8

Network Slicing

Alexandros Kaloxylos[1], Christian Mannweiler[2], Gerd Zimmermann[3], Marco Di Girolamo[4], Patrick Marsch[5], Jakob Belschner[3], Anna Tzanakaki[6], Riccardo Trivisonno[7], Ömer Bulakçı[7], Panagiotis Spapis[7], Peter Rost[2], Paul Arnold[3] and Navid Nikaein[8]

[1] University of Peloponnese, Greece
[2] Nokia Bell Labs, Germany
[3] Deutsche Telekom AG, Technology Innovation, Germany
[4] Hewlett Packard Enterprise, Italy
[5] Nokia, Poland (now Deutsche Bahn, Germany)
[6] National and Kapodistrian University of Athens, Greece and University of Bristol, UK
[7] Huawei German Research Center, Germany
[8] EUREKOM, France

8.1 Introduction

The 5th generation (5G) network is promising to upgrade not only the well-known mobile broadband services, but also enable the support of services for the so called "vertical industries" (e.g., health, transportation, factories, energy). An extensive list of 5G use cases can be found in Chapter 2. All these verticals have their own requirements and needs which may be highly divergent. Their operational requirements are translated into different key performance indicators (KPIs) such as user experienced data rate, end-to-end (E2E) latency, reliability, communication efficiency, availability, and energy consumption. These have to be satisfied in specific environments characterized by different parameters such as mobility, expected data traffic, density and types of network nodes, position accuracy, etc.

As discussed in Section 5.2.2, network slicing is introduced as one of the key enablers to support the required level of flexibility in 5G networks. Network slices are essentially multiple logical networks deployed over the same physical infrastructure. During the past years, there has been a lot of debate to reach a commonly accepted and concrete definition of network slicing. At a first glance, the notion of slicing may seem to be very similar to well-established solutions that essentially support logical networks, such as:

- Virtual local area networks (VLANs) where different hosts are logically brought under the same broadcast domain;
- Virtual private networks (VPNs) that are used to connect multiple hosts through private and secure tunnels providing logically closed groups;

5G System Design: Architectural and Functional Considerations and Long Term Research, First Edition.
Patrick Marsch, Ömer Bulakçı, Olav Queseth and Mauro Boldi.

- Dedicated mobile core networks (CNs) with standardized solutions (e.g., DECOR, eDECOR), as discussed in Section 8.2.1.

This observation would indeed be correct if the target of 5G network slicing was only to separate nodes or share resources and apply specific security and policies per service. However, these solutions do not address the need to support a number of use cases with different KPIs. This suggests that the network functions (NF) will not necessarily be the same in all slices. The formal specification of slices related to the enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC) and massive machine-type communications (mMTC) families of use cases, as described in Section 2.2, is currently underway. Already, there are hints to new functions that need to be introduced (e.g., for slice selection) or others that need to be defined per use case (e.g., for session management or mobility management).

In the latest specifications, the 3rd Generation Partnership Project (3GPP) considers a network slice to be “A logical network that provides specific network capabilities and network characteristics” [1]. In [2], it is mentioned that it is a “network created by the operator customized to provide an optimized solution for a specific market scenario which demands specific requirements with end to end scope”. Also, 3GPP has defined in [3] that a network slice is implemented by “slice instances”, which in turn are created from a “network slice template”, being a template of a logical network including the NFs and the corresponding resources. A similar definition is also provided in [4], where the use of common functions and sharing of resources among slices is possible. Thus, these definitions suggest the ability of deploying multiple logical networks possibly over the same physical infrastructure. This level of flexibility is needed to support the diverse requirements and KPIs of the 5G use cases as well as to reduce the cost for network deployment and operation. Thus, network slicing is expected to be one of the key features of 5G networks, realized by introducing solutions based on softwarization, virtualization and functional modularization.

The key services provided by network slicing and a comparison with legacy cellular systems are illustrated in Figure 8-1 [5]. On the left side, where legacy systems are depicted, one can see that the same NFs over monolithic network elements are used to support all telecommunication services

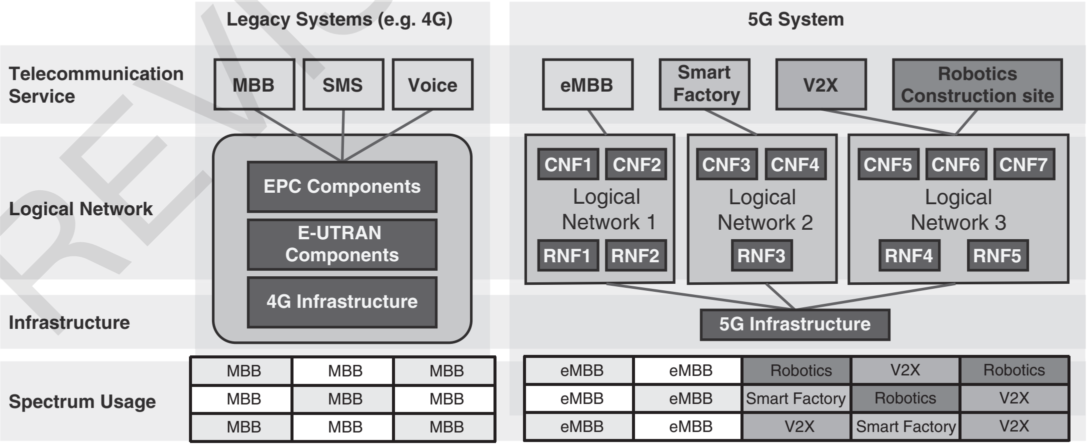


Figure 8-1. Key principles of network slicing.

(e.g., voice, SMS, web browsing, video streaming). Such homogeneous treatment of all services, even with the definition of different Quality of Service (QoS) classes, is a non-acceptable compromise for 5G services. This is because KPIs like ultra-high reliability and ultra-low delay cannot be deployed for all vertical services in a technically and economically viable way. In 5G networks, multiple verticals will be supported by dedicated or shared logical networks running on top of the 5G infrastructure. These logical networks will be a composition of core network functions (CNF) and radio network functions (RNF) and will run over the same physical network components. Note that dedicated spectrum may be allocated for each slice, or several slices may share the same spectrum, but manage to meet the service-level agreements (SLAs) with the verticals using the specialized NFs, as detailed further in Section 8.2.3.

The key enablers for the dynamic deployment of slices are considered to be a) network function virtualization (NFV), allowing the virtualization of sets of NFs and their organization into virtual blocks that may be connected together to create communication services [6] and b) software defined networking (SDN) used for separating control plane (CP) and user plane (UP) and allowing for full programmability of the network [7]. These enablers are extensively presented in Section 10.2. Moreover, contrary to the evolution of previous generations of mobile networks, it is foreseen that 5G will require not only improved networking solutions but also a sophisticated integration of massive computing and storage infrastructures into the different network domains (i.e., access, transport, and core network) to support the different use cases and services.

Although slicing of 5G networks is a rather new topic, the research area is growing rapidly. In [8] and [9], several solutions for the slicing of shared resources as well as the virtualization of NFs are discussed. This chapter contains a comprehensive summary and analysis of the latest 3GPP specifications and also findings from several 5G Public-Private Partnership (5G PPP) research projects. Section 8.2 provides detailed information for each network domain, as well as for the support of slicing across different operator administrative domains. It also discusses a realistic E2E example of network slicing. In Section 8.3, several slice operation aspects such as slice selection and isolation, context transfer, slice orchestration and management are presented. Finally, Section 8.4 summarizes the key findings and also lists the technical challenges that still remain open.

8.2 Slice Realization in the Different Network Domains

The network slicing concept refers to E2E logical networks and is meant to provide flexibility to each component of the communication system (i.e., access, core, transport). However, technological specificities of each domain require addressing different key issues for the realization of the slicing concept. For this reason, the problem of network slicing is addressed on a domain basis in the following sub-sections.

8.2.1 Realization of Slicing in the Core Network

To achieve the intended flexibility and adaptability for future 5G services, a modularization of NFs, based on detailed functional decomposition and use in dedicated slices, is a prerequisite especially in the CN [10], [11].

In 4th generation (4G) networks, there are monolithic network elements within the Evolved Packet Core (EPC), i.e., Serving Gateway (S-GW), Packet Gateway (P-GW), and Mobility Management Entity (MME), which aim to integrate hardware/software (HW/SW) implementation in physical nodes. Nevertheless, slicing approaches are initially available also in 4G. One example is a multi-operator core network (MOCN), which allows several operators to share a common radio access network (RAN), while running separated CNs with proprietary services [12]. However, 5G slicing not only enables sharing of the underlying infrastructure (core and access) by multiple logical networks, it also allows for a different configuration of these logical networks.

With a dedicated core (DECOR), initially introduced in 3GPP Release 13, operators can deploy multiple dedicated CNs (DCNs) within a single operator network [13]. A DCN may consist of one or more MMEs and one or more S-GWs/P-GWs, each element potentially featuring different characteristics and functions. The introduction of DECOR was triggered by the problem of enhancing mobile networks with architecture flexibility and enabling either resource sharing or resource isolation among specific groups of subscribers (e.g. for MTC/CIoT subscribers, subscribers belonging to specific enterprises or to separate administrative domains, etc.). The design of DECOR aimed at having no impact on legacy user equipments (UEs) and at allowing different DCNs to share the same RAN, but has some drawbacks with respect to increased initial access time and signaling efforts.

