E2E Network Slicing Requirements
Version 4.0
10 July 2024

Security Classification: Non-Confidential

Copyright Notice

Disclaimer

Compliance Notice

The information contain herein is in full compliance with the GSMA Antitrust Compliance Policy.

This Permanent Reference Document is classified by GSMA as an Industry Specification, as such it has been developed and is maintained by GSMA in accordance with the provisions set out GSMA AA.35 - Procedures for Industry Specifications.

Table of Contents

1 Introduction

1.1 Overview

Network slicing is a concept of running multiple logically customized networks on a common infrastructure, for different customers in different industries and for different required functions, based on an agreed Service Level Agreement (SLA). Network slicing was outlined in the NGMN 5G White Paper [1] as a vision of 5G capabilities that drive value creation. After that, it has become one of the key capabilities specified by 3GPP to be supported in 5G systems (5GS) [2]. Network slicing allows mobile operators in the 5G era to provide customized network services to their customers using a common network infrastructure. It is expected to have great commercial potential for the operators.

However, what should be taken into account is that, from the customer's perspective, the technical requirements and specifications are specified not only for mobile networks, as specified in 3GPP, but also for transport networks, management systems, and (user) devices, i.e., from an End-to-End (E2E) perspective. These network and technology domains must be well coordinated.

The GSMA has published a white paper on this topic [3]. It describes the network slicing architectural blueprint designed from an E2E perspective, spanning different technology domains (e.g., devices, access networks, core networks, transport networks, and network management systems). In addition, this whitepaper describes the technical aspects required for network slicing, and the gaps between the ongoing work and what is expected to be achieved by external organizations such as Standards Development Organizations (SDOs) and fora, as a snapshot of the current state of the art per 3GPP Release 17.

The previous work is continued in this Permanent Reference Document (PRD). This document proposes technical requirements for E2E network slicing. In addition to technical requirements, it also maps the necessary specifications to specific external organizations to achieve the requirements, and identifies gaps between this PRD and these specifications, for information. This mapping creates the potential for liaising between GSMA and external organizations asking them to fill the gaps.

1.2 Scope

This document proposes technical requirements for E2E network slicing with the aim to align all technologies and their specifications for the correct creation of E2E Network Slices (NSs). The document proposes network slicing requirements for operators from an E2E perspective and relates different architectures.

This document covers the following areas:

- Technical requirements for E2E network slicing
- Architectures, functions, and roles
    - Reference architectures envisioned for E2E network slicing
- Standardization and external organizations such as SDOs and forums, specifically:

- Gap analysis of standards: This PRD includes an analysis of the gaps in current standards and identifies the SDOs that are suitable to complete the E2E network slicing architecture.
- Review SDO progress and submit liaison statements to ensure that the E2E system is defined consistently across these organizations.

Technical requirements mentioned in this PRD includes requirements to enable Network Slice as a Service (NSaaS), which is the ultimate goal of operators as mentioned in the GSMA white paper [3].

1.3 Definitions

Term	Description
E2E network slicing	Slicing concept for mobile network which include UE, RAN, CORE and Transport.

1.4 Abbreviations

Term	Description
3GPP	3rd Generation Partnership Project
5GS	5G System
AMF	Access and Mobility Management Function
AN	Access Network
API	Application Programming Interface
B2B2X	Business to Business to everything
CN	Core Network
CNF	Containerized Network Function
CU	Central Unit
DNN	Data Network Name
DRB	Data Radio Bearer
DU	Distributed Unit
E2E	End-to-End
E2EO	End-to-End Orchestrator
GA	General Availability
GBR	Guaranteed Bit Rate
GST	Generic network Slice Template
IaaS	Infrastructure as a Service
IMS	IP Multimedia Subsystem
KPI	Key Performance Indicator
MARO	Multi-Aspect Resource Optimization

