In the following, estimates for the required transport capacity for interconnecting macro and small cell sites with an edge cloud for the given split options are presented. Table 6-6 and Table 6-7 compile transport capacities for 20 MHz / 100 MHz component carriers. The case of 3 sector macro sites distinguishes between frequency bands up to 2.6 GHz where single multi-band antenna per sector are deployed, realizing four antenna ports and medium frequency bands where M-MIMO shall be applied using an extra antenna panel per sector realizing 64 antenna ports.

Small cells are differentiated between cells at low and medium frequency bands equipped with four antenna ports and cells at high bands applying M-MIMO with 64 antenna ports.

The required mid-haul and backhaul capacity scales with peak data rates provided by the component carriers assuming a best-case modulation and coding scheme (MCS). In addition, the number of transmitted and received layers influence peak data rates that may be experienced by one device under ideal conditions. If more than one carrier is deployed per site transport capacities simply cumulate.

In case of CPRI front haul (split 8) the carrier bandwidth as well as the number of antenna ports influence the resulting site transport bandwidth in a linear manner.

The overhead introduced by the upper radio protocol layers is negligible hence with respect to transport capacity there is no big difference between mid-haul split (split 2) and backhaul (S1). As can be seen from Table 6-6 and Table 6-7, the fronthaul option requires the most transport capacity.

[12] Unfortunately, currently no statistical data available.

**Table 6-6: Required transport capacity for different functional splits per component carrier (macro cells) [5GN-D22]**

| Carrier Bandwidth | No of antenna ports | Required transport capacity | | |
|---|---|---|---|---|
| | | *Mid-haul (split 2) [Gbps]* | *Fronthaul (split 8) [Gbps]* | *Backhaul [GBps]* |
| **20 MHz @ sub 1 GHz and low bands** | 4 | 0.4 | 4.9 | 0.4 |
| **20 MHz @ medium bands** | 64 | 0.4 | 79 | 0.4 |

**Table 6-7: Required transport capacity for different functional splits per component carrier (small cells)**

| Carrier Bandwidth | No of antenna ports | Required Interface Bandwidth | | |
|---|---|---|---|---|
| | | *Mid-haul (split 2) [Gbps]* | *Fronthaul (split 8) [Gbps]* | *Backhaul [GBps]* |
| **20 MHz @ low and medium bands** | 4 | 0.4 | 4 | 0.4 |
| **100 MHz @ high bands** | 64 | 6.0 | 1259 | 6.0 |

Table 6-8 summarizes the required x-haul transport capacities for collocated macro sites. The calculation considers component carriers are derived from Table 6-5 considering a minimum granularity of component carriers of 10 MHz. In the last column, the total required transport capacity in case of mid-haul option is given. Concluding, it can be observed that interconnecting collocated macro sites in a phase of full blown spectrum deployment by front haul may be challenging not only in terms of latency requirements but also in terms of required transport capacity. Mid- and backhaul require moderate transport capacities in the range of 12 Gbps. Transport technology described in Table 6-4 may enable sharing between InP and hence CAPEX and OPEX savings. Due to the granularity of component carriers the maximum spectrum per site can be deployed in case of 2 collocated spectrum owners (InP).

**Table 6-8: Required x-haul bandwidth for collocated macro sites**

| **No. of InPs** | **Car-riers** | **Car-riers** | **Sectors** | **Required transport capacity** | | | | | | **Mid-haul Total [Gbps]** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | **Mid-haul [Gbps]** | | **Fronthaul [Gbps]** | | **Backhaul [Gbps]** | | |
| ***Band [GHz]*** | ***≤2.6*** | ***3.5*** | | ***≤2.6*** | ***3.5*** | ***≤2.6*** | ***3.5*** | ***≤2.6*** | ***3.5*** | **-** |
| 1 | 4.0 | 1.0 | 3 | 4.8 | 1.2 | 58.8 | 237.0 | 4.8 | 1.2 | 6.0 |
| 2 | 4.0 | 1.0 | 3 | 9.6 | 2.4 | 117.6 | 474.0 | 9.6 | 2.4 | 12.0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2.5 | 1.0 | 3 | 9.0 | 3.6 | 110.3 | 711.0 | 9.0 | 3.6 | 12.6 |
| 4 | 1.5 | 1.0 | 3 | 7.2 | 4.8 | 88.2 | 948.0 | 7.2 | 4.8 | 12.0 |

As already discussed above even sharing of small cell sites between InP would make sense from performance point of view. If up to 60 MHz spectrum are deployed the required mid- and backhaul capacity is in the range of 1.2 Gbps for small cells at low and medium frequencies. If high user data rates and according small cell deployments at high frequency bands are introduced, the backhaul requirements grow to 6 Gbps. Interconnecting these sites by front haul may be excluded for economic reasons (cf. Table 6-8).

### 6.4.2.3 Input for economic evaluations

#### 6.4.2.3.1 Regulatory situation in Europe

As already discussed, multi-tenancy via the 5G NORMA platform has the potential to deliver performance, functional and operational benefits. These will be of interest to existing mobile network operators (MNOs) in terms of opportunities to increase revenues and/or reduce costs compared with today's Evolved Packet Core (EPC) 4G networks. Multi-tenancy support in 5G NORMA will also be of interest to potential new entrants to the mobile industry in the form of new infrastructure providers, mobile service providers and tenants.

Regulators will be keen to ensure that the right environment exists for the mobile industry, and its knock-on benefits to consumers and the wider economy, to grow. As such multi-tenancy is a feature that will attract attention from regulators, two key questions arise:

(1) Should existing mobile network operators be permitted to pool spectrum and network resources to a level similar to mobile operator core networks approach (MOCN) proposed in 4G networks?
(2) Should spectrum sharing, trading and leasing be encouraged more in existing licensed spectrum and should more low power shared access spectrum be made available to enable potential new entrants to make the most of the reduced barriers to entry that multi-tenancy networks offer?

The implications of 5G NORMA for regulators, including the above two issues, will be discussed in [5GN-D23]. However, it is worth briefly noting the current situation on the above two items.

Spectrum sharing has been acknowledged by the European Commission as an enabler for cost savings for MNOs and affordable connectivity [EC2012]. It is also described as supporting innovation and market entry [EC2016]. However as noted in [PL2016], across Europe the regulatory status on MOCN is quite fragmented due to concerns regarding maintaining a competitive environment within the industry. In Hungary, for example, two MNOs, Telenor and Magyar Telekom, have a sharing arrangement whereby one provides the mobile sites and infrastructure for the west of the country and the other the east (excluding Budapest). A leasing arrangement for the 800MHz spectrum band is in place between the two. Hungary has benefitted from this approach in having rapidly achieved a high level of LTE coverage nationally. In Germany whilst sharing of infrastructure is permitted this is only on the basis that the MNOs retain independence as competitors. MOCN however is not permitted as the spectrum sharing element is thought to compromise this independence.

Regarding our central London study area, the UK regulator Ofcom is quite forward looking in terms of supporting spectrum trading and spectrum sharing and has already rolled out a database based system for reusing TV white spaces. Joint ventures, such as MBNL and Cornerstone, have already been set up between existing MNOs in the UK to facilitate network sharing up to a MORAN level. MOCN has not been commercially taken up in the UK as yet but there are no

obvious regulatory barriers to this other than that any pooling of spectrum amongst operators would need to be sanctioned by the regulator from a competition perspective.

#### 6.4.2.3.2 Resource savings due to traffic diversity

Multi-tenant dynamic resource allocation exploits the potential of multi-tenant traffic diversity. The achievable gains increase with increasing portion of the shared network functions. Of course, for MBB services only, traffic behaviour will be more correlated between tenants at central network entities. In case of similar services most gain is to be expected at the distributed macro and small cell sites.

Unfortunately, currently there is no information on multi-tenant / multi-service traffic diversity available which may complicate techno economic analysis.

#### 6.4.2.3.3 Identification of single / multi-tenant cost differences

Analysing results described so far with respect to techno economic impact the following sources for cost savings can be identified:

- Site cost reduction (passive sharing)
- OPEX reduction due to common RAN functions in edge clouds as well as at antenna sites
- CAPEX reduction due to common RAN functions in edge clouds as well as at antenna sites
- CAPEX reduction due to joint antennas and PNFs at antenna sites
- Postponed capacity extension due to spectrum sharing and multi-connectivity (lower virtual cell load, shorter air times for elastic services, more spectrum due to traffic multiplexing gains)
- Extended EMF limitations due to reduced number of antenna planes and hence increased average (macro) antenna height
- postponed acquisition of new sites due to more flexible spectrum utilization
- Reduced number of small cell sites due to higher spectrum efficiency (less interference with frequency reuse >1 by small cell spectrum sharing)
- Joint utilisation of x-haul

## 6.4.3 Multi-service evaluation

### 6.4.3.1 Performance requirements

#### 6.4.3.1.1 mMTC performance

Performance, traffic demand and coverage requirements for mMTC services are detailed in Table A-2 and Table A-3. As daily payload volumes, devices densities and latency requirements of

- Environmental monitoring, waste management, and congestion control
- Smart meters sensor data, meter readings, individual device consumption
- Smart grid sensor data and actuator commands
- Logistics senor data for tracking goods

are quite moderate these connectivity could even be provided by legacy networks like NB-IoT. From capacity point of view the data volumes are negligible and could be provided at sub 1 GHz frequencies. As up to 40 dB penetration loss may arise with smart meter applications legacy RATs at sub 1 GHz frequency bands will not be able to provide the required deep indoor coverage of 99 %. 5G NORMA can help improving coverage probabilities by providing multi-connectivity

and / or virtual base station densification[13]. But even these gains might not be sufficient to fulfil this challenging coverage KPI.

Compared to NB-IoT, novel 5G technology will be more flexible. Semi static configuration (cf. Figure 6-9) frequency configuration could be replaced by flexible traffic and context aware frequency assignments but this would be applicable only for new appearing applications and devices (sensors) because available sensors are bounded to certain frequencies and would not be exchanged in a short-term manner.

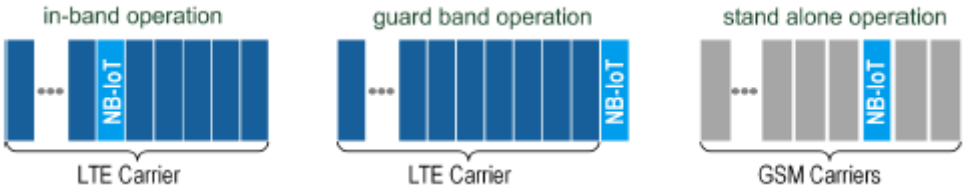



**Figure 6-9: NB-IoT spectrum deployment options (Source R&S).**

#### 6.4.3.1.2 V2I performance

According to serviced definition the V2I slice has to provide service components for

- V2I – Infotainment,
- V2I - driver information, and
- V2I – assisted driving.

Performance requirements of V2I-Infotainment are similar to those of eMBB. Mobility requirements are slightly higher and with 325 cars per km$^2$ the device density is much less than those in case of eMBB - consumer portable devices. Applying Eq. 6-1 and assuming that 1/7 of the data volume is to be processed in the network busy hour, traffic density becomes approx. 3 Gbit/s per km$^2$. Compared to the 46 Gbit/s per km$^2$ required by eMBB this can easily be provided by single tenant mobile networks. As up to three persons per car should be able to consume 4k video streaming user experienced data rates of 10 Mbps are mandatory. In urban areas FD-MIMO with in the range of 3 times enhanced spectrum efficiencies on the one hand and superior suitability for low to medium mobility may support these kinds of services sufficiently.

According to Table A-2 performance requirements for V2I – driver information are quite moderate and basically could even be fulfilled by legacy NB-IoT systems. Steadier and lower latency can be achieved by 5G NORMA enablers like multi-connectivity and user centric connection area. In terms of traffic demand the 1.7 GB data volume per day in connection with 325 cars per km$^2$ lead to moderate capacity demand of 188 Mbps/km$^2$ (application of Eq. 6-1). However the condition that V2I – driver information should be provided at sub-1 GHz bands exclusively (cf. Section 6.3.3) would be difficult to fulfil[14]. Hence similar as the infotainment component even the driver information service would have to resort on spectrum at sub-1 GHz, low and medium frequency bands.

V2I – assisted driving is the only service component that has been assigned to the category uMTC which means that for this service high reliability is required. Reliability can be improved significantly by multi-connectivity. In addition, the site grids of available RAN-InP should be integrated as much as possible. The traffic demand described in Table A-3 leads to a required capacity density of 166 Mbps/km$^2$ (cf. Eq. 6-1). Due to the higher reliability requirements, this

[13] The integration of antenna sites of different RAN-InP is denoted as virtual base station densification.

[14] The total capacity density of a single operator with average inter-site distances of 500 m amount to 900 Mbps/km$^2$

capacity demand is to be multiplied by a factor considering multiple transmission of identical data packets.

#### 6.4.3.1.3 Multi-connectivity throughput evaluation

With respect to multi-connectivity, there exist two major architecture options for integration of small cells into the network (cf. Figure 6-10):

- *Approach 1- Dedicated aggregation for small cells:* A data centre is used between the core cloud and the (macro and small) cells. The small cells are further organized into clusters such that small cells of the same geographical region belong to the same cluster. For each cluster, the data layer is connected to dedicated aggregation points, while the control layer remains in the macro cell. The connection between the macro cell and the small cells for both the data and control layer is established via the aggregation point (for each cluster of small cells) and the data centre.
- *Approach 2 – Centralized scenario:* The macro and small cells use a common aggregation point. Such topology corresponds to a generalized version of the cloud-RAN scenario, where both the data layer and the control layer are centralized.

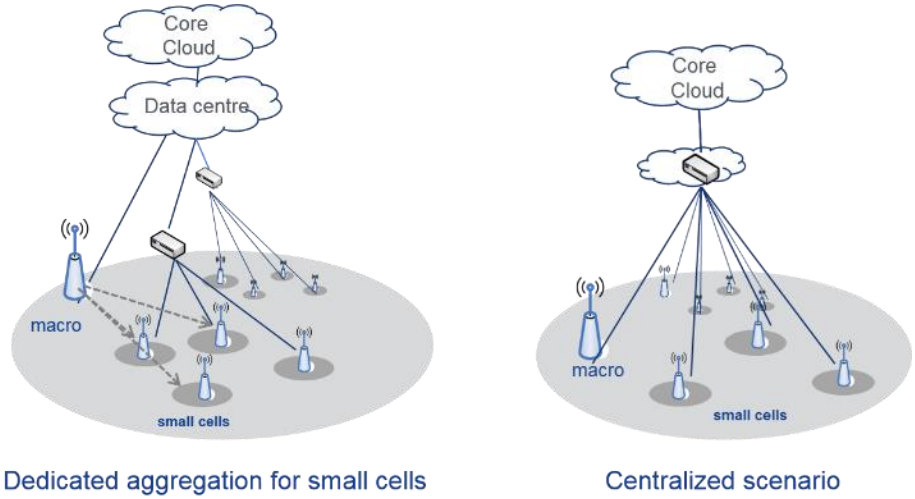



**Figure 6-10: The small-cell aggregation and centralized scenario architectural approach**

The above approaches are associated to different performance features due to the different topologies involved. Here, an evaluation of the two considered approaches in terms of the throughput that is achievable by the corresponding multi-connectivity scenario is provided.

*Simulation scenario*: Simulations have been conducted which involve a HetNet deployment consisting of 21 macro cells and 84 small cells; the small cells are organized in clusters of four, such that the within the coverage area of each macro cell four small cells are used with overlapping coverage. The macro cells are assumed to be typical wide-area 5G macro cells, operating at a central frequency of 5.9 GHz with a 20 MHz bandwidth. The small cells are assumed to operate in the millimetre-wave band, with central frequency of 28 GHz and 100 MHz bandwidth. Each UE within the overlapping area of a small cell and a macro cell is connected to both access points, in a dual-connectivity fashion.

*Traffic flow control considerations*: A traffic flow control mechanism is used, which is assumed to be located a) at the macro cell for approach 1; b) at the aggregation point for approach 2. This flow control mechanism is used to ensure that enough traffic is sent to both links, depending on the buffer status of both types of cells and the channel quality of the links involved.

The throughput performance of the two considered approaches is depicted in Figure 6-11 and Figure 6-12, obtained from the simulation scenario described above. The fronthaul delay values correspond to the delay between the small cell clusters to the data centre (which is assumed co-

located with the aggregation point in approach 2), as well as the delay between the macro cell and the data centre. The throughput is shown in the cumulative distribution function (CDF) form, such that the curves correspond to the probability that the throughputs achieved are at maximum the projection values to the x-axis.

With such illustration, it is easy to identify the percentage of time that the network performs above a given throughput requirement. For instance, for approach 1 and for the ideal case of 0 ms delay, 50 % of the time the achievable application-layer throughput is above 300 Mbps, whereas for 1ms fronthaul delay such percentage drops to approximately 20 %. Interestingly, approach 2 seems more robust to fronthaul delay as can be deduced by comparing Figure 6-11 and Figure 6-12. For the example discussed above, the percentage of time that the throughput is larger than 300 Mbps is again approximately 50 % for (the ideal case of) 0 ms delay, however for 1 ms delay the corresponding value is approximately 40 %, i.e., twice as much as for approach 1.

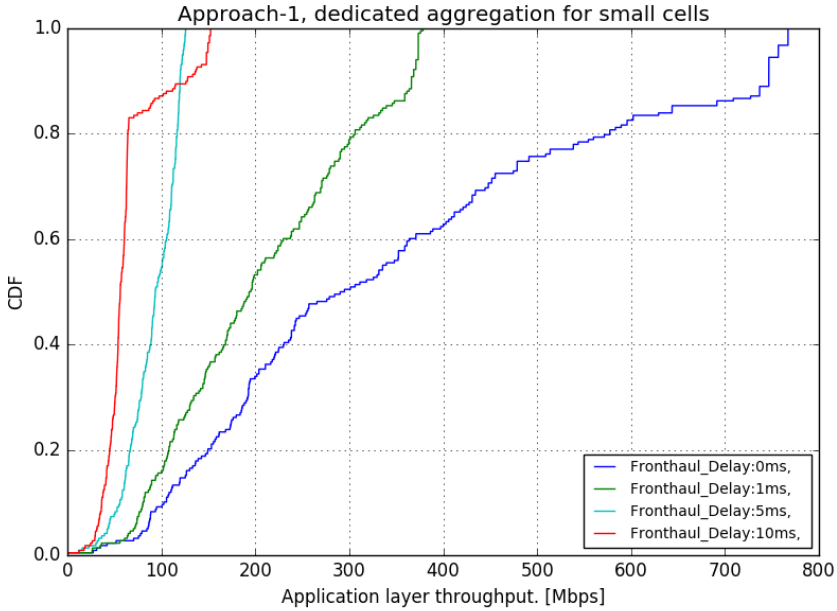



**Figure 6-11: Throughput performance (at application layer) of approach 1**

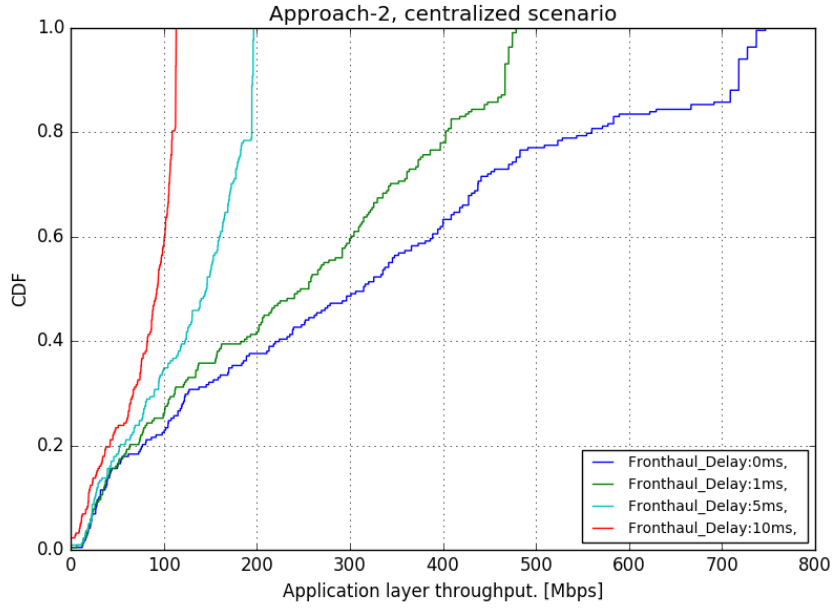



**Figure 6-12: Throughput performance (at application layer) of approach 2.**

The above analysis reveals that a centralized architectural approach offers a relative robustness to multi-connectivity against delays that may occur in the link between the access points, where RAN is realized, and the network aggregation points. Such architectural considerations are important when assessing the performance of multi-connectivity, since, together with the design of the RAN, they represent a crucial limitation factor to the effectiveness of the network to meet its requirements.

For the London study area, this would mean that in order to make sure that traffic flow control is centralized at least RAN split option D (cf. Figure 6-7) should be realized. The data planes of antenna sites hosting overlapping cells should be back- / mid-hauled to identical aggregation points located at the CU where PDCP control is hosted. From operator point of view benefit and additional cost of course will have to be balanced.

#### 6.4.3.2 Multi-service RAN benefits

Common spectrum shared by multiple tenants and deployed on shared infrastructure, facilitates very specific and/or extreme services and use cases. The reason is that these services can reuse the ubiquitously deployed RAN according to their specific spatially and temporally limited demand without the need to in advance deploy a dedicated RAN for it, which then most of the time would be unused. Besides such specific services, a multi-service RAN, when compared to multiple single-service RANs, each with dedicated (static) spectrum, helps improving the performance of each service through exploiting statistical multiplexing.

In the following, three examples are given:

- Improved peak data rates and peak number of concurrent users,
- Extreme coverage of services and
- Phase-in/phase-out as well as temporary and/or spatially limited provisioning of slices.

The most obvious benefit is the improved peak data rates for single (eMBB) users and the larger number of concurrently supported (mMTC) users during traffic bursts. A prominent use case of the former is instant download of purchased multimedia (HD movies, TV series, computer games), which regularly have tens of gigabyte. The latter speeds up periodic collection of sensor measurements (environmental, utility), making the data earlier available for post processing and presentation, i.e. improving their timeliness and real-time characteristics. Here, SDM-X respectively the Multi-tenant scheduling application running on top provides the means to dynamically reassign unused resources from other slices to the eMBB respectively to the smart city mMTC slice.

With respect to specific services, especially uMTC, it may become affordable to offer extreme coverage by exploiting virtual base station densification and pre-emptive use of radio resources. For example, this ensures that V2I – assisted driving slice experiences uninterrupted coverage under all circumstances, something that would be cost-wise prohibitive if radio resources would be (semi-)statically assigned. The enablers in 5G NORMA are its multi-service radio access, cf. Section 3.2 in [5GN-D42], which provides means for highly dynamic radio resource reassignments among slices (in RAN slicing Option 2 and Option 3), and MAC Scheduling, which, by being a distributed control layer function, allows fast enough control to realize pre-emption even under the stringent latency and reliability requirements of uMTC.

The phase-in of new and phase-out of legacy RATs (using RAN slicing Option 1) can take place very smoothly, depending on their specific co-existence capabilities, i.e. their ability to share spectrum in time and frequency with other RATs. In case of limited robustness against in-band interference (from neighbouring sites) and suppression of out of band interference (from neighbouring frequency bands), additional guard resources are used. For example, utility companies that still have legacy 2G/GSM smart meters may request a GSM telecommunication service in a specific neighbourhood once in a while for a short period in time to read out these meters. Similarly, NB-IoT spectrum can be reduced with equal pace as NB-IoT devices will be replaced by NR MTC devices at the end of their 10-year battery lifetime. Such “breathing” of

slices over time and shorted-lived temporary slices are enabled by the fast on-demand slice creation and adaptation provided through SDM-O.

#### 6.4.3.3 Investigation of mobility concepts

Dynamic service aware selection of mobility schemes as described in [5GN-D52] bears the potential of significant savings with respect to resources (storage, compute, transmit) and protocol overhead. In Annex A.2.3.6 resource consumption and protocol overhead appearing with service aware selection of fitting mobility schemes is compared to those appearing with a one fits all solution. Figure 6-13 shows savings in relative signalling overhead (compared to payload) and savings with respect to storage, compute and transmit over time reflecting the service mix defined in Section 6.2.1.

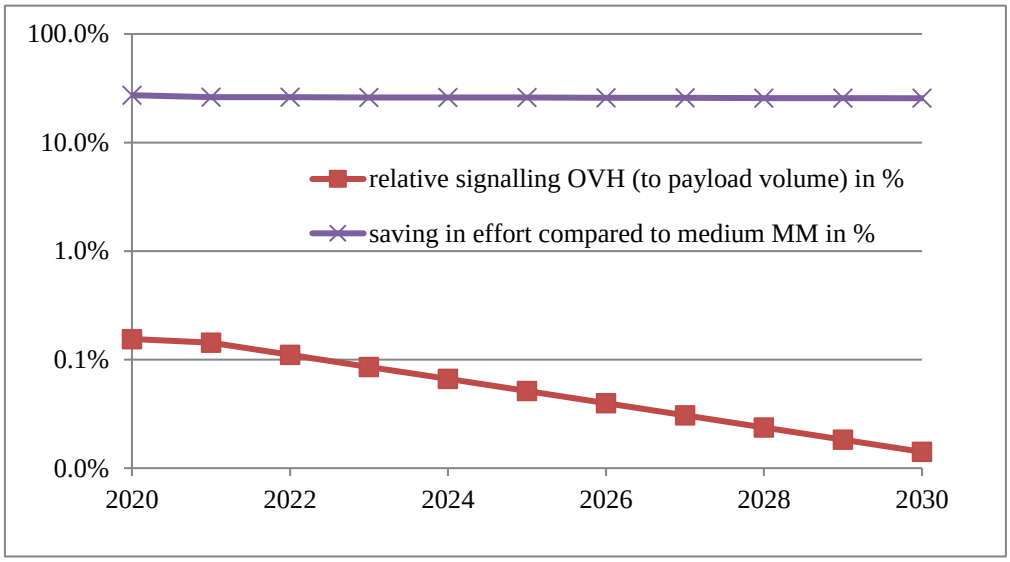



**Figure 6-13: Expected development of MM signalling and processing effort for specifically tailored MM per year.**

Due to increasing payload data volumes, the relative gain in signalling overhead decreases over time and is negligible at the end observation time. Saving of resources (storage, compute, transmit) needed for mobility management are in the range of 25-30%.

#### 6.4.3.4 Soft KPI

##### 6.4.3.4.1 Internal and external interfaces, comparison 4G/5G interfaces

When compared to 4G, the 5G NORMA system design significantly increases the number of interfaces and reference points. This is an inherent consequence of two design choices in 5G NORMA: First, the layering of the architecture into four layers, in particular the strict split of control and data layer, has resulted in several vertical or cross-layer interfaces to be specified. Second, the decomposition of functions in both access and non-access stratum brought along an increased number of horizontal interfaces. While these two design choices, from a superficial perspective, increase system complexity, they were mandatory to reap the benefits of the 5G NORMA enablers.

Therefore, a typical 3GPP-style specification for all 5G NORMA interfaces would be unfeasible due to the tremendous effort that would be required. However, by re-composing elementary functions to meaningful sets (cf. [5GN-D42], [5GN-D52]), such effort can be reduced to specifying how such sets of functions shall interact. In order to overcome such weaknesses, an alternative approach consists of the emergence of quasi-standard interface solutions that get established by means of wide-spread dissemination, e.g., as result of an early mover advantage or a superior implementation (“best practice”) and usability. Such developments have been observed

in the IT industry and could therefore, reinforced by the increased popularity of IT solutions within the telecommunications industry, become more likely for systems as architected in 5G NORMA as well.

The most crucial vertical interfaces within the architecture design have been the controller (SDM-C and SDM-X) interfaces, both northbound towards the controller applications and southbound towards the controlled NFs. The difficulty has been in maintaining the degree of programmability of NFs as high as possible while at the same abstracting from technology intricacies and implementation characteristics as much as possible. In particular, the project has shown that the design space for southbound interfaces primitives easily explodes when faced with the heterogeneity of decomposed radio access and core network functions.

The southbound interfaces are more likely to be vendor-specific and therefore would require a vendor-specific plug-in for the controller, particularly for vendor-specific, differentiating features of NFs, such as, radio schedulers or end-to-end QoS control. However, many of the northbound controller interfaces could, for example, be realised in a restful manner (cf. Section 3.3.2) and therefore be standardised or at least harmonized more easily. In summary, this means that at least northbound interfaces of SDM-X and SDM-C should be multi-vendor capable, thus allowing for interoperability between control application from vendor A and controlled functions from vendor B.