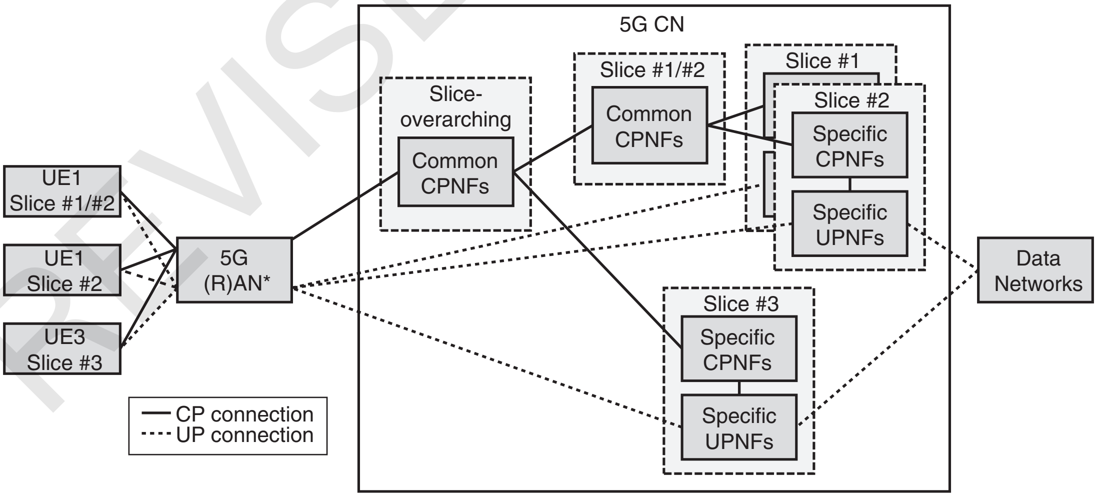
Evolved DECOR (eDECOR) [14] was introduced in 3GPP Release 14. Its aim was to support all 3GPP RANs while being backward-compatible to DECOR. DCN selection and allocation procedures, as well as slice isolation among DCNs, are improved. Also, the required CP signaling is minimized by reducing or avoiding the occurrence of redirection procedures. eDECOR introduced the concept of UE-assisted DCN selection, which is unfortunately not applicable with legacy UEs, and a network assigned DCN-ID. The DCN-ID, permanently stored at a UE, is included in Radio Resoure Control (RRC) messages piggybacked in Non-Access Stratum (NAS) signaling (e.g., Attach Request, Tracking Area Update). This allows the Network Node Selection Function (NNSF) in the RAN to directly select the proper DCN towards which the NAS signaling needs to be forwarded. eDECOR further addressed congestion control for DCN types to cope with CP NAS signaling congestion which may occur at MME serving (and hence be shared amongst) multiple DCNs. In addition, it also optimized load balancing among MMEs.

These early attempts can be considered as precursors of 5G network slicing. They even provided some customization for different use cases (e.g., mMTC). However, they were lacking the flexibility expected to be in place to support the diverse 5G use casess since their customization was always limited by the functionality of the EPC architecture. To progress towards next generation networks, the full advantage of network enablers such as SDN, NFV [6], mobile edge computing or multi-access edge computing (MEC) [15] and cloud computing technologies needs to be exploited, in combination with the introduction of the network modularization design principle as well as the service-based architecture (SBA) model. These new features target the support of multi-tenancy, the minimization of service delivery through flexible network deployment, nodes' reconfiguration and the empowerment of third parties to customize the network slices according to their needs.

Considering the 5G modularization approach, as introduced in Section 5.4.1, all relevant CN NFs should be broken down to a suitable fine-grained level. Modularization does not mean that it is always required to go down to an "atomic granularity", but to combine basic NFs in a way that they create together a functional building block with a self-contained dedicated task. This block is addressable and configurable within a network slice instance (NSI). Following the SDN principle, there will additionally be a strict separation between CP and UP NFs in the 5G CN, which is not originally the case in 4G. In 4G networks, P-GW and S-GW include both CP and UP NFs, whereas the MME entirely relates to the CP. This separation in 5G networks will enable an independent scalability of resources for CP and UP and the introduction of new NFs for both parts. It has to be noted that 3GPP is also taking care of that fact in the current standardization process by identifying and separating those NFs as well as by defining interfaces in between [16]. Based on the results of research projects and standardization organizations, orthogonal sets of CN NFs have been defined for 5G CP and UP. A more detailed description of those NFs and procedures in between is given in Chapter 5 and can be found in [1], [3] and [17]. It is still an ongoing discussion if the granularity of those NFs is sufficient to achieve the targeted flexibility for 5G, or if a finer level of separation among functions is still needed.

Slices in the 5G CN for dedicated business or service purposes are instantiated by a concatenation of selected NFs taken from available repositories based on special network slice templates (NST) [3][17]. Dependent on requirements with respect to isolation (resource, security, etc.), NSIs may consist of fully separated CP and UP NFs, but it may be also possible that some of the NFs are shared by several NSIs. Control plane network functions (CPNFs) have been classified as common control NFs (CCNFs) and slice-specific control NFs (SCNFs). Sharing of CPNFs is especially linked to Access and Mobility Management Function (AMF) instances serving UEs which are simultaneously connected to more than one NSI. In that case, the corresponding AMF instance should be common to all NSIs serving one UE. Among CCNFs, there is also one NF dedicated to slice selection, hence denoted as Network Slice Selection Function (NSSF). In Figure 8-2, an example with three NSIs is given where two of those share CCNFs. Also, the figure illustrates a slice overarching CPNF block where slice generic functions like NSSF may reside. As depicted in this figure, the 5G CN is being designed to support not only radio access networks (RANs), but any type of access networks (AN), including even fixed networks.

As NSIs have to act as separate domains, trustworthiness with respect to security aspects during the slice lifecycle is of extreme importance. Any potential cyber-attack on one NSI must have no impact on another one running on the same infrastructure or sharing common NFs [18].



* Slicing inside the (R)AN not explicitly shown

Figure 8-2. Exemplary implementation of network slices in the 5G CN with common and slice-specific NFs.

In addition to the legacy 4G point-to-point interface model, an alternative SBA model has been defined by relevant standardization organizations like 3GPP [1] for CPNFs in the 5G CN. Within that approach, each CPNF exposes a service-based interface (SBI), by which the authorized NFs can access the services it provides. SBIs provide higher flexibility in the interaction among NFs and allow multiple alternative interconnections among those to define slice-tailored architectures.

Another important aspect with respect to the differentiation against 4G is the Network Exposure Function (NEF) that provides the means to expose services and capabilities of CPNFs to third parties and application functions (AFs) such as MEC, monitoring, policy/charging or data analytics. AFs will provide information to the 5G CN, e.g., for packet flow handling (routing, QoS, etc.) and policy control. Trusted AFs may directly interact with CPNFs using application programming interfaces (APIs), while untrusted AFs have to apply the NEF framework.

Due to typically asynchronous timing behavior of NF processing in the 5G CN with respect to radio framing, CN NFs may usually be implemented as virtual NFs (VNFs) [6] in cloud infrastructures, e.g. in front- or backend data centers of operators or third parties (central clouds). In contrast to such centralized approach, some of the CN NFs as well as AFs may be located closer to the access network in so-called edge clouds, e.g. to support low latency use cases by caching, local break outs or MEC. Edge clouds may be placed, e.g., at central offices of operator networks or at larger campus locations (enterprise premises, factory halls, sport stadiums, etc.) [11][17]. This flexibility w.r.t. the placement of functions on a per-slice basis is expected to be one of the key characteristics of 5G networks.

8.2.2 Slice Support on the Transport Network

The 5G heterogeneous transport network is envisioned to rely on the convergence of a variety of technologies including wireless and optical networking, and to support a variety of services including backhaul (BH) and fronthaul (FH) services offering efficiency, scalability and management simplification, as discussed in detail in Chapter 7. In this context, transport solutions enhanced with advanced features such as slicing and virtualization will allow a pool of network and compute HW and SW resources to be shared and accessed remotely without the prerequisite of ownership. This will allow to create infrastructure slices that integrate heterogeneous technologies. These slices can transport FH services corresponding to various functional splits as well as BH services adopting novel approaches such as the notion of service chaining (SC). Network slicing and SC can be facilitated adopting and integrating architectural models such as the SDN reference architecture and the European Telecommunications Standards Institute (ETSI) NFV standard, as described in Section 10.2.

In the *highly heterogeneous 5G transport*, a critical challenge that needs to be addressed is that of “cross-domain” slicing. In this context, some fundamental incompatibilities associated with technology heterogeneity need to be addressed, such as the separation of CP and UP for some technologies (e.g., optical/wireless SDN) and close coupling between CP and UP for others. This introduces challenges in defining the relevant interfaces not only across domains, but also between the slicing systems and the orchestrators responsible for the composition and the provisioning of SCs over the transport slices. More specifically, interfacing between technology domains, including isolation of flows, flexible scheduling schemes and QoS differentiation mechanisms *across domains*, plays a key role and can be achieved by adopting flexible HW functions. These functions will be exploited to enable dynamic and on-demand sharing, partitioning and grouping of resources as required to form *independent transport network slices* with guaranteed levels of isolation and security. In this context, programmable Network Interface Controllers (NICs) that are commonly used to bridge different technology domains at the UP can have an instrumental role. These controllers have a unique ability to provide HW-level performance exploiting SW flexibility and can offer not only network processing functions (i.e., packet transactions), but also HW support for a wide variety of communication protocols and mechanisms.

5G transport slicing can be implemented through the adoption of a hierarchical architectural approach that supports management of network elements and abstracted resources by different layers [19]. Each network domain may host multiple SDN UP elements and expose its own virtualized resources through an SDN controller to the upper layer SDN controllers. A hierarchical SDN controller approach can assist in improving network performance and scalability as well as limit reliability issues. The top network controller will manage network resource abstractions exposed by the lower-level controllers that are responsible to manage the associated network elements. Orchestration of both computational and network resources can be performed by the NFV orchestrator and can be used to support multi-tenant chains, facilitating virtual infrastructure provider operational models. This will also be responsible to interact with third party operations and support systems (OSS).