Term	Description
MDT	Minimization of Drive Tests
MNO	Mobile Network Operator
NF	Network Function
NFV	Network Function Virtualization
NFVO	Network Function Virtualization Orchestration
NRM	Network Resource Model
NS	Network Slice
NSaaS	Network Slice as a Service
NSC	Network Slice Customer
NSMS_C	Network Slice Management Service Consumer
NSMS_P	Network Slice Management Service Provider
NSP	Network Slice Provider
NSSAI	Network Slice Selection Assistance Information
OAM	Operation Administration and Maintenance
OSS	Operation Support System
PaaS	Platform as a Service
PDU	Protocol Data Unit
PoC	Proof of Concept
PRB	Physical Resource Block
PRD	Permanent Reference Document
RAN	Radio Access Network
RCS	Rich Communication Service
RRC	Radio Resources Control
SDN	Software Defined Networking
SDO	Standards Development Organization
SLA	Service Level Agreement
SLS	Service Level Specification
SMS	Short Message Service
TN	Transport Network
UE	User Equipment
URLLC	Ultra-Reliable and Low Latency Communications
URSP	UE Route Selection Policy
VNF	Virtualized Network Function

1.5 References

Ref	Doc Number	Title
[1]	NGMN 5G White Paper (2015)	NGMN 5G White Paper (2015) https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf
[2]	3GPP, TS23.501	System architecture for the 5G System (5GS) https://www.3gpp.org/DynaReport/23501.htm
[3]	GSMA, NG.127	E2E Network Slicing Architecture Version 1.0, June 2021. https://www.gsma.com/newsroom/wp-content/uploads//NG.127-v1.0-2.pdf
[4]	RFC 2119	"Key words for use in RFCs to Indicate Requirement Levels", S. Bradner, March 1997. http://www.ietf.org/rfc/rfc2119.txt
[5]	RFC 8174	Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words https://www.rfc-editor.org/info/rfc8174
[6]	GSMA NG.116	Generic Network Slice Template. Version 6.0, November 2021 https://www.gsma.com/newsroom/wp-content/uploads/NG.116-v6.0.pdf
[7]	ETSI GS NFV-IFA 013	Network Functions Virtualisation (NFV) Release 4; Management and Orchestration; Os-Ma-nfvo reference point Interface and Information Model Specification https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/013/04.03.01_60/gs_NFV-IFA013v040301p.pdf
[8]	3GPP TS28.530	Management and orchestration; Concepts, use cases and requirements https://www.3gpp.org/DynaReport/28530.htm
[9]	Sensors	Wichary, T.; Mongay Batalla, J.; Mavromoustakis, C.X.; Żurek, J.; Mastorakis, G. Network Slicing Security Controls and Assurance for Verticals. Electronics 2022, 11, 222. DOI: 10.3390/electronics11020222. https://www.mdpi.com/2079-9292/11/2/222
[10]	GSMA, White Paper	An Introduction to Network Slicing. White Paper 2017. https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf
[11]	NGMN Alliance	Security Aspects of Network Capabilities Exposure in 5G v1.0. White Paper 2018. https://ngmn.org/wp-content/uploads/Publications/2018/180921_NGMN-NCEsec_white_paper_v1.0.pdf
[12]	3GPP TS22.261	Service requirements for the 5G system https://www.3gpp.org/DynaReport/22261.htm
[13]	3GPP TS28.531	Management and orchestration; Provisioning https://www.3gpp.org/DynaReport/28531.htm
[14]	3GPP TS28.550	Management and orchestration; Performance assurance https://www.3gpp.org/DynaReport/28550.htm
[15]	3GPP TS28.545	Management and orchestration; Fault Supervision (FS) https://www.3gpp.org/DynaReport/28545.htm
[16]	3GPP TS23.222	Common API Framework for 3GPP Northbound APIs

Ref	Doc Number	Title
		https://www.3gpp.org/DynaReport/23222.htm
[17]	Sensors	Ordonez-Lucena, J.; Ameigeiras, P.; Contreras, L.M.; Folgueira, J.; López, D.R. On the Rollout of Network Slicing in Carrier Networks: A Technology Radar. Sensors 2021, 21, 8094. DOI: 10.3390/s21238094. https://www.mdpi.com/1424-8220/21/23/8094
[18]	HEXA-X project	https://hexa-x.eu
[19]	GSMA, White Paper	Network slicing Use case Requirements, 2018. https://www.gsma.com/futurenetworks/wp-content/uploads/2018/07/Network-Slicing-Use-Case-Requirements-fixed.pdf
[20]	3GPP TS28.533	Management and orchestration; Architecture framework https://www.3gpp.org/DynaReport/28533.htm
[21]	3GPP TS28.541	Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3 https://www.3gpp.org/DynaReport/28541.htm
[22]	3GPP TS24.526	Telecommunication management; Study on the Self-Organizing Networks (SON) for 5G networks https://www.3gpp.org/DynaReport/24526.htm
[23]	3GPP TR 28.861	Study on the Self-Organizing Networks (SON) for 5G networks https://www.3gpp.org/DynaReport/28861.htm
[24]	GSMA NG.141	Guidelines for URSP https://www.gsma.com/newsroom/wp-content/uploads//NG.141-v1.0-8.pdf

1.6 Conventions

"The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [4] and clarified by RFC8174 [5], when, and only when, they appear in all capitals, as shown here.