Regarding external interfaces, the most important novelty of 5G NOMRA consists of exposing management and orchestration capabilities towards external tenants, such as, vertical industries. Access control, comprising entity authentication and action authorisation, is a further important aspect to any MSP/MNO exposing system capabilities externally. While it has not been in the scope of the project to develop detailed access control solutions for resource management, orchestration, and control interfaces, the consortium believes that state-of-the-art mechanisms can be adapted in a meaningful manner to assure a high quality of access control. For identity confirmation, such mechanisms include multi-factor authentication, where multiple so-called factors are used to prove one's identity, e.g., a factor you own, a factor you know, and a biometrical factor. Extensions include multi-layer or continuous authentication approaches. For authorization, two schemes seem to be applicable in the 5G NORMA context: role-based access control, where actions are authorized based on a (rather static) pre-defined tenant role or profile that explicitly depicts allowed/non-allowed actions, or context-based access control, where authentication of an action is derived from the current (and therefore dynamic) context.

#### 6.4.3.4.2 Demonstrator learnings

Besides the WP3 verification activities, WP6 demonstrators contribute to the proof-of-concept of the 5G NORMA key innovations. A brief summary describing these demos can be found in Annex A.3, and, for more detailed information, in [5GN-D61]. Description of practical learnings and status of activities is compiled in A.2.4. In summary, most of the activities are still ongoing and it is too early to draw conclusions. A status overview of WP6 activities is briefly summarized in Table 6-9.

**Table 6-9: WP6 activities and status.**

| Demo | Activities |
|---|---|
| Native Multi-Service Architecture | • Based on a communication protocol between SDM-C and eNB service aware placement of core network functions have been successfully demonstrated.<br>• E2E latency and user throughput requirements have been adapted by automatic reconfiguration under control of SDM-C |

| | |
|---|---|
| Multi-slice service aware orchestration | • Validation of an ETSI NFV MANO implementation for orchestration of demo VNF as well as verification and selection of fitting MANO platforms is still in progress<br>• The demo will focus on low latency and MBB requirements<br>• Functional blocks already available in ETSI NFV MANO framework will be implemented<br>• The purpose is to demonstrate the practical movement of VNF between central and edge cloud |
| Secured Multi-Tenancy Virtual Network Resources Provisioning via V-AAA | • A graphical user interface will visualize tenant information retrieved from local and remote databases<br>• Tenant as well as data isolation will be demonstrated |
| Online Interactive 5G NORMA Business Cases Evaluation | • Interactive web based tool allows users to explore the economic benefits of 5G-NORMA architecture<br>• Results of the different evaluation cases will be demonstrated<br>• Early versions of the online demonstrator released in autumn 2016 and spring 2017 |

#### 6.4.3.4.3 Charging and lawful interception for multi-service case

Charging and lawful interception (C&LI) for multi-service cases refers to the possibility to employ non-uniform, differentiated charging (or lawful interception) policies, e.g., tailored according to requirements of specific user groups, the law enforcement agency, or the charging objectives of the operator or mobile service provider.

The technical feasibility of such service- or tenant-specific C&LI is ensured by the availability of slice-specific controllers, i.e., the respective SDM-C. Dedicated applications for both charging and lawful interception ("Charging Control" and "Lawful Interception" application, respectively, cf. Section 3.4) run on top of the SDM-C and provide the appropriate enforcement rules to the according functions in the 5G NORMA data layer. For charging, this is the "Traffic Reporting" function, whereas for lawful interception it is the "Intercepting Element" function. Service-specific policies for C&LI are maintained by the respective Policies Management functions.

Specifically, considering the multi-service case in the London study area, and under the assumption that each service type (eMBB, mMTC and V2X) is deployed as a separate slice instance per tenant or per MSP (in case the MSP deploys own network slice instances for its subscribers), this means that C&LI for multi-service cases is technically feasible. In particular, a service-specific and programmable pair of charging and lawful interception rule sets can be applied to each service and per tenant/per MSP. Considering Table 6-2, there does not seem to be major obstacles to the deployment of the selected services.

## 6.5 Verification summary

This final report concludes the architecture design verification activities from a technical point of view. For verification of the 5G NORMA architecture design a network roll-out in an urban London study area has been emulated for three different evaluation cases. The evaluation cases are based on a clear defined set of services including KPI on performance and coverage as well as functional requirements. Services have been selected by WP2 based on a socio economic

benefit and revenue analysis published in [5GN-D22]. Besides evaluation criteria mentioned before the evaluation is extended by considering operational and security requirements as well as soft-KPI that measure in a qualitative way the feasibility of envisioned network flexibility, complexity and standardisation effort. Interrelations with techno-economic evaluations have been worked out in specific subsections in order to highlight input to work done by WP2.

### 6.5.1 Baseline evaluation

The first evaluation case denoted as baseline case has been intended mainly setting basic assumptions on the network roll-out as an input for techno economic evaluations where cost differences between legacy eMBB deployments and 5G networks for single operator networks are intended to be identified.

Enablers developed in 5G NORMA contribute to DL/UL traffic demand assimilation (virtual cell concept) and improvement of user experience (virtual cell concept, multi-connectivity, flexible allocation of network functions). Interaction between SDM-O, SDM-C and SDM-X provides a flexible means of resource management that allows for improved interference management as well as inter-cell resource allocation (spectrum, compute, storage, transport) and hence improves user experience. According to evaluations done in [5GN-D32], DL traffic demand requirement of 46 Gbit/s per $km^2$ for eMBB in the time span between 2020 and 2030 can be served under assumptions taken for baseline evaluation. Extreme traffic demand as appearing with new services like outdoor AR/VR cannot be realized in an area covering manner but at points of interest deployment of mmWave nodes will allow for those applications.

Separation between RAN central units (CU) in edge clouds and distributed units (DU) located at antenna sites will introduce mid-haul and provide more flexible functional splits that allow for on demand network configurations. The split to be deployed will be conditioned by the characteristics of the available transport network (capacity, delay and jitter). In order to adapt the network to the emergence of 5G services the transport networks need to provision capacity on demand through automatic elastic connectivity services in a scalable and cost-efficient way. Those requirements for flexibility and dynamicity across different network domains, along with the need for efficient consumption of resources, reinforces the demand for network programmability that transport networks already face.

A flexible antenna site to edge cloud mapping realized by interworking of 5G NORMA SDM-O with a WAN Network SDN orchestrator of the transport network InP will enable an optimized deployment of edge clouds. Because of the highly distributed nature of edge clouds and the required hardware overhead a minimum usage of compute and storage resources may not be dropped meaning that aggregation of antenna site traffic must be sufficient. In addition, it would be helpful if the InP of transport and edge cloud NFVI would be identical in order to avoid cumbersome gateways for demarcation [5GN-D52].

### 6.5.2 Multi-tenant evaluation

Multi-tenant evaluation has investigated opportunities, benefits and feasibility of multi-operator deployments for eMBB. For this purpose, it is assumed that up to four traditional MNO split their roles into MSP, InP and tenants. Besides highlighting of performance and operational opportunities input for economic evaluation has been generated.

Performance benefits in terms of virtual network capacity extensions and improved user experienced data rates can be achieved by more flexible spectrum deployment and multi-connectivity. Virtual network capacity extension in this context means that inter-slice spectrum sharing enabled by SDM-X will facilitate shifting the allowed load in the cells more to the saturation point. Multi-tenant scheduling maximizes the monetarisation of the infrastructure (spectrum, compute, storage, and transport) enabling the establishment of new network slices including the adherence of existing SLA. Resource savings due to traffic diversity and multiplexing gains could incur at more central aggregation points (edge and central clouds).

However, these benefits may be limited due to correlated traffic behaviour within the different eMBB slices. However, there are currently no verification sources in terms of measurement results or simulations available.

Whereas current drivers for RAN sharing are quick roll-out for new entrants including cost savings even for incumbents and – in urban areas – in addition the lack of suitable site locations future drivers are to be seen in new 5G services requiring better coverage, reliability and availability. Integration of RAN infrastructure of different MNO where needed will enable fulfilment of these requirements and unlock the potential to deliver non-eMBB more niche services with potentially new revenue streams.

5G NORMA enablers for RAN sharing are RAN slicing, more flexible spectrum deployment and more flexible functional RAN protocol split (also covered by 3GPP).

RAN slicing provides a flexible split of RAN InP specific and commonly used infrastructure including spectrum. Sharing antenna panels at macro sites not only reduces cost: By deploying less planes of sector antennas with identical construction engineering effort the panel balance points may be moved more beyond roof tops. EMF limitations can be mitigated and the maximum deployable spectrum per site may be incremented.

Multiplexing gains with respect to common resources may be more attractive deploying slices with different traffic behaviour instead as here assumed identical services. A significant benefit is provided by the more flexible spectrum usage under SDM-X control. Allowing for flexible spectrum deployment of several RAN-InP could increase capacity per macro site, site densities, and reduce costs.

The performance of small cell deployments can be improved by small cell spectrum sharing: Small cell spectrum pools allow for a flexible, SON-based frequency reuse factor >1 between densely arranged small cell sites.

Regarding spectrum sharing regulatory status in Europe is quite fragmented but regulators will be keen to ensure that beneficial conditions exist for the mobile industry. Most important sources for cost savings by multi-tenant networks are

- Site cost reduction (passive sharing)
- OPEX reduction due to common RAN functions in edge clouds as well as at antenna sites
- CAPEX reduction due to common RAN functions in edge clouds as well as at antenna sites
- CAPEX reduction due to joint antennas and PNFs at antenna sites
- Postponed capacity extension due to spectrum sharing and multi-connectivity (lower virtual cell load, shorter air times for elastic services, more spectrum due to traffic multiplexing gains)
- Extended EMF limitations due to reduced number of antenna planes and hence increased average (macro) antenna height
- postponed acquisition of new sites due to more flexible spectrum utilization
- Reduced number of small cell sites due to higher spectrum efficiency (less interference with frequency reuse factor >1 by small cell spectrum sharing)
- Joint utilisation of x-haul

### 6.5.3 Multi-service evaluation

For multi-service evaluation, the London baseline network has been extended by network slices for mMTC and V2I so that the whole set of generic 5G service can be deployed on a common infrastructure. Besides discussion of performance, operational and security aspects the fulfilment of soft-KPI checking for convertibility of proposed solutions has been checked and learnings from demonstrator set up have been concluded.

Regarding capacity, throughput, and device density, the selected mMTC service components for London do not make high demands and may be provided even with legacy technologies like NB-IoT. In order to comply with increased coverage requirements slices for these services should be deployed at sub-1 GHz spectrum which due to low required data volumes should be feasible. Exploiting 5G NORMA enablers like multi-connectivity and virtual base station densification[15] coverage indoor and outdoor can be improved without really increasing the base station density. The required 99 % indoor coverage for smart meters might be difficult to attain.

V2I – mMTC service components require distinct higher data volumes than the mMTC service components treated in the last paragraph. Due to increased reliability requirements especially provisioning of V2I- assisted driving should include flexible frequency assignments at sub 1 GHz low and medium bands as well as exploitation of multi-connectivity and virtual base station densification. At this point the flexibility introduced by 5G NORMA network slicing and multi-tenant scheduling will provide significant benefits over legacy LTE-V technology realising semi-static frequency assignments.

Regarding mMTC and uMTC, not only flexibility in terms of context aware multi-connectivity but foremost flexible slice specific integration of multi-operator antenna site grids will enable improved coverage and reliability with feasible economic effort that cannot be attained with legacy technologies.

Realising multi-connectivity, it would be beneficial to have the flow control and data aggregation points at the same location. By this aggregated throughput becomes more robust against front- or mid-haul delays. Practically, this means that PDCP layer (flow control) and data aggregation points for different transmission path (macro and small cells) should be located ideally at central units (CU) by application of flexible functional splits. This could be an argument motivating the deployment of mid- instead of backhaul in urban areas.

An integration of multi-operator antenna site grids will superpose statistically independent service coverage patterns and hence increase the coverage probability compared to single site grids. This mechanism will be an important enabler for improved coverage, reliability and availability as demanded by new 5G services.

The most attractive application of network slicing appears if slices with very different requirements and traffic patterns share the same resources (spectrum, compute, storage and transport). Significant increase of multiplexing gains in multi-service RANs enable services that would not be economically feasible otherwise.

Security aspects have already been addressed in [5GN-D32]. These aspects are mainly independent from the services and hence from this roll-out study. Details on 5G NORMA work on security concepts may be taken from Section 4.4, Section 5, and [5GN-D42].

When compared to 4G, the 5G NORMA system design significantly increases the number of interfaces and reference points. This is mainly due to splitting the architecture into four layers (vertical interfaces) as well as due to the decomposition of functionalities into a set of control and data layer network functions in both access and non-access stratum (horizontal interfaces). The effort specifying the interfaces by SDO's can be reduced by recomposing functions to meaningful sets and how these sets shall interact introducing quasi standards by early movers. Most crucial are controller (SDM-C, SDM-X) interfaces where northbound interfaces to the applications should be standardized and southbound interfaces due to their exploding design space might be vendor specific. Access control, comprising entity authentication and action authorisation, is a further important aspect to any MSP/MNO exposing system capabilities externally. Specification of horizontal interfaces by SDOs can be limited to inter-stakeholder interfaces (cf. Figure 4-4),

[15] The integration of antenna sites of different RAN-InP is denoted as virtual base station densification.

where it is most likely that NFs from different vendors interconnect, while the remaining data layer interfaces may well be proprietary.

With the 5G NORMA demonstrator's developments practical feasibility of network function mobility, software implementation of RAN protocol stack and ETSI NFV MANO architecture extensions as proposed by WP3 could be proofed. In addition, the project partners gained practical experience applying different MANO platforms.

# 7 Migration Paths towards 5G NORMA Networks

After presenting the 5G NORMA concepts in Chapters 2-5 and validating them in Section 6, this chapter addresses the question how these concepts can be put into practice.

Already from Figure 2-2, it becomes obvious that migration from today's mobile networks to 5G NORMA-based networks is not about replacing 4G / LTE by 5G NR, but about moving from single-service monolithic network elements to cloud-based multi-service / multi-tenancy platforms. In this sense, even the use of an LTE air interface on a cloud-based multi-service / multi-tenancy platform is an option for migration to a 5G NORMA-based network.

Since the 5G NORMA architecture has been extended by several function blocks for controlling and managing multiple services or tenants, a 5G NORMA-based network for a single service will most likely be more expensive than a legacy single-service network. Hence, the main driver for a migration to 5G NORMA-based networks will be the possibility of a multi-service / multi-tenancy operation and the opportunity to faster and more flexibly serve new customer and develop new markets in a cost-effective way. For this reason, considerations on possible migration paths have to take into account not only infrastructure and deployment aspects, but also functional aspects. In the following, we will start with a review of the existing network infrastructure, then look at the 5G NORMA target architecture and then identify potential migration steps.

## 7.1 Operators' network infrastructure assets

According to the topology shown in Figure 2-6, a 5G NORMA network comprises the following main infrastructure components:

- Mobile access networks: base station sites
- Fixed-line access networks and Operator Points of Presence (PoP), e.g., central offices
- Transport and aggregation networks
- Data centres (in the centre of the network as well as at the network edge)

**Mobile access network**

The radio access network (RAN) of representative mobile network operators nowadays consists of a mixture of 2G (GSM), 3G (UMTS) and 4G (LTE) platforms. Within the anticipated timeframe for 5G NORMA deployment, only the 4G infrastructure is considered as relevant starting point. Figure 7-1 shows the current architecture of 4G networks. The 4G mobile network is optimised for mobile broadband services (MBB) and consists of a flat architecture including one or more 4G evolved packet cores (EPC) in central locations and distributed radio nodes (eNodeB / eNB).

Virtualisation of the EPC has been intensively studied in recent years, among others by ETSI NFV, and commercial vEPC solutions are available. Among others, the EPC comprises two important control functions, namely the MME for Radio Bearer management and the PCRF for QoS management.

Most eNBs today are deployed as fully integrated base stations, each serving a single antenna site. As alternatives, so-called centralized-RAN (C-RAN) concepts have been proposed initially by China Mobile and studied later e.g. in ETSI NFV. However, at present, these concepts are rarely used in practice. In these concepts, the RF components remain at the antenna site, while digital processing tasks of multiple antenna sites are executed in a central location on a pool of virtualized open hardware, like x86/ARM CPU based servers. Antenna sites are geographically distributed depending on coverage and capacity demands. In areas with particularly high capacity demands, currently an increasing number of small cells are deployed, leading to so-called heterogeneous networks (HetNets).

The aggregation network interconnects eNBs among each other (*X2* interfaces in Figure 7-1) and with the EPC (*S1* interfaces in Figure 7-1). In practice, it consists mostly of Ethernet links, where transport technology is mostly optical fibre or, more rarely, microwave links. Physically these links are mostly organised in some kind of star topology from the eNB towards the central EPC nodes. Logically the X2 links directly interconnect two eNBs, but physically these links are usually established from one eNB via a central node to the another eNB.

Typically, a mobile network is owned and operated by a single network operator. However, due to the distributed nature of the RAN, a RAN is costly to build and operate. To reduce these costs, network sharing solutions have been developed allowing multiple operators to share RAN infrastructure in areas with low capacity demands as well as the EPC infrastructure (provided this is permitted by regulation authorities).

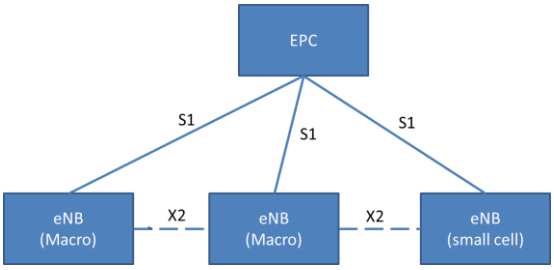



**Figure 7-1: Current RAN architecture status**

**Fixed-line access network**

Although the fixed-line access networks have been developed earlier than mobile networks, they show a similar organization as mobile networks with Central Offices, containing nodes for the production of network services and certain control functions, and distributed operator Points of Presence (PoPs) providing the access link to the customers' premises.

Fixed access lines are usually costly to deploy and thus operated over a long period of time. This leads to a mix of multiple access technologies in fixed-line networks, ranging from copper-based VDSL2 via G.FAST to PON deployments with optical fibres. Depending on the access technology, different PoP types exist: Multi-Service Access Nodes (MSAN), Optical Line Terminations (OLT), LAN-to-LAN Switches (SWL2L). The mix of deployed access technologies shapes the network topology as well as the characteristics of PoP deployments.

Central Offices comprise nodes like Broadband Remote Access Servers (BRAS) and Broadband Network Gateway (BNG) for control functions such as AAA, as well as nodes for the production of residential services like voice, video, Internet. It has been typical to have different cores dedicated to different services.

With the advent of NFV technologies, fixed line operator are also evolving their networks to include VNFs based on x86 hardware deployed in their Central Offices or Points of Presence. Centralised control layer nodes such as AAA or IMS are already being virtualized, but the trend is to also virtualise some distributed data layer functions of the IP layer once the required performance can be achieved.

**Transport (aggregation) network**

Traditionally, the transport network (or aggregation network) and fixed-line access networks are closely integrated and often managed by the same operator. The aggregation network typically provides data transport for a combination of different services:

- Residential Business: Broadband, CMTS (HFC) service and voice access,
- Mobile services: 2G, 3G, 4G, 5G, WiFi,

- Enterprise Business: MPLS-VPN access, VPLS/L2L access, Managed WiFi, Internet access,
- Wholesale Business: IP transit / IPVPN services, Broadband connectivity, Interconnection for LEC (Local Exchange Carriers) / CUG (Closed User Group), TV Links, L2L.

Aggregation networks are organised hierarchically in different network segments, which aggregate traffic from the lower segments to gain advantage of statistical multiplexing.

On the Physical Layer, transport network operators deploy optical fibres that provide pure capacity transport across the operator Points of Presence. Parts of this optical layer are the fibre rings and the ROADMs (Reconfigurable Optical Add-Drop Multiplexer) used to inject traffic into these rings. In order to prepare the signals to be injected into the fibre rings and links, C/DWDM transponders and multiplexers are used to interface the customer feeds into the optical layer. The transport capacity obtained with this optical layer is offered to end-customers, such as business customers or vertical industries, while at the same time it is used by the network operator itself to build on top of that an IP services layer, cf. Figure 7-2.

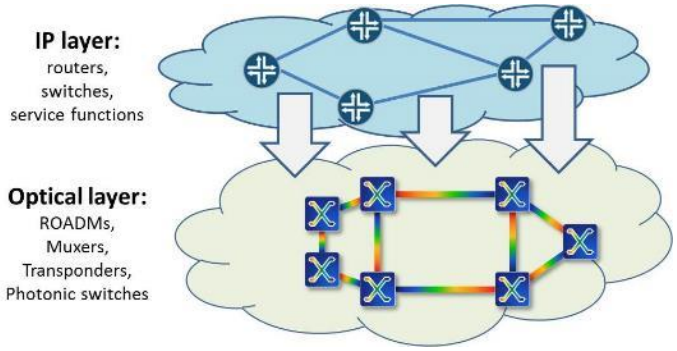



**Figure 7-2: Optical and IP layers of a fixed line network operator**

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

**Data centres**

The data centre business is becoming a noteworthy component in contemporary economies providing critical support infrastructure [CGL17]. As a whole, data centres are commonly used in three main segments:

- Enterprise or in-house data centres.
- Multi-tenant (colocation) data centres.
- Hyper-scale data centres.

An enterprise or in-house data centre is typically a unit within a public or private sector organization that houses and maintains back-end IT systems and data stores. They are commonly operated for the benefit of the controlling enterprise. The primary benefit of these data centres is that firms retain complete control of their infrastructure. They are commonly used for strategic and mission-critical applications which may contain sensitive data non-susceptible to be shared with third parties. Anyway, while many organizations with enterprise data centres operate their

own data centres, for larger global organizations the operation and maintenance of the data centre facility may be outsourced.

According the Uptime Institute 2016 Survey [UI16], about 71% of IT assets were located in enterprise-owned data centres; from the rest, the bigger part is owned by multi-tenant data centre operators, while the smallest part goes to Cloud Computing/Hyper-scale data centres. However, although Enterprise Data centres continue to play a very important role in the IT market, analysts suggest that over the coming years more and more enterprise IT workloads will move to multi-tenant or hyper-scale data centres [CGL17].

Multi-tenant data centres (or Colocation Data centres as they are also referred) are data centres operated by third parties for the benefit of multiple enterprise tenants (this would be the case for 5G NORMA, where the InP leases part of the infrastructure to different tenants). Basically, there are two main types of multi-tenant data centres:

- **Wholesale colocation data centres**, i.e., commercial data centre that typically sells warehouse-scale or large sub-components of physical space to a small number of big enterprises. In this case the data centre operator is just responsible of the physical buildings, access to utilities (e.g., power and communications), and the maintenance of the physical space and associated utilities, although it is not responsible for the hardware or the software running on that hardware.
- **Retail colocation data centres.** This is a variant of a wholesale data centre but with a greater number of tenants. They can also be split in two categories:
    - Managed hosting data centres. In this case the data centre operator also manages and/or rents compute, network and storage infrastructure to the customers.
    - Shared hosting data centres. These are similar to the previous category, but here clients are allowed to share services at a higher level (e.g., these data centres are commonly used for website hosting that can be used by businesses and consumers). 5G NORMA falls into this category.

Main driver for colocation data centres is the significant cost savings when compared to building, operating and maintaining enterprise data centres. They are also valuated as an asset for disaster recovery (even when an organization uses an enterprise data centre, they may use a colocation data centre for disaster recovery services).

The Uptime Institute's 2015 Data Centre Survey [UI15] reports that colocation data centre investment increased of up to 74%. Also, the Structure Research 2016 report [SR16] suggests that the colocation segment in 2016 was about 25% of the market, while the retail colocation accounting was the remaining 75%.

Finally, hyper-scale data centres are defined as "large-scale data centres often architected for a homogeneous scale-out greenfield application portfolio using increasingly disaggregated, high-density, and power-optimized infrastructures. They have a minimum of 5,000 servers and are at least 10,000 sq. ft. in size but generally much larger" [CGL17].

In general, hyper-scale data centres are based on a single compute architecture which is massively scalable. Such architectures are built on infrastructure and systems made up of hundreds of thousands of individual servers offering compute and storage resources and connected by high performing networks.

The magnitude and complexity of scalability in modern hyper-scale data centres also serves to distinguish them from other data centres. Hyper-scale data centres have diverse workload requirements and high service level expectations from end-users and customers. The emergence of new types of specialized processor architectures (e.g.: GPUs, MICs and FPGAs) makes them more efficient for both: to manage more complex workloads and to consume less energy, resulting more cost efficient. In practice, only a handful of companies can afford the infrastructure necessary for hyper-scale computing; they are global companies such as Microsoft, Apple, Google, Amazon, IBM, Twitter, Facebook, Yahoo!, Baidu, eBay and PayPal amongst others.

Forecasts on the size of the hyper-scale data centre market vary, but according [Cisco] growth from 259 in 2015 to 485 by 2020 with a quintupling of traffic within these data centres in the same period is expected. The same survey claims that by 2020 these hyper-scale data centres will represent:

- 47% of all installed data centre servers
- 53% of total traffic
- 68% of all data centre processing power
- 57% of all data stored in data centres

As a general conclusion, we can tell that big changes are on the horizon since probably, most of the IT workloads will reside off-premise in the near future. According to [Cisco], it is expected that global data centre traffic can firmly reach the zettabyte scale. But not only the data centre traffic is growing, it is also getting streamlined with innovations such as NFV and SDN; these innovations (on which 5G NORMA relies) can offer new levels of optimization for data centres. Also, a rapidly growing segment of data centre traffic is cloud traffic, which will nearly quadruple over the forecast period and represent 92 percent of all data centre traffic by 2020. This is because an important traffic enabler for cloud computing is the data centre virtualization.

Moreover, according [Cisco] private clouds will have significantly more workloads than the public cloud initially, but public cloud will grow faster in a second stage. Many enterprises will adopt a hybrid approach to cloud as they transition some workloads from internally managed private clouds to externally managed public clouds. Also, all three types of cloud service delivery models—IaaS, PaaS, and SaaS—will continue to grow as more and more businesses realize the benefits of moving to a cloud environment.

Additional trends influencing the growth of data centres include the wide adoption of multiple devices, the growth of mobility and the IoE phenomenon (and extraordinary amount of data in the range of about 600ZB is expected to be generated by IoE applications). Also, over time, more and more of the data resident on client devices will move to the data centre. Other consumer focused applications (video streaming, social networking) or enterprise focused (compute and collaboration, Data Base/Analytics and IoT) will contribute to this rapid growing.

Beyond IT, Edge-Cloud data centres can become an important complement to the traditional IT services listed above, because they combine the Enterprises' demand for keeping control on their data with the trend towards outsourcing. Also, Data Base / Analytics and IoT can benefit from 5G NORMA's flexibility to allocate functions to the network edge or the central network.

## 7.2 Target functional architecture and network infrastructure

### 7.2.1 Target functional/logical architecture

The architecture design of 5G NORMA allows for a service-specific composition of modularized network functions as well as a more flexible deployment of virtualised network functions (VNFs), i.e., a requirement-aware selection of the instantiation location of VNFs. This subsection elaborates on reasonable design choices for these two aspects by depicting the functional architecture and function placement of two specific services: 3GPP eMBB and cross-operator V2I service.

#### 7.2.1.1 eMBB network slice

The 5G NORMA functional architecture instantiation of an eMBB slice in cellular networks is depicted in Figure 7-3. It comprises the required functions in data, control, and MANO layer, distinguishes between dedicated and shared functions, and features a set of functional characteristics that particularly support the eMBB performance and functional requirements.

In the NAS part of the data layer, the network slice comprises only dedicated functions. Since they are expected to be very service- or even application-specific, the figure only shows two mandatory functions, "Policy Enforcement" and "Charging". For eMBB, various other functions can be added in a slice-specific manner, independent from other slices. Examples include, but are not limited to, Deep Packet Inspection (DPI), Lawful Interception, or forwarding and gateway functions. For efficiency reasons, the access stratum (AS) is characterized by a certain level of centralization, e.g., to realize multiplexing gains and allow for different multi-connectivity options. This means that PDCP instances are located in the edge cloud, e.g., at a tier 1 aggregation point. RLC as well as lower layers of the protocol stack need to be co-located at the radio access site. This is due to their real-time operation, implying a synchronous interaction between one another and thus a low latency inter-layer communication. Moreover, to enhance overall throughput, multi-connectivity (MC) would be employed in the "distribution" mode, i.e., data packets are distributed across several legs. Moreover, MC would be realized by both PDCP split bearer mode and "carrier aggregation" mode. In case multiple slice instances share the same RAN, transmission point and user specific part is shared across network slices, and the service (or bearer) specific part is implemented in each network slice. The RLC/PDCP function blocks in the data layer belong to the dedicated part, i.e., they are service/bearer-specific (RLC/PDCP for user-specific control signalling belong to the common part).