8.2.3 Impact of Slicing on the Radio Access Network

Slicing support for the RAN has been vigorously researched during the past years. In this section, we capture the latest status of 3GPP standardization activities and also elaborate on the key principles. As currently reported by 3GPP [20], slicing in the RAN can be realized by Medium Access Control (MAC) scheduling and by providing different configurations for the NFs. Thus, traffic for different slices is handled by different Protocol Data Unit (PDU) sessions, for instance at MAC or Packet Data Convergence Protocol (PDCP) level. The different treatment among slices can be achieved by using specific identifiers in signaling messages to indicate specific slices. The configuration of the NFs to support the different slices is considered as implementation detail by 3GPP. The selection of the RAN part for an E2E network slice is done by assistance information, provided by the UE or the CN entities, that identifies one or more preconfigured network slices. The system supports policy enforcement between slices as per SLAs, and is able to apply the best radio resource management (RRM) policy to support the slice-specific SLA. Since the RAN can support multiple slices, resource isolation mechanisms have to be in place. These mechanisms are mainly RRM policies including scheduling schemes as well as protection mechanisms that are currently considered as implementation details in the context of the standardization activities. Nevertheless, the MAC-layer scheduling requires to be aware of slice definition and user membership in order to apply the RRM policies. Note that it is possible to fully dedicate resources (i.e., spectrum) to a specific slice and thus isolate it from other slices. The following paragraphs elaborate on these main principles.

One aspect about network slicing that is especially relevant in the RAN is the notion of **sharing the same radio resources and physical infrastructure** (e.g., processing capabilities) among multiple slices, ideally to the largest possible extent, as described in Chapters 11 and 12 (see for instance Section 12.6). Many envisioned 5G use cases are expected to be only economically viable if they can exploit significant synergies with other use cases and do not require dedicated infrastructure. However, there may be cases where some physical separation of slices is requested by involved stakeholders or even mandated by the law or some other administrative domain. For example, for highly

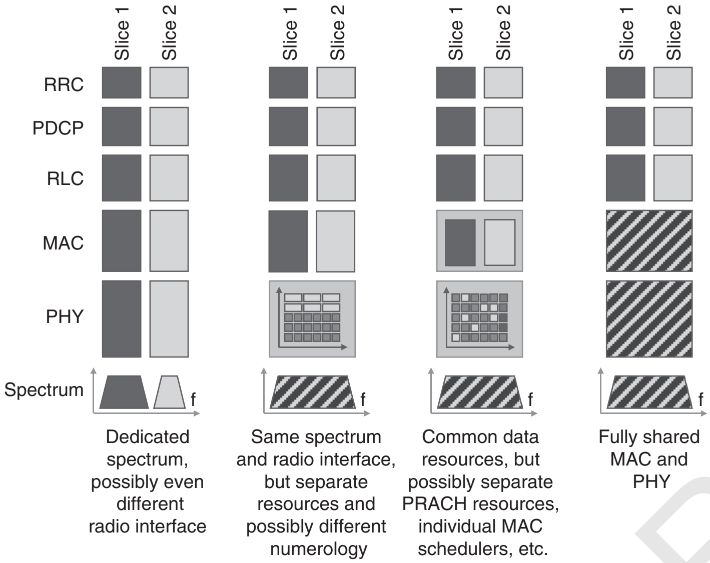

Slice 1
Slice 2
Slice 1
Slice 2
Slice 1
Slice 2
Slice 1
Slice 2
RRC
PDCP
RLC
MAC
PHY
Spectrum
f
f
f
f
Dedicated spectrum, possibly even different radio interface
Same spectrum and radio interface, but separate resources and possibly different numerology
Common data resources, but possibly separate PRACH resources, individual MAC schedulers, etc.
Fully shared MAC and PHY

Figure 8-3. Representative RAN slicing scenarios with different level of resource sharing and isolation.

safety-critical use cases, it may be required by related regulators to keep some physical separation of the radio access for different slices.

Figure 8-3 illustrates some representative RAN slicing scenarios with different levels of resource and infrastructure reuse among slices that are thinkable, from full slice separation (left) to maximum multi-slice RAN integration (right). On the very left, we see the case where two slices use dedicated spectrum and possibly different radio interface specifications; this may be seen as the legacy case where for instance Long Term Evolution – Advanced (LTE-A) is used for an MBB slice. Then, one could have the case where slices share the same spectrum, but still use strictly separated physical resources therein, possibly involving different numerologies, being interleaved or overlaid to each other in some form. In this case, one would likely also have dedicated MAC instances and MAC schedulers for the different slices. Resembling a further extent of reuse, multiple slices could share the same spectrum, radio numerology and most of the resources, but still have some dedicated resources, such as dedicated Physical Random Access Channels (PRACH), for instance to guarantee stringent slice-specific QoS service requirements. In the case on the very right, one would have a fully shared and integrated MAC and physical layer (PHY) for both slices. Note that if the PHY is largely shared, one could further consider having individual MAC instances per slice, or a common MAC instance across multiple slices, for instance enabling multi-slice MAC scheduling. Many different flavors are possible in this respect, i.e., one may for instance assume that two slices use a same high-level MAC scheduler that allows for some flexible resource split among slices, while the slices involve some finer-granular dedicated schedulers per slice.

Figure 8-4 illustrates a slice-aware scheduler architecture with a resource visor that abstracts and shares the physical resources among slices according to the enforced RRM policies, and a slice resource manager (SRM) that allocates resources for UEs belonging to its slice according to the applied scheduling algorithm (e.g., proportional fair - PF, round robin – RR, priority-based, delay-based) [21]. It can be seen from the figure that scheduling is performed in two levels,

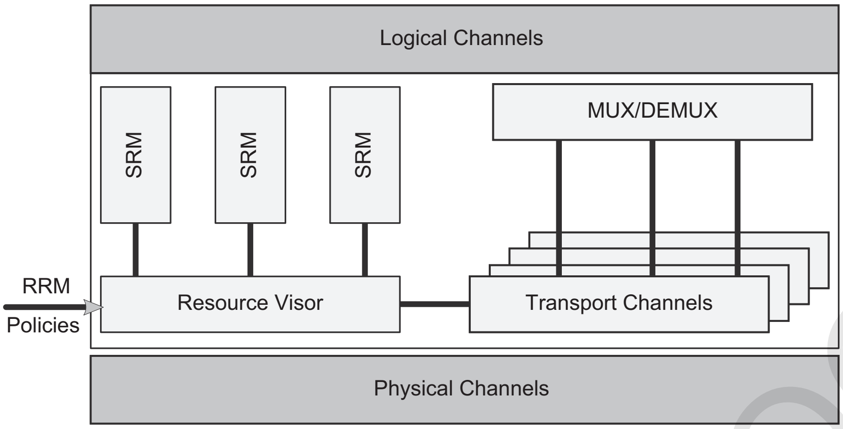


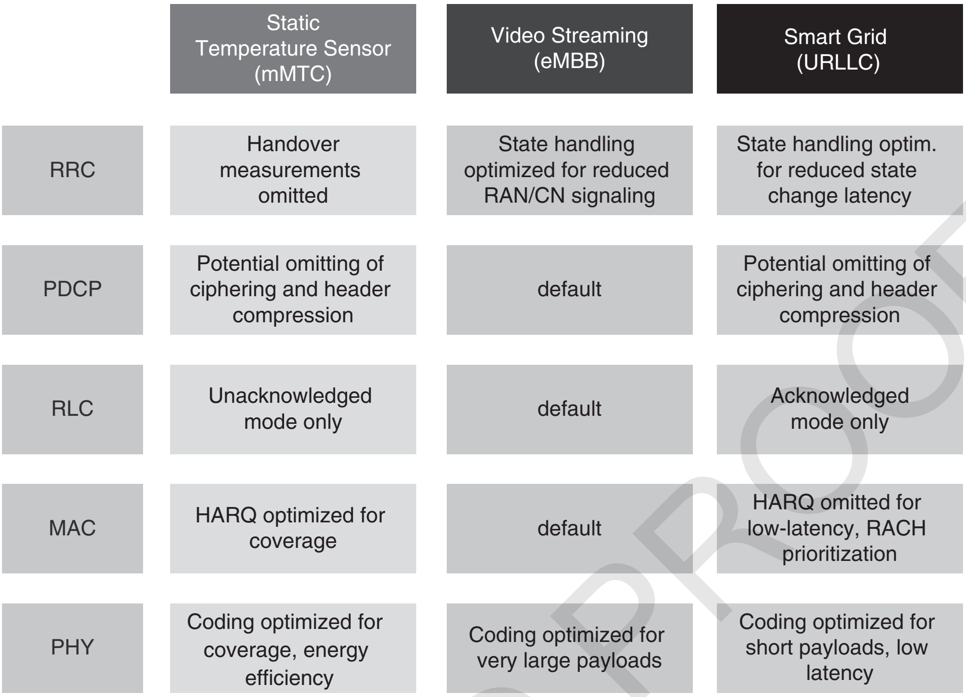
Figure 8-4. Slice-aware MAC scheduling architecture.

namely intra-slice and inter-slice, to decouple how UEs are served and how the resources are granted and mapped to the physical channels.

Another important aspect of network slicing in the RAN relates to the question of whether different services or slices use the **same or different specification and/or implementation of NFs**, and, in the former case, to which extent NFs may be **configurable** and chained to reflect service-specific or slice-specific needs. We explicitly refer here also to services and not only to slices, as even in the context of a single slice encompassing multiple services one may use different NFs or different configurations thereof to obtain service-tailored treatment. This is to some extent independent of the extent of physical resource and infrastructure sharing discussed before. For instance, two services or slices may use separated spectrum, but still reuse the same implementation and possibly configuration of, for instance, PHY network functions. On the other hand, two services/slices may use the same spectrum and radio resources, but use different implementations of some PHY functions, like for instance service- or slice-tailored encoding/decoding functions.