2 High-level Architecture

2.1 High-level Reference Architecture

This PRD's primary goal is to provide technical requirements for E2E network slicing to align all technologies and their specifications for the correct creation of E2E NSs. For communication service customers, it will enable comprehensive network slicing services.

Considering an E2E perspective, multiple actors such as network operators and service providers may also need to interact with each other to provide E2E communication services and to share resources. For this reason, a common way of the interworking is particularly important to be specified so as to be consistent with standard specifications in each technical domain.

In order to satisfy the above, the E2E reference architecture in this PRD is based on the architecture illustrated in GSMA Whitepaper NG.127 [3]. Figure 1 shows the architecture in GSMA Whitepaper NG.127, which is a high-level architecture with three different stratums - infrastructure stratum, network and application function stratum, and O&M stratum. The detailed architecture specification of technology domains in each stratum is defined by different SDOs and fora, per GSMA Whitepaper NG. 127. As a general principle, it is important to fit within an established ecosystem. Therefore, existing and established standards should be reused as much as possible.

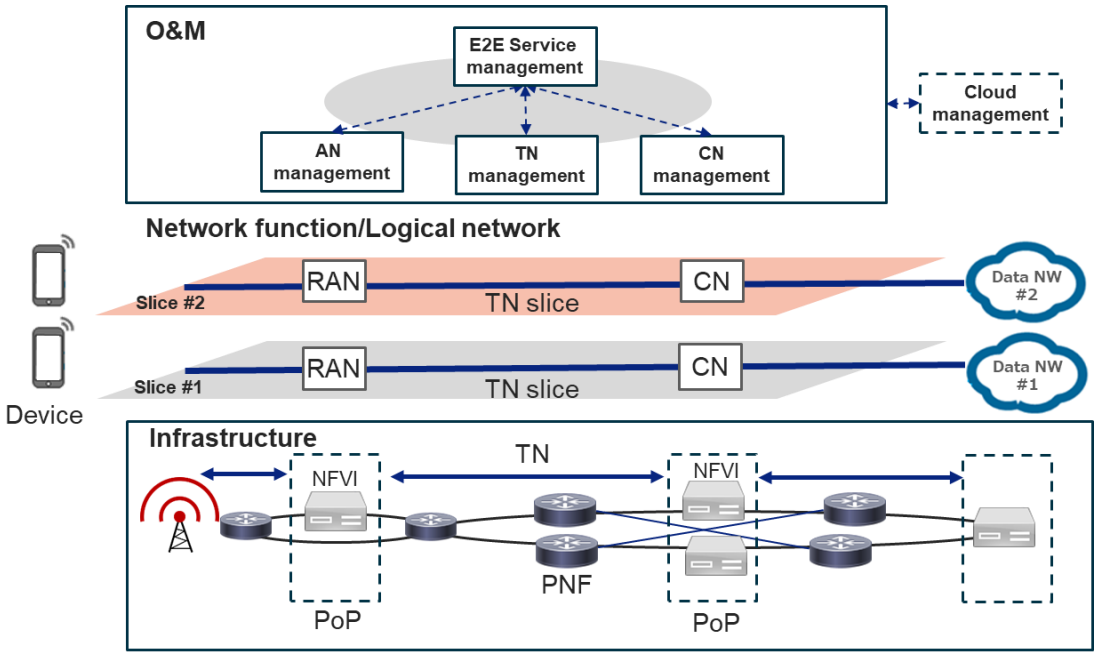

Figure 1 Overall Network Slicing Architecture example for mobile network [3]

3 Technical Requirements

3.1 General

Technical requirements relating to E2E network slicing are addressed in this PRD, while GSMA Whitepaper NG.127 [3] indicates technical aspects expected to be achieved by E2E network slicing to guide the requirements. In addition to those requirements addressed in this PRD, there are other network slicing related requirements for each domain specified by other SDOs and fora.