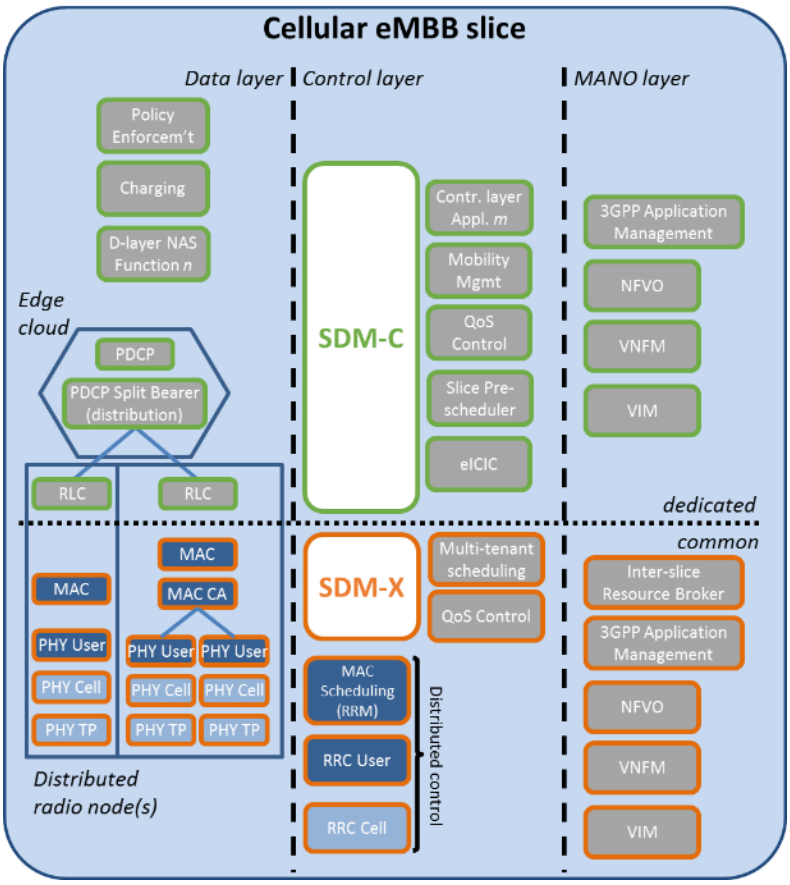



**Figure 7-3: Functional architecture of a cellular eMBB network slice**

The NAS part of the control layer is completely service- or slice-specific and includes customized functionality for mobility management, QoS and charging control, session control, etc. All these functions are executed as applications on top of the slice-specific controller (SDM-C), which logically constitutes a centralised entity, but can be implemented in a distributed manner, e.g., running on multiple edge clouds. Southbound, SDM-C controls all slice-specific NFs of the data layer, incl. AS functions PDCP and RLC. For synchronization reasons, RRC User, RRC Cell, and MAC scheduling (RRM) functions are deployed in a distributed manner, i.e., they are executed

at (or very close to) the antenna site. In contrast, multi-tenant scheduling and QoS control run as applications on top of the centralized common controller (SDM-X). This means, as an example, that the QoS Control function block receives reports of pre-defined events from the SDM-X and analyses the data, enforcing new configuration actions (e.g., resource re-scheduling, new resource assignments) on the network function blocks to meet the QoS requirements of each mobile service according.

The MANO layer for the eMBB network slice comprises ETSI NFV MANO functions as well as 3GPP application management functions for both the dedicated and the common control/data layer functions. These MANO layer functions can run on general-purpose hardware, i.e., depending on the requirements of the managed function and the scope of the MANO layer function, the respective MANO layer function can be placed at any network aggregation level between antenna site and central offices. The complexity of MANO layer functions is significantly higher in the common part of the network since they have to accommodate, balance, and prioritize the requirements of multiple slice instances. For example, 3GPP network management functions have to configure RAN functions such as multi-tenant scheduling or MAC scheduling in an SLA-compliant manner. The full overview of resource supply and demand and the according matching policies are available at the Inter-slice Resource Broker, which should therefore be deployed in a rather central infrastructure node.

Finally, requirements on fronthaul capacity are moderate since the depicted scenario employs a higher layer split of the radio protocol stack (PDCP-RLC).

#### 7.2.1.2 V2I network slice

Figure 7-4 depicts the the 5G NORMA functional architecture instantiation of a V2I slice. Since both infrastructure nodes and mobile nodes (i.e., vehicles) could be associated with different network operators, this network instance has to be designed as a cross-operator slice. Like the eMBB slice, it comprises the required functions in data, control, and MANO layer, distinguishes between dedicated and shared functions, and features a set of functional characteristics that particularly support the V2I performance and functional requirements. As defined in Section 6.2.1, V2I includes the following services: infotainment and advertising to passengers, non-critical driver information services on road and driving conditions, navigation, and assisted (automated) driving services. The focus here is on the assisted driving services, which require reliability and, in selected cases, reduced latency in addition to the requirements of the eMBB services above.

In the data layer of the V2I slice, the NAS part is realized by a mixture of common and dedicated network functions. In the RAN, distributed radio nodes (master nodes) implement the full protocol stack at the antenna site, thus helping to satisfy the latency requirements of specific V2I services. For increased reliability on the radio link, MC is utilized in the so-called “duplication mode” [5GN-D42]. In the given example, a secondary radio node implements a second transmission leg, i.e., each packet is sent via both connections. This secondary node can operate in the same or in a different frequency band (intra- vs. inter-frequency MC). Similarly, diversity can be improved further if the radio nodes are not collocated (same antenna mast) but are placed in different sites. In order to further emphasize reliability over latency, RLC operates in acknowledged mode (AM), i.e., the outer Automatic Repeat request (ARQ) is explicitly included [5GN-D42].

In the control layer, all AS-related functions, among them *RRC User* and *RRC Cell*, *MAC* and *Multi-Tenant Scheduling*, and *eICIC* are realised in a distributed manner to avoid any additional latency in the control-data-layer signalling (for a detailed description of the interfaces between control and data layer in the AS, the reader is referred to [5GN-D42]). Consequently, the SDM-X (and according control applications) only control common NAS data layer functions in a direct manner and configure distributed control layer functions. Mobility Management and QoS control are placed in the common functions domain since the according procedures should be handled in a uniform manner for subscribers (vehicles) from different operators, in particular for coordinated

or autonomous driving applications. However, these functions are not time-critical and thus can be executed at edge cloud or even central cloud locations.

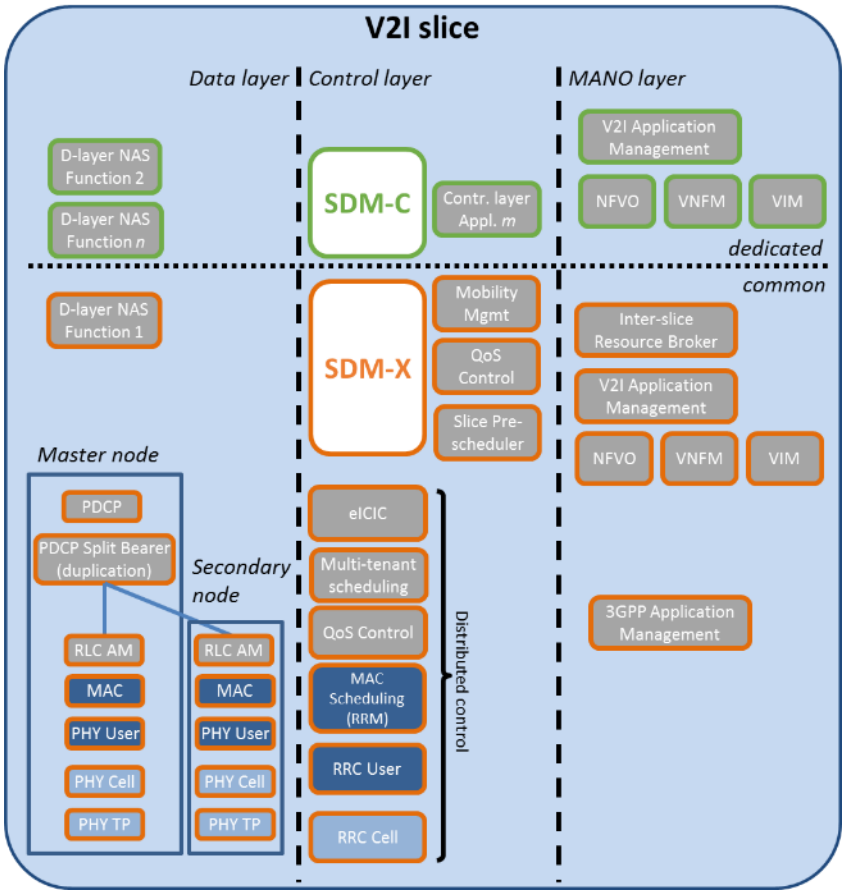



**Figure 7-4: Functional architecture of a (cross-operator) V2I network slice**

In the MANO layer, only 3GPP application management functions are required for the AS since all NFs are implemented using dedicated, application-specific hardware, i.e., no ETSI-style lifecycle management of VNFs is required. In the NAS domain, the full set of MANO layer functions is present, including a V2I domain application management function responsible for specific applications, e.g., navigation or coordination of autonomous cars. Since such applications can be hosted in both the common and the dedicated part, multiple (and possibly different) of such V2I application management function instances are required. In a more extreme case, the entire NAS domain in both control and data layer could be non-3GPP.

There are no requirements for fronthaul capacity for the V2I network slice since radio nodes are implemented in a fully distributed manner. However, in case of multi-connectivity ("duplication mode"), capacity requirements can be moderate since PDCP-PDUs (and RLC-SDUs, respectively) have to be exchanged between master and secondary node.

## 7.2.2 Required infrastructure resources and topology

The infrastructure of a 5G NORMA network is expected to change in comparison to a legacy network for three reasons:

(1) Increase of mobile traffic will result in a higher demand for network capacity;
(2) Adaptive decomposition and allocation of network functions and
(3) Multi-tenancy.

The increase in mobile traffic and the higher demand for network capacity will impact the radio network planning. Densification of the macro cell network to yield additional capacity has its

limits: For practical reasons (limited site availability) as well as technical reasons (interference with multi- sector antennas mounted above roof-top) a minimum inter-site distance of 200 m should not be undercut for macro cell sites [5GN-D32]. If capacity demand will continue to increase, the introduction of small cell layers cannot be avoided. However, the number of small cells that can be deployed per macro cell is limited by intra-layer interference. Propagation conditions for mm waves are more favourable in terms of spatial separation and allow to deploy a larger number of mmW nodes per macro cells than small cell nodes at carrier frequencies below 6 GHz.

5G NORMA's multi-tenancy capability strongly suggests the sharing of the network infrastructure by multiple MNOs. While this is possible in principle and economically highly beneficial, some limitations should be kept in mind: For EMF safety reasons, the acceptable Tx power per antenna site is restricted. If multiple MNOs share a single antenna site and together have many frequency bands licenced, it may be necessary to reduce the amount of spectrum that each MNO could deploy (cf. Table 6-5). If traffic demand in such situations exceeds network capabilities, a densification of the antenna sites and possibly deployment of additional small cells may be unavoidable.

Tx power per frequency band is limited, as otherwise the total Tx power at the site would exceed the permitted limit. To overcome such situations, a densification of the antenna sites and possibly deployment of additional small cells may be unavoidable.

*Adaptive decomposition and allocation of NFs* allows a more flexible allocation of RAN functions to antenna sites, edge clouds and central clouds. Multiple function splits have been investigated, cf. Annex A.2.1.1.1, Figure A-1. Relevant options for function splits are a lower physical layer split, a split at the MAC or RLC layer and a split above the PDCP layer. Based on these split options, RAN-related processing will happen at three locations: at the antenna mast, at distributed units (DU) in close vicinity to the antenna mast (or even at the antenna mast as well), and at centralized units (CU) in edge cloud data centres. The selection of a suitable split option depends on the available transport infrastructure as well as the traffic demand at the respective sites. As shown in Section 7.2.1, at least the PDCP layer should be located at a central point within the network, for macro sites as well as for small cell sites. In order to deploy 5G NORMA VNFs, central and edge cloud infrastructure resources have to be present on which the aforementioned VNFs can be executed.

In the case of the central cloud locations, these can be designed according to state of the art IT datacentre architectures. IT datacentre architectures, as shown in Figure 7-5, usually employ a leaf (also called Top-of-Rack, ToR) and spine switching fabric to connect the servers hosted in the datacentre. Connections from the datacentre to external networks are achieved by deploying another hardware equipment, called the Data Centre Gateway (DC-GW), in a redundant configuration that connects the cloud datacentre infrastructure to the network operator equipment, usually an MPLS Provider Edge node. This back-to-back connection of these 2 pieces of equipment (DC-GW and PE) allows to have an administrative demarcation point between the cloud provider and the network provider. On the other hand, the overhead implied by this DC-GW hardware equipment is shared by the high number of servers and VNFs present in a central cloud location.

Due to the level of capillarity required in terms of the edge cloud locations, fixed/transport network providers are very good candidates to become edge cloud providers, since they already own and run distributed locations with the capillarity required. In the case of a network provider becoming an edge cloud provider, using the same kind of IT data centre architectures for edge cloud deployments imposes an excessive overhead because of the DC-GW hardware required for the connection of the edge cloud to external networks and other remote cloud locations. This overhead can be very high if the number of edge VNFs in a distributed location, and correspondingly the required compute power (number of servers) is very low Moreover, in the long run, the need for purpose-built hardware (e.g. PEs in traditional IT datacentre architectures)

will disappear, and the edge cloud locations will become the location where virtualised network functions are executed on general purpose compute nodes.

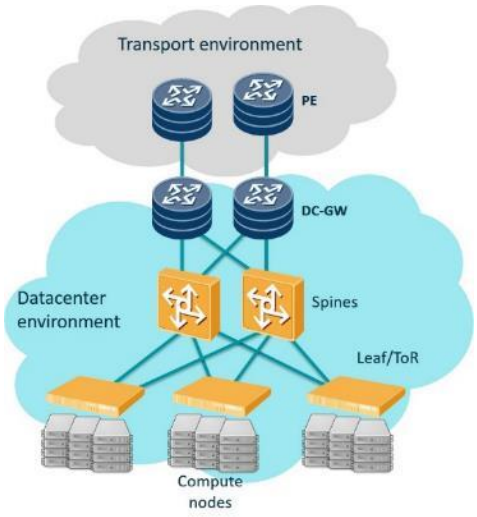



**Figure 7-5: Central Cloud location network architecture**

As a result, network providers are in the process of defining new ways to deploy NFV infrastructure as distributed as possible in an efficient and optimized way in order to become edge cloud providers. This evolution towards a cloud-"enabled and enabling" network is sometimes called in the network industry the Telco Cloud.

One example of this evolution to a Telco Cloud is the CORD (Central Office Rearchitected as a Data centre) project [PAB+16] of the ON.Lab initiative, led by AT&T. Another example of the move to the Telco Cloud is the ongoing work of the Broadband Forum (BBF) to define the Cloud Central Office (Cloud-CO) that has started with the edition of the Working Text WT-384 [BBF384].

In terms of network layout, as shown in Figure 7-6, the different Telco Cloud proposals from the industry share a target network reference architecture that just includes the leaf and spine switching fabric, general purpose servers and some Physical Network Functions (PNFs) hosting access-facing and network-facing I/O cards connected directly to the switching fabric. There is no hardware specifically devoted to the DC-GW and PE functions as in a traditional data centre architecture, to reduce the overhead and to become self-contained, being the necessary functions to interact with external networks distributed across the other elements in the architecture (fabric, servers and I/O cards). This is the final target of a transition to a Telco Cloud for traditional network operators to become edge cloud providers.

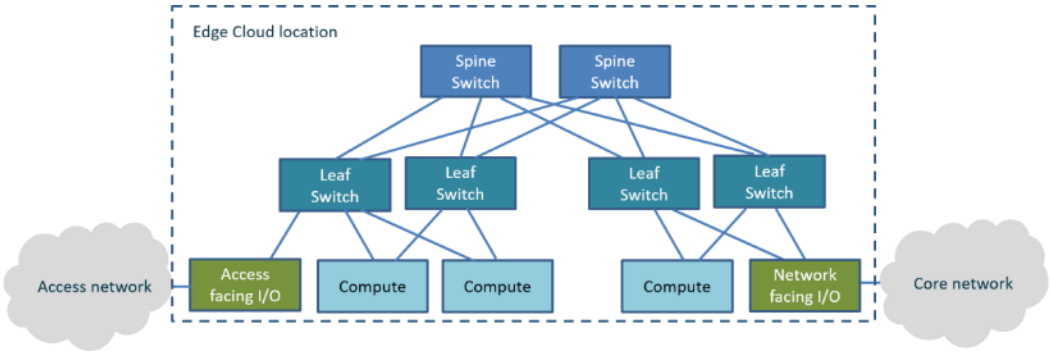



**Figure 7-6: Edge Cloud location network architecture**

## 7.3 Migration steps towards 5G NORMA networks

### 7.3.1 Requirements on the migration process

Deploying a mobile radio infrastructure for E2E network slicing requires high investments. The same holds for the reconstruction of an existing network.

As long as only a single service is provided on a network infrastructure, it is more economical to use a network infrastructure with conventional architecture instead of an architecture that supports network slicing, due to the additional costs for the slicing of the network functions and for the corresponding MANO functions needed to manage the slices. This results in a 'chicken and egg' problem: Without slicing capabilities no second service can be offered on a common infrastructure, and **without** a second service the deployment of a slicing-capable infrastructure is economically not attractive.

The most promising way out seems to be the introduction of network slicing in incremental steps, with low upfront investments and low economic risks. Ideally, after each introduction step, multiple customers start using the new technological possibilities and thus motivate the next incremental introduction step. In this way, the speed of the technical rollout has to adapt to the speed at which the commercial usage of sliced network infrastructure evolves.

In summary, this yields the following requirements on the migration concept:

- Deployment in incremental steps, in line with the speed at which the new technical possibilities are adopted by end customers and tenants;
- Minimum additional upfront investments per deployment step;
- Maximum benefit for the end customers and tenants by the new technical possibilities of each deployment step;
- Coexistence of legacy and sliceable network infrastructures;
- Migration of network functionality and infrastructure has to be synchronised with penetration of different UE generations;
- Technical migration steps need to support and match the expected stakeholder model as described in Section 4; and
- The migration process needs to be planned and implemented such that major security requirements, in particular across administrative domains, are reflected

### 7.3.2 Migration of infrastructure

At the infrastructure layer, several migration steps can be envisaged to allow for a phased transition of mobile access networks on the one hand as well as data centres and transport & aggregation networks on the other.

**Migration of mobile access networks**

Multi-service / multi-tenancy and flexible allocation of network functions are two key innovations of the 5G NORMA concept presented in Section 0 above. These properties can also be found in the Single RAN concept and the use of the CU/DU split discussed in Annex A.2.1.1.1. Therefore, these two key innovations may be suitable as migration steps towards a 5G NORMA based network platform.

*Single RAN:* The term "Single RAN" indicates a RAN technology that allows mobile telecommunications operators to support multiple mobile communications standards (GSM, WCDMA, LTE) on a single network infrastructure. Single RAN can be understood as a form of network slicing according to Option 1 in Figure 2-10: The radio signals of multiple RATs are (frequency-) multiplexed and share the same RF equipment. All processing above the RF processing is done specifically for each mobile network according to the networks' communication standards.

Single RAN solutions are commercially available today. They have the advantage that they can be applied to stand-alone base stations. Thus, no change in network planning is needed, and deployment can start without building edge data centres first. Sharing processing capabilities beyond different communication standards allows to adapt flexibly to changes in mobile user traffic, e.g. the shift of traffic from WCDMA to LTE. In this way, deployment of Single RAN equipment can be a useful step towards the virtualization of network functions and to prepare the 5G NORMA network slicing concept.

*CU/DU split:* Adaptive decomposition and allocation of NFs is one of 5G NORMA's innovative enablers and an essential design principle. An obvious prerequisite for function decomposition is the existence of suitable interfaces between the network functions. Several 5G PPP projects, like 5G XHAUL and 5G NORMA, as well as 3GPP have investigated possible function splits and identified several options, cf. Annex A.2.1.1.1, Figure A-1.

In rural areas, stand-alone macro base stations are likely to persist in the future for several reasons:

- Distances between antenna sites in rural areas are typically large, in order to ensure mobile radio coverage all over the country in an economically feasible way.
- Accordingly, connections from a central location to multiple antenna sites will be very long and thus costly.
- On the other hand, joint processing of radio signals from multiple base stations has no major benefits in a rural environment: While Coordinated MultiPoint (CoMP) processing can improve the service quality at the cell edge in principle, this has only low relevance in practice for sparsely populated areas.

Hence, for rural areas the CU/DU split may be rarely used and stand-alone base stations may be preferable in most situations.

In urban areas, the substitution of stand-alone base stations by Cloud-RAN solutions appears to be more attractive. However as explained in the previous section, due to the high upfront investment needed for a nationwide rollout, a nationwide rollout of a Cloud-RAN solution is unlikely. Instead rollout of Cloud-RANs seems to be more likely in the context of a) local network capacity extensions and b) swapping HW equipment at its end of life.

For local capacity extensions, new additional antenna sites can be equipped with RRHs and DUs and connected to a CU in an edge cloud. Similarly, a HW swap of an end of life base station to shift the CU to an edge cloud, while maintaining RRHs and DU at the antenna site. Then the CU can effectively perform functions like eICIC and CoMP to reduce inter-cell interference and improve the service quality at the cell edge.

In this way, small data centres / clouds will emerge at the network edge. How such clouds can grow from a few servers to a large central office location will be addressed below. An important prerequisite are effective management mechanisms for small decentralized data centres: Decentralised data centres should be completely manageable from remote and have efficient redundancy mechanisms. Ideally, it should be sufficient to physically visit a data centre e.g. once in a year to replace defective components, and in between two site visits, fault management by remote reconfiguration should be sufficient.

Multi-connectivity may become another driver for the deployment of Cloud-RAN data centres at the network edge and for a CU/DU split. Multi-connectivity will be beneficial to improve the resilience against radio link failures and to increase the data rate. As shown in [5GN-D42] (, the traffic should be split (in DL direction) resp. aggregated (in UL direction) at the PDCP layer. Then it will be advantageous if the traffic flow controller and the traffic aggregation points are in the same edge cloud location, to achieve a traffic flow as stable as possible.

**Migration of data centres and transport & aggregation networks**

The following steps outline a possible migration process for a fixed network provider with central data centres to become also a provider of edge clouds, where the final target would be the Edge Cloud network architecture described in Section 7.2.2:

(1) Step 1: Compute-only NFVI-PoPs

In this step, only compute servers are connected directly to the network provider existing equipment. All connectivity, between servers in the edge cloud location, if required, and towards central cloud locations is provided by the existing network equipment in the network location. No separate physical switching fabric and DC-GW are deployed. This results in an NFVI-PoP consisting of only compute nodes (compute-only NFVI-PoP), cf. Figure 7-7.

This evolution phase will be justified for distributed deployments where the number of servers per location is very small and for which the deployment of any other NFVI infrastructure such as switches, or a local VIM, is a huge overhead. Also, this phase is applicable if the amount of service chaining across the local compute servers is small, because they host mainly unrelated network functions (e.g. vBNG and vEPC), and there is no benefit in adding a local switch for local connectivity or traffic aggregation.

The orchestration of network services and their required connectivity in this kind of NFVI-PoP is driven by the NFV MANO: In the virtualised infrastructure (e.g. vSwitches in compute nodes) this is done by means of an NFVI-SDN controller (NFVI-SDNc) or, in existing network equipment, by already deployed Network SDN "master" orchestrator, respectively. This Network SDN master orchestrator can have as clients both traditional Operations Support Systems (OSS) and the NFV MANO, and arbitrates between their requests, as shown in Figure 7-7.

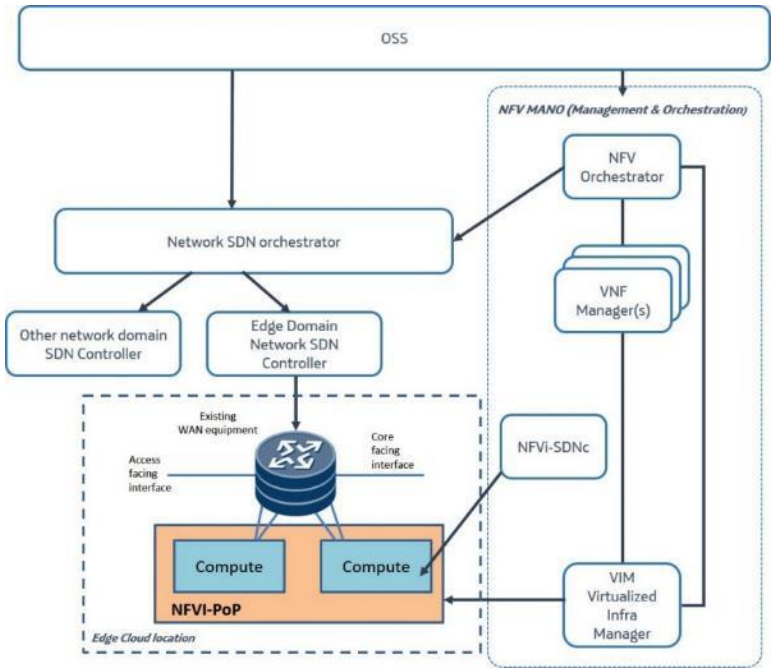



**Figure 7-7: Edge Cloud MANO and Network SDN**

In parallel to the evolution to become an edge cloud provider, network operators are undertaking the evolution of their networks to be SDN-enabled. The use of these SDN capabilities will simultaneously benefit from NFV applications deployed at the Edge Cloud, and by other traditional legacy OSS applications. As such, a hierarchy of SDN controllers (devoted to different technology and geographical domains) will be deployed that can be invoked from a top-level Network SDN "master" orchestrator as needed, as shown in Figure 7-7.

(2) Step 2: Local switching fabric

Several factors can justify the addition of a local switching fabric in an Edge Cloud location, cf. Figure 7-8:

- The number of compute servers. The Leaf Switch can interconnect compute servers with cheaper ports than a router serving also high-rate connections to Access and Core Network. In this way traffic aggregation towards the existing network equipment can result in a lower TCO as the number of compute servers raises.
- Amount of traffic between local compute servers. If traffic local to the Edge Cloud location is high, it can be more cost-effective provide local connections using a local switching fabric.
- The number of subscribers. As the number of subscribers served by one Edge Cloud location increases, High Availability schemes become inevitable to limit the impact of failures. This as well as the possibility to share VNFs (e.g., non-generalized Value-Added Services such as parental firewall) across a higher number of subscribers demands for local traffic switching and for redundant connections across the location, that are best served by a local switching fabric.
- Out-of-band local management. For security or architectural reasons, compute servers may need to be managed on physically separate Ethernet ports, forcing the inclusion of a switching fabric, at least for management purposes.

The local switching fabric (leaf switch) is connected to the existing network equipment that still holds the access and network I/O as shown in Figure 7-8.

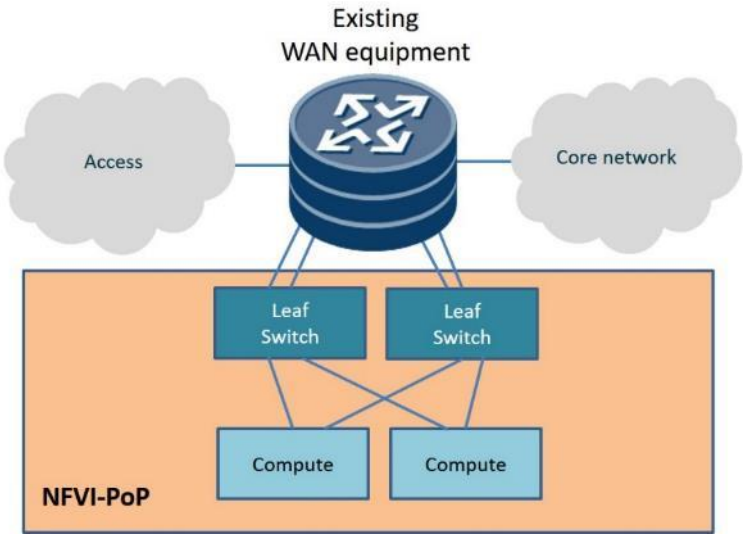



**Figure 7-8: Edge Cloud location with local switching fabric**

(3) Step 3: Partial migration of I/O

The next step towards the target architecture is shifting the access facing and network facing I/O functionalities off the existing network equipment.

In many network deployments, such as the fixed residential broadband networks, the access facing I/O interfaces are Layer 2 based. Subscribers are received on VLANs (either single tag or double tagged) and the access I/O functionality is to steer the subscriber to the appropriate VNF. However, injecting this kind of traffic to a traditional Ethernet switching fabric may not be possible because of the required support in the number of MAC addresses or the handling of Q-in-Q traffic with the required granularity. As such two possibilities arise:

- Inject the access facing interfaces to an evolved switching fabric (Option 1 in Figure 7-9). The switching fabric needs to support the requirements in terms of MAC addresses or

granularity level as commented. One approach can be the support of granular tunnelling over IP (e.g. VXLAN) to avoid MAC address learning for this kind of Layer 2 traffic.