When it comes to actual slice implementation, the maximum reuse of resources and functions should be targeted so as to achieve the optimum use of the available resources and maximize the multiplexing gain. On the other hand, given the fact that different services with different requirements are targeted, the slices will not necessarily have the same functions or functions' configurations. Definitely, some functions need to have at least some common parts such as RRM functionalities for slices that share the same physical resources (e.g., for ensuring the slice protection), or RRC (e.g., for enabling the initial slice selection). Other functions, such as Hybrid Automatic Repeat reQuest (HARQ), random access, ciphering, etc. may be differently configured or even omitted if they are not needed.

The common understanding is that for the sake of a swift standardization process, simplified implementation and also less complexity, chip space etc., one should strive for a maximum reuse of RAN network functions among services and network slices. However, it is generally envisioned to allow the configuration of NFs such that they can be tailored to the specific needs of a given service or slice [22][23]. As an example, one could configure an RRC state machine to work very differently, depending on whether this is used in the context of an eMBB or an mMTC slice. For the former, the



	Static Temperature Sensor (mMTC)	Video Streaming (eMBB)	Smart Grid (URLLC)
RRC	Handover measurements omitted	State handling optimized for reduced RAN/CN signaling	State handling optim. for reduced state change latency
PDCP	Potential omitting of ciphering and header compression	default	Potential omitting of ciphering and header compression
RLC	Unacknowledged mode only	default	Acknowledged mode only
MAC	HARQ optimized for coverage	default	HARQ omitted for low-latency, RACH prioritization
PHY	Coding optimized for coverage, energy efficiency	Coding optimized for very large payloads	Coding optimized for short payloads, low latency

Figure 8-5. Examples for service- or slice-specific network functions or configurations thereof [22].

minimization of CN/RAN signaling could be an important objective. For the latter, the device power consumption could be the main issue of interest, as discussed further in Section 13.2. Further, it is envisioned that certain NFs, or elements of these, could be turned on or off for certain services and slices. For instance, the PDCP implementation for different services may in general be the same, except that for some services header compression is activated, while for others it is not. The stated and other examples of service- or slice-specific configuration or activation/deactivation of NFs is illustrated in Figure 8-5, and this topic will be further illustrated alongside the description of a single E2E slice example in Section 8.2.5. Note that beyond a different selection and configuration of NFs for different services and slices, some service-specific processing optimizations may be applied, as detailed in Section 6.4.2.

Please note that some slice- and service-specific processing will also inherently be facilitated by the new QoS framework that 3GPP has decided upon, as described in detail in Section 5.3.3. As a part of this, 3GPP has specified in [24] a new sublayer called Service Data Application Protocol (SDAP) to operate on top of PDCP, see also Section 6.4.2.1. The main services and functions it provides include the mapping between a QoS flow and a data radio bearer, and the marking of QoS flow identifier. The new information can be used by RRM functions like scheduling and admission control to offer customized support for slices. As will be detailed in Section 8.3.4, SDAP can also be utilized for inter-slice RRM functions in an attempt to fulfil the SLAs of all network slices. Additional information on this topic can also be found in Section 12.6.

Finally, one aspect that is likely specific to the implication of network slicing on the RAN is the introduction of **functionality that is specifically designed to facilitate the operation of network slices** with diverse and stringent requirements in a common radio infrastructure. Clearly, if one aims at a strong reuse of spectrum, radio resources and infrastructure resources such as processing capabilities, one needs other means to enable aspects such as slice protection, i.e., the guarantee that issues in one slice such as excessive load, possible malfunctioning caused, e.g., by erroneous devices or security attacks, do not have an impact on other slices. One further specific example for novel multi-slice functionality that will likely be introduced in the 5G context are means for multi-slice QoS and resource management, as outlined in Section 8.3.2 and described in more detail in Section 12.6.

8.2.4 Slice Support Across Different Administrative Domains

The slicing concept has been consistently identified so far with *single-provider* slicing. This means that a given service provider can create and deploy slices within the boundaries of its own administrative domain, using the resources at his own disposal. A key step ahead towards 5G objectives is the ability to define and provision cross-provider slicing, orchestrating resources offered by different administrations into E2E multi-provider services. This is paramount to maximizing the overall resource usage, avoiding the need of hinge overcapacity to reach the quality and performance levels demanded by 5G services.

There is no known solution to this problem yet. Resource orchestration and slice composition have until recently only been considered at intra-provider level. Main standardization working groups have recently started to take into account this scenario extension. Two relevant examples are given in [6] and [30].

From the research domain, there are also key findings for the support of cross-provider slices through the introduction of appropriate multi-provider orchestrator entities. Cross-provider slicing means that the slice can be made up by individual resources (e.g., virtual computation, storage and connectivity resources) which are partially or fully located in administrative domains different from the one spawning the slice creation and provisioning process, and owning, using and terminating the slice itself. This is also referred to as *resource slicing*. Such cross-provider slices can in turn be orchestrated with other slices, VNFs and connectivity resources to create more complex cross-provider services, also referred to as *service slicing*. The key point to make this happen is that every needed component must be searched, selected and provisioned according to a *service* model. A visiting domain entity, looking for the best resources to orchestrate and provision a given slice, must never have detailed visibility and access inside a visited domain. Instead, the visited domain must expose its available resources through a proper service catalog, where the resources themselves are presented in an abstract, mutually understood description. The orchestrating provider must be able to select the resources for its slice and access them for their requested usage, keeping a full separation from the rest of the visited domain.

Figure 8-6 shows how slicing is envisioned in a recent research project [25] tackling cross-provider orchestration, considering both resource slicing and service slicing. This architecture is based on a clear layer separation. Individual domains continue operating their own infrastructure, low-level controllers and intra-provider resource orchestrators. The I5 interface shown in the diagram is the legacy intra-domain interface between local orchestrators and infrastructure resources.

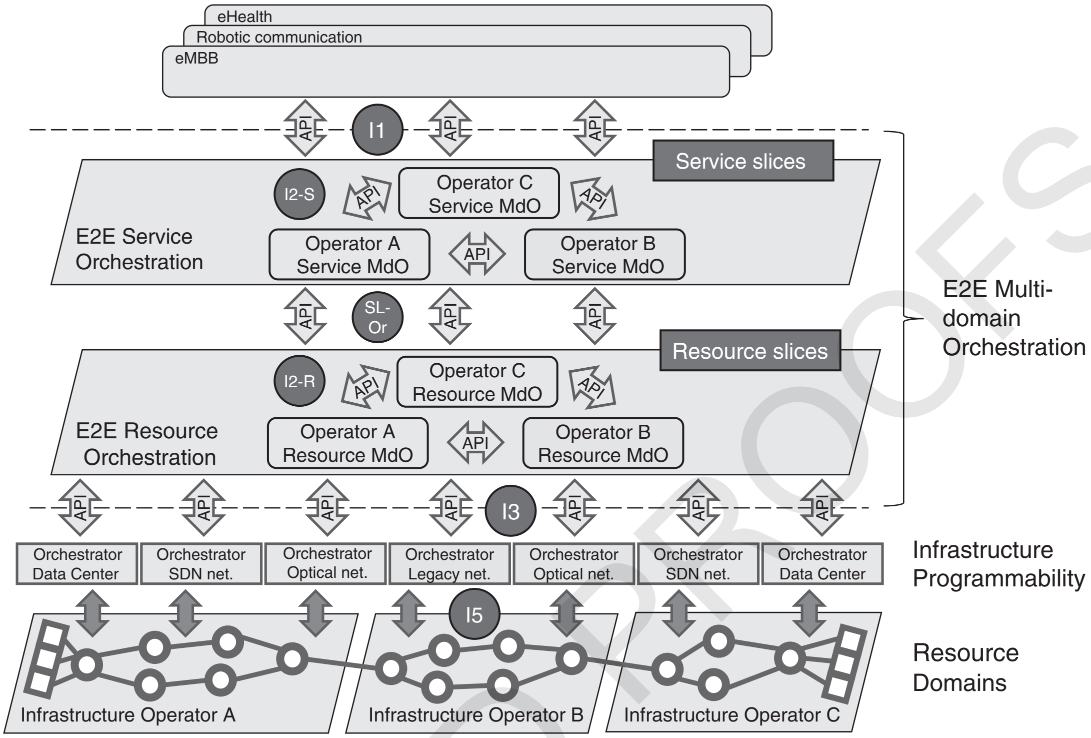


Figure 8-6. Slicing in the context of cross-provider orchestration [26].

On top of the legacy intra-provider layer, a multi-domain orchestrator (MdO) is positioned, communicating with the domain orchestrators via the interface labeled I3. Each MdO is logically decomposed into two submodules: a Resource MdO and a Service MdO, reflecting the architectural evolution currently undergoing in the ETSI NFV Management and Orchestration (MANO), as detailed in Section 10.2, and in charge of implementing resource slicing and service slicing as described above. MdO orchestrators in different domains communicate through the east-west interface generally labelled I2, used to expose the available resources from each provider, negotiate resource/service inter-provisioning, exchange information aimed at slice lifecycle management, and share a business layer (not highlighted in the picture). The I2 interface specifies a number of sub-interfaces, each one fulfilling a specific duty. Figure 8-6 highlights two such sub-interfaces:

- I2-R allows to share the resource-level details (computational resources and network topologies) available to the cross-provider ecosystem;
- I2-S facilitates service lifecycle management operations (instantiation, termination, and so forth).

The resource orchestration and service orchestration layers are interconnected through the Sl-Or interface, exposing to the Service MdOs a view of the available resources.