The following sub-sections state the technical requirements to support E2E network slicing.

3.1.1 Co-operation of application layer and transport layer

3.1.1.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to co-operation of application layer and transport layer:

1. Radio Access Network (RAN), Core Network (CN) and Transport Network (TN) should co-operate, to fulfil SLA for E2E network slicing.
2. TN, as well as RAN, and CN should identify and differentiate a traffic for certain slice in order to fulfil SLA for E2E network slicing.
3. All factors in E2E network slicing should co-operate to achieve an SLA for E2E network slicing.

3.1.1.2 Description

At this moment, RAN and CN already share common identifier, Network Slice Selection Assistance Information (NSSAI), which is standardized at 3GPP. But RAN and CN should also share the identifier with TN.

There are multiple possible technical solutions to co-operate

Solution 1: Automation of an identifier(s) translation between RAN and CN and TN.
Solution 2: Embedding S-NSSAI used in RAN and CN, into identifier used by TN
Solution 3: Depend on operator's deployment.

Example of Solution 1 could be a VLAN identifier used at the boundary of RAN and TN, and at the boundary of TN and CN. This VLAN<->S-NSSAI mapping, tracking that VLAN 'X' corresponds to S-NSSAI 'Y' at particular RAN/CN to TN handoff point, must be automated/coordinated.

Example of Solution 2 could be S-NSSAI (32 bits) used as part of (embedded into) TN ID (e.g. IPv6 address 128 bits) used in TN. With this solution, no assignment coordination (no mapping table between S-NSSAI and TN ID) would be required between e.g. RAN-NSSMF and TN-NSSMF, as the S-NSSAI is used across two subnets (embedding S-NSSAI into TN specific - here example IPv6 address - ID).

One of those solutions is required to achieve SLA which is assigned to E2E network slicing.

3.1.2 SLA assurance in E2E

3.1.2.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to the SLA assurance in E2E:

1. The E2E network slicing shall be able to maintain high quality of E2E network services, including network and applications based on customer requests.
2. The E2E network slicing shall be able to provide E2E service for Business to Business to everything (B2B2X) partners with coordinated network and applications.
3. The E2E network slicing shall be able to do orchestration to achieve E2E alignment in order to reflect the slice-specific requirements of network domains based on the E2E SLA.
4. The E2E network slicing shall be able to reserve appropriate amounts of resources (e.g., radio rand compute resources) and to deploy network functions at appropriate places in all network domains.

3.1.2.2 Description

A customized network requires performance requirements to be a fundamental concept for network slicing. An SLA is a commitment of provisioned network services between an operator and a consumer. The consumer declares communication service(s) requirements to the operator. Network performance attributes such as throughput, latency and reliability could be part of the technical specification of an SLA. GSMA PRD NG.116 [6] specifies such attributes.

3.1.3 Deterministic and operator-controlled UE behaviour

3.1.3.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirement related to deterministic and operator-controlled User Equipment (UE) behaviour:

1. The 3GPP system should enable the network control of the UE behavior to address the issues described below. The UE should receive from the network the intended control information.

3.1.3.2 Description

Currently, the network can provide the UE with configuration of the NSs the UE can use. In addition, URSP can offer the operator the opportunity to associate applications to connectivity provided by a NS and the Data Network Names (DNNs) that are supported in the NS.

However, it is for UE implementation on how it decides when to register with the NSs (it may do so based on the configured NSs list, or trigger registration only at detection of the need to use a NS). Likewise, it can deregister with a NS or tear down a Protocol Data Unit (PDU) session that are unused at a time of its discretion. There is therefore a need for the network to improve the control it has on the UE behaviour.

Mobile Network Operators (MNOs) may want to enforce the registration to NSs only when the need to use them is detected (e.g., to minimize the number of registered users, or to enable the choice of the optimal Access and Mobility Management Function (AMF) (used for the NSs) or based on configuration (e.g., to reduce the times a UE register and deregisters with a NS). This degree of network control is missing and also it is not possible to specify the time a UE should de-register from a NS after the last PDU session stops using it (i.e., after the last PDU session established in the NS by the UE is torn down).