- Inject the access interfaces to some compute nodes with a specialized VNF that provide the traffic steering functionality (Option 2 in Figure 7-9) to the appropriate serving VNF (e.g. vBNG, vPE, vEPC). These access VNFs would implement the aforementioned tunnelling approach across a traditional Ethernet switching fabric.

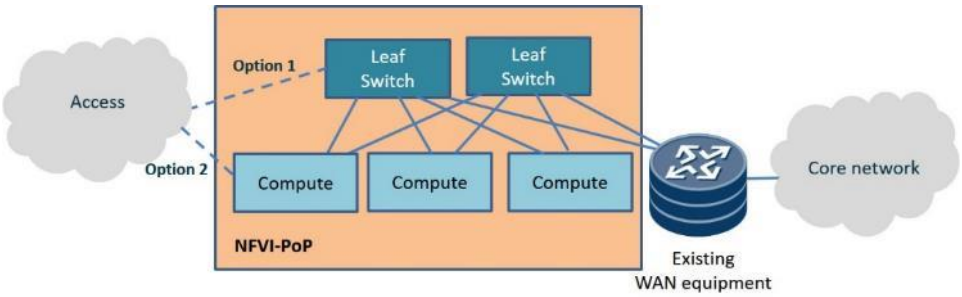



**Figure 7-9: Migration options for access facing I/O**

Network facing, and in some cases also the access facing, I/O interfaces are typically Layer 3-routed interfaces. To be capable of taking off these interfaces and their corresponding functionalities from the existing network equipment and into the NFVI realm, both data layer and routing functionalities have to be considered. Again, two different alternatives are possible:

- Move the network facing interfaces to the switching fabric (Option 1 in Figure 7-10). To be able to have a switching fabric simpler (and accordingly less costly) than the existing network equipment, the routing control layer interacting with the rest of the network is likely to be run as an application that would propagate the appropriate forwarding entries to the switching fabric.
- Move the network facing interfaces to the compute nodes that would be controlled by a routing control layer application running on a node in a Central Office location (Option 2 in Figure 7-10) or in location-wise routing control layer running in another compute node of that Edge Cloud location (i.e., control and data layer are collocated and run on a local compute node as a VNF, Option 3 in Figure 7-10).

Step 4: Target architecture

When both access and network I/O have been shifted off the existing network equipment and a switching fabric is in place, the network operator completes its network evolution towards the target architecture described in Sec. 7.2.2 that enables it to become an Edge Cloud provider in addition to its role as network provider.

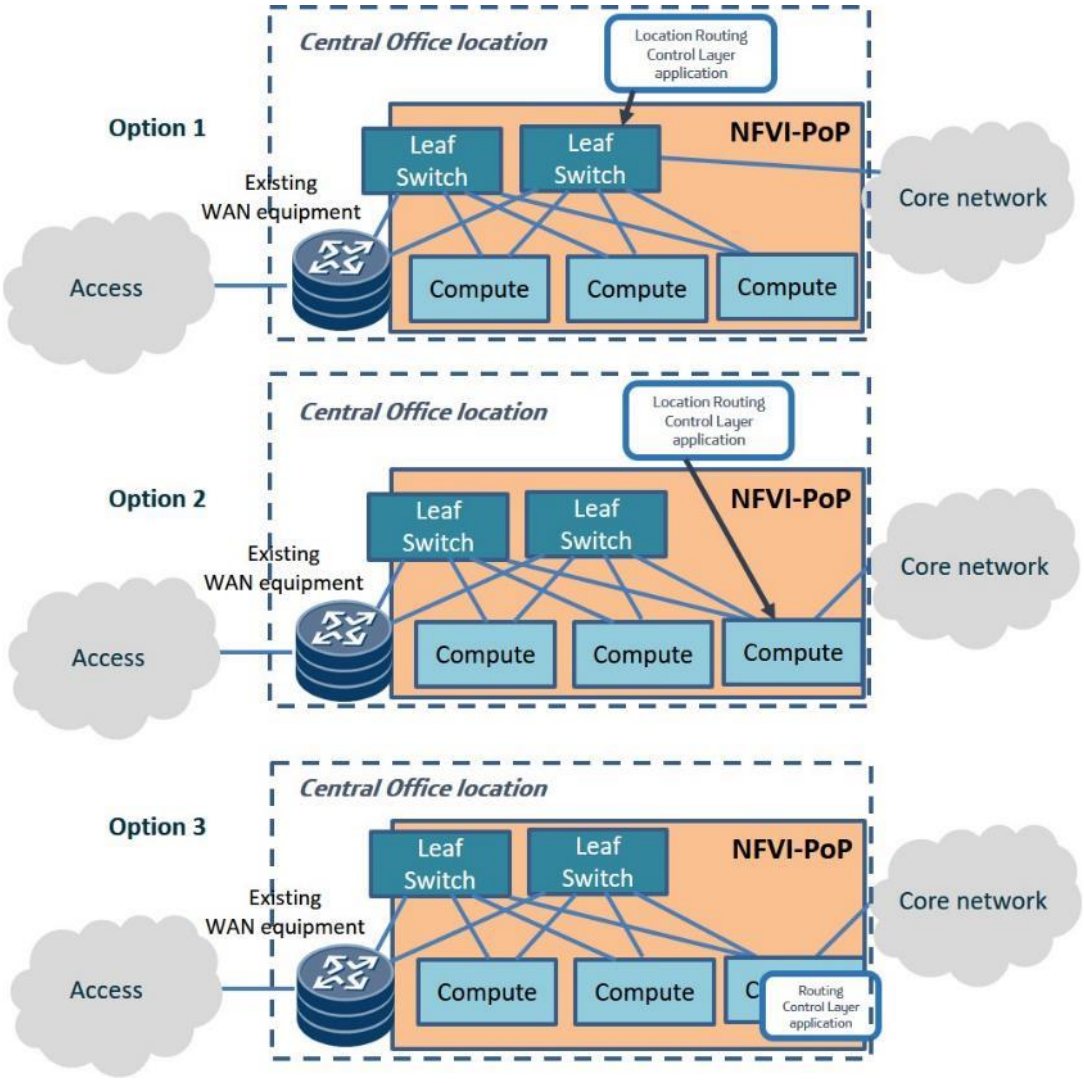



**Figure 7-10: Migration options for network facing I/O**

### Migration of UEs

Obviously, end-to-end network slicing is possible only if it supported also by the UEs. Basically, three types of UEs can be distinguished:

(1) UEs that do not support network slicing at all
(2) UEs that support network slicing, but only for a single slice
(3) UEs that fully support network slices, even for multiple slices.

Type (1) will be most common when deployment of 5G NORMA-like networks starts. For this type of UEs, it may be beneficial to create a "default slice" that is accessible by UEs without requiring any network-slicing specific functionality and thus offers backwards compatibility for such UEs without slicing capabilities.

Type (2) may be attractive for MTC in various Vertical Industries. Such UEs can benefit from network slicing, in particular from the isolation of network slices against their surrounding and other slices, but also from the possibility to design network slices with slice-specific network functions and forwarding graphs. Modifications in the UEs will be needed mainly for RRC, to control and manage the access of the UE to the slice that it is intended for. The effort for these modifications is expected to be comparatively small.

Type (3) may be relevant to both Vertical Industries as well as consumers. E.g. for automotive usage, separate network slices may be created for safety-critical communication between cars,

non-safety critical driver support (e.g. navigation), and entertainment. A UE built into a car then has to be connected to multiple slices in parallel. For consumers, it may be attractive if slicing-capable UEs allow to separate private and business usage of a smartphone device. Obviously, the impact on UE design is much greater for this type of UEs than Type b) above.

### 7.3.3 Migration of network functionality

Network slicing impacts data layer, control layer and management and orchestration layer. Currently, the extent of technical possibilities for network slicing in these areas differs:

**Data layer**

Main issues of network slicing in the data layer are the mutual isolation of traffic flows and customization of network functions. Various tools for traffic isolation have been identified in 5G NORMA (e.g. virtualisation, multiplexing, multitasking) and exist already in legacy networks (e.g., RAN sharing, synchronous / asynchronous multiplexing of radio bearers). Customization of network functions is less popular in legacy networks; it requires the definition of appropriate interfaces between network functions.

Migration could proceed based on incremental steps such as the following:

(1) Multi-service/multi-tenancy is possible in the CN by virtualizing the EPC and applying network sharing, e.g. MOCN or GWCN, and also in data centres. All necessary technologies are available today, thus minimizing the economic risk. The main drawback is the restriction to CN and data centres; this solution is not feasible to RAN and transport network.
(2) Greater flexibility for the network topology of single-service networks can be achieved by deployment of Mobile Edge Computing servers, allowing to direct user traffic either to a network edge cloud or a central cloud. This solution is possible today, although not widely practiced. It can be deployed locally according to business needs, which minimizes the business risk.
(3) For slicing in the RAN, base stations need to be enhanced. Parallel processing of different radio technologies (e.g. GSM, UMTS, LTE, NB-IoT) in a single base station is state of the art, but for multiplexing multiple slices on MAC level, appropriate products comprising features like separate PDCP and RLC instances per network slice and differentiation between MAC PDUs of different slices have to be developed. There is no impact on aggregation network assets and sites; limitation of impact on base stations to a software upgrade would further motivate this migration step.
(4) As a further migration step, base stations could be deployed as Cloud-RAN base stations where the upper parts of the RAN protocol stack are executed in an edge data centre. This requires the commercial availability of Cloud-RAN components as well as fronthaul fibre connectivity, thus increasing the economic risk compared to the previous steps.
(5) In the CN, a network slicing solution as specified by 3GPP SA could be implemented.

**Control layer**

The control layer comprises, among others, functions for resource management within and between network slices, QoS/QoE management, mobility management. Similar functions exist also in legacy networks, but as separate controllers (e.g. radio schedulers for resource allocation to radio bearers (not slices), PCRF for QoS, MME for radio bearer setup and mobility management).

As first migration step towards a SDMC solution, these controllers could be integrated in a single entity. The evolution to SDM-C / SDM-X-like controllers according to the 5G NORMA concept however requires further standardisation and product development efforts.

**MANO layer**

As explained in Section 2 above, 5G NORMA's MANO concept builds on the ideas of ETSI NFV MANO. Implementation of such tools is progressing, partly even as open source projects. However necessary extensions such as integrated management of VNFs and PNFs, resource management for multiple slices, function chaining, and last but not least the capability to provide tenants with their own management functions are currently missing in commercially available systems.

In a step-wise migration, design of MANO systems should start from today's O&M processes:

(1) Initially, when
- the number of network slices is small,
- network slices are covering only mall regions (e.g. the production plant of an industrial tenant), and
- there are almost no slice modifications during the lifetime of a slice,

manual operations seem feasible in practice. This avoids upfront costs in terms of CAPEX for an automated MANO system (at the price of higher OPEX for operational efforts) and allows to collect practical experience with network slice operations.

(2) Automated slice orchestration for the Network PaaS model could follow, where an operator manages the slice, i.e. a tenant cannot yet orchestrate his slice on his own. Then in contrast to the situation shown in Figure 2-16, no separate t-MANO stack is needed for the tenant, thus greatly simplifying the MANO layer.

(3) Finally, automation of slice orchestration for the Network IaaS model, where tenants can manage network resources assigned exclusively to them, could follow as third step. This leads to the situation shown in Figure 2-16 where MNO as well as the tenants have their own c-MANO resp. t-MANO stacks.

## 7.3.4 Co-existence of 5G NORMA and 4G networks

In Section 7.3.1, two important requirements for the migration from 4G networks to 5G NORMA networks have been identified: (i) progressing in incremental steps and (ii) coexistence between 4G networks and 5G NORMA-based networks. While the previous two subsections have addressed requirement (i), and some possible migration steps have been identified for the infrastructure as well as for the network functionality. This section examines the aspect of co-existence, i.e., a possible step-wise integration of and migration towards the 5G NORMA architecture is described.

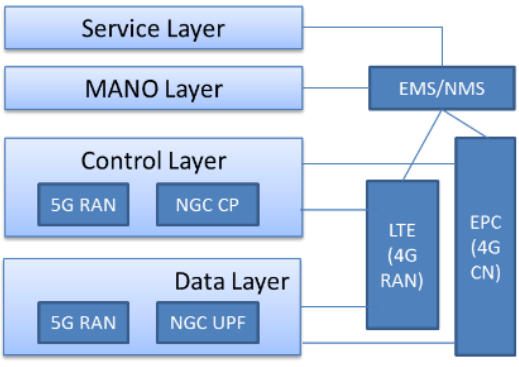
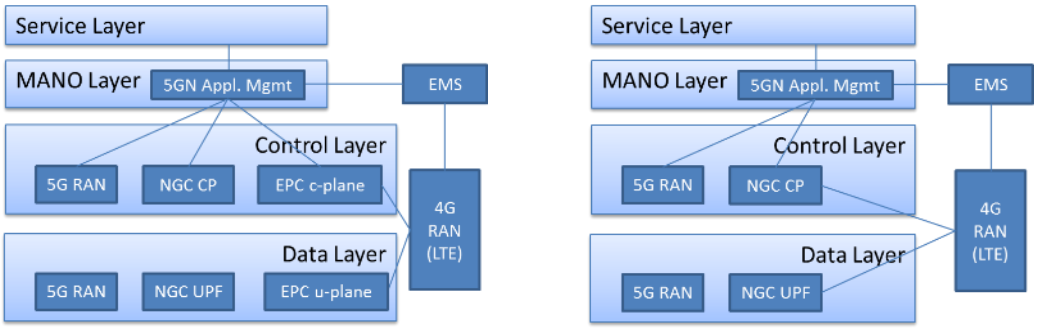



**Figure 7-11: First step of architecture integration - integration at service layer**

The lifetime of UEs, infrastructure HW, and VNF software differs significantly. In practice, a 5G UE shall be able to use existing 4G equipment and network functions in areas where 5G is not yet available, and legacy 4G UEs must be able to access 4G network functions, even when these are already integrated into a 5G network. For this kind of backwards compatibility, some

interconnections (i.e. interfaces, Service Access Points, or APIs) between 5G / 5G NORMA network functions and existing 4G network elements are needed. In the following, it will be described how 5G NORMA functions and legacy 4G functions can grow together and how the interconnections between these functions will change during this process.

Figure 7-11 uses an abstracted variant of Figure 2-4 and shows the first step of the architecture integration, where 5G and 4G RAN (LTE) and core network (CN, also referred to as Evolved Packet Core (EPC)) would operate in parallel. An interfacing between both would be required in order to support, for instance, inter-RAT mobility (e.g., inter-RAT measurements) or multi-RAT connectivity. However, in this first step, the main integration of 5G and 4G mobile networks would happen at EMS/NMS (element/network management system) level (i.e., on the 5G NORMA MANO layer. While 5G would be deployed initially in "islands", e.g. for enterprise tenants, legacy 4G terminals would still dominate and only few 5G terminals would be used in these "5G islands". Due to the utilisation of a single service layer, multi-tenancy would be offered across 4G and 5G networks; in 4G networks, using technologies such as eDECORE, MOCN, and multi-PDN connectivity. However, 5G NORMA would still be more flexible and could offer service-specific enhancements as described in [5GN-D22].




a) Integration EPC into architecture b) Connecting 4G RAN through NGC

**Figure 7-12: Second step of architecture integration – integration of core network and MANO**

The next step, as shown in Figure 7-12a), would integrate the user and control plane of the 4G core network into the 5G NORMA control and data layer. Furthermore, the NMS would be replaced with the 5G NORMA application management for managing 5G NORMA NFs (in RAN and CN) as well as legacy EPC NFs/NEs. This would enable network slicing for 4G as well, with the 4G RAN as common network function, while virtualised 4G CN functions would share data centre resources with the 5G network functions. The 4G RAN would be maintained and still, on the level of individual NEs, managed by the 4G EMS. The integration within the CN may happen faster than an integration in RAN because the number of nodes is smaller and EPC virtualization already progressed. Such an integration would be further facilitated by data center integration of 4G and 5G mobile networks.

As a consecutive step, the 4G RAN would be connected with 5G Next Generation Core (NGC) as shown in Figure 7-12b), which implies that all RAN functionality (both 4G and 5G) would then be connected to the same core network. This allows for integrating 5G NORMA and 4G at NAS level and efficient controlling of the access network. Furthermore, an integration of the CN would improve the MANO layer (only EMS for legacy RAN NEs left).

Finally, as shown in Figure 7-13, the 4G RAN would be integrated in the 5G NORMA control and data layer, e.g., as PNFs and VNFs. This would happen when 4G access points are updated towards 5G, which may also be capable of serving 4G terminals (single RAN for legacy and 5G terminals). Similar to 5G RAN functions, 4G RAN would be decomposed into control and data layer as well as PNFs and VNFs, which may be collocated with 5G PNFs/VNFs (cf. also [5GN-D42]). Since the replacement of RAN equipment implies the highest cost and usually takes place

only about every eight years (cf. [5GN-D22]), the integration at RAN level will happen in the last step.

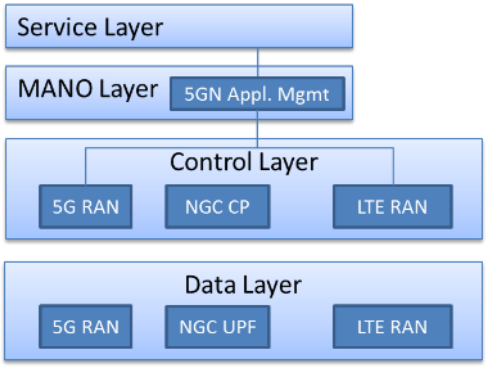



**Figure 7-13: Final step of architecture integration – RAN integration**

### 7.3.5 Conclusions

The previous section can be summarized in a few recommendations:

- All complexity must be avoided unless it provides obvious or provable benefits. Otherwise it will be a cost driver, and costs that cannot be justified by benefits for the network tenants are hindering the evolution of network slicing.
- Networks for network slicing should be introduced in small steps, in time as well as geographically.
- It will take some time to evolve to fully-fledged, sliced network infrastructures. In the meantime, coexistence and parallel operation of legacy and sliced network infrastructures is a must. A default slice for legacy terminals can easily realize that parallel operation.
- Existing functions for legacy networks should be executable in network slices as well, to minimize the migration efforts for tenants.
- Good candidates for initial deployment of sliced networks are
    - well-known and proven network components, to minimize the risks from novel technologies and simplify the search for early faults;
    - 5G NORMA components and capabilities that provide a clear business benefit, to minimize the impact of unavoidable upfront investments;
    - components in a network that exist in small quantities (e.g. core network components), as their benefits reach potentially many customers, while the small quantity minimizes the deployment effort.

# 8 Conclusions

## 8.1 Summary of 5G NORMA architecture design

This deliverable concludes the technical part of the third iteration of the 5G NORMA architecture design. Based on the available results, WP2 will perform the final socio-economic and business analysis, to be documented in [5GN-D23]. The final architecture integrates the final results of WPs 4 and 5 (cf. [5GN-D42] and [5GN-D52]) with the latest updates of WP3 work items into an overall and harmonised multi-service mobile network architecture. This architecture replaces today's network of monolithic entities by a flexibly composable network of functions, incorporating the three 5G NORMA innovative enabling technologies *adaptive (de)composition and allocation of mobile network functions*, *Software-Defined Mobile network Control (SDMC)*, and *joint optimization of mobile access and core network functions*. These enablers, together with further novel enablers of 5G NORMA, realise the two innovative functionalities of the new architecture, namely *multi-service and context-aware adaptation of network functions* and *mobile network multi-tenancy*.

For this purpose, 5G NORMA has, to the largest extent, followed a clean-slate design approach, resulting in an architecture design that, in selected areas, significantly deviates from SDO (such as, 3GPP) design approaches. More specifically, WP3 has achieved the following results and outcomes:

- Modular architecture design: the functional architecture comprises four independent and largely decoupled layers, each fulfilling a well-defined set of tasks. Modularisation is further supported by highly de-composed functions that can be re-composed based on service requirements and network infrastructure capabilities.
- Split of control and data layer: a further reduction of inter-dependencies is realised by a stringent split of control and data layer throughout radio access network (RAN) and core network (CN)
- Software-Defined mobile Network Control (SDMC) decouples control *logic* from control *agents* (enforcement points): slice-specific and slice-dedicated controllers (SDM-C and SDM-X, respectively) hide technology- and implementation-related details of NFs from the control applications. Control logic can easily be modified by replacing an SDM-C/-X application with a new one, the agent (controlled NF) does not necessarily have to be changed. Hence, SDMC not only enables control and data layer split, but also is an important technology for enhanced network programmability. Mobility management, QoS/QoE control, and multi-tenant radio resource management include some of the functionalities that have been designed and evaluated following the SDMC approach.
- Service Management function: It facilitates the automated mapping of service requirements (as defined by tenants) to network slice templates. These templates are further annotated with service-specific parameters that allow for instantiation and activation of a network slice instance.
- Extension of the ETSI NFV MANO framework to achieve integrated lifecycle management for network instances ("network slices") composed of both physical and virtualised network functions (PNFs/VNFs): The 5G NORMA has re-designed the ETSI NFV MANO architecture to include application management functions for managing the application logic of an NF, thus also integrating orchestration and lifecycle management with fault, configuration, and performance management. The novel Inter-slice Resource Broker allows for efficient and SLA-compliant resource sharing across network slices.
- MANO-as-a-Service (MANOaaS): The 5G NORMA MANO layer enables the commissioning of tenant-specific MANO stack instances that allow each tenant to orchestrate and manage allocated resources and functions according to own policies and in a multi-administrative-domain environment.

- Novel security concepts for 5G NORMA networks: Based on a thorough analysis of the security requirements in novel multi-service and multi-tenant mobile networks. innovative security concepts to mitigate identified threats have been developed. This includes a virtualised authentication, authorisation, and accounting (V-AAA) component, a novel RAN (access stratum) security concept, and the Trust Zone specification, allowing for secure operation of temporarily isolated 5G "islands".
- The 5G NORMA ecosystem: The defined stakeholder roles and network slice "Offer Types" have been applied in the context of industrial communications. The analysis has collected recommendations how to operate network slices using network infrastructure from both public mobile network operators and private companies ("verticals"). Moreover, solutions for achieving different levels of isolation and security have been described.
- Architecture verification analysis: The final architecture has been evaluated both quantitatively and qualitatively against the requirement groups as defined by WP2. The evaluation has been performed in the concrete context of the London study area (cf. WP2), where three evaluation cases have facilitated the successful verification of most requirement groups.
- Finally, the work package has outlined possible migration paths from 4G networks to 5G NORMA networks, taking into account both technical and economic considerations. A successful introduction of 5G NORMA networks can only be realised by a gradual transition. Moreover, it requires a longer time period of co-existence with legacy networks, in particular LTE/EPC, that will incrementally be integrated with and finally substituted by 5G NORMA MANO, control, and data layer functions.

WP3 continues to disseminate latest results in both scientific articles/papers and academic/industry events. It has contributed to several SDOs, including 3GPP and IETF. Final statistics on these efforts and activities will be contained in the final deliverable of WP7.

## 8.2 Open issues and future research

5G NORMA has introduced several novelties to mobile network architecture that significantly deviate from previous designs. While one of the most important benefits of the 5G NORMA results comprise a substantial increase in flexibility and adaptability of 5G networks, such benefits also come at the cost of increased complexity in various dimensions

Therefore, based on the experience gained in the 5G NORMA project, important future research questions include:

- How can the increased number and complexity interface relations, resulting from the modular designed, be efficiently coped with?
- Multiplexing gains versus SLA compliance
- What are promising directions to resolve the conflict between the design goal of (logically) centralised control functions via SDMC and the requirement for distributed control functions, e.g., due to latency constraints in RAN?
- How can heterogeneous lifecycle management, application management, and orchestration functions be further integrated and harmonised in order to reach higher levels of automation in service and network management & orchestration, as envisaged by e.g., the Open Network Automation Platform (ONAP)?
- How can inter-network slice control and cross-domain management mechanisms become more efficient and yield the full potential of network slicing, e.g., in terms of reaping multiplexing gains or increased reliability of SLA fulfilment?
- How can the architecture design evolve towards more steady, continuous development and optimisation, e.g., in order to allow for more frequent architecture updates towards a truly cloud-native design?

- What would natively cloud-enabled functions, procedures, and protocol stacks in mobile networks look like and what are generally applicable design criteria and key characteristics?
- How can features for specific vertical sectors, such as increased resilience, customisable levels of security, or increased resource elasticity of network functions, be integrated into the architecture in a more seamless and on-demand manner?
- How can the feasibility and scalability of the SDMC and SDMO concepts be validated in practical, large-scale deployments?
- How can the migration to 5G NORMA networks be realised in ways to not overburden operators and service providers by the experienced cost and technological complexity when rolling out heterogeneous services?

Clearly, some of these research questions and open issues are already tackled by some of the 5G-PPP Phase 2 projects, but many of them will require continuous analysis throughout Phase 3 as well.

# References


[23.203] 3GPP, “TS 23.203, V14.0.0; Technical Specification Group Services and System Aspects, Policy and charging control architecture (Rel. 14)”. June 2016; http,//www.3gpp.org/ftp/Specs/archive/23_series/23.203/23203-e00.zip

[23.251] 3GPP, “TS 23.251 v14.1.0, Technical Specification Group Services and System Aspects; Network Sharing; Architecture and functional description (Rel. 14)”. May 2017; http,//www.3gpp.org/ftp/Specs/archive/23_series/23.251/23251-e10.zip

[23.401] 3GPP, “TS 23.401 v15.1.0, Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Rel. 15)”. Sept. 2017; http,//www.3gpp.org/ftp/Specs/archive/23_series/23.401/23401-f10.zip

[23.501] 3GPP, "TS 23.501 V1.4.0, Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 Rel. 15)". Sept. 2017; http,//www.3gpp.org/ftp/specs/archive/23_series/23.501/23501-140.zip

[25.433] 3GPP, "TS 25.433 v14.1.0; Technical Specification Group Radio Access Network; UTRAN Iub interface Node B Application Part (NBAP) signalling (Rel. 14)." June 2017, http,//www.3gpp.org/ftp/Specs/archive/25_series/25.433/25433-e10.zip

[28.801] 3GPP, "TR 28.801 V15.0.0; Technical Specification Group Services and System Aspects; Telecommunication management; Study on management and orchestration of network slicing for next generation network (Rel. 15)". Sept. 2017; http://www.3gpp.org/ftp//Specs/archive/28_series/28.801/28801-f00.zip

[32.762] 3GPP, “TS 32.762 V11.7.0, Technical Specification Group Services and System Aspects; Telecommunication management; Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS) (Rel. 11) http://www.3gpp.org/ftp/Specs/archive/32_series/32.762/32762-b70.zip

[32.692] 3GPP, "TS 32.692 V11.0.0, Technical Specification Group Services and System Aspects; Telecommunication management; Inventory Management (IM) network resources Integration Reference Point (IRP); Network Resource Model (NRM) (Rel. 11)", http://www.3gpp.org/ftp/Specs/archive/32_series/32.692/32692-b00.zip

[33.107] 3GPP, “TS 33.107, V14.1.0, Technical Specification Group Services and System Aspects, 3G security; Lawful interception architecture and functions (Rel. 14)”. March 2017; http,//www.3gpp.org/ftp/Specs/archive/33_series/33.107/33107-e10.zip

[33.501] 3GPP, "TS 33.501 V0.3.0, Technical Specification Group Services and System Aspects; Security Architecture and Procedures for 5G System (Rel. 15)." Aug. 2017; http,//www.3gpp.org/ftp/Specs/archive/33_series/33.501/33501-030.zip

[38.801] 3GPP, "TR 38.801 v14.0, Study on new radio access technology, Radio access architecture and interfaces (Rel. 14)". March 2017; http,//www.3gpp.org/ftp/Specs/archive/38_series/38.801/38801-e00.zip

[38.913] 3GPP, "TR 38.913 v14.3.0, Study on Scenarios and Requirements for Next generation Access Technologies (Rel. 14)". August 2017, http,//www.3gpp.org/ftp/Specs/archive/38_series/38.913/38913-e30.zip

[5GEx] EU H2020 5GEx, "5G Exchange – enabling cross-domain orchestration of services", https://www.5gex.eu/wp/, retrieved August 2017

[5GCM] Aalto University, BUPT, CMCC, Nokia, NTT DOCOMO, New York University, Ericsson, Qualcomm, Huawei, Samsung, INTEL, University of Bristol, KT Corporation, University of Southern California, "5G Channel Model for bands up to100 GHz"

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

[5GN-D22] EU H2020 5G NORMA, "D2.2: Evaluation methodology for architecture validation, use case business models and services, initial socio-economic results", October 2016

[5GN-D23] EU H2020 5G NORMA, "D2.3: Socio-Economic Findings for 5G NORMA", December 2017

[5GN-D32] EU H2020 5G NORMA, "D3.2: 5G NORMA network architecture – Intermediate report", January 2017

[5GN-D41] EU H2020 5G NORMA, "D4.1: RAN architecture components – preliminary concepts", Nov. 2016.