In the slice provisioning flow, one of the participant providers acts as *front-end provider*, and is the one who directly interacts with the user requesting the slice provision, through the interface labeled I1. Such provider is the ultimate owner of the user (customer) relationship, though of course, there must be mechanisms in place ensuring that all the providers cooperate to properly manage the slice, and business models duly apportion liabilities and responsibilities among the different providers.

The provisioning and usage of a slice built across different administrative domains requires clarifying a number of questions, among which the following should be stressed:

- **Specification**: a common data model must be defined, allowing to create slice abstraction descriptors or templates that can be automatically mapped and provisioned to external providers. This can be done by extending legacy data models (e.g., the ETSI NFV Network Service Descriptor) to incorporate the needed cross-provider add-ons;
- **Exposure**: the client MdO (i.e., the one residing at the front-end provider premises) must be able to search, in a service catalogue type repository, a directory of slice templates made available by other providers, to be selected, purchased along with a related SLA, and provisioned;
- **Slice control**: the client MdO must be given, beside UP endpoints to interconnect the slice with other service components, an additional control endpoint to internally access its assigned slice, configure its internal resources, and integrate them with others residing at different providers. This Fault Configuration, Accounting, Performance, Security (FCAPS) path typically goes through a service-specific component like the Element Management (EM) of the ETSI NFV MANO architecture;
- **Isolation**: the client MdO must be able to access the slice and its composing resources, while at the same time being fully isolated and unable to access any other resource or object inside the visited domain. Security provisions must hence be integrated in the design of the multi-provider orchestration framework;
- **Slice lifetime management** (see also Section 8.3.5): the client MdO must be able to monitor some given slice KPIs, detect possible failures and shortcomings, and trigger due actions when needed. Again, this must happen while safeguarding the privacy and non-accessibility of every off-slice resources in the visited domain. This is realized by sharing the relevant local providers' metric measures through the sub-interface I2-R, with the front-end provider collecting and combining all the partial metrics into E2E slice-level metrics. Monitoring of these latter metrics triggers recovery actions in case of need (e.g., when a given resource of the slice needs to be scaled out), enacted by the front-end provider, again through the proper I2 sub-interface conveying the due action requests to the external peer providers.

8.2.5 E2E Slicing: A Detailed Example

This section elaborates on an illustrative example for deploying multiple network slice instances in a mixed environment consisting of public, i.e. mobile network operator (MNO) owned, networks and private network infrastructure owned by a vertical enterprise. The infrastructure is used to commission an Internet of Things (IoT) network slice and two eMBB network slices.

5G Services on Factory Premises

As an exemplary deployment, a process automation use case from industrial manufacturing is considered. Traffic is composed of sensor readings, actuator control signaling, and eMBB services providing access to local as well as remote applications (e.g., augmented reality for machine maintenance). In a process automation environment, IoT devices include actuators, such as pumps, valves, etc., and sensors for capturing heterogeneous physical and logical quantities. The latter may for instance

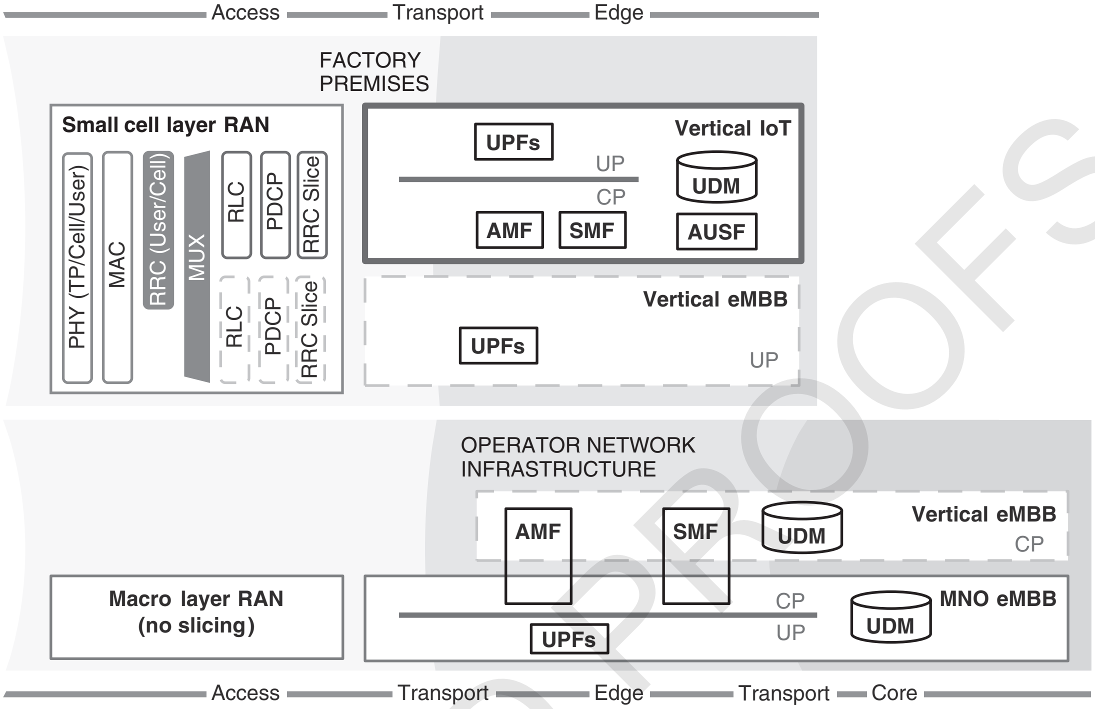


Figure 8-7. E2E network slice example for 5G services on factory premises.

include sensors for "reverse engineering" supporting maintenance processes or for critical safety applications, aiming to improve the overall operational efficiency and safety of the factory. When connecting such IoT devices, latency and bandwidth requirements can be very diverse. Such a setup requires two types of network slices, one covering machine-type communications for IoT devices and one covering eMBB traffic from smartphones, tablets and similar terminals. Figure 8-7 shows an example scenario with two E2E network slices (IoT and vertical eMBB) for the factory owner (also referred to as "vertical") as well as an eMBB network slice for the MNO. The network slices run on MNO infrastructure as well as the vertical's telecommunication infrastructure on the factory premises.

Ownership of Infrastructure, Spectrum, and Subscriber Data

In the given scenario, the vertical provides the small cell layer RAN equipment for all network slices that require coverage on the factory premises, i.e., both the IoT slice and vertical eMBB slice. Beyond the RAN, this includes the transport network and edge cloud resources, such as general purpose hardware for computing and storage. For operating the small cell layer RAN, the vertical rents dedicated spectrum resources from the MNO, such as higher frequency spectrum (e.g., above 6 GHz) with coverage strictly limited to the factory premises. For the vertical eMBB service, both small cell layer RAN and, if required, the MNO-provided macro layer RAN are utilized. RRM functions for the macro layer strictly remain under control of the MNO, which further owns a 5G-compatible network infrastructure consisting of centralized datacenters and distributed edge clouds comprising general-purpose as well as application-specific hardware. Subscriber information data including long-term security credentials are in possession of the vertical for the IoT devices. This assures full isolation of the vertical's IoT subscriber information from the MNO. For the eMBB subscribers, the MNO holds the corresponding data for own eMBB subscribers as well as vertical eMBB subscribers.

Domain-specific Network Slice Deployment and Operation Incl. CP and UP Considerations

For the deployment of the individual network slices, multiple options exist. As depicted in Figure 8-7, the IoT network slice deploys all functions in the domain of the vertical, and it is only used by IoT devices that are registered in the vertical's home subscriber system (HSS) or Unified Data Management (UDM) and Authentication Server Functions (AUSF). These devices are mostly stationary and never leave the factory premises. The small cell layer RAN as well as the IoT-specific User Plane Processing Functions (UPFs) and control plane functions, in particular Access and Mobility Management Function (AMF) and Session Management Function (SMF), are operated locally under full control of the vertical. Since also the security mechanisms are strictly realized locally (i.e., access stratum security, optional "over-the-top" security), the entire network slice operates in the shielded factory environment without exposure of any data to the MNO. In contrast, the vertical eMBB slice is deployed in an inter-domain manner, see also Section 8.2.4. The CN control plane is shared with the MNO eMBB network slice and operated by the MNO outside the factory, including AMF and SMF as well as AUSF and UDM for authentication towards the core network and Non-Access Stratum (NAS) ciphering and integrity protection, respectively. Regarding the transport network, independent slices with guaranteed levels of isolation and security are used by both the vertical and the MNO, see also Section 8.2.2. In the vertical's small cell layer RAN, on the factory premises, PHY and MAC in the UP and RRC in the CP are shared by both slices. This approach limits the complexity because resource multiplexing is implemented across all network slices on MAC level, forcing each network slice to make use of the same efficient flexible RAN implementation. On the other hand, each network slice may still customize the operation through configuration and parameterization of Radio Link Control (RLC), PDCP, and RRC-Slice functions. RRM for the small cell layer realizes resource allocation according to the defined SLAs for the IoT and eMBB slices of the vertical. Further, the UP (i.e., the UPFs) is realized in a completely local manner if only access to local services is required, for instance provided by enterprise application servers in the factory's edge cloud. If a UE requests a "remote" service, such as a national voice call or Internet access, UPFs are realized on the MNO's infrastructure nodes, comparable to the regular MNO eMBB slice.

Security and Isolation

The vertical's IoT slice is completely isolated from the operator network by using own resources and hosting all functions locally. Transmitted user data and security termination points are strictly kept locally, sensitive subscriber data is maintained by the vertical, and the vertical has full control over the network.