Similarly, MNOs may want certain PDU sessions to be established at all times and some instead to be just established when usage is detected/requested by applications. This degree of control is missing, and also it is not possible to specify the time a UE should release a PDU session after the applications that trigger its establishment stop using it (i.e., all the applications that were bound to the connectivity are shut down or do not require a specific connectivity). Given that it is a PDU Session that make use of resources, keeping a PDU session established for longer than needed can cause unnecessary resource shortage. On the other hand, tearing a PDU session down too soon may cause more Session management signalling if this PDU session is needed again within a short lapse of time. Hence, it is important that the network takes additional measures to control the establishment and release of PDU Sessions, which in turn can also allow detection of when there is a need to register or deregister from a NS.

Lastly, it is important to associate specific standardized categories of applications traffic with the specific connectivity defined by an operator to serve these. These standardized categories defined in this document are described in section 3.1.9.

3.1.4 Multi-aspect resource optimization

3.1.4.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to multi-aspect resource optimization:

1. The provider of Multi-aspect Resource Optimization (MARO) shall consume network resource-related provisioning, assurance and supervision services on the different network dimensions including radio resources and virtual resources; the deployment environments including virtual machines or cloud instances, as well as the slices and subnets utilizing these functions, to jointly optimize resource utilization across multiple network dimensions
2. The provider of MARO shall offer provisioning services on the resource parameters configuration, to jointly optimize resource utilization across multiple network dimensions.

3.1.4.2 Description

The resource management in current mobile networks is handled in a network dimension-specific way, i.e., there are specific procedures and parameters for optimizing resource utilization for each dimension of network resources. The network dimension in this context represents the type and scope of network resource, for example, the resource types include physical radio resources, or virtualization and cloud resources, or the different network

management scopes like individual NS instances. The different resource optimization mechanisms (e.g., as illustrated by Figure 2) have visibility of resource utilization limited to specific dimensions, e.g., only to the RAN resources or to the virtual infrastructure resources, but not to both simultaneously. However, the 3GPP management system responsible for RAN slice subnet instances will have an interface towards RAN Network Function (NF) management, but also an interface towards Network Functions Virtualization (NFV)/cloud management in case parts of the network is virtualized, and in this case example, Os-Ma-nfvo (Network Function Virtualization Orchestration) [7] is used. Furthermore, this means that resource management is limited to a specific dimension. In Figure 2, for example, the number labels indicate the different areas where resource related actions may be performed, specifically: changing radio parameters and policies at (1) and (2); changing slice configurations at (2) or (5); changing virtualization resource configurations by undertaking load balancing among Virtualized Network Functions (VNFs) at (3) or scaling the virtual resources at (4).