[5GN-D42] EU H2020 5G NORMA, "D4.2: RAN architecture components – final report", June 2017.

[5GN-D51] EU H2020 5G NORMA, "D5.1: Definition of connectivity and QoE/QoS management mechanisms – intermediate report", November 2016.

[5GN-D52] EU H2020 5G NORMA, "D5.2 Definition and specification of connectivity and QoE/QoS management mechanisms, Final Report", June 2017

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

[BBF384] Broadband Forum, Technical Work in Progress. Available from: https://www.broadband-forum.org/standards-and-software/downloads/technical-work-in-progress

[CGL17] M. Callan, A. Gourinovich, T. Lynn, "The Global Data Centre Market". Market Briefing, July 2017

[Cisco] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2015–2020". Available from: http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf, 2016

[CPRI] Common Public Radio Interface (CPRI), Interface Specification. V7.0. October 2015.

[CROSS-D21] H2020 5G-Crosshaul, "Deliverable D2.1: Study and assessment of physical and link layer technologies for Crosshaul". V2.1, April 2016

[CSC14] A. Csoma, B. Sonkoly, L. Csikor, F. Nemeth, A. Gulyas, D. Jocha, J. Elek, W. Tavernier, and S. Sahhaf, "Multi-layered service orchestration in a multi-domain network environment," in Software Defined Networks (EWSDN), 2014 Third European Workshop on. IEEE, 2014, pp. 141–142.

[CSS+16] S. Costanzo and R. Shrivastava and K. Sarndanis and D. Xenakis and X. Costa-Pérez and D. Grace, "Service-oriented resource virtualization for evolving TDD networks towards 5G", in Proceeding of WCNC'16 – IEEE Wireless Communications and Networking Conference 2016, Doha, Qatar, Apr. 2016.3

[CSZ+14] Chen F., Shan Y, Zhang Y., Wang Y., Franke H., Chang X., and Wang K, "Enabling FPGAs in the cloud". In Proceedings of the 11th ACM Conference on Computing Frontiers (CF '14), Cagliari, 2014

[DOCSIS3.1] CableLabs Consortium, "Data-Over-Cable Service Interface Specifications. DOCSIS 3.1. MAC and Upper Layer Protocols Interface Specification". CM-SP-MULPIv3.1-112-170906. September 2017.

[EC2012] European Commission, "Promoting the shared use of radio spectrum resources in the internal market", https://ec.europa.eu/digital-single-market/en/promoting-shared-use-europes-radio-spectrum, 2012.

[EC2016] European Commission, "5G for Europe: An action plan", https://ec.europa.eu/digital-single-market/en/news/communication-5g-europe-action-plan-and-accompanying-staff-working-document , September 2016.

[ERI2016] Ericsson, "Ericsson mobility report for November 2016", https://www.ericsson.com/assets/local/mobility-report/documents/2016/ericsson-mobility-report-november-2016.pdf

[ES 201 158] ETSI ES 201 158, "Telecommunications security; Lawful Interception (LI); Requirements for network functions". European Telecommunications Standards Institute; April 2002.

[ETSI TC LI] ETSI: "Lawful Interception". European Telecommunications Standards Institute, web site, http://www.etsi.org/technologies-clusters/technologies/lawful-interception

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

[G.9701] ITU-T Recommendation G.9701: Fast access to subscriber terminals (G.fast) - Physical layer specification, December 2014

[GSMA] GSMA report: Mobile Infrastructure Sharing

[Han17] Bin Han, Stan Wong, Christian Mannweiler, Mischa Dohler and Hans D. Schotten: "Security Trust Zone in 5G Networks". 24th International Conference on Telecommunications (ICT), Limassol, Cyprus, May 2017.

[HFC] Nokia, "DOCSIS for LTE small cell backhaul". Nokia Bell Labs consulting White Paper, Nokia, 2016. Online document: resources.nokia.com/asset/180498. Retrieved August 2017

[HKL17] Fang Hao, Murali Kodialam, T. V. Lakshman and Sarit Mukherjee, "Online Allocation of Virtual Machines in a Distributed Cloud", IEEE Transactions on Networking, Vol. 25., No. 1, February 2017

[ICNIRP] "ICNIRP Guidelines for Limiting Exposure to Time- Varying Electric, Magnetic and Electromagnetic Fields (up to 300GHz)", http://www.icnirp.org/cms/upload/publications/ICNIRPemfgdl.pdf

[iJOIN-D53] EU FP7 iJOIN, "D5.3: Final definition of iJOIN architecture", May 2015.

[Intel] Intel, "Distributed multi-domain policy management and charging control in a virtualized environment an etsi* poc reference architecture," Intel, Technical White Paper, Tech. Rep., 2015.

[JCL] https://jclouds.apache.org

[KS16] Kachris C. and Soudris D., "A survey on reconfigurable accelerators for cloud computing". 26th International Conference on Field Programmable Logic and Applications (FPL), Lausanne, 2016, pp. 1-10

[KGI] Khronos Group Inc., "Open CL Project". Website: https://www.khronos.org/opencl/

[KHC14] S. Khatibi, L.M. Correia, “Modelling of Virtual Radio Resource Management for Cellular Heterogeneous Access Networks”, in Proc. PIMRC’14 – IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications, Washington, USA, Sep. 2014.

[KNE16] K. Katsalis, N. Nikaein and A. Edmonds, “Multi-Domain Orchestration for NFV: Challenges and Research Directions”, 15th International Conference on Ubiquitous Computing and Communications,2016, and 8th International Symposium on Cyberspace and Security, 2016.

[LGL+13] Y. Lin, Y. Gao, Y. Li, X. Zhang and D. Yang, “QoS aware dynamic uplink-downlink reconfiguration algorithm in TD-LTE HetNet”, in Proceeding of IEEE Globecom Workshops (GC’13 Wkshps), Atlanta, GA, USA, Dec. 2013.

[MAG-D14] EU H2020 mmMAGIC, "Deliverable D1.4: Use case description, spectrum considerations and feasibility analysis", June 2017

[MAG-D41] EU H2020 mmMAGIC "Deliverable 4.1 Preliminary radio interface concepts for mm-wave mobile communications”, June 2016

[MGB+17] Montori F., Gramaglia M, Bedogni L., Fiore M., Sheikh F., Bononi L., Vesco A., “Automotive Communications in LTE: a Simulation-based Performance Study”, to appear on Proceedings of IEEE VTC-Fall 2017, Toronto, Canada

[MKG+16] Magaki I., Khazraee M, Gutierrez L.V., Bedford Taylor M., “ASIC Clouds: Specializing the datacenter”, IEEE, 2016.

[MUA+15] Muñoz, A., Urueña, M., Aparicio, R., & Rodríguez de los Santos, G. (2015), "Digital Wiretap Warrant: Improving the security of ETSI Lawful Interception". Digital Investigation, 14 (March 2011), 1–16. https://doi.org/10.1016/j.diin.2015.04.005

[mWT] ETSI GS mWT 002 V1.1.1, " Group Specification: millimetre Wave Transmission (mWT); Applications and use cases of millimetre wave transmission". August 2015

[NFV_003] ETSI GS NFV 003 V1.1.1, “Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV”. Oct. 2013, http://www.etsi.org/deliver/etsi_gs/NFV/001_099/003/01.02.01_60/gs_NFV003v010201p.pdf

[NFV_IFA001] ETSI GS NFV-IFA 001 V1.1.1, “Network Functions Virtualisation (NFV) Acceleration Technologies, Report on Acceleration Technologies & Use Cases”, December 2015, http://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/001/01.01.01_60/gs_nfv-ifa001v010101p.pdf

[NFV_IFA002] ETSI GS NFV-IFA 002 V2.1.1, "Network Functions Virtualisation (NFV); Acceleration Technologies; VNF Interfaces Specifications", March, 2016, http://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/002/02.01.01_60/gs_NFV-IFA002v020101p.pdf

[NFV_IFA004] ETSI GS NFV-IFA 004 V2.1.1, “Network Functions Virtualisation (NFV) Acceleration Technologies Management Aspects Specification”, April, 2016,

http://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/004/02.01.01_60/gs_NFV-IFA004v020101p.pdf

[NFV-IFA011] ETSI GS NFV-IFA 011 V2.1.1, "Network Functions Virtualisation (NFV); Management and Orchestration; VNF Packaging Specification". Oct. 2016, http://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/011/02.01.01_60/gs_NFV-IFA011v020101p.pdf

[NFV-IFA014] "ETSI GS NFV-IFA 014 V2.1.1, ""Network Functions Virtualisation (NFV); Management and Orchestration; Network Service Templates Specification "". Oct. 2016, http://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/014/02.01.01_60/gs_NFV-IFA014v020101p.pdf "

[NFV-MAN001] ETSI GS NFV-MAN 001 V1.1.1, "Network Functions Virtualisation (NFV); Management and Orchestration." Dec. 2014, http://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_NFV-MAN001v010101p.pdf

[NFV-SOL004] "ETSI GS NFV-SOL 004 V2.3.1, ""Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; VNF Package specification"". July 2017, http://www.etsi.org/deliver/etsi_gs/NFV-SOL/001_099/004/02.03.01_60/gs_nfv-sol004v020301p.pdf "

[NGFI] "Next Generation Fronthaul Interface". Web page, accessed August 2018 http://standards.ieee.org/develop/wg/NGFI.html

[NGMN] NGMN, "NGMN 5G White Paper". March 2015, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[NGPON2] ITU-T Recommendation G.989.2 (2014) – Amendment 1. 40 "Gigabit-capable passive optical networks 2 (NG-PON2): Physical media dependent (PMD) layer specification"

[NKG+15] Nikaein N., Knopp R., Gauthier L., Schiller E., Braun T., Pichon D., Bonnet C., Kaltenberger F., and Nussbaum D., "Demo: Closer to Cloud-RAN: RAN as a Service". In 21st Annual International Conference on Mobile Computing and Networking. ACM, 2015

[NOKIA] NOKIA: "Nokia Bell Labs first to show use of ultra-low latency 10G PON for mobile fronthaul". Press release, https://www.nokia.com/en_int/news/releases/2017/06/20/nokia-bell-labs-first-to-show-use-of-ultra-low-latency-10g-pon-for-mobile-fronthaul. Accessed August 2017.

[NOMOR] "5G RAN Interfaces and eCPRI". Nomor Research Whitepaper. September 2017

[Obs06] Open Base Station Architecture Initiative (OBSAI), "BTS System Reference Document, Version 2.0", http://www.obsai.com/specs/OBSAI_System_Spec_V2.0.pdf, 2006

[OPB] http://openbaton.github.io

[OS4J] http://www.openstack4j.com

[OSM] https://osm.etsi.org

[PAB+16] L. Peterson, A. Al-Shabibi, T. Anshutz, S. Baker, A. Bavier, S. Das, J. Hart, G. Palukar, W. Snow, "Central Office Re-Architected as a Data Center", IEEE Communications Magazine, October 2016.

[PL2016] Plum Consulting, “Review of efficiencies with Multi-Operator Core Network (MOCN) technology”, http://plumconsulting.co.uk/review-efficiencies-multi-operator-core-network-mocn-technology/, November 2016.

[PLS15] V. Pauli, Y. Li, and E. Seidel, “Dynamic TDD for LTE-A and 5G”, Whitepaper – Nomor Research GmbH, Munich, Germany, Sep. 2015.

[Prab17] Prabhu H., “Hardware Implementation of Baseband Processing for Massive MIMO”, PhD Thesis, The Department of Electrical and Information Technology, Lund University, 2017

[RAH+17] M. Rates Crippa, P. Arnold, D.v. Hugo, V. Friderikos, O. Holland, S. Wong, B. Gajic, B. Sayadi, C. Guerrero, I. Labrador Pavon, Ignacio, V. Sciancalepore, F.Z. Yousaf, “Resource Sharing for a 5G Multi-tenant and Multi-service Architecture”, Proceedings of the 23rd European Wireless Conference, Dresden, Germany, May 2017

[RFC6749] Dick Hardt, “The OAuth 2.0 Authorization Framework”, Internet Engineering Task Force (IETF), Request for Comments: 6749, October 2012.

[RFC7519] M. Jones J. Bradley N. Sakimura, “JSON Web Token (JWT)”, Internet Engineering Task Force (IETF), Request for Comments: 7519, May 2015.

[SCS15] B. Sonkoly, J. Czentye, R. Szabo, D. Jocha, J. Elek, S. Sahhaf, W. Tavernier, and F. Risso, “Multi-domain service orchestration over networks and clouds: a unified approach,” in Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. ACM, 2015, pp. 377–378.

[SR16] Structure Research, "Marketshare Report: Global Data Centre Colocation". Available from: https://structureresearch.net/product/marketshare-report-global-data-centre-colocation, 2016

[SRM15] M.K. Samimi, T.S. Rappaport, G.R. MacCartney, “Probabilisitic Omni directional path loss models for millimeter wave outdoor communications”, IEEE Comms. Letters, Vol. 4, Issue 4, Aug. 2015

[SSP+14] K. Samdanis, R. Shrivastava, A. Prasad, P. Rost, and D. Grace, “Virtual Cells: Enhancing the Resource Allocation Efficiency for TD-LTE”, in Proceeding IEEE 80th Vehicular Technology Conference (VTC14Fall), Vancouver, Canada, Sept. 2014.

[SSP+16] K. Samdanis, R. Shrivastava, A. Prasad, D. Grace, X. Costa-Perez, “TD-LTE Virtual Cell: An SDN Architecture for User-Centric Multi-eNB Elastic Resource Management”, Elsevier Computer Communications Journal, Vol. 83, Jan. 2016.

[SW1] www.spreadsheetweb.com

[TAC] Tacker, https://wiki.openstack.org/wiki/Tacker

[TENOR] https://github.com/T-NOVA/TeNOR

[TOSCA-NFV] OASIS, "TOSCA Simple Profile for Network Functions Virtualization (NFV) Version 1.0", http://docs.oasis-open.org/tosca/tosca-nfv/v1.0/csd04/tosca-nfv-v1.0-csd04.html

[UI15] Uptime Institute, "Uptime Institute’s 2015 Data centre Survey". Available from: https://uptimeinstitute.com/uptime_assets/08200c5b92224d561ba5ff84523e5fdefeec6b58cbf64c19da7338e185a9c828-survey15.pdf, 2015

[UI16] Uptime Institute, "Uptime Institute’s 2016 Data centre Survey". Available from:

https://uptimeinstitute.com/uptime_assets/10605ff30621660fd68cebfee7a8d407831ad4113d896cea1d5a33e4ac331b56-Survey16.pdf, 2015

[XHAUL-D22] 5G-XHAUL Deliverable D2.2. Dynamically Reconfigurable Optical-Wireless Backhaul/Fronthaul with Cognitive Control Plane for Small Cells and Cloud-RANs. July 2016

[XHAUL-D23] 5G-XHAUL Deliverable D2.3 Architecture of Optical/Wireless Backhaul and Fronthaul and Evaluation. February 2017

[YGF+17] F.Z. Yousaf, M. Gramaglia, V. Friderikos, B. Gajic, D.v. Hugo, B. Sayadi, V. Sciancalepore, M. Rates Crippa, "Network slicing with flexible mobility and QoS/QoE support for 5G Networks". Proceedings of the *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1195-1201, 2017.

[ZTE] ZTE, "5G-oriented Optical Transport Network Solution". Whitepaper; online document http://get.knect365.com/zte-whitepaper-ngon2017/, accessed August 2017

# Annex A. Details of Verification Analysis

## A.1. Definitions

### A.1.1. KPI definitions

#### A.1.1.1. Capacity and traffic density

Area wide capacity density describes the ability of a network to provide an amount of data volume per service area during the hour with the highest traffic load (busy hour). As the performance of mobile networks depends on many influencing factors for simplicity it is assumed that during busy hour all radio resources are occupied (fully loaded system). The spectral efficiency is defined as the aggregate uplink / downlink cell full buffer cell throughput per spectrum block assignment bandwidth. Hence cell capacity [Mbit] can be calculated from spectral efficiency by multiplication with system bandwidth and time duration (1 hour).

As macro cells provide full coverage in the whole cell range (except for small percentage described by coverage probability) the capacity contribution by macro cells is sufficiently characterised by macro cell capacity as described above.

Heterogeneous networks consist of different network layer (macro, small cell, WiFi) where some of those network layers are not available in the whole cell (e.g. small cell coverage is assumed to be only a small fraction of the macro cell). Hence the amount of data volume carried by those layers in addition of the node capabilities depends on the positioning of nodes and user distribution (a node at a location without traffic demand cannot contribute to network capacity). The capacity provided by layers with spotty coverage (small cells, Wifi) depends on the number of nodes within the service area and may be determined by measurements or by expert estimation.

Traffic density characterises the demand of data volume during busy hour per service area.

#### A.1.1.2. User plane Latency

UP latency is defined as the one-way transmission time of a packet between the transmitter and the availability of this packet in the receiver. The measurement reference is the MAC layer on both transmitter and receiver side. The analysis must distinguish between UP latency in infrastructure-based communications and in device-to-device (D2D) communications [PA-5GPPP].

#### A.1.1.3. E2E latency

Different types of latency are relevant for different applications. E2E latency, or one trip time (OTT) latency, refers to the time it takes from when a data packet is sent from the transmitting end to when it is received at the receiving entity, e.g., Internet server or another device. Another latency measure is the round trip time (RTT) latency which refers to the time from when a data packet is sent from the transmitting end until acknowledgements are received from the receiving entity. The measurement reference in both cases is the interface between Layer 2 and 3 [PA-5GPPP].

#### A.1.1.4. Peak data rate

The peak data rate is the highest theoretical single user data rate, i.e., assuming error-free transmission conditions, when all available radio resources for the corresponding link direction are utilised (i.e., excluding radio resources that are used for physical layer synchronization, reference signals or pilots, guard bands and guard times). Peak data rate calculation shall include the details on the assumed MIMO configuration and bandwidth [PA-5GPPP].

### A.1.1.5. User Experienced Data Rate

Data rate requirements are expressed in terms of user experienced data rate, measured in bit/s at the application layer. The required user experienced data rate should be available in at least 95% of the locations (including at the cell-edge) for at least 95% of the time within the considered environment. The user experienced data rate requirement depends on the targeted application/use case. It is set as the minimum user experienced data rate required for the user to get a quality experience of the targeted application/use case [NGMN].

### A.1.1.6. Mobility

Mobility refers to the system's ability to provide seamless service experience to users that are moving at a certain speed. Mobility requirements may be specified by a maximum percentage decrease of user throughput that is caused by increasing the device velocity [NGMN].

### A.1.1.7. Device density

Device density denotes the number of devices per service area that are connected to the network. For connection-oriented services devices must be in active or idle mode. For connectionless services devices just have to be within coverage area of the network.

### A.1.1.8. Reliability

The reliability of a communication is characterised by its reliability rate, defined as follows: the amount of sent packets successfully delivered to the destination within the time constraint required by the targeted service, divided by the total number of sent packets. Note that the reliability rate is evaluated only when the network is available [NGMN].

### A.1.1.9. Availability

The availability in percentage is defined as the number of places (related to a predefined area unit or pixel size) where the QoE level requested by the end-user is achieved divided by the total coverage area of a single radio cell or multi-cell area (equal to the total number of pixels) times 100.

(Note: FANTASTIC-5G defines availability as equal to (1 – service blocking probability), where service blocking probability is due to lack of enough resources to access, grant and provide the service, even in case of adequate coverage) [PA-5GPPP].

### A.1.1.10. Coverage

Coverage probability refers to geographical locations and indicates the percentage of locations with respect to the whole service area where a certain service can be provided.

### A.1.1.11. Description of functional requirements according to D2.1

**Table A-1: Identified groups of functional requirements**

| Requirement group | Group name |
|---|---|
| **RG#1** | Fast network reconfiguration within a network slice |
| **RG#2** | Fast network reconfiguration between network slices |
| **RG#3** | Device duality |
| **RG#4** | Separation and prioritization of resources on a common infrastructure |

| | |
|---|---|
| **RG#5** | Multi-connectivity in access and non-access part of the 5G system |
| **RG#6** | Massive scalability of protocol network functions |
| **RG#7** | Highly efficient transmission & processing |
| **RG#8** | QoE/QoS awareness |
| **RG#9** | Adaptability to transport network capabilities |
| **RG#10** | Low latency support |
| **RG#11** | Security |

The idea for the Requirement Groups is to facilitate the design process, so requirements that are similar enough, or that can be addressed with the same technological solutions, are not considered separately.

The different Requirement Groups (RGs) are defined as follows:

(1) RG#1: Fast network reconfiguration within a network slice

This group covers functional requirements which are related to a fast reconfiguration of the network and its NW functions, respectively, within a dedicated NW slice during running operation. Fast reconfigurations may happen e.g. in case of failures on NW elements or on the links between the elements belonging to the same slice instance.

(2) RG#2: Fast network reconfiguration between network slices

In contrast to RG#1 this group covers functional requirements which are related to a fast reconfiguration of NW and NW functions between different slices (i.e. in a multi-tenancy operation). Slices are – from a logical perspective – independent virtual networks, but as they will be operated on a common infrastructure, changes in one slice may have impact onto other slices (e.g. w.r.t resource utilization).

(3) RG#3: Device duality

This group is related to functional requirements addressing aspects of device duality in a future 5G system. Device duality means that a device can act both as "usual" end user device (incl. sensor types) and as a network node extending the infrastructure part of the system [MET15-D64]. Examples are e.g. devices acting as cluster head for sensor nodes in their neighbourhood or devices acting as relays for other devices covering also self-backhauling functionalities with radio resource sharing (not for P2P). Moreover, requirements related to D2D communication aspects belong to RG#3 (both for NW controlled D2D and D2D without impact from infrastructure, e.g. V2V communication outside the radio cell coverage or links via several devices in emergency situations).

(4) RG#4: Separation and prioritization of resources on a common infrastructure

RG#4 addresses functional requirements w.r.t. separation and prioritization of resources on a common infrastructure. A 5G system architecture based on SDN/NFV concepts allows a flexible management of network, storage, and computing resources, but has to consider the separability of the resources for operational and security purposes, e.g. running different slice instances in parallel. In addition to separation the resource usage has to be prioritized according to needs of the different services/slices, so a joint resource management is expected to be implemented with further differentiation on available infrastructure layers.

(5) RG#5: Multi-connectivity in access and non-access part of the 5G system

This group covers functional requirements related to multi-connectivity in access and non-access part of the 5G system. Multi-connectivity via different radio access technologies (RATs) or via different links of the same RAT (e.g. via different sites and/or frequency bands) will result in

better performance (e.g. data throughput) and/or increased reliability for services to be offered within 5G. This is also true for the non-access part, e.g. utilizing the redundancy in transport network links.

(6) RG#6: Massive scalability of protocol network functions

RG#6 is related to functional requirements for massive scalability of protocol NW functions. The 5G system has to support different services with strongly diverging demands. To fulfil those demands on a common infrastructure the NW functions in the communication protocol stacks have to be flexibly scalable and adaptable.

(7) RG#7: Highly efficient transmission & processing

This group covers functional requirements which are related to highly efficient transmission & processing of data. Examples for that feature are realizations of NW functions inside the radio protocol stack allowing e.g. fast access of devices for mMTC with extremely low overhead in C-Plane signalling.

(8) RG#8: QoE/QoS awareness

RG#8 addresses functional requirements for QoE/QoS awareness, i.e. the adaptability of the NW and its NW functions, respectively, to the demands of the services offered to the customers. This covers not only the processing during NW operation, but also during instantiation of slices as QoE/QoS demands of associated services limits flexible placement of NW functions on available elements of the NW infrastructure.

(9) RG#9: Adaptability to transport network capabilities

Functional requirements related to the adaptability to transport network capabilities are covered by RG#9. The set-up and operation of virtual NWs (slices) has to take care of the available transport NW capabilities between NW elements. Different deployment scenarios w.r.t. the RAN can be addressed dependent on suitability of transport capabilities for ideal or non-ideal backhaul/fronthaul, resulting in distributed or centralized placement of radio NW functions (D-RAN, C-RAN).

(10) RG#10: Low latency support

In RG#10 all functional requirements are collected related to the support of low latency service creation. They are addressing architectural solutions like mobile edge computing, i.e. placement of NW functions and their operation nearest to the access link.

(11) RG#11: Security

This group covers all functional requirements which are related to security aspects in a 5G system, i.e., to all aspects of how to secure the network and the traffic in it against cyber-attacks.

## A.1.2. Quantitative and qualitative service requirements

The following tables provide an overview of the service components considered in 5G NORMA, the associated requirements, and the reference to the respective section where the results are evaluated.