For the eMBB service, the vertical and the MNO have a "roaming" agreement established that assures that vertical eMBB UEs (e.g., smartphones and tablets) that leave the factory premises connect to the MNO macro layer RAN to assure service continuity in the vertical's eMBB slice. In contrast, eMBB UEs subscribed directly to the MNO are continuously served by the macro layer RAN also when entering the factory premises. For vertical eMBB UEs, NAS ciphering and integrity protection is provided by the MNO, while access stratum security is terminated in RAN equipment owned by the vertical. Such a setup requires a minimum level of trust between the MNO and the vertical, since the operator owns the subscriptions including long-term credentials for vertical eMBB UEs. Therefore, the vertical can additionally employ over-the-top security (e.g., based on IPsec or TLS) to protect UP traffic from the MNO. However, this would require additional security functions to be maintained by the vertical (not shown in Figure 8-7).

8.3 Operational Aspects

After having described the E2E view for network slices, we now provide additional details on a number of slice operational aspects such as slice selection, connectivity to multiple slices, inter-slice RRM functions, and the overall management of network slices.

8.3.1 Slice Selection

Slice selection refers to the mechanisms used to identify the NSIs for a UE. The type of network slicing can affect the slice selection. In case of hard slicing, including the transmission and reception points, the initial access procedures can be very similar to legacy networks, such as LTE. Nevertheless, when the RAN supports multiple NSIs and these are sharing the same base stations, the slice selection can also influence the initial UE access procedures.

The subscription of a UE to NSI(s) can be determined via slice identifiers (IDs), such as the configured Single Network Slice Selection Assistance Information (S-NSSAI) stored in the subscription database. The S-NSSAI is used to identify a slice and thus to assist the 5G network in selecting a particular NSI [1]–[3]. It comprises a slice/service type (SST) referring to the expected slice behavior (i.e., features, services, etc.) and a slice differentiator (SD). The SD optionally allows further differentiation for selecting a dedicated NSI from potentially multiple NSIs complying with the indicated SST. The S-NSSAI can have standard values or Public Land Mobile Network (PLMN) specific values. In the latter case, S-NSSAIs are associated to the PLMN identifier of the PLMN that assigns it. The E2E slice selection consists of the selection of the CN [3] and the selection of the RAN part of the NSI [2].

Different alternatives for slice selection can be considered. Firstly, the slice IDs that a UE can be associated to can be configured a priori. In such a case, during the initial access, e.g., an RRC connection request, the slice IDs of the UE can be sent to the RAN. These slice IDs are utilized to determine the RAN policies for the associated NSIs as well as the selection of the CN part of the NSIs, such as the slice-specific CP NFs at the CN. If the slice IDs are not configured a priori, the UE subscriber information can be retrieved by the default NSI at the CN, e.g., by the CCNF, and the slice IDs of the UE will then be configured by the slice-specific CP NF over NAS messaging. Alternatively, the information about the available slices can be broadcasted in order to save signaling exchange and time during the attach procedure [27]. The penalty one has to pay for such an approach is the waste of radio resources used for broadcasting this information.

Providing the slice IDs by a UE during the initial access procedures can have the advantage of applying slice-specific RAN policies right away. This can be particularly advantageous in case of mission-critical NSIs with strict latency requirements. Yet, for non-mission-critical NSIs, determining slice IDs from the UE subscription information via the default NSI can reduce the signaling overhead during initial access procedures.

8.3.2 Connecting to Multiple Slices

Simple devices, such as sensors within the mMTC framework, will typically be associated to a single slice. For more complex devices, due to the mixed service needs, multiple slice associations can be realized. A good example for a multiple-slice UE are vehicles that will require multiple 5G services. For example, these may require both infotainment services, which are related to eMBB, and services for autonomous driving, which are mission-critical and thus related to URLLC. Another example that indicates the need for multi-slice connectivity is provided in Section 8.2.5. Some devices like tablets or smartphones may run factory-related applications that require to have access only to the local slice inside a factory over secure links, while other applications may require typical access to the Internet.

If slices are logically separated, a device will have to somehow be associated with all of them. Having UEs being associated to multiple networks may increase their complexity as well as the overall signaling considerably. When the same NFs (e.g., RRC or mobility management) are implemented for multiple slices, signaling overhead and UE complexity can be substantially reduced. However, the customization level per NSI is reduced in this case. For instance, mobility management procedures can be significantly different for an eMBB slice as compared to a URLLC slice, as discussed in Section 8.2.3. ~~It is worth noting that in NR Release 14 [2], one signaling connection is foreseen on the RAN side, while a UE can access to multiple slices simultaneously. Yet, the details of implementation options are to be analyzed in the normative phase of NR in Release 15.~~

In case a UE has simultaneous access to multiple slices, and depending on the use case requirements, it is possible to consider a context transfer from one slice NF to another. For example, the information about the current location of a vehicle in the URLLC slice may be transferred to the corresponding entity (e.g., the AMF) in the eMBB slice, thus reducing the signaling overhead that is required during a location update process. Similarly, subscription information related to a smartphone operating inside a factory can be transferred from the eMBB slice to the local URLLC slice minimizing the need for the local network operator to manage duplicated information already available in the typical operator's NFs.

Finally, the plethora of available network slices will require some adaptability from the UE side, especially for general purpose devices such as smartphones, tablets or laptops. This indicates that UEs will have to be to some extent open for programmability and reconfigurability to meet the slicing needs. Such abilities will constitute the UEs to be part of an E2E slice. However, further investigations are needed to clarify how UEs can be flexibly adapted to different slices without the need of extended operating system updates.

8.3.3 Slice Isolation

Network slicing targets the facilitation of different businesses that use the same infrastructure but possibly have diverse requirements. The deployment of multiple slices over the same infrastructure inside the network of one operator should enable the reuse of resources such as physical or software resources, which brings very big benefits compared to a hard splitting of the resources among slices. In [28], examples are provided that demonstrate that hard splitting of resources among slices can lower significantly the overall (busy hour) capacity of the network.

On the other hand, specific slice instances should be protected from the performance of other slices deployed over the same infrastructure. This feature is called *slice isolation* and relates to the ability of the network to minimize negative inter-slice effects. In the RAN, 3GPP already supports a partial protection of certain services under an extensive load of others through mechanisms such as Access Class Barring and extensions of such schemes [29], but such solutions cannot ensure the serviceability of one slice instance under excessive load of other instances.

Especially in case of slice instances that share radio resources such as common radio channels or common mobility management functions, congestion in one slice should not have a negative impact on another slice instance. A simple example of such case could be two slices that share the same Random Access CHannel (RACH) for initial access and also the same preambles. In this case, if one slice is overloaded, it will have a tremendous negative effect on the performance of the other, due to the large number of collisions.

In general, slice isolation may be achieved by:

- **Horizontal separation of resources**, based on the separation of physical resources for the different slices. This approach may lead to inefficient resource usage as explained above;
- **Efficient scheduling/coordination mechanisms**, where slice-overarching functions such as scheduling functions, QoS schemes and initial access mechanisms ensure that the service requirements of each slice are met. In any case, a soft separation of resources is required so as to prioritize certain slice instances over others.

The isolation of slice instances may also be indicated by regulations which necessitate the different treatment of certain slices. One potential example for such case could be autonomous driving slice instance(s) which may need to have their own spectrum resources. Moreover, another requirement for separating slices is security. As explained in the example of Section 8.2.5, specialized local networks such as factories or medical facilities require that even the serving operator will not have any access to local data of sensitive nature. For this reason, it is considered that even some of the core NFs will have to be re-deployed in the local network. Some context transfer may take place, for instance related to subscription information, but only from the operator network towards the local network.

8.3.4 Radio Resource Management Among Slices

With respect to the RAN, the management of the scarce radio resources is a critical issue. Thus, pooling and sharing these resources among network slices in an efficient manner is an important target. The RRM is also responsible for allocating the resources in a way that the SLAs of all network slices are fulfilled.

The basis for allocating resources in a slice-aware manner is to monitor the status of the network slices with respect to their SLAs. This could take place in a new entity of the RAN, e.g., an access controller, which has to be aware of the existing network slices and their SLAs, as well as which data stream belongs to which network slice. Corresponding information can be obtained via signaling from the CN. The enforcement of the network slice specific requirements can be realized with different levels of complexity as shown in Figure 8-8. On the right side of the figure, a lightweight implementation of multi-slice resource management is depicted. Based on the outcome of the SLA monitoring, the QoS Class Identifiers (QCIs) of the individual data streams are adjusted and traffic steering is executed. An enforcement of SLAs happens by adapting the QoS classes of individual data streams. If, for example, the SLA of a network slice guaranties a certain latency, any data stream of this slice could be mapped to a corresponding QoS class. In the case of 5G NR, this functionality can