**Table A-2: Performance requirements for selected services**

| Service component | User Experienced Data Rate | Latency | 5G NORMA improvements against legacy | Link to result discussion |
|---|---|---|---|---|
| eMBB – consumer portable devices | 10 Mbps DL/UL | 100 ms | Improved QoE | Section 6.4.1.1 |
| V2I – infotainment | 10 Mbps DL | 100 ms | Improved QoE | Section 6.4.3.1.2 |

| | | | | |
|---|---|---|---|---|
| (eMBB) | | | | |
| V2I – assisted driving<br>(uMTC) | 0.5 Mbps DL/UL | <100 ms | More steady and lower latency, improved service coverage | Section 6.4.3.1.2 |
| V2I – driver information service (mMTC) | 0.5 Mbps DL/UL | <100 ms | More steady and lower latency, improved service coverage | Section 6.4.3.1.2 |
| Environmental monitoring, waste management, and congestion control (mMTC) | 2 bps UL | > 50 ms | Improved outdoor coverage | Section 6.4.3.1.1 |
| Smart meters - sensor data, meter readings, individual device consumption (mMTC) | 2 bps UL | > 50 ms | Improved indoor coverage, more efficient protocols for small data packages | Section 6.4.3.1.1 |
| Smart grid<br>sensor data and actuator commands (mMTC) | 2 bps UL | > 50 ms | Improved outdoor coverage | Section 6.4.3.1.1 |
| Logistics – sensor data for tracking goods (mMTC) | 2 bps UL | > 50 ms | Improved outdoor coverage | Section 6.4.3.1.1 |

**Table A-3: Traffic demand and coverage requirements of selected services**

| Service component | Data volume | Number of devices | Coverage | Link to result discussion |
|---|---|---|---|---|
| eMBB – consumer portable devices | On average, each device consumes 0.23 GB per day in 2020 growing to 2.85 GB by 2030 (29% CAGR) | 2020:<br>43k per sqkm<br>2030:<br>47k per sqkm | 95% Outdoor | Section 6.4.1.1.3 |
| V2I – infotainment (eMBB) | 1 GB-25 GB per day per car (2020-2030) | 325 vehicles per sqkm | 95% (vehicles, outdoor) | Section 6.4.3.1.2 |
| V2I – assisted driving (uMTC) | On average 52 MB consumed per day per car in study area in 2021 growing to 1,503 MB per day per car in 2030 (45% CAGR) | 325 vehicles per sqkm | 99.9% (Vehicles, outdoors) | Section 6.4.3.1.2 |

| | | | | |
|---|---|---|---|---|
| | Note 0% uptake in 2020 so on average 0 demand per car for this service. | | | |
| V2I – driver information service (mMTC) | On average 52 MB consumed per day per car in study area in 2020 growing to 1,711 MB per day per car in 2030 (42% CAGR) | 325 vehicles per sqkm | 95% (vehicles, outdoor) | Section 6.4.3.1.2 |
| Environmental monitoring, waste management, and congestion control (mMTC) | On average 229 bytes per day per roadside item (i.e. traffic lights, road signs, bins etc.) in 2020 growing to 1,516 bytes per day per roadside item by 2030 (21% CAGR) | 100 devices per sqkm | 95% (Outdoors) | Section 6.4.3.1.1 |
| Smart meters - sensor data, meter readings, individual device consumption (mMTC) | 1,600 bytes per smart meter per day i.e. 200 byte messages, 8 messages per day | 30k per sqkm<br>100% uptake assumed from 2020 so no growth over time. | 99% (Indoors) | Section 6.4.3.1.1 |
| Smart grid sensor data and actuator commands (mMTC) | 60kbytes per smart grid neighbour area network (NAN) gateway based on 20 byte commands, 10 messages per day per smart meter device being controlled. | 1.3 smart grid neighbour area network gateways per km2 in 2020 growing to 25.6 per km2 by 2030 based on smart grid uptake in IR2.2 (each controlling 300 smart meter devices). | 95% (Outdoors) | Section 6.4.3.1.1 |
| Logistics sensor data for tracking goods (mMTC) | 4 MB per day per equipped vehicle based on 200 byte messages, 100 messages per day (i.e. updates every approx. 15 mins) per sensor. | On average 9 smart logistics vehicles per km2 in 2020 growing to 58 by 2030 i.e. a 21% CAGR.<br>Each vehicle assumed to have 200 tracked items so sensor density growing from 1,800 to 11,600. | 95% (vehicles, outdoor) | Section 6.4.3.1.1 |

**Table A-4: Functional requirements for selected services**

| Service component | Requirements | Link to result discussion |
|---|---|---|
| eMBB – social media | Application awareness | [5GN-D32] Section 6.2.2.1.3 |
| | Multi-layer and multi-RAT connectivity | [5GN-D32] Section 6.2.2.1.4 |
| | Efficient backhaul | [5GN-D32] Section 6.2.2.1.2 |
| | User privacy and security | Section 6.4.1.2.1 |
| | Capacity for uplink and downlink | |
| V2I – infotainment (eMBB) | Application awareness | [5GN-D32] section 6.2.2.1.3 |
| | Multi-layer and multi-RAT connectivity | [5GN-D32] section 6.2.2.1.4 |
| | Efficient backhaul | [5GN-D32] section 6.2.2.1.2 |
| | User privacy and security | Section 6.4.1.2.1 |
| | Capacity for uplink and downlink | |
| V2I – assisted driving (uMTC) | Targeted dissemination of safety messages | [5GN-D41], page 49 |
| | Optimizations for control plane and data plane functions | [5GN-D32] Section 6.2.2.3 |
| | The system should guarantee the coexistence of safety and non-safety vehicular applications operating over the same scenario. | |
| | Very high network availability and therefore superior robustness against attacks, in particular DoS attacks, is required. This includes strong authentication between devices and network in order to prevent unauthorized communication. Moreover, integrity protection and encryption is required for the signalling traffic and – unless the applications build on application layer security mechanisms – also for the user plane. Security mechanisms must be robust against loss of network nodes; security mechanisms must be available also in RAN parts that are isolated from central components. Security aspects are of high importance for the use case. | Section 6.4.3.1.2 |
| V2I – driver information service (mMTC) | The system should guarantee the coexistence of safety and non-safety vehicular applications operating over the same scenario | Section 5 |
| | The mobility management should support stationary, nomadic, and highly mobile devices and should consider also roaming across network boundaries. | Section 6.4.3.3 |
| Environmental monitoring, waste management, and congestion control (mMTC) | Depending on device type the network access should be applicable via dedicated RATs and frequency bands or in a flexible way | Section 7.3.2 |
| | The mobility management should support stationary, nomadic, and highly mobile devices and should consider also roaming across network boundaries. | Section 6.4.3.3 |
| | The system should support both unidirectional as well as bidirectional communication between sensors and other radio nodes. | |
| | The network should provide flexible security and authentication procedures for mMTC as well as means for easy security credential provisioning for massive number of devices. | Section 5 |
| Smart meters - sensor data, meter readings, individual device | Depending on device type the network access should be applicable via dedicated RATs and frequency bands or in a flexible way | Section 6.4.3.1.1 |
| | The mobility management should support stationary, nomadic, and highly mobile devices | Section 6.4.3.3 |

| | | |
|---|---|---|
| consumption (mMTC) | and should consider also roaming across network boundaries.<br>The system should support both unidirectional as well as bidirectional communication between sensors and other radio nodes.<br>The network should provide flexible security and authentication procedures for mMTC as well as means for easy security credential provisioning for massive number of devices. | Section 5 |
| Smart grid sensor data and actuator commands (mMTC) | The 5G system should support appropriate authentication for low power devices/sensors.<br>The 5G system should be able to accept unsolicited information from large numbers of sensor devices without the need for bearer establishment or mobility signalling (mobility signalling is not required because it is not "connected" to any particular node)<br>The system should support an infrequent uplink data transfer in a "non-connected" mode<br>The system should be able to deactivate the service and sensors, possibly for future use. | Section 5<br><br>Section 6.4.3.1.1 |
| Logistics – Sensor data for tracking goods (mMTC) | The protocol stack (access/core) should allow the management of a massive number of devices w.r.t. ID management and addressing.<br>The access network should handle the network resources in C-Plane and U-Plane in a highly efficient way, it should especially only require a low signalling overhead.<br>The system should support the use of sensors-type devices with very low cost and long battery lifetime.<br>Depending on device type the network access should be applicable via dedicated RATs and frequency bands or in a flexible way.<br>The mobility management should support stationary, nomadic, and highly mobile devices and should consider also roaming across network boundaries.<br>The system should support both unidirectional as well as bidirectional communication between sensors and other radio nodes.<br>The network should provide flexible security and authentication procedures for mMTC as well as means for easy security credential provisioning for massive number of devices. Such security aspects are of high importance for the use case. | |

**Table A-5: Service overarching functional requirements**

| Requirement | Link to result discussion |
|---|---|
| Network programmability | Annex A.2.3.2 |
| QoE based routing and network agility | Annex A.2.3.3 |
| Edge function mobility | Annex A.2.3.4 |
| Slice and service specific mobility concepts | Annex A.2.3.6 |

**Table A-6: Operational requirements**

| Requirement | Link to result discussion |
|---|---|
| Multi-tenant dynamic resource allocation (D3.1) | [5GN-D32] Section 6.2.3.1 Section 6.4.2.2 |
| Saving of operational and capital expenditures (D3.1) | [5GN-D32] Section 6.2.3.2 |
| Service specific and context-aware derivation of service requirements, adaptation and placement of VNF (D3.1) | [5GN-D32] Section 6.2.3.3 |
| Flexible vertical-specific and service-specific detection of traffic and dynamic network monitoring (D3.1) | [5GN-D32] Section 6.2.3.4 |
| Adaptation and placement of VNF (D3.1) | Annex A.2.3.1 |
| Capability of spectrum sharing or reuse (NGMN) | Section 6.4.2.2.3 |

**Table A-7: Security requirements [5GN-D31]**

| Requirement | Link to result discussion |
|---|---|
| Tenant isolation | [5GN-D32] Section 6.2.4 |
| Secure Software Defined Mobile Network Control | [5GN-D32] Section 6.2.4 |
| Physical VNF separation | [5GN-D32] Section 6.2.4 |
| Flexible security | [5GN-D32] Section 6.2.4 |
| Support of reactive security controls | [5GN-D32] Section 6.2.4 |
| Security orchestration | [5GN-D32] Section 6.2.4 |
| Reliable fallback | [5GN-D32] Section 6.2.4 |
| Service related requirements as identified as part of functional requirements | [5GN-D32] Section 6.2.4 |

**Table A-8: Soft KPIs**

| Requirement | Link to result discussion |
|---|---|
| Interfaces between Service Management and Management and Orchestration | [5GN-D32] section 6.2.5.1 |
| Scalability of centrally arranged management and control functions | [5GN-D32] Section 6.2.5.2 |
| Feasibility of growing number of slices | [5GN-D32] Section 6.2.5.3 |
| Roles of external & internal interfaces | Section 6.4.3.4.1 |
| Demonstrator learnings | Annex A.2.4 |
| Feasibility of charging and lawful interception (C&LI) | Section 6.4.3.4.3 |

# A.2. Evaluation details

## A.2.1. Baseline evaluation

### A.2.1.1. Backhaul aspects

To enable the most challenging services targeted by 5G, 5G networks are being specified so that they will permit to increase performance over current mobile networks with figures as challenging as:

- mobile data volume per geographical area, around hundreds of Gbps/km²,
- number of connected devices, over a million 40,000 terminals/km²,
- 5x improvement in end-to-end latency, targeting figures lower than 5 ms.

These figures represent significant improvements in network capacity, peak bandwidth, and latency over current deployed mobile networks. The increase in performance targeted for 5G standards will not only impact Radio Access Network performance requirements, but will also lead to a similar increase in requirements for the transport networks. Capacity-wise, backhaul pipes will need to grow (tens of Gbps optical and wireless e.g. at mmWave frequencies). Latency-wise, the backhaul needs fast and resilient forwarding as well as intelligent leveraging of the edge networking and computing infrastructure.

Cost-wise, there is a need for using less fibres where possible (e.g. DWDM and/or fibre-like wireless). Programmability of the backhaul control and its centralization become essential to add flexibility and agility for shorter service deployment times and traffic-aware backhaul network management so that one can achieve higher (energy) efficiency.

#### A.2.1.1.1. Backhaul and fronthaul requirements for heterogeneous functional splits

Current eNodeB architectures already permit to separate Remote Radio Unit (RRU) RF functionality from baseband central processing Baseband Unit (BBU), by a transport segment called fronthaul, in order to improve deployment and operational efficiency, simplifying complexity of the equipment installed near tower masts. Also, separation of baseband processing and RF functionality permits baseband pooling, also known as C-RAN (Centralized RAN), with additional benefits, such as sharing BBU resources to improve efficiency, joint processing among base stations to reduce interference, introducing mobile edge computing (MEC) at the network edge, etc.

These benefits come at the expense of a high bandwidth consumption in the fronthaul segment, and limitations on the maximum distance between RRU and BBU.

The increase in bandwidth and improved latency make these architectures much more challenging for the new 5G standards. Realistic deployments will need to take into account the costs of this infrastructure.

It is likely that a combination of decentralized and centralized architecture, leveraged also on virtualization technologies, will be finally deployed, as a result of a case-by-case network planning process for each particular scenario, taking the form of what it is currently being named as Cloud RAN, which is supported by 5G NORMA through the flexible and dynamic slicing of network resources, for both centralized and decentralized network deployments.

As commented, extreme cases of Cloud RAN deployments are those of centralized C-RAN based on CPRI (Common Public Radio Interface) [CPRI] or OBSAI (Open Base Station Architecture Initiative) [Obs06].

In Centralized RAN (C-RAN), HARQ is processed centrally, which renders a maximum distance between BBU and RRH due to HARQ stringent latency requirements. Typically, it is specified a value between 50-75 μs one-way delay propagation to accommodate also packet processing in the BBU hardware). Given typical propagation speed in an optical fibre, maximum distance is around 10-15 km.

Above all, whereas C-RAN permits the construction of very simple remote heads, it entails the challenge of requiring high bandwidth (e.g. a 1 sector 20 MHz 4 antenna LTE eNodeB, would require a dedicated (dark fibre[16]) link of 4,9 Gbps in the fronthaul segment). Peak data rate increases targeted for 5G systems represent a challenge in bandwidth that current or near-future

[16] Dark-fibre provides access to the entire physical medium, or at least an optical band within the physical medium i.e 82 50GHz channels.

transport technologies may not offer, and thus different functional splits with reduced fronthaul requirements are being investigated.

Project H2020 5G XHAUL [XHAUL-D23] derives theoretical results for data rate in the transport for different RAN functional splits, corresponding to 5G scenarios defined by 3GPP [38.913]. 5G XHAUL focuses on the evaluation of certain splits (defined as A, B, C, and D by the project), considered as the most promising ones by the project, as shown in Figure 6-7. Those functional splits are then benchmarked between them and against CPRI as shown in Table A-9.

**Table A-9: Backhaul data rates for several functional splits [XHAUL-D23]**

| Parameter | Symbol | LTE | Sub-6 | Low mmWave | High mmWave |
|---|---|---|---|---|---|
| Carrier Frequency [GHz] | $f_C$ | 2 | 2 | 30 | 70 |
| Channel Size [MHz] | $BW$ | 20 | 100 | 250 | 500 |
| Sampling Rate [MHz] | $f_S$ | 30.72 | 150 | 375 | 750 |
| # Antennas | $N_A$ | 4 | 96 | 128 | 256 |
| # ADC/DAC chains | $N_P$ | 4 | 16 | 12 | 10 |
| # Layers | $N_L$ | 4 | 16 | 12 | 10 |
| Overhead | $\gamma$ | 1.33 | 1.33 | 1.33 | 1.33 |
| Quantizer resolution time domain | $N_{Q,T}$ | 15 | 15 | 12 | 10 |
| Quantizer resolution frequency domain | $N_{Q,F}$ | 9 | 9 | 8 | 7 |
| Modulation order | $M$ | 64 | 1024 | 256 | 64 |
| Max. code rate | $R_C$ | 0.85 | 0.85 | 0.85 | 0.85 |
| Frame duration [ms] | $T_F$ | 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 | $\mu$ | 1 | 1 | 1 | 1 |
| Formula data rate split CPRI | $D_{CPRI} = 2 \cdot N_A \cdot f_S \cdot N_{Q,T} \cdot \gamma$ | | | | |
| Formula data rate split A | $D_A = 2 \cdot N_P \cdot f_S \cdot N_{Q,T} \cdot \gamma$ | | | | |
| Formula data rate split B | $D_B = 2 \cdot N_P \cdot N_{SC,act} \cdot N_{Sy} \cdot N_{Q,F} \cdot T_F^{-1} \cdot \mu \cdot \gamma$ | | | | |
| Formula data rate split C | $D_C = N_L \cdot N_{SC,act} \cdot N_{Sy} \cdot R_c \cdot \log_2 M \cdot T_F^{-1} \cdot \mu \cdot \gamma$ | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Peak data rate split CPRI [Gbps] | $D_{CPRI}$ | 4.9 | 574.6 | 1532.2 | 5107.2 |
| Peak data rate split A [Gbps] | $D_A$ | 4.9 | 95.8 | 143.6 | 199.5 |
| Peak data rate split B [Gbps] | $D_B$ | 1.6 | 34.9 | 49.8 | 72.6 |
| Peak data rate split C [Gbps] | $D_C$ | 0.46 | 16.5 | 21.2 | 26.5 |

As shown in Table A-9, the most demanding 5G scenarios would require a fronthaul capability ranging from 26.5 Gbps in split C to more than 5 Tbps in CPRI that not any current transport technology may currently provide cost-effectively[17].

3GPP is currently analysing several functional splits [38.801] alternatives, shown in Figure A-1. For higher layer splits, 3GPP has favoured Option 2. For lower layer splits, that permit efficient interference management, enhanced scheduling or joint transmission, the decision in [38.801] is left for further study. However, Option 6 (split C in 5G XHAUL) and Option 7 (splits A and B in 5G XHAUL) are considered as most promising candidates.

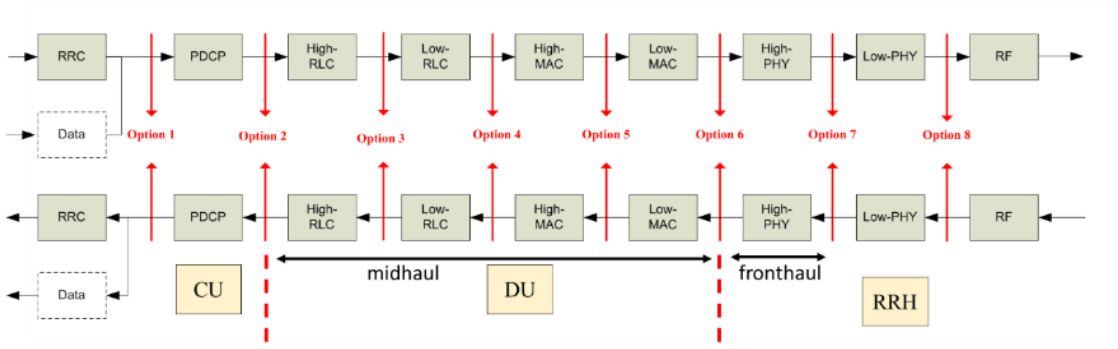



**Figure A-1: RAN functional split separation alternatives**

In any case, the split to be eventually deployed will be conditioned by the characteristics of the transport network available (capacity, jitter, delay). 3GPP considers preferable that several options are supported to accommodate to the specific transport network characteristics.

In fact, it has been proposed that the traditional CPRI BBU is split [ZTE] [NOMOR] into two separate entities CU (processing higher layer functions) and DU (processing lower layer functions), splitting fronthaul into two separated segments or domains, one between the CU and DU (*midhaul*), and one between DU and RRU (*fronthaul*).

#### A.2.1.1.2. Opportunities for multiservice and mobile backhaul/fronthaul convergence

Current C-RAN fronthaul deployments are based on CPRI, using microwave technologies (E-band) or fibre deployments. Current fibre deployments do not use any optical technology, but grey interfaces.

[17] Long distance transport link bandwidth is of Tbps but challenge for fronthaul is costs.

CPRI does not allow for statistical multiplexing, aside including a limited number of different sectors of the same RRU, and is limited to point to point (dark-fibre) connections, synchronization and bandwidth being some of the technical reasons behind these limitations.

When the number of RRUs increase, the operator must deploy optical transmission to reach the BBU site. In 5G, targeted data rates imply the need for massive number of antenna elements (e.g. massive MIMO) and large spectrum e.g. mmWave in the access, leading to waveform samples for fronthaul that would lead to data rates exceeding 5 Tbps.

The adoption of heterogeneous functional splits with reduced fronthaul bandwidth requirements, open the possibility of developing technologies allowing also for statistically multiplexing and coexistence with classical backhaul and the rest of the services in the network.

H2020 project 5G Crosshaul [CROSS-D21] is developing such a 5G integrated fronthaul and backhaul transport network solution, enabling a flexible interconnection of the radio access with the core network by software-defined reconfiguration of all network elements.

In the standardization area, IEEE "Next Generation Fronthaul Interface" [NGFI] working group is defining a standard in which Ethernet, as a common switching protocol, is being used as the supporting technology for fronthaul, paving the way for a converged fronthaul and backhaul, i.e., a Crosshaul, featuring common data, control, and management planes. Also, CPRI is defining eCPRI, an open specification for 5G fronthaul with a new split point, announced to enable a ten-fold reduction of the required bandwidth, with Ethernet as underlying technology.

This possibility will be key for the success of 5G deployments, since it will enable the deployment of a single converged transport network, in which resources which are efficiently shared between 5G, and the rest of enterprise and broadband existing services in the network with the consequent benefits in infrastructure investment and operational efficiency.

#### A.2.1.1.3. Available technologies for fronthaul/backhaul

Among the different currently existing technologies available (optical wireless, fibre, wireless, PON, xDSL, etc.), it can be assumed that fronthaul and backhaul will mainly consist mainly of fibre deployments, complemented for specific scenarios (e.g. small cells deployments) with wireless technologies.

- Wireless Technologies
  5G XHAUL [XHAUL-D22] is proposing the use of mmWave (E-Band) point-to-point links for small cell fronthaul links also complemented by sub-6GHz mesh transport for the backhaul. It needs to be noted that, since those same frequency bands are also being considered for next generation RAN, backhaul availability will be conditioned by its use in RAN. To overcome this limitation 5G-XHAUL proposes that "the same wireless BH technology can be used for both BH and access links, making more efficient use of spectrum resources as they can be shared dynamically".
  H2020 project 5G CROSSHAUL also proposes the microwave band for moderate bandwidth scenarios, especially for compressed CPRI or reduced bandwith functional splits scenarios, and mmWave band in the the V-band (57- 66 GHz) and E-band (71-76 GHz and 81-86 GHz) for large data rates, where large unused continuous bands (several Ghz each) exist that could be used for fronthaul and backhaul applications, following the proposals from the ETSI Millimetre Wave Transmission Group [mWT]. 5G CROSSHAUL [CROSS-D21], provides a state of the art of commercial technologies showing commercial products capable of providing well over 1 Gbps bandwidth, and 40-50 μs latency figure, which are well suited for last hop of backhaul/fronthaul of 5G small cells.
- Optical Technologies
  Current optical backhaul networks are deployed using IP/MPLS equipment on top of ring physical topologies. This physical infrastructure utilizes ring ducts from the SDH deployment, where the number of fibres was limited. The IP/MPLS equipment evolved

from 1G to 10G technologies, but now they must migrate from 10G interfaces to 100G. Such evolution must consider not only the packet layer, but also the underlying infrastructure with its limitations.

Packet layer typically consists of a cascade of routers in different segments to aggregate the traffic. At the lower part of the hierarchy, the traffic from different base stations (BS) and from the DSLAMs and OLTs is aggregated until the IP edge is reached. There are rural areas where extra hierarchical levels are needed due to the low capacity demands. The increment of the traffic in the back-haul and the synergy with fixed-subscribers, forces to have more capacity at the infrastructure level. At the packet level, the operator may consider changing the 10G rings into 10G stars or 100G rings. However, the deployment of optical technologies allows creating logical stars even when the fibre is deployed forming ring structures, using in a more efficient way the deployed fibre plant. Both network segments (fronthaul and backhaul) face similar challenges as the operator may decide whether to deploy optical technologies and when. The initial point at both network segments is that the operator has dark fibre deployments.

However, when facing fronthaul, especially in low layer functional splits, there will be the need for increased requirements for time and frequency synchronization, latency and bandwidth.

The best solution for both scenarios, it is to use coloured interfaces at the router equipment, using simple passive filters for the sake of adding/dropping signals at the fibre topology. There are some limitations to this solution: the number of passive filters and the distances. When the number of hops is high or the distances are significant, the only solution is to deploy CWDM or DWDM technology with amplification. In this way, the distance is a limiting factor due to the latency restriction of the services.

Latency, synchronization and jitter requirements may discard some technologies like current generation of OTN, for front-haul or long-haul links to reach remote areas.

In OTN, aside latency introduced by FEC implementations, there are some considerations derived from the tight frequency and time synchronization requirements for technologies such as CPRI [G.SUP56]. However, some public announcements are appearing from industry [ZTE], that introduce optimizations in OTN technologies to satisfy latency with a combination of frame and FEC optimization, more powerful DSP and the introduction of time synchronization protocols such as IEEE 1588v2.

Some transport equipment vendors are proposing packet-optical solutions. This equipment has not only the transponders for CWDM/DWDM, but also a packet matrix that enables the aggregation of traffic at the transport equipment. The use of packet-optical technologies may end up in a reduction on the number of ports at the routers, as the traffic is by-passed and aggregated at the transport layer. The drawback of this approach is that the network operation is more complex. Instead of using L3 services from the edge, the transport equipment and the routers must be jointly configured.

#### A.2.1.1.4. Opportunities for a convergent fixed-mobile deployment in the last mile

To finish with the analysis, it is included a specific mention to the possibility of using already deployed broadband networks for the mobile backhaul, either based on copper or fibre

In general, copper networks will not satisfy the required bandwidth for mobile backhaul in 5G. Most of xDSL technologies do not satisfy the required bandwidth for mobile backhaul in LTE or 5G systems, although there might be an opportunity for the use of G.FAST [G.9701], able to provide 150 – 300 Mbps DL in short links (50-100m) for the deployment of small cells backhaul. There could also be an opportunity of using cable networks (DOCSIS3.1) able to provide 10G/1G in a point to multipoint architecture for mobile backhaul or midhaul of small cells, although there is no experience of its use in commercial networks for LTE backhaul [HFC].

In any case, latency, synchronization and jitter requirements for 5G mobile backhaul or midhaul should be carefully checked in any case prior to the deployment of the solution. It should be also

noted that the strong asymmetry DL/UL bandwidth of this solutions (typically 1:10) would greatly limit the applicability of the solution

Fibre technologies offer higher bandwidth and lower latency than copper, and are in general more suited to backhauling applications. However, for 5G scenarios, they are still somehow limited in bandwidth for the general 5G targeted scenarios, due its point to multipoint nature in which available bandwidth (2,5G/1,25G in GPON, and XGS-PON (10G/10G) would be shared between a number of fixed users and base stations.

Future NGPON2 [NGPON2] technology increases available bandwidth to 40Gbps symmetrical, which increases its applicability and includes an optional PtP WDM PON, able to map host OTU2 or CPRI option 7 (10G) line rates, which could represent a real alternative for small cells deployments of 5G backhaul in a convergent fixed-mobile architecture

#### A.2.1.1.5. SDN and Network Slicing

The variety in service requirements for 5G and the necessity to create network slices on demand will also require an unprecedented flexibility in the transport networks, which will need to create dynamically connections between geographically distributed sites (likely across different network domains), network functions or even users, providing resource sharing and isolation.

The versatile consumption of resources and the distinct nature of the functions running on them can produce very variable traffic patterns on the networks, changing both the overlay service topology and the corresponding traffic demand. In order to adapt the network to the emergence of 5G services it is required the provision of capacity on demand through automatic elastic connectivity services in a scalable and cost-efficient way.

Those requirements for flexibility and dynamicity across different network domains, along with the need for efficient consumption of resources, reinforces the demand for network programmability that transport networks already face.

SDN decouples network control and forwarding planes, and places control in a (logically) centralized controller. Northbound, SDN controller provides an API to higher layer control applications, abstracting network resources to them. Southbound, it controls connectivity of forwarding nodes, typically through their embedded SDN agents or NETCONF interfaces. Also, the centralized control plane capabilities provide E2E visibility of network resources for establishing and maintaining and optimized connectivity.

Standardization of interfaces will be key in order to have a common way of controlling transport infrastructures, vendor agnostic and multi-domain.

For many years, a combination of SDN and non-SDN enabled elements, physical and virtual elements will coexist. To facilitate an E2E view of the network, network resources need to be treated as generic resources, leaving the specifics of each technology to specific domain controllers during this coexistence period.

## A.2.2. Multi-tenant evaluation

### A.2.2.1. Current RAN sharing options

Today's network sharing is typified by different levels ranging from passive sharing of cell sites and masts to sharing of radio access networks (RANs) and other active elements such as network roaming and the core [GSMA_1]. Reasons and drivers for site sharing differ between countries and markets according to level of market maturity.

**Site (co-location) and mast sharing**

In early phases of network deployment network sharing is most commonly site sharing to facilitate fast network roll-out at lower cost by new entrants.

Another reason motivating this option is that multiple local operators and increased coverage requirements make site acquisition more complex. In urban areas, sites are often located on rooftops and other high structures. As there is limited availability of such locations, operators may have little choice other than collocating sites. In rural areas, construction costs such as power supplies and access roads constitute a significant percentage of the total site costs. In such cases, operators may be highly motivated co-locating their sites. Shared elements in case of site sharing are given Figure A-2. Operators share the same physical compound but install separate masts, antennas, cabinets and backhaul.

Mast sharing (cf. Figure A-3) is a step up from simple site sharing. As assembly and construction of the mast base forms a major percentage of the overall construction costs mast sharing has the potential to significantly reduce operators cost during the network roll-out phase. Operators install their own antennas, cabinets and backhaul but in addition to site sharing they share the antenna masts.

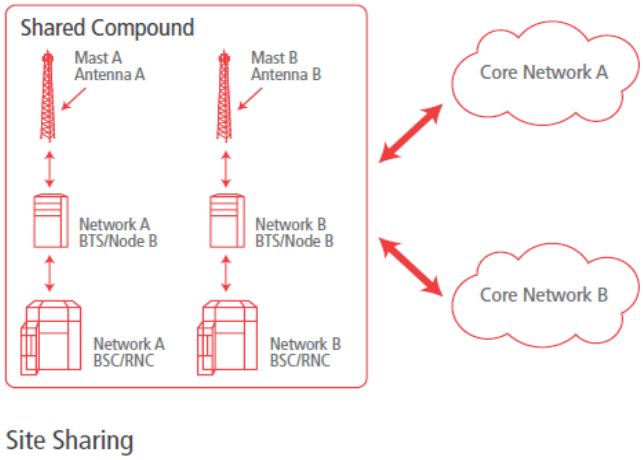



**Figure A-2: Site sharing [GSMA_1]**

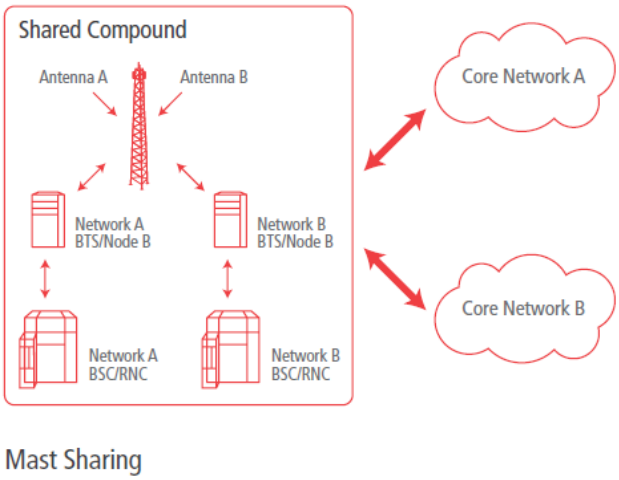



**Figure A-3: Mast sharing [GSMA_1]**

### RAN sharing

One of the key drivers of RAN sharing (Figure A-4) in mature markets is to reduce operational network costs. Operators in this case share radio equipment, masts, antennas, site compounds and backhaul. The implementation may vary between different operators, but normally at the interconnection point to the core network each operator splits out the traffic from its respective customers for processing by its own core network elements. Challenges appear due to different

independent developed architectures and interworking of equipment purchased from different vendors as well as different operational mechanism and control.

RAN sharing may also be commercially attractive in rural and peripheral areas with lower subscriber density and low ARPU users. Cost savings for typical European operators are in the range of 20% of the cash flow. Whereas site sharing allows certain degree of freedom to competing operators RAN sharing demands much more readiness for collaboration.

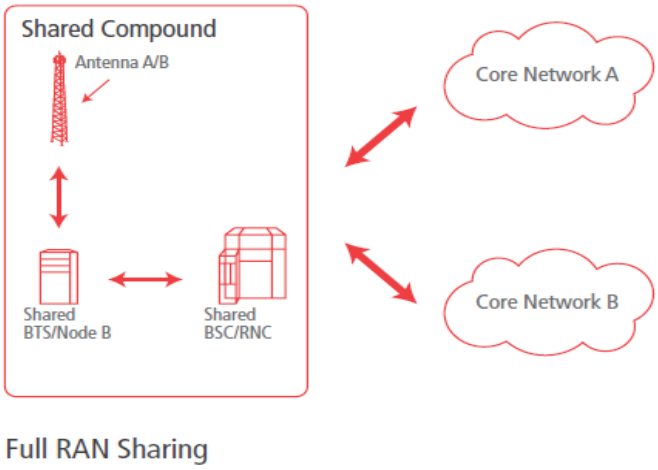



**Figure A-4: Full RAN sharing [GSMA_1]**

**Core network sharing**

Core network sharing has two levels. Operators may share either the core transmission rings and/or core network logical entities. Transmission ring sharing is mainly attractive for new entrances that are lacking their own ring. Core network logical entity sharing represents a much deeper form of infrastructure sharing and refers to permitting the partner operator to access all parts of the core network. Hence as any service or function that one operator implements can be replicated by the other the partners become more aligned which might be unfavourable in a competitive environment. The benefits of core network sharing are not as clearly defined as sharing the RAN.

**Drivers and regulatory aspects**

In early phases of network deployment site sharing and roaming allows new entrants quick network roll-out at lower cost. Even incumbent operators benefit from cost reductions. If networks mature the focus shifts from deployment to service innovation and revenue optimization becomes more important. In urban areas in addition site densification becomes more difficult and builds another reason for site sharing. Full RAN sharing today is complex and demands a lot of readiness for collaboration in a competitive environment. Core network sharing is not seen as providing substantial saving potential.

Regulators interest in infrastructure sharing is threefold. Generally, infrastructure sharing is seen as enabler for improved quality of service at lower prices and having a positive environmental impact. However, the decrease of network competition has to be weighed against.

## A.2.2.2. Limitation of spectrum deployments

### A.2.2.2.1. Macro sites

Due to high transmit power and significant antenna gains radiated power at macro antenna sites is rather high. Hence keeping the required safety distances introduces the main limitations for spectrum deployment. For our evaluations, we are facing the problem that conditions are site specific very variable. Antenna masts have different height, they are differently arranged and

often sites are collocated two or more operators. Therefore, basically it is not possible to give general rules for maximum spectrum deployment.

For our evaluations in the London study area we assume a typical site arrangement that allows exemplary for calculation of safety distances. The antenna arrangement depicted in Figure A-5 is based on our assumptions on single operator spectrum availability at the end of our investigation window in 2030 which is

- 25 MHz @ sub 1 GHz bands
- 60 MHz @ low bands
- 20 MHz @ medium bands.

For sub-1 GHz and low bands we deploy multi-band sector antennas realizing 4 antenna ports for each band. At medium frequency bands M-MIMO with up to 64 antenna ports allowing for 3D beamforming shall be applied. As beamforming performance is wave length sensitive for this frequency bands we deploy an extra antenna panel. The two-operator arrangement therefore consist of 2x2 antenna planes with each time three sector antenna panels where the M-MIMO antennas are arranged on top of the multi-band antennas. A total length of the antenna mast of 10 m can be assumed as feasible in most of the cases.

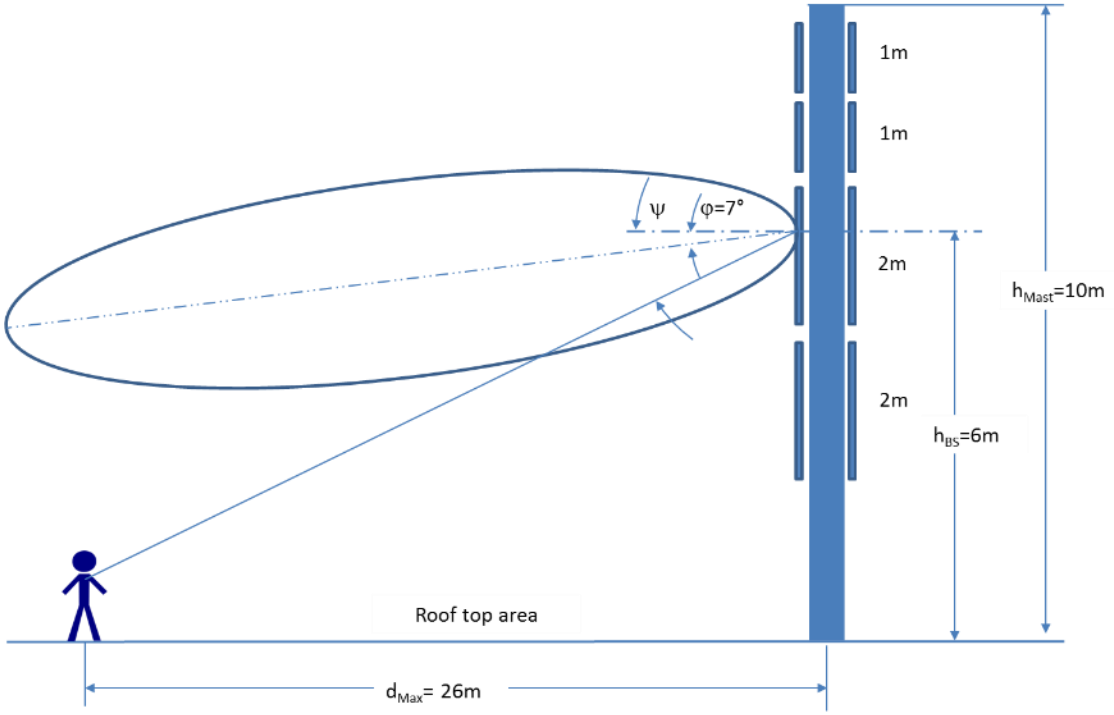



**Figure A-5: Two operator antenna arrangement at macro sites.**

According to [ICNIRP] the maximum allowable power flux density depends from frequency and can be calculated by

$$P_{max} = \frac{f\,[MHz]}{200}\ [\frac{W}{m^2}], \text{ for f} \leq 2{,}000 \text{ MHz} \qquad \text{(Eq. A-1)},$$

$$P_{max} = 10\ \frac{W}{m^2} \text{ for f} > 2{,}000 \text{ MHz}.$$

For calculation of the power flux density we select in each plane one of the sector panels shift it to an antenna panel balance point $h_{BS}$ of 6 m and assume a typical antenna tilt of 7°. For the vertical antenna pattern, we take the antenna pattern used for system level simulations from [3GPP 36814] with a half beam width of 10°.

It is assumed that the roof of buildings are public places where human beings are allowed to move without restriction. The worst cast distance from antenna mast is the location where power flux density at maximum (in our example at a distance of 26 m).

EIRP is assumed to be 64 dBm for sub-1 GHz bands and 72 dBm at medium bands, due to higher antenna gain with M-MIMO.

To calculate the total exposure quotient $E_{\text{total}}$, the exposure quotients of each band is cumulated.

$$E_{total} = \sum_i \frac{P_i}{P_{max,i}} \qquad \text{(Eq. A-2)}$$

The total exposure $E_{\text{total}}$ must not exceed 1.

The maximum spectrum that may be deployed under this assumptions as function of the $h_{BS}$ is compiled in Table A-10. As can be seen the height of the antenna mast is a very critical influencing factor.

**Table A-10: Spectrum limit at macro sites as function of antenna panel balance point.**

| $h_{BS}$ | Spectrum limit |
|---|---|
| 4 | 70 MHz |
| 6 | 210 MHz |
| 8 | 340 MHz |

#### A.2.2.2.2. Small cells

According to [5GN-D32] the spectrum available for a single operator dedicated for small cells amounts to

- 60 MHz at low and medium frequency bands and
- 200 MHz at high frequency bands.

Applying the methodology for calculation of exposure quotients described above (Eq. 2) we find out that for antenna deployments at 3 m and 8 m height the exposure does not reach critical limits.

This is based on the fact that for economic reasons, the transmit power of small cell nodes must not exceed a certain limit e.g. 5 W for SC < 6 GHz and 1 W for SC @ mmW [MAG-D14]. Maximum antenna gains are to be expected in the range of 2-9 dBi for small cells < 6 GHz and 25 dBi for small cells > 6GHz.

Hence, link budget restrictions in terms of minimum EIRP power density will restrict the amount of spectrum to be deployed per radio node. For the above assumptions on maximum transmit power and antenna gains of small cell equipment, the cell range can be calculated by link budget as described in [MAG-D14] (cf. Table A-11).

**Table A-11: Link Budget for the outdoor small cell hotspots**

| Frequency Band | DL 3.5 GHz | UL 3.5 GHz | DL 28 GHZ | UL 28 GHz |
|---|---|---|---|---|
| **Transmitter** | | | | |
| Tx Power [dBm] | 37 | 23 | 30 | 23 |
| Antenna Gain [dB] | 5 | 0 | 25 | 6 |
| Cable Loss [dB] | 2 | 2 | 2 | 2 |
| EIRP [dBm] | 40 | 21 | 53 | 27 |
| **Noise and signal level** | | | | |
| Thermal noise [dBm/Hz] | -174 | -174 | -174 | -174 |
| Operable Bandwidth [MHz] | 60 | 20 | 400 | 120 |
| White noise addition [dB] | 78 | 73 | 86 | 81 |

| | | | | |
|---|---|---|---|---|
| Rx Noise Figure [dB] | 5 | 2,5 | 5 | 2,5 |
| Target SNR level [dB] | 6 | 6 | 6 | 6 |
| Interference margin [dB] | 3 | 3 | 2 | 2 |
| Minimum Signal level [dBm] | -82 | -89 | -77 | -85 |
| **Receiver** | | | | |
| Antenna Gain [dB] | 0 | 5 | 6 | 25 |
| Cable Loss [dB] | 2 | 2 | 2 | 2 |
| **Penetration Loss [dB]** | 20 | 20 | 0 | 0 |
| **Total Gains [dB]** | -22 | -17 | 4 | 23 |
| **Maximum Path loss [dB]** | 100 | 93 | 134 | 135 |
| **Possible Cell Range [m]** | | 90 | 165 | |

In the link budget calculation assumptions, the limiting maximum path loss (MPL) appears at low and medium frequencies in UL. The reason for this is that the lower bandwidth (20 MHz) in UL cannot compensate for lower transmit power on the device side. As small cells at those frequency bands shall cover also indoor locations, a (fix) indoor penetration loss of 20 dB is assumed.

Different conditions appear at high frequency bands, which have been exemplarily investigated at 28 GHz. Lower UL transmission bandwidth compensates for the lower EIRP and antenna gain at devices and hence the link budget is nearly balanced between DL and UL.

In order to get an idea of the magnitude of possible cell ranges for both frequency bands propagation models according to [SRM15] [5GCM] have been applied. Reference distance path loss models with frequency-dependent path loss exponent (CIF models) for LOS and LNOS have been combined probabilistically using an empirical formula describing the distance dependency of LOS probability. Assumed model parameters are given in Table A-12.

**Table A-12: Parameter of applied probabilistic path loss model**

| **Model Parameter** | | |
|---|---|---|
| Path Loss Model Parameter | 3.5 GHz | 28 GHz |
| LOS/NLOS breakpoint -dBP (m) | 27 | 27 |
| Decay parameter -α (m) | 71 | 71 |
| Wavelenth [m] | 0,86 | 0,11 |
| Path Loss Exponent $n_{LOS}$ | 1,98 | 2,1 |
| Path Loss Exponent $n_{NLOS}$ | 3,19 | 3,4 |
| Shadow Fading Standard Dev. s LOS[dB] | 3,1 | 3,6 |
| Shadow Fading Standard Dev. s NLOS[dB] | 8,2 | 9,7 |

Results for mean weighted path loss applying this parameterisation are depicted in Figure A-6. According to these investigations, it should be possible to deploy the whole spectrum per operator at small cell sites. Possible cell ranges are in the order of 90 m to 150 m for small cells at low and medium as well as high frequency bands respectively.

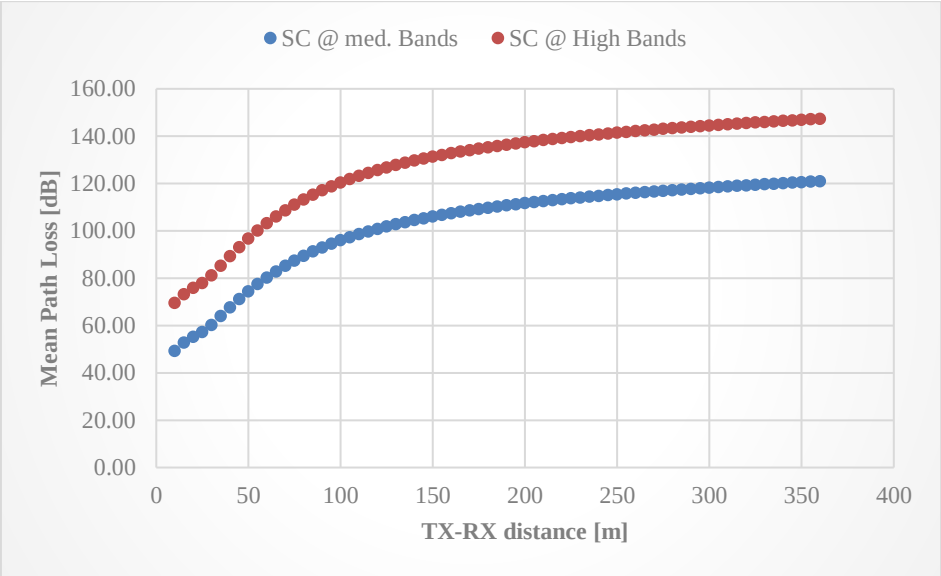



**Figure A-6: Estimated mean weighted path loss as function of TX-RX distance**

## A.2.3. Multi-service evaluation

### A.2.3.1. Adaptation and placement of VNFs

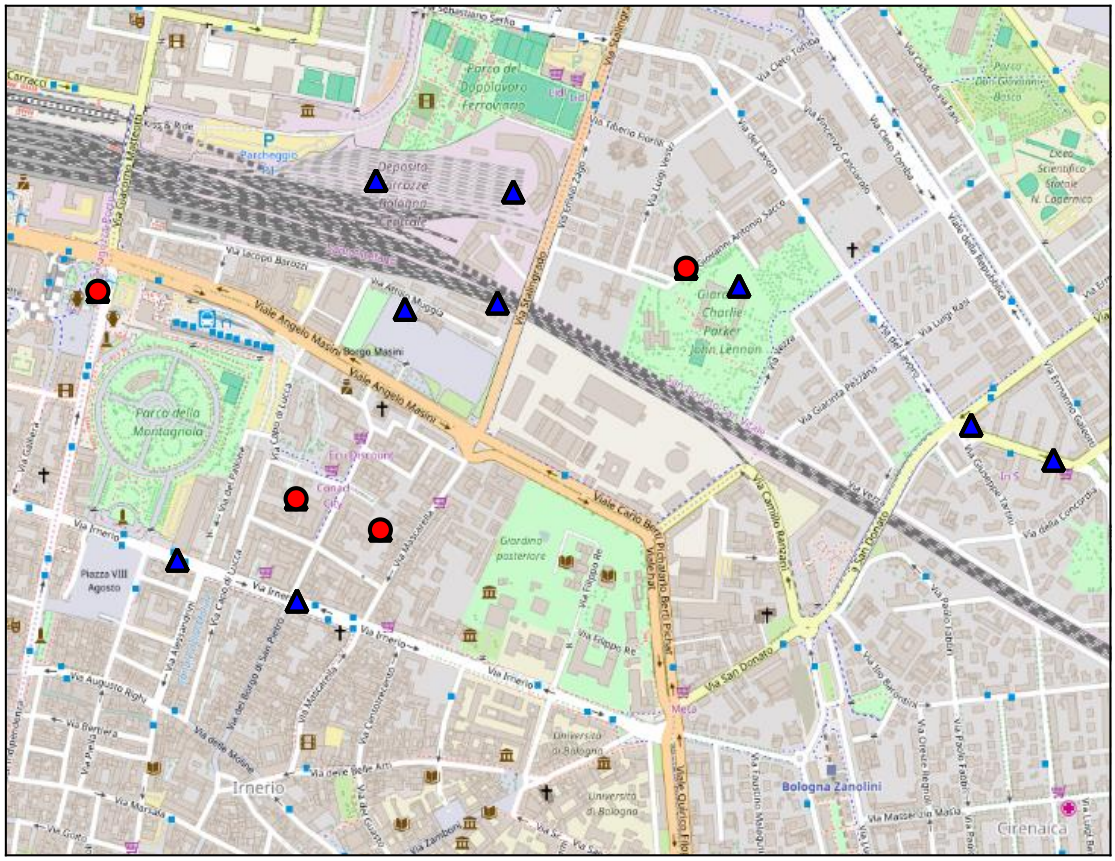



**Figure A-7: (P)NF placement in the simulated scenario [5GN-D52]**

The flexible placement of NF belonging to a VNF chain that are used to implement a certain service is certainly one of the enabler of enhanced functionality in future 5G networks.

Throughout the deliverable we discussed, especially in Section 2.3.5 why this is important and in [5GN-D52] we detail how to enable this functionality by means of intelligent algorithms that are implemented in the orchestration.

In the following, we discuss on the operational requirements associated to this feature, by taking as example the multi-service use case described in Section 6.3.3, in particular focusing on the V2I use case. Due to the extensiveness of the London scenario described above, we leverage on the simulations results available in [MGB+17] to assess the impact of VNF placement, that target a smaller scenario that is feasible for our implementation.

The adaptability of Network Slice orchestration to the real requirements of the network is one of the fundamental characteristics that flexible VNF adaptation allows for. Current network planning strategies are rather "static" and usually try to accommodate traffic peaks of any kind. Only small adjustment to the initial deployment are allowed, such as cell power adaptation or switching on and off towers.

Driven by the availability of denser wireless networks and by the spreading of small cell connectivity, the available spectrum has to be split across different slices in an efficient way (cf. Annex A.2.3.2 for more details). This operation is facilitated by the adoption of enhanced controllers such as SDM-C and SDM-X, but how to optimally place VNFs across different levels of cloud is an important design decision that needs to be taken.

As seen in [MGB+17] a distributed and fixed deployment of resources at the antenna side tends to be very inefficient (cf. Figure A-7). Typically, V2I applications such as collision warning systems have an uneven load distribution and generate more traffic when a collision is more likely to happen. Therefore, those areas with traffic jams will need to have a more consistent set of resources dedicated to the V2I network slice. Also, this amount of resources is time-varying and should be automatically adaptable to the load. 5G NORMA defines algorithms and procedures to automatically process resource re-orchestration across slices.

The V2I one is a network slice that do not need session continuity as it mostly involves the geo-fenced broadcasting of safety massages. It has stringent requirements in terms of delay, though, so all the VNF devoted to the RAN should be in the antenna site, consuming resources just when needed by the changing vehicular traffic conditions.

For the eMBB service, the requirements are less stringent on the delay perspective. Hence, RAN functions can be co-located at edge clouds or even at the central cloud, if the operator would like to maximize multiplexing gains at the cost of a higher bandwidth on the backhaul links.

There, the operational requirements are mostly related to the chosen RAN functional split [iJoinD5.2]. Then, depending on the fronthaul / backhaul links available in the city of London, the feasibility of a flexible decomposition of VNFs different placements will be feasible or not.

As described in Section 7.2.1.2, the V2I slice deployment will integrate different RAN-InP in order to provide the most ubiquitous coverage. Therefore, to both improve the end to end latency of V2I communication and to ease the RAN sharing, it is expectable that most of the lower RAN will (densely) be co-located at the edge, where more capacity is needed. Indeed, deploying network slices over hierarchical network function is one of the 5G NORMA MANO functional requirements. Moreover, several orchestration algorithms were provided in [5GN-D52]: they optimize network metrics such as multi-path exploitation, which are relevant to the V2I.

### A.2.3.2. Investigation of network programmability

Based on the above description, we focus on the London city scenario to explain the benefits of the deployment of a softwarised network running different services such as eMBB, mMTC, and uMTC. These services have very different service requirements that need tailored configuration, as we thoroughly discussed throughout this document. Still, this configuration should be achieved in a flexible and reproducible way, as done e.g., for different releases of software. By applying

the network softwarisation concepts explained in Section 2.1.3, it is expectable that the Service Creation Time will be reduced.

The softwarised controllers described in the 5G NORMA architecture allow for a better network performance mostly along these dimensions: i) fast network slice onboarding, due to standardized software procedures, ii) Scalable network slice management and iii) efficient shared resource control. In the following, we use explain how a V2X service running on a shared infrastructure, benefit from this concept.

As explained in Section 2.3.4, by allowing software-based management and control of both network functions and orchestration, future networks will be more flexible allowing thus an optimized operation while reducing costs. In this section, we elaborate on the functional requirements of the three controllers (SDM-C, SDM-X and SDM-O) that embody the network programmability concept in 5G NORMA architecture. We use as specific example the V2X deployment as above.

**SDM-C**

The SDM-C is in charge of the control of dedicated network functions. These are likely to be core functions for this kind of scenario. However, as V2X services require very limited core functionalities: authentication, billing and geo-broadcasting, SDM-C is going to control a very limited number of network functions. Therefore, movement of this function is very unlikely, allowing thus for higher multiplexing gains.

**SDM-X**

The V2X network slice is probably going to share the RAN with other slices such as eMBB. Therefore, the relevant RAN component such as the MAC scheduler should deal with different assignments of resource blocks to different slices. For further details, we refer the reader to Section 3.5.2. The software defined approach adopted by SDM-X makes possible to handle the heterogeneity of service requirements introduced by 5G network. SDM-X Applications will deal with changing resource assignments (especially in the available spectrum) that are expectable when dealing with fluctuating demands

**SDM-O**

The SDM-O has different software algorithms that affect the V2I slice deployment. First, the admission control algorithm (cf. Section 2.4.2) that decides whether there are enough resources for the V2I network slices and how this network slice resources should be spread across antenna site. Then, the NFV-O algorithm that onboards the network slice, assigning VNFs to specific location or warns the SDM-X of a new slice in the controlled VNFs. Finally, it reacts to possible resource re-orchestration triggers by eventually reassigning resources.

### A.2.3.3. Investigation of QoE based routing and network agility

One of the main improvements that 5G NORMA will bring when compare to legacy systems is an end-to-end improvement of QoE. In order to achieve that goal, it's necessary to enhance the current software-defined routing with QoE considerations. A new routing scheme using Q-routing is proposed in [5GN-D52]. Simulations demonstrated that this routing scheme has comparable performance as an ideal global routing scheme, improving over routing scheme that don't take QoE into account.

When it comes to the multi-service evaluation case, it would be the task of the MNO to check if using this new routing scheme makes sense for the different services (this might also be part of a SLA).

Two services would benefit the most from this innovation:

(1) eMBB: Improving QoE for eMBB is one of the most direct benefits of 5G NORMA. For this service, some form of metric for a generic QoE will be necessary. Ideally, this should not require direct user input.
(2) V2I – Infotainment: This service presents a different challenge for deployment of the new routing scheme, since there are some proposed models for calculating QoE for audio and video services without the input from the users. However, the higher mobility of the users might present a challenge to the new routing scheme, since Q-routing requires some time to adjust itself and provide the best routes.

### A.2.3.4. Investigation of Edge function mobility

If a group of users is being served by an edge cloud, and then move to another remote edge cloud, when should the network functions in the original cloud be migrated/replicated? In [5GN-D52], a placement decision method was described, using Markov Decision Processes, which determines if an edge network function should remain in the same edge cloud, be moved to the serving edge cloud, or perhaps moved to the central cloud. Simulations showed that the edge cloud change of user groups has deep impact on performance, and that the decision can be made in reasonable time, considering the mobility model used.

The service that can most used this placement method is V2I, since having group of cars using the same service moving around the city will not be unusual. Ideally, the decision and reaction by the network of any changes in serving edge cloud should be made quickly, allow for seamless service continuity.

### A.2.3.5. Multi-connectivity protocol overhead

Section 6.4.3.1.2 and 6.4.3.1.3 have discussed advantages of multi-connectivity, namely coverage improvement and throughput increase. A drawback of multi-connectivity is a higher signalling load and accordingly a enlarged reconfiguration delay, e.g. in case of handovers. The amount of signalling depends on two architectural aspects, namely the protocol layer in which the data flows are split and the allocation of protocol layer functions to network elements.

Several architecture options for multi-connectivity have been assessed in [5GN-D42] and compared by means of a protocol overhead analysis. An evaluation scenario has been defined consisting of a 5G low-band access point (5G LB AP) and two 5G mmWave access points. All signalling traffic is carried by the 5G LB AP to guarantee high reliability. The user traffic is distributed onto multiple 5G mmW APs within the coverage of a terminal to achieve a higher throughput despite the limited coverage range of mmWave APs. For this scenario, three different protocol structures, each with two options for the allocation of protocol functions to network elements, have been assessed. For all options, the necessary number of signalling messages for a reconfiguration has been evaluated based on message sequence charts.

Aside the number of signalling messages, also the impact of backhaul traffic at the 5G LB AP has been taken into account in the overall assessment. This resulted in a recommendation for a particular architecture option with a well-balanced relation of signalling load and backhaul traffic. For details, the reader is referred to [5GN-D42].

### A.2.3.6. Investigation of mobility concepts

Aim of this section is to evaluate the outcome of an application of dynamic selection of mobility concepts as described in [5GN-D52] in terms of advantages and potential drawbacks to the selected London study area described in Figure 6-2 and in Annex A.1.2 where the requirements and characteristic performance parameters of the exemplary services shown in Table 6-1 are detailed. Aim is to estimate the gain in terms of effort and resource consumption (radio transmission and processing at network nodes) achieved by application of slice and service specific selection of mobility schemes according to the actual requirements instead of a ‘one-size-fits-all’ approach.

#### A.2.3.6.1. Approach

The deployment of modular flexible MM schemes (implemented within 5G NORMA SDMC architecture as SDN applications and VNFs) allows for tailored and selected MM service per use case.

A differentiation of MM services means that to fulfil specific requirements as session continuity and seamlessness of handover efficiently some dedicated functionalities are needed not per default but only for some use cases.

The effort required per service function is estimated in terms of amount and complexity of functional entities, signalling effort (e.g. message overhead volume) and efficiency (i.e. ratio of normalized invested means vs. resulting performance), and the correspondingly resulting resource consumption (storage, compute, transmit) which would be demanded. This aspect has been partly already addressed in [5GN-D52] and the concepts will be shortly summarized in the next section.

As a reference, the effort for a 'full mobility' or 'average level of mobility support' as standard approaches for all services is determined. The latter approach is not effective, as some use cases would not be supported satisfactorily whereas the prior one is not efficient since it means waste of resources for a large amount of new usage scenarios. A rough estimation of the expected gains is given after the following section.

#### A.2.3.6.2. Considered mobility concepts and use cases

The different mobility concepts as described in [5GN-D51] allow for fine granular differentiation in network functionality per logical slice, e.g. for HO seamlessness and success rate probability. The selected use cases for the London study area demand for e.g. high uplink data rates (eMBB), improved coverage (mMTC and uMTC), or very low latency (V2I). The general KPIs/requirements, which have also to be considered in selection the optimum MM scheme per use case, thus typically include

- large device connection density and huge number of data packets,
- protocol efficiency,
- low/ very low latency,
- high reliability,
- high availability,
- support of architectural features e.g. allowing Local break out at the radio nodes, and
- support of RAN features including e.g. NR air interfaces enabling low latency and high reliability.

While some of them are assured by topological decisions (MEC instead of central cloud), radio technology deployment (small cell sizes), the impact of the right choice of MM scheme especially impacts protocol efficiency (in terms of signalling overhead) and reliability and availability issues (to reduce handover induced service outage).

#### A.2.3.6.3. Investigation results

According to assumed London study area the distance between MECs and BSs amounts to between 500 m and 3 km at maximum. The cell radius can even become as low as 200 m for a dense urban scenario as present in London study area. From this type of topology, no impact on latency should be experienced and the requested availability figures together with LTE-A and NR technology should be achieved.

According to [5GN-D52], five different mobility schemes have been identified, characterized by their capability in terms of type of Handover (HO) to support and effort in terms of functional entities involved and number of messages to be exchanged. The HO type ranges from stationarity/nomadicity (i.e., no HO) via simple mobility (Horizontal or intra-RAT HO) and vertical (or inter-RAT) HO to x-D (cross-domain) HO across network boundaries and concurrent multi-connectivity support with e.g. full redundancy (2xHHO). A maximum message overhead

of 114B has been identified depending on the underlying (3GPP or IETF) control protocol assumed. The resulting effort is mainly a function of cell size and (allowed maximum) terminal speed, whereas the performance in terms of latency also depends on distance between control plane entities as a function of their location. Since only distributed MEC near cell towers and small cell sizes are assumed this criterion is non-critical here.