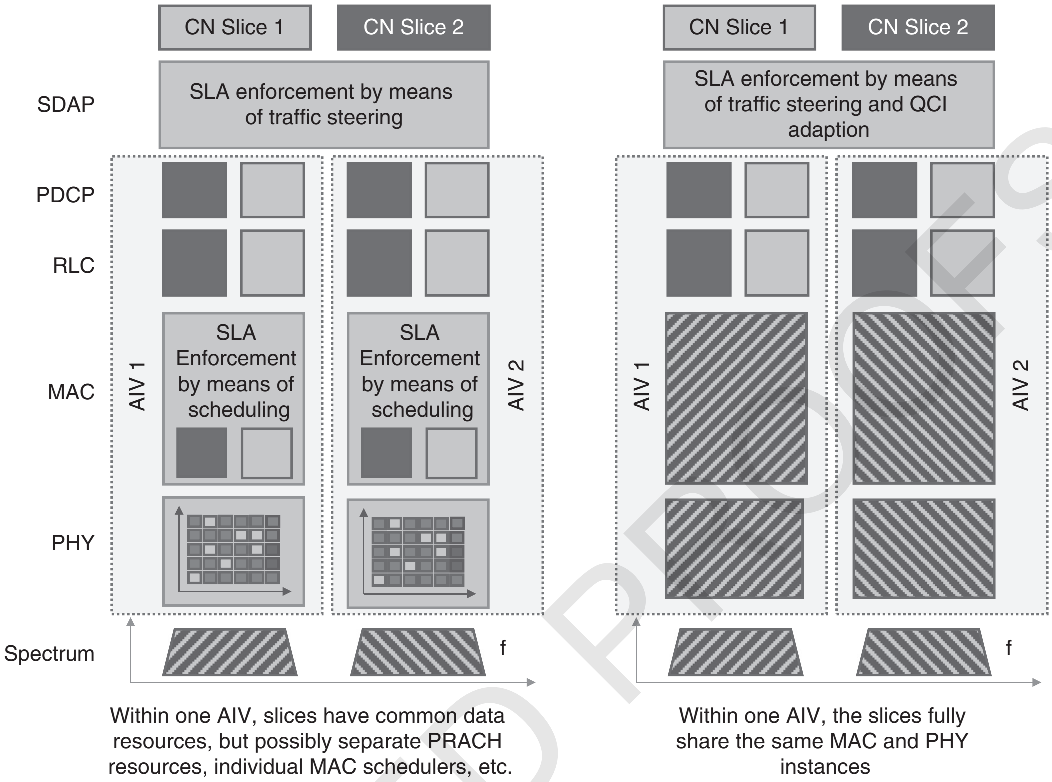

CN Slice 1
CN Slice 2
SDAP
SLA enforcement by means of traffic steering
PDCP
RLC
MAC
PHY
AIV 1
AIV 2
SLA Enforcement by means of scheduling
SLA Enforcement by means of scheduling
Spectrum
f
Within one AIV, slices have common data resources, but possibly separate PRACH resources, individual MAC schedulers, etc.
CN Slice 1
CN Slice 2
SLA enforcement by means of traffic steering and QCI adaption
AIV 1
AIV 2
f
Within one AIV, the slices fully share the same MAC and PHY instances

Figure 8-8. Implementation options for multi-slice resource management.

be part of the new SDAP sublayer. In this implementation, the individual AIVs can operate in a slice-agnostic way, but have to fulfil the QoS defined.

On the left side of the figure, the functionality of SDAP is enhanced by a slice-aware real-time RRM that also performs the scheduling in a slice-aware manner based on the status of the SLA monitoring. The QoS mapping or the slice-aware scheduling, respectively, is a dynamic process, which is supposed to solve conflicts between network slices in a way that all SLAs can be fulfilled. More details on multi-slice resource management can be found in Section 12.6.

8.3.5 Managing Network Slices

Current business support systems (BSS) and operating support systems (OSS) of communication service providers (CSPs) expose service management capabilities to customers. In contrast, network management and infrastructure management functions, such as the 3GPP Network Management System (NMS) and Element Management System (EMS) are typically not exposed to customers. With network slicing, CSPs will need to extend the current service management offers to a certain level of management exposures for both network and infrastructure. The extent of such exposure depends on the level of expertise of the network slice instance customer and the CSP's readiness to open internal systems to tenants. Generally, three levels of exposure can be differentiated: In the *“monitoring”* option, the CSP operates the network slices on behalf of the tenant (e.g., an OTT application provider) and only provides slice-specific KPIs. The *“limited control”* option gives the tenant (e.g., a vertical industry enterprise) the possibility to (re-) configure selected parameters of NFs associated to network slices. In the *“extended control”* option, the tenant (e.g., a virtual CSP) can rather independently operate its own network slices and use own management systems. Based on these constraints, the following key questions and challenges related to network slicing management have been identified and will be investigated in upcoming activities in both the research community and standardization processes:

- Design of network slices to host communication services supported by the infrastructure: How can the specified service requirements be supported by a network slice instance?
- Network slice instance management: How can FCAPS management be used for network slicing management?
- Conflict resolution: How can conflicts from policies created by different service requirements be resolved?
- Orchestration and lifecycle management of network slices: What are the different compositions of a network slice and how are they orchestrated?
- Multi-domain network slice orchestration: How can a slice be created and deployed across multiple administrative domains?
- Automation for network slice management: How can, e.g., evolved self-organizing network (SON) concepts and cognition be applied to network slicing management?
- Shared network slice instance management: How shall a network slice instance (or parts of it) be shared across multiple services?

The following sub-sections will detail a subset of the listed challenges and sketch potential solutions.

8.3.5.1 Managed Objects and Network Slice Instance Lifecycle

Next Generation Mobile Networks (NGMN) [4] has introduced the concepts of ‘Network Slice Blueprint’ and ‘Network Slice Instance’ which have largely been taken over by 3GPP as ‘network slice template’ and ‘network slice instance’, respectively [30]. Instance-specific policies and configurations are required when creating a network slice instance from network slice templates. A network slice is composed of one or multiple ‘network slice subnets’ which in turn contain one or multiple physical or virtualized network functions.

The lifecycle management of these NFs (both CNFs and RNFs) comprises both 3GPP domain-specific FCAPS management as well as domain-agnostic lifecycle management and orchestration. Regarding the lifecycle of a network slice instance, three distinct phases have been defined [30], as depicted in Figure 8-9: (A) commissioning phase, (B) run-time phase, and (C) decommissioning phase. A so-called Network Slice Management Function (NSMF) oversees the respective tasks of each phase, and is detailed in the following sub-section.

8.3.5.2 Network Slice Management Function (NSMF)

The NSMF [30] is responsible for managing the lifecycle of a network slice instance. Provided that the preparatory tasks such as network slice design, network slice pre-provisioning, template on-boarding and the general preparation of the network environment have been completed, the NSMF is ready to process incoming requests for communication services. It does so by selecting a network slice template that can provide the agreed service requirements including the required levels of isolation, security, and management exposure, as shown in (Figure 8-10).

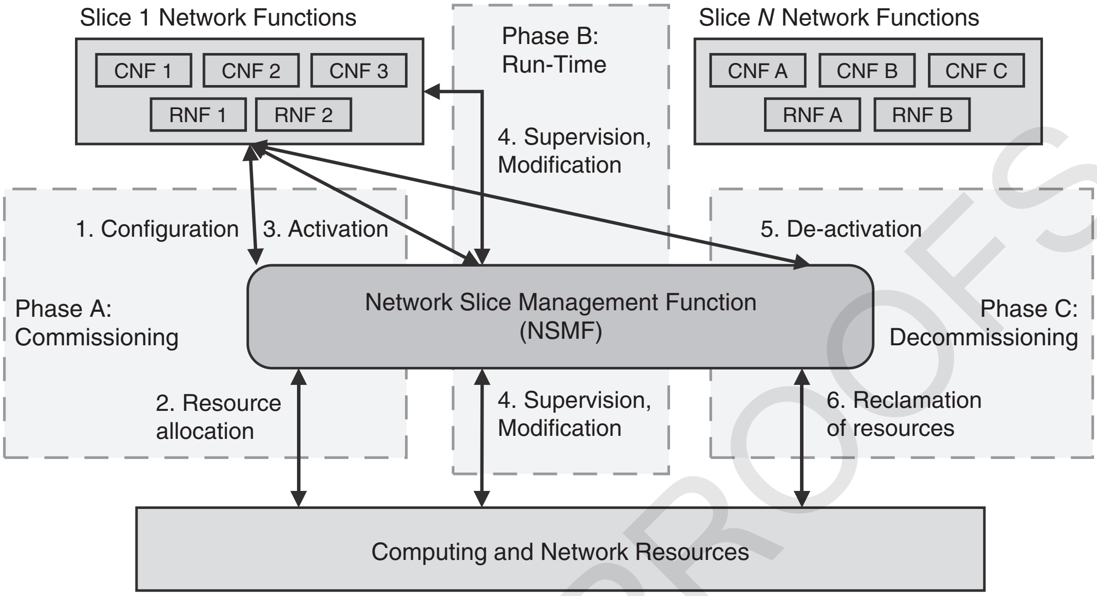

Slice 1 Network Functions
CNF 1
CNF 2
CNF 3
RNF 1
RNF 2
Phase B: Run-Time
Slice N Network Functions
CNF A
CNF B
CNF C
RNF A
RNF B
4. Supervision, Modification
1. Configuration
3. Activation
5. De-activation
Phase A: Commissioning
Network Slice Management Function (NSMF)
Phase C: Decommissioning
2. Resource allocation
4. Supervision, Modification
6. Reclamation of resources
Computing and Network Resources

Figure 8-9. Phases of network slice lifecycle management [30].

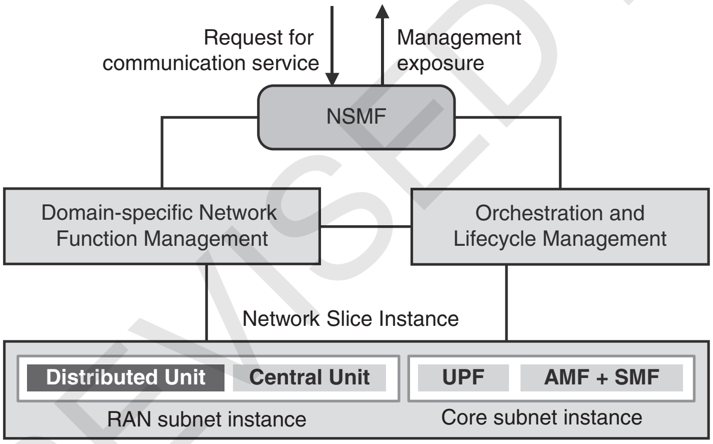


Figure 8-10. Domain-specific FCAPS management and lifecycle management of a network slice.

The NSMF commissions a network slice instance consisting of shared and/or slice-specific RAN and CN functions. Knowing which network slice subnet instances (NSSI) are associated with each network slice instance, the NSMF determines to completely reuse an existing (operating) NSI or create a new NSI. For the latter case, the NSMF instantiates and configures slice-specific VNFs and subnet instances, and re-configures already operating and shared NFs and NSSIs. Subsequently, the NSI is activated by activating all necessary CN, RAN and other functions.

For performance supervision during run-time, the NSMF creates performance management jobs for the NFs in each NSSI to generate performance data of a network slice instance and monitor thresholds for selected performance parameters. The data is collected and provided to the NSMF by the respective management functions of the NFs, e.g., the EMS. The NSMF compiles and monitors faults and performance on the level of a network slice instance and ensures that agreed service requirements are met, including generating the management data separately for different customers. For this purpose, the NSMF receives alarm notifications for slice instances, both from the shared and the dedicated NSSIs. The NSMF also triggers necessary upgrade, reconfiguration and scaling actions. For decommissioning of a network slice instance, the NSMF requests deactivation and subsequent termination of NSS instances at the respective FCAPS and lifecycle management functions.

8.3.5.3 Shared Network Functions and Automation of Network Slice Management

Network slices are instantiated from a common underlying infrastructure, resulting in the need for rule-based allocation of hardware, software, and radio resources and functions. Available sharing technologies include virtualization, but also more established techniques such as multi-tasking and multiplexing. Such sharing rules define how resource commitment schemes (e.g., static allocation, dynamic demand-based allocation) are to be applied and how resource requests are prioritized if demand exceeds available capacity. They are derived from management policies which are maintained by the resource owner. Management functions such as the EMS and NMS, together with other NFs such as radio schedulers, apply and enforce these rules in order to fulfil the SLAs associated to each network slice. On one hand, such cross-slice interdependencies considerably increase the complexity of management and orchestration tasks. On the other hand, the massive number of managed objects also significantly raises the requirement on the scalability of network management procedures. 5G networks therefore require a higher level of management automation. Here, cognition and more autonomous decision making processes constitute key enablers, as elaborated in detail in Section 10.7.

8.4 Summary and Outlook

Slicing is one of the key characteristics of 5G networks. Its main goal is to support the diverse 5G use cases and their requirements in a very flexible way, and to run them cost-efficiently over a common network infrastructure. Slicing allows the selection of NFs from a pool, their configuration, synthesis and deployment to form logical networks. Previous 3GPP releases have attempted to introduce network customization mainly in the core network. Compared to these attempts, 5G network slicing is not bound to a single logical architecture, as it was the case for 4G, and new levels of flexibility also appear in the transport and access domains. Obviously, such flexibility introduces some new architectural issues that need to be solved, such as slice selection, computation and network resources virtualization and sharing among slices, slice isolation, and support for end devices that are able to connect to multiple slices. Moreover, the 5G architecture will enable the bi-directional communication of NFs with application functions provided by the content and application service providers. This will create the ability for further tailor-cut solutions for these providers, improving multi-tenancy support, but also introducing the need to properly define the required northbound interfaces for programmable interaction with the tenant.

Although the standardization community has provided the first version of specifications for the 5G release, several issues remain open for future releases. Thus, network slicing is far from being a thoroughly studied feature, and further work on this topic is expected during the next years. In Table 8-1, the key points analyzed in the previous sections are grouped and summarized, and remaining open issues are listed.

Table 8-1. Key points and main open issues related to the support of network slicing.

Technical Area	Key Points	Main Open Issues
Core Network	● Decomposition of CN NFs ● Further separation of control and user plane ● Common or slice-specific NFs (which will belong to which category, how will this affect issues like security) ● Service-based interfaces ● Service exposure to 3rd parties	● Clear categorization is needed between common or slice-specific NFs. Impact on slice isolation ● Find a balance point between complexity and flexibility that NF modularization introduces ● Introduce slice policy conflict resolution and slice prioritization that may occur via the exposure of NFs to 3rd parties
Transport Network	● Convergence of highly heterogeneous transport technologies ● More efficient per-slice QoS support and on demand resource allocation, adapted to dynamic workloads	● Identify potential standardization of open interfaces to support convergence in a multi-vendor environment ● Identify the need for potential enforcement of policies and slice prioritization in the transport domain
Radio Access Network	● Flexibility to allow full separation of resources among slices or their sharing ● NFs on different radio protocol stack layers should be configurable to support 5G use cases and allow flexible placement on physical or logical network nodes (e.g., central or distributed units)	● Identify the common functions among slices ● Allow different specification and implementation of NFs without increasing the complexity of the specification ● Clarify the impact of flexible NF placement on the transport domain (interfaces, latency/bandwidth requirements, etc.)
Multi-operator Slicing	● Slice support over multiple network providers through the introduction of appropriate orchestrators and their interworking ● Limited exposure of available resources to other domains	● Appropriate models and interfaces need to be standardized
User Equipment	● Slice selection ● Concurrent connectivity to multiple slices ● Openness of UEs for programmability to slicing needs (UEs as part of the E2E slice)	● Identify how to minimize unnecessary signaling and UE complexity ● Clarify how UEs can be flexibly adapted to varying needs of slicing use cases without the need of extended operating system updates
Slice Management	● Dynamically translate customer needs and instantiate slice instances ● Improve network management to support a multi-slice environment (sharing of resources, functions etc.)	● Standardize a northbound interface between NMS and 3rd party application and services ● Introduce schemes for conflict resolution among slice specific SON functions that operate over shared resources

References

1 3GPP TS 23.501, “System Architecture for the 5G System; Stage 2 (Release 15)”, Version 15.0.0, December 2017

2 3GPP TR 38.801, “Study on New Radio Access Technology; Radio Access Architecture and Interfaces (Release 14)”, Version 14.0.0, March 2017

3 3GPP TR 23.799, “Study on New Radio Access Technology; Radio Access Architecture and Interfaces (Release 14)”, Version 14.0.0, December 2016

4 NGMN Alliance, “Description of Network Slicing Concept”, January 2016

5 5G PPP 5G NORMA project, Deliverable D3.3, “5G NORMA network architecture – final report”, Oct. 2017, see: https://doi.org/10.5281/zenodo.1120246

6 ETSI, Industry Specification Group (ISG) Network Functions Virtualization (NFV), see http://www.etsi.org/technologies-clusters/technologies/nfv

7 B. Blanco et al., “Technology pillars in the architecture of future 5G mobile networks: NFV, MEC and SDN”, Computer Standards and Interfaces, January 2017

8 M. Richart J. Baliosian, J. Serrat and J-L Gorricho, “Resource Slicing in Virtual Wireless Networks: A Survey”, IEEE Transactions of Network and Service Management, vol. 13, no. 3, Sept. 2016

9 C. Liang and F.R. Wu, “Wireless Network Virtualization: A Survey, some research issues and challenges”, IEEE Communication Surveys and Tutorials, vol 17, no. 1, Q1 2015

10 NGMN Alliance, White Paper, “5G White Paper”, Version 1.0, March 2015

11 5G PPP Architecture Working Group, White Paper “View on 5G Architecture”, Version 1.0, July 2016

12 3GPP TS 23.251, “Network Sharing; Architecture and functional description (Release 14)”, Version 14.0.0, March 2017

13 3GPP TR 23.707, “Architecture Enhancements for Dedicated Core Networks; Stage 2 (Release 13)”, Version 14.0.0, Dec. 2014

14 3GPP TR 23.711, “Enhancements of Dedicated Core Networks selection mechanism (Release 14)”, Version 14.0.0, Sept. 2016

15 ETSI ISG Mobile Edge Computing (MEC), see http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing

16 3GPP TS 23.214, “Architecture enhancements for control and user plane separation of EPC nodes; Stage 2 (Release 14)”, Version 14.3.0, June 2017

17 3GPP TS 23.502, “Procedures for the 5G System; Stage 2 (Release 15)”, Version 15.0.0, December 2017

18 3GPP TR 33.899, “Study on the security aspects of the next generation system”, Version 1.2.0, June 2017

19 Anna Tzanakaki et al, “Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services”, IEEE Communications Magazine (to appear)

20 3GPP TS 38.300 “NR and NG-RAN Overall Desciption”, Release 15, Version 1.2.0, May 2017

21 Adlen Ksentini and Navid Nikaein, “Toward enforcing network slicing on RAN: Flexibility and resources abstraction”, IEEE Communications Magazine, vol. 55, no. 6, June 2017

22 5G PPP METIS-II project, White Paper, “Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design”, March 2016

23 P. Marsch et al., “5G radio access network architecture: design guidelines and key considerations”, IEEE Communications Magazine, vol. 54, no. 11, pp. 24–32, Nov. 2016

24 3GPP TR 38.912, "Study on New Radio (NR) access technology", Release 14, Version 14.0.0, March 2017

25 5G PPP 5GEx project, White paper, "5GEx Multi-domain Service Creation - from 90 days to 90 minutes", March 2016

26 5G PPP 5GEx project, Deliverable D3.1, "Description of protocol and component design", July 2016

27 X. An, C. Zhou, R. Trivisonno, R. Guerzoni, A. Kaloxylos, D. Soldani and A. Hecker, "On end to end network slicing for 5G", Transactions on Emerging Telecommunications Technologies, Wiley, June 2016

28 I. da Silva et al., "Impact of network slicing on 5G Radio Access Networks", European Conference on Networks and Communications (EuCNC 2016), June 2016

29 3GPP TS 22.011, "Service Accessibility (Release 15)", Version 15.1.0, June 2017

30 3GPP TR 28.801, "Study on management and orchestration of network slicing for next generation networks", Release 15, Version 15.0.0, September 2017