The mapping of the required HO performance per each of the eight use cases is derived from service descriptions in Annex A.1.2 and the weight of each use case from the overall number of devices per km² given in Table A-3.

Based on the results of [5GN-D52] a classification of mobility schemes and corresponding types of HO supported is derived and given in Table A-13.

**Table A-13: Classification of mobility schemes investigated in [5GN-D52]**

| Mobility scheme as HO type | Amount of involved entities | Amount of messages to exchange per HO | Comment |
|---|---|---|---|
| **No HO** | 3 | 3 | Paging only |
| **Simple HHO** | 4 | 7 | intra RAT |
| **VHO** | 5 | 9 | Inter RAT |
| **X-D HO** | 6 | 17 | Inter domain |
| **2xHHO** | 7 | 18 | Multi-link |

The mapping of service components to mobility schemes is provided in Table A-14 together with the weights.

**Table A-14: Mapping of mobility schemes given in Table A-13 to London study area service components (use cases) together with weights of each one**

| Service component | Mobility requirements | HO scheme applied | Weight according to scenario 2020 | Weight according to scenario 2030 |
|---|---|---|---|---|
| eMBB – consumer | Adaptable mobility per service | HHO / VHO / x-D HO | 58,30% | 60,39% |
| V2I – infotainment (eMBB) | High mobility | HHO / VHO | 0,44% | 0,42% |
| V2I – assisted driving (uMTC) | Highly reliable | 2xHHO | 0,44% | 0,42% |
| V2I – driver info service (mMTC) | Highest mobility | Simple HHO / VHO /x-D HO | 0,44% | 0,42% |
| Environmental monitoring, etc. (mMTC) | - no mobility | No HO | 0,14% | 0,13% |
| Smart meters – etc. (mMTC) | simple mobility | Only HHO | 40,67% | 38,54% |
| Smart grid (mMTC) | no mobility | No HO | 0,00% | 0,03% |
| Logistics - tracking goods (mMTC) | High mobility | HHO / VHO | 0,01% | 0,07% |

Assuming the figures as given in Table A-13 above for eight service components mapped to a set of one to maximum three out of five different mobility schemes and the % of devices (assuming here up to three flows per device in eMBB – corresponding to three different services and 70 % of stationarity – as well as up to three passengers per car with 10 % stationarity in V2I/eMBB) requiring own mobility handling the effort in processing at network entities and the signalling

overhead was estimated. The actual effort assuming specifically tailored MM as compared to an effort expected for a uniform medium (maximum) MM scheme for each service results in saving of 27.3 % (76.0 %) for 2020 – slightly decreasing to 25.6 % (75.4 %) in 2030. The overall relative MM signalling OVH compared to data traffic volume amounts to 0.16 % (ranging from 0.13 % for eMBB to more than 1000% in mMTC - due to the IoT low data volume) for 2020 figures – decreasing to 0.01 % for high traffic volume in 2030.

Assuming the expected linear CAGR in device and traffic volume figures the required effort for MM (in terms of relative processing at network functions as product of number of messages to be exchanged and entities involved as denoted in [5GN-D52]) is shown as saving compared to assuming a medium MM for each service component over the year timeline in Figure A-8. The figure also shows the relative signalling effort in terms of traffic volume compared to actual data traffic. With increasing traffic volume those figures are decreasing.

According to results reported in [5GN-D52] for an average cell size of 500 m a signalling overhead of between 1.3% and 5.4% at 30km/h can occur whereas for 3 km it remains below 0.9% - on the other hand for 70 km/h up to 12.5% overhead for 500 m is estimated for full mobility. Extending the corresponding calculations down to 200 m cell range up to 13.4% even at 30 km/h can occur. For the London study area and V2I an average cell size of 1.6 km (ranging from 200 m to 3 km) has been assumed and an average vehicular speed (taking into account highways with 70 km/h and city streets with 30 km/h only) of 50 km/h resulting in 675 HOs per day. For eMBB a pedestrian speed of 5 km/h is expected in areas with lower cell sizes of 0.5 km with 72 HOs per day only.

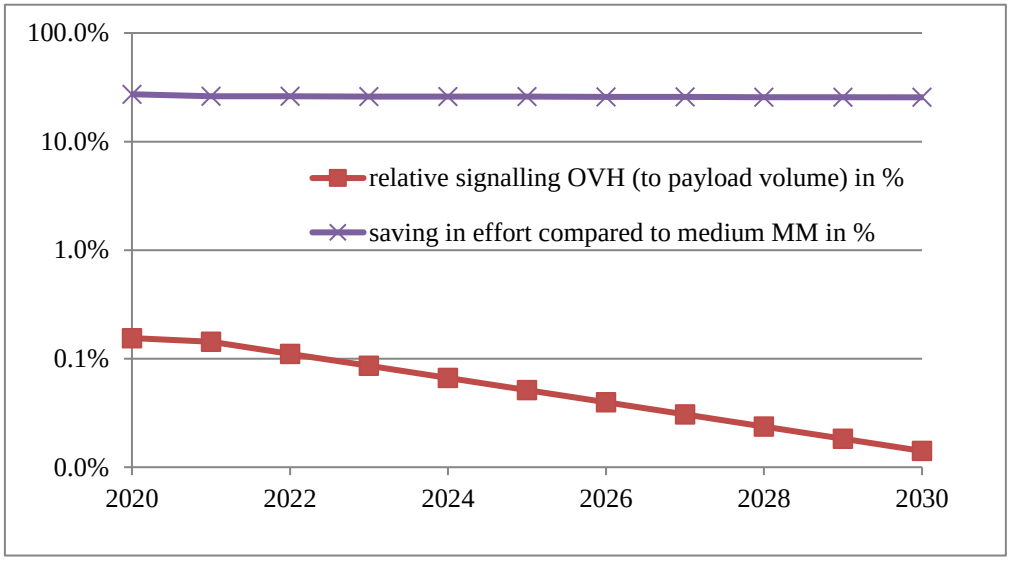



**Figure A-8: Expected development of MM signalling and processing effort for specifically tailored MM per year.**

As can be seen from the estimation above the concept of flexible assignment of MM schemes to different services or use cases and network slices results in saving of both processing effort in the (core) network and transmission resources in the (radio) access network – where the latter saving which with future growth in service usage will even increase.

## A.2.4. Demonstrator learnings

**Learnings in Demo #1**

The main learnings from this demonstrator are the following: in the HW part we've evaluated three KPIs, which we designate as "Hard" (the 1st couple) and "Soft" (the last one). The hard KPIs are basically used to estimate the overall system performance; they are:

- average session throughputs,
- End-to-end latency.

We have seen that these KPIs are affected for both the two real users which are experiencing two different services: a LL service and a MBB service, respectively. This is aligned with our assumptions, because the E2E latency and the throughput can benefit from the SDM-C reconfiguration and the flexible functions allocation.

Regarding the soft KPI, it is mostly related to the new design proposal based on the SDM-C concept defined in 5G NORMA: the network can adapt itself to the demands and requirements of frequently varying network conditions by means of network programmability. The hardware part of this demo has showcased that a novel design of the network architecture with SDM-C produce gains also over legacy network architectures like LTE and that this concept is technically feasible and can be implemented with reasonable efforts.

On the other hand, regarding the software part of this demo, it has been shown that the introduction of the SDM-C and its interfacing with all eNodeBs results in a considerable performance improvement for different kinds of services. For this purpose, two different services were investigated with different requirements for user satisfaction, i.e.

- Video streaming user with high throughput requirement
- MTC user with low latency requirement

It has been observed from the simulation results that a multitude of diverse service requirements cannot be met if the entire network is configured statically in either edge- or in central-cloud configuration. However, with a more dynamic network configuration using an intelligent SDM-C, the service requirements of both services can be met. The novel network architecture has been used to demonstrate an improvement in system performance over legacy LTE. This concept of dynamic and service-aware network reconfiguration can be extended, in future 5G Networks, to more than two services and even to a mix of services under the same eNB.

**Learnings in Demo #2**

This demo evaluates two types of KPIs:

- KPIs regarding the RAN deployment and
- cloud infrastructure KPIs.

Regarding the 1st set, we have seen that SW implementations of the LTE stack are a relatively new topic in the mobile network ecosystem and are usually undergoing high maintenance tasks (e.g., adding new features, supporting new hardware or simply refactoring the code are sources of new inconsistencies). Also, different branches that are usually present in the software versioning platform are other sources of inconsistencies, as different features may be disabled or superseded by source code updates.

The implementation approach for the access part includes the use of two software implementations of UE and eNB. While this approach is allowing us to fully understand and control the behaviour of both the communication ends, we've seen that aligning two different pieces of software with different releases cycles proved to be difficult (e.g., we had to develop patches for the authentication procedures to make both software packages fully compliant).

Also, we've seen that the selection of the used band proved to be fundamental. To minimize the interference to and from the licensed spectrum, we set up a wired scenario for connecting the UE with the eNB.

On the other hand, regarding the cloud infrastructure, we've evaluated the following KPIs up to now:

**KPI 1a**: Demonstrate the feasibility of the ETSI NFV MANO architecture, i.e., verify that it can be implemented as a proof of concept.

In order to do this, we have to implement and validate a specific ETSI NFV MANO implementation able to manage and orchestrate the set of VNFs in the demo, as well as to integrate with the external HW components. This work is still in progress, so there are no definitive results at the moment regarding this. At the moment, we have an initial limited-scope implementation which is expected to fulfil the demo requirements, but some important issues remain open (e.g., the migration of network functions).

Anyway, the main lesson learned is regarding the TeNOR orchestration and management platform [TENOR] that was initially selected to be used in this demo; this component has demonstrated not to be suitable for our requirements (at least in its current version).

**KPI 1b:** Demonstrate the applicability of the ETSI NFV MANO architecture, i.e., verify that the ETSI NFV MANO implementation can be actually used for the specific purposes of this demo. In this sense, the MANO functionality should include the possibility of moving the VNFs between different specific nodes representing a hypothetic operator's network.

Same situation as the one described for the previous KPI 1a, i.e., no definitive results at the moment, and, the results are expected to be available until the end of the project. Anyway, we've been performing some initial testing regarding the possibility of moving VNFs using the OpenStack live-migration functionality, but performance was insufficient (time was in the range of minutes). However, this is just an initial prospection testing and further research is being performed on this.

Apart of this, we have promising results using the OpenStack4j library [OS4J] to implement the basic ETSI NFV MANO abstractions. However, there are still a number of open tasks in order to have the required complete functionality using this approach.

**KPI 1c:** Evaluate different MANO platforms. Although the demo will be performed on a specific ETSI MANO implementation, we consider it also necessary to evaluate and test different software platforms.

Different ETSI compliant NFV MANO platforms were evaluated. Main focus was on the TeNOR orchestration platform, but as already mentioned, we found this platform was not applicable in the context of our demos (although it was possible to deploy basic dummy VNFs and Network Services, we did not succeed in attempting to deploy a more realistic service, as we needed for the demo). Besides TeNOR, we also evaluated other three platforms: Open Baton [OPB], Open Source Mano [OSM] and Tacker [TAC]. All of them were also discarded; the main reason is that we considered that greater chances of success could be achieved using an ad-hoc approach specific for this demo, instead of a general purpose out-of-the-box solution like those ones. The effort to learn using a new platform could consume considerable resources. Also, depending on a third party to solve unexpected issues could be very limiting.

This decision was also supported because, beyond the specific problems due to TeNOR, the perception after evaluating the other similar possibilities (Tacker, OpenBaton and OSM) was that the state of the art regarding this type of general-purpose platforms is still not mature enough. A recent reference supporting this affirmation is the 1st ETSI NFV Plugtests Report [ETSI.NPR-17] released by the ETSI Centre for Testing and Interoperability, which evaluates different MANO solutions and NFV platforms. The report actually shows high rates of test execution and interoperability for simple features such as Network Service on-boarding, instantiation and termination, but results are still limited for more complex operations (e.g., scaling or NS updates. In addition, no results are reported beyond basic functional testing (e.g., test cases about performance or migration of NFs are not reported).

Besides the above-mentioned MANO platforms, we also evaluated different solutions to implement the ad-hoc solution for the demo. The main three options were jClouds [JCL], the OpenStack REST API and OpenStack4j [OS4J]. jClouds is an open source multi-cloud toolkit for Java that provides the possibility to create applications that can be portable across different clouds (e.g., AWS, OpenStack, CloudStack, Azure…). However not all of the OpenStack functionalities

are supported (e.g. the Ceilometer service for monitoring). OpenStack4j, although limited to OpenStack, offers all the functionalities supported by this platform. It also offers a higher level of abstraction that the plain REST API. Hence it was considered the most reasonable option to fulfil the requirements of this demo.

**KPI 1d:** Prove the ETSI NFV MANO implementation efficiency, i.e., the implementation can be achieved using reasonable amount of time and resources.

We consider it is still too early to evaluate this KPI. It is necessary to have the full NFV MANO implementation first.

**KPI 1e:** Explore the ETSI NFV MANO implementation limitations in order to identify possible drawbacks in the architecture.

As before, it is still too early to evaluate this KPI.

**KPI 2a:** Validate the technical requirements and KPIs coming from WP2.

[5GN-D22] groups requirements around three axes:

(1) Very low latency and reliability, for critical machine-type communications.
(2) High throughput (compared to legacy networks) for massive MBB communication.
(3) The ability to support high volumes of devices for massive MTC (mMTC).

We consider that addressing each single KPI defined in the WP2 is out of scope for this specific demo, which has a much more limited scope that those general purpose KPIs defined in that WP; the main objective here will be on the lessons learned regarding those specific KPIs which are clearly in the scope for this demo. It is expected that WP6 demonstrators will contribute to mainly check the Soft-KPIs and the Performance Requirements (although some of performance related information especially in context with novel air interfaces will be taken from other R&I projects)[18]. In any case, this demo will mainly focus on LL and MBB requirements (points 1 & 2 above), although at this point in time it is still premature to say about fulfilment of these performance requirements.

**KPI 2b:** Evaluate the 5G NORMA architectural principles described in WP3. Although not all the functional building blocks described in WP3 will be implemented for this demo, some of the main components (those already present in the ETSI NFV MANO framework) can be demonstrated.

The architectural principles described in WP3 have been carefully evaluated, especially those regarding the implementation of the MANO layer. In fact, from the WP6 we have actively collaborated in the definition of these architectural principles in the scope of this WP3. Since the architectural principles described in WP3 are mainly based in the ETSI NFV MANO framework (WP3 defines specific ETSI MANO stacks for each slice in the network), we consider there is a good alignment between the demo and the definitions in this WP3.

Due to the usage of a new Java based implementation for the NFV management and orchestration blocks, we consider this demo could go beyond the initial intention of this KPI by exploring some of the 5G-NORMA specific architectural elements described in the WP3 (if time and resources are available). An interesting example could be the implementation of a kind-of Inter-Slice Resource Broker component.

**KPI 2c:** Evaluate adaptive allocation of functions to different network nodes (WP4).

[18] Functional requirements (defined in Deliverable D2.2 [5GN-D22]) are to be checked by tools like protocol and protocol overhead analysis outside the WP6 context (although most of the comparisons are to be executed by qualitative analysis). The same applies for operational and security requirements defined in Deliverable D3.1, [5GN-D31].

As already mentioned, initial research and testing has been performed regarding this; at the moment results are not positive regarding performance using the OpenStack live-migration functionality. An initial approach has been performed and hence further trials are still being performed.

**KPI 2d:** Evaluate QoS/QoE mapping and monitoring control processes, orchestration functions and VNF life-cycle management as described in WP5.

No results regarding this KPI at the moment.

**KPI 2e:** Provide feedback to the other WPs from the results obtained.

Some feedback has been provided to WP3 (providing ideas about the Management and Orchestration architecture) and to WP5 (defining the QoS/QoE framework). More complete and specific feedback will be provided when more specific results are available.

**KPI 3:** Demonstrate one of the project's key principles, i.e., the possibility of dynamic relocation of network functions between the edge of the network and a centralized cloud infrastructure, thereby enabling low latency communication.

This is one of the main focuses of this demo. Both slices (MBB & LL) are designed considering the movement of VNFs between edge and central clouds. In any case, there are no conclusions available yet.

**KPI 4a:** The demo should serve as platform to be exhibited in public conferences and events.

An initial version of this demo has been already presented in the IEEE ICC'17 conference [5GN-IR61]. Although the work presented was just an early version of the demonstration, it served as an initial approach to the final demo set-up, and triggered some logistic-related issues, such as the transportation of certain HW elements and the possibility to access certain services in a remote manner. At the moment we can consider this KPI is only partially fulfilled, being necessary to better specify certain aspects to face the final demo and the presentation in other relevant events.

**KPI 4b:** Since for this demo we are extensively using third party Open Source projects, another important objective for this demo is to release the software components specifically developed for this demo to the Open Source community.

There are no results available for this KPI yet. At the moment, all the software that Atos is developing to implement the NFV management and orchestration functions are proprietary software. Atos has to evaluate if the final implementation will be released as Open Source software.

**Learnings in Demo #3**

Demo #3 is about Secured Multi-Tenancy Virtual Network Resources Provisioning via V-AAA. The main KPIs we are considering up to know are:

- Using graphical user interface (GUI) to visualize the tenant information that are retrieved from the local database (Tenant's distributed database) and the remote database (Mobile network operator hierarchical database),
- Using GUI to show tenant data isolation and tenant data consolidation,
- Using GUI to show all tenant's data at the mobile network operator's database,
- Using tokenization technique to provide secure techniques.

**Learnings in Demo #4**

Demo #4 is basically a web graphical interface allows users to explore the economic benefits of 5G-NORMA architecture. The main results include the following:

- An interactive web-based GUI to display the 5G-Norma economic model data elaborated in the context of the WP2. This tool enables users to see key outputs through a simple

dashboard of a complex cost model. Users can interactively see how certain parameters in the model can affect each other in a graphical and intuitive way.

- Although in Deliverable D6.1 [5GN-D61] we anticipated that the web tool would be developed by using the PHP (Hypertext Pre-processor) and HTML (Hypertext Mark-up Language) web-programming languages, we finally encounter more convenient approach based on Spreadsheetweb [SW1]. We consider this approach is simpler and good enough to show the results of the 5G-Norma model.
- Early versions of the online demonstrator were released in autumn 2016 and spring 2017, these versions allowed users to select a range of inputs such as:
  - Traffic Growth,
  - User Density,
  - Use Case

Once a user makes a selection of desired inputs, the outputs are shown in terms of traffic in Mbps/km$^2$ and indicative figure of the network cost in GBP/km$^2$. This early version is a key prototype with the following functionalities:

- Select input combinations
- Interrogate the abstraction data
- Present results on the web-interface

Based on these functionalities, final release of the demo will have an improved design of the web-interface and additional results. Regarding the KPIs that were defined in Deliverable D6.1 [5GN-D61], these are reported below:

- Show operating and capital expenditure
- Show service revenues
- Show TCO savings

The KPI "Operating and capital expenditure" is shown in the latest version of the demonstrator in terms of TCO over the study period and network reduction in expenditure. Service revenues will be shown in the final release of the demo.

The key outputs from the demonstrator is to show the benefits of the innovations in 5G NORMA architecture. These benefits can be in terms of costs or revenues. The outputs from demo 4 show the economic benefits of the 5G NORMA innovations for three evaluation cases, in particular:

(1) Evaluation case 1: Results of DRAN vs. flexible CRAN are shown in terms of cost benefits
(2) Evaluation case 2: Benefits in terms of cost saving per tenant when multi-tenancy is applied
(3) Evaluation case 3: Benefits of multi-service support via network slicing in 5G NORMA which are two-fold:
    a. Cost savings per service due to 5G NORMA architecture providing multiple services from one shared platform vs. multiple single service legacy networks
    b. Providing extra revenue from non-eMBB services

Note that the demonstrator is a taster of WP2 results. It illustrates the key messages and the benefits of 5G NORMA innovations, for a full set and detailed results readers can refer to WP2 documents.

A specific learning using this infrastructure is the advantage of CRAN vs. DRAN (cf. Figure A-9) and site sharing. The results show the benefits of CRAN vs. DRAN with one or two operators taken into consideration. In addition, benefits of multi-tenancy due to 5G NORMA architecture is shown in the web-based demo.

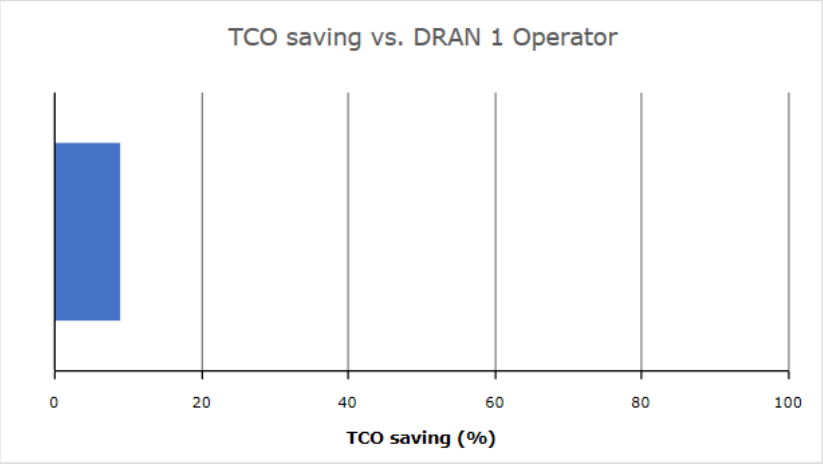



**Figure A-9: Example results of site sharing gain of CRAN and DRAN.**

## A.3. 5G NORMA Demonstrators

5G NORMA develops four specific demonstrators to showcase the main concepts of the architecture.

The first demonstration (Demo 1) is named "Native Multi-Service Architecture". This demo introduces the concept of Software-Defined Mobile Network Control (SDM-C), one of the core components of the whole project, and Multi-service and context-aware adaptation of network functions. The main goal is to study the SDM-C behaviour and the interaction with eNBs in two different scenarios. The first one wants to show the gain produced by the SDM-C connected to a hardware legacy eNB. The second one aims to show the SDM-C behaviour in a more realistic 5G environment, where is connected to tens of simulated software eNBs. In both the scenarios SDM-C and eNBs communication is done with a dedicated communication protocol, which is deployed for this demo with reasonable efforts, through the south bound interface (SBI), cf. [5GN-D51]. The output of the demo is to show how the SDM-C can react when different users require different services. Based on the feedback of the eNB and the SDM-C can both trigger reconfiguration of core network function from edge or central cloud and manage the scheduling policy of the eNB to guarantee the QoS required by every user. As a result of such intelligent placement of eNBs into central and edge-cloud configuration QoS of every user is satisfied. This concept of dynamic and service-aware network reconfiguration is a great improvement in system performance over legacy LTE network.

The second demonstration (Demo 2) is the Multi-slice service aware orchestration and it is focused on the MANO layer. It introduces the concept of Adaptive (de)composition of network functions with Software Defined Mobile Network Orchestrator (SDM-O) and novel concepts of network sharing and network slicing. Its objective is to simulate a 5G NORMA-like architecture, as defined in WP3, [5GN-D32] and built slices for different network services including different network functions, i.e. LL and MBB. Both the different requirements must be satisfied, i.e. low and controlled end to end latency, high throughput and support of high volumes of devices for the two services respectively. These different requirements are dealt by means of the dynamically deployment of the VNFs, depending on the service and the infrastructure requirements. This means moving NFs from edge cloud, near the antenna side and consequently to the user, and central cloud, near the server, consequently very far from the user, cf. [5GN-D31] for more detail, and between different central clouds, showing how the 5G NORMA innovations let to satisfy different requirements.

The third demonstration (Demo 3), Secured Multi-Tenancy Virtual Network Resources Provisioning via V-AAA, shows PoC of tenant data isolation, secured virtual network resources provisioning and a hierarchical and distributed V-AAA approach for the flexible 5G network architecture. This demonstration illustrates the most important security work in any flexible RAN or system, which are tenant isolation and tenant data isolation, tenant's data replication in many-to-one manner and maintaining the central governance in the mobile network operator (MNO) core network, as defined in [5GN-D31]. Moreover, wants to show how to protect and prevent network slices and network entities from being compromised during the network resources provisioning and deployment which take place. In this demo is used a tokenization technique for securely identifying, accessing, terminating, provisioning and deploying the network resources and services via a provisioning platform. The output is a web application in which it is wanted to show to the user all the procedures and that the protection is consistent

The fourth demonstration (Demo 4) is the Online Interactive 5G NORMA Business Cases Evaluation Tool, which presents the economic benefits of a 5G network, which is one of the key principles in 5G NORMA project using a graphical interface. Demo 4 shows the economic feasibility of the 5G NORMA architecture by looking at three different evaluation cases which are described in Deliverable D3.1 [5GN-D31]. The network cost modelling is based on real-life areas with different geographic features, i.e. population density, clutter, and others, called *geotypes*. This graphical interface gives users the ability to quickly check the benefits of the new architecture by looking at different evaluation cases, cf. [5GN-D61] for more details. The types of parameters that the user is able to change in the model include among others:

- The difference between 5G and LTE-A Pro ARPU (Average Revenue Per User) by service, i.e., varying from pessimistic to optimistic assumptions of the additional value delivered by 5G network capabilities.
- The growth rate of service demand in broad lower, central and higher sensitivities.
- Up to 3 different configurations of antenna, edge cloud and central cloud sites reflecting variations in the distribution of processing load between the edge and core and the split cost between them.
- Total “Present Value” cost, is the total cost of the network for the study period by taking into consideration the present value of the money.

The platform for web-development lets to share and disseminate results of a research project, helping end users to see key outputs through a simple console of a complex cost model. The output of this demo is the costs variations in term of revenues, profits or cash flows and social benefits over time in different combination of the input.

**SDMC Description**

As previously mentioned, Demo #1 (Native Multi-Service Architecture) implements a version of one of the core functional blocks in the 5G-NORMA architecture: the SDM-C component. This implementation of the SDM-C is connected to the eNB (in particular to L2) and provides two main functionalities:

- Monitoring and controlling of the radio transmission key parameters of the connected users, for example, the received signal strength, the active service-type and load in both downlink and uplink directions;
- Reconfigure the placement of the network functions in the edge-cloud, i.e., deployed at eNodeB with low latency and limited processing resources, and central-cloud, i.e., placed in the core network with relatively higher latency and more processing resources, if the resources are enough or trigger the SDM-O if new resources are needed.

The SDM-C is used in two different scenarios. The first one in a legacy network connected to a hardware eNB. In the second is connected to tens of software eNB a in a simulated scenario, to show its behaviour in a 5G-like scenario. SDM-C and eNBs communicate by means of a dedicated protocol, which is out of scope of 5GNORMA, and defined only for this demo, through the SBI. This let the SDM-C to monitor the network, and on the reports generated by the eNBs, i.e.

wideband channel quality indicator (WCQI), buffer occupancy, number of resource block (RB), and modulation and coding scheme (MCS), the QoS/QoE of users, the SDM-C trigger or start the reconfiguration of the network. On the SDM-C in the Demo 1 is also implemented a service-aware scheduling decision logic algorithm. The decision logic let the SDM-C to control and modify the UE-specific scheduling for specific users based on the feedback parameters, the same already mentioned before, the type of service running at the user, its pending load, the RSRP and MCS, etc. In the particular of Demo 1 first scenario when an interference signal is added and the QoS of the MBB user decrease the SDM-C based on the feedback messages of the eNB, has to redefine the policy scheduling for this user. Then it sends a control message to the eNB through the SBI, with a new scheduling policy which guarantee the QoS to be in target again for both the users, since the LL user must not be affected from the decision done for the MBB user. This feature is roughly 1000 times slower than the scheduler on the eNB, so it is not running takes in real-time, and it is activated only when the requirements of the service are not satisfied. The refresh rate is kept slow to avoid unnecessary overhead, since the commands do not change frequently. On the other hand, the scheduler on the eNB works in real time, guaranteeing the correct scheduling policy is applied every TTI.

In the second scenario of the Demo1 SDM-C can reconfigure of network functional elements into central or edge-cloud based on the network parameters and SDM-C's decision logic. In this scheme, SDM-C can intelligently make the decision to place the network functional elements of each individual eNBs into central- or edge-cloud. This choice is taken considering the traffic load on eNBs or the type of dominant service running in the specific eNB. User with LL services require to be placed in the edge cloud, near the user, on the other side the service require high computational load are placed in the central cloud. The SDM-C can detect the concentration of different service-requirements in different cells based on feedback messages it receives from the connected eNBs. SDM-C detects what cells can serve a particular requirement of the users that is serving and place its functional elements in the best cloud, central or edge. Using the service-aware decision-logic, the SDM-C places the remaining eNBs in edge or central cloud configuration respectively. As a result of such intelligent placement of eNBs into central and edge-cloud configuration, the UEs can achieve the desired requirements depending on the service required, improving the overall system performance and can satisfy a mix of different services in the network.