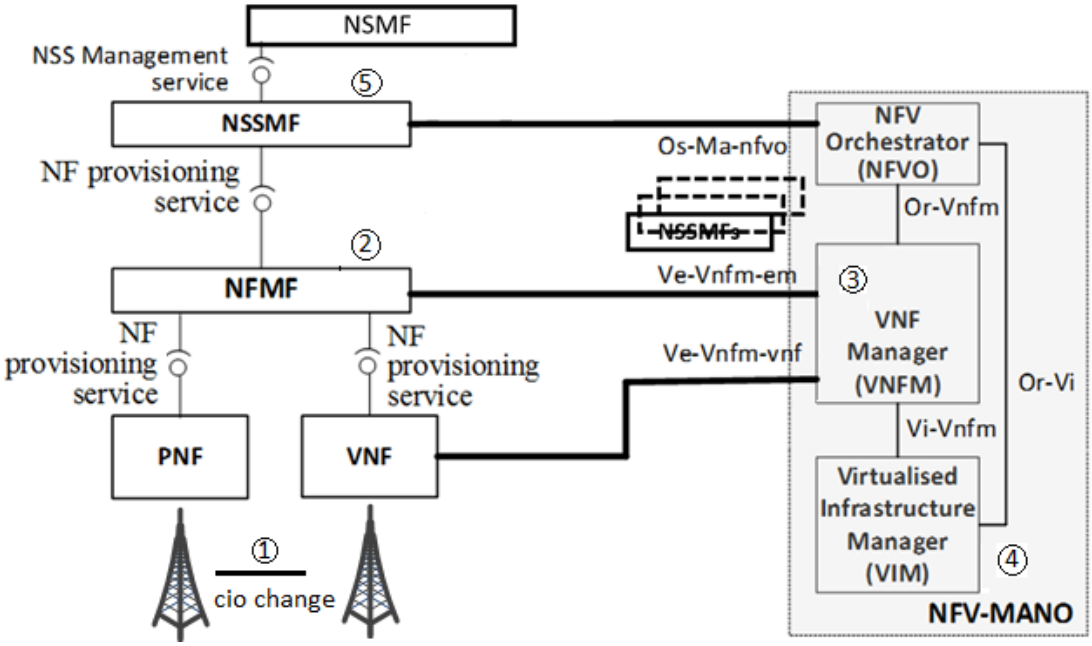

Figure 2 Resource optimization (Load balancing) options labelled 1 through 5.

The impact of dimension-specific actions to other potentially influenced dimensions and the benefits of cross-dimensional decisions are not considered. Such multi-dimensional awareness becomes critical especially with advent of virtualization support in network deployment where network functions in a NS can be implemented on different infrastructure domains, e.g., partially on physical and partially on virtualized resources such as in case of Central Unit (CU)/Distributed Unit (DU) split of gNB. Besides, multi-dimensional awareness can be leveraged to maximize the available resources, e.g., if core network latency drops from 100ms to say 60ms, the extra 40ms can allow the Operation Administration and Maintenance (OAM) to use RAN functions that would otherwise violate the E2E latency budget (this is already in scope for cross-domain management as part of NS management in 3GPP). Moreover, as network slicing enables the cases of network function sharing, the same network function can belong to different administrative domains corresponding to different managed NS/subnet instances. Therefore, it is of high importance to be able to

manage such interdependencies among the different network dimensions in order to perform adequate resource management.

3.1.5 Feasibility check and resource reservation

3.1.5.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to resource reservation and feasibility check:

1. The E2E network slicing architecture shall achieve alignment with an End-to-End Orchestrator (E2EO) in order to perform NS provisioning and modification without failure.

3.1.5.2 Description

According to 3GPP TS 28.530 [8], there are use cases of provisioning including feasibility checks, and the derived requirements for a 3GPP management system.

3.1.6 Automatic provisioning

3.1.6.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to automatic provisioning:

1. External interface to receive requests from customers/3rd party directly and to provide network services agilely to customers.
2. Each automatic provisioning function in E2E manner, e.g., automatic activation, automatic slice creation, and closed loop.

3.1.6.2 Description

The lifecycle phases of a NS include service design, provisioning, deployment, operation, and removal. In order to serve customers using network slicing, a network service provider is required to map customer requirements into typical values of GSMA PRD NG.116 [6] Generic network Slice Template (GST) attributes and to deploy resources and network functions properly so as to satisfy customer's SLAs. As network slicing is spanning over several network domains, E2E orchestration is a key capability to provide NS service offerings from the perspective of lifecycle management. E2E orchestration can control management functions in each domain to provide lifecycle management service, such as provisioning, and can make services available faster to the Network Slice Customer (NSC) by reducing offering time. Toward the goal of NSaaS, automatic provisioning function in an E2E manner is one of the expected functions to bring benefits, e.g., cost advantages, by removing manual interventions in the lifecycle management. process.

3.1.7 Isolation

3.1.7.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to isolation:

1. The NSC shall be able to express the Service Level Specification (SLS) for a service to be hosted in a NS.
2. The Network Slice Provider (NSP) shall be able to segregate resources of an NS instance from other NS instances.
3. The NSC shall be able to express that wants its management data to be safe and separated from the rest of NSCs.
4. The NSP shall provide means to guarantee the management data is securely stored, and accessible only for authorized NSC.
5. The NSC shall be able to perceive the NS as a self-contained, dedicated network.
6. The NSP shall provide means to allow for multi-tenancy support (controllability separation in the network), based on the definition of separate yet tailored management spaces for different NSCs.
7. The NSP shall have the ability to allow for the on-demand and controlled exposure of capabilities to the NSC, according to the settings of associated management space.
8. The NSP shall be able to make capabilities available for consumption through Application Programming Interfaces (APIs).
9. The NSP shall have mechanisms to manage interactions between NSC and NSP, including API registry & discovery, access control, and authorization.

3.1.7.2 Description

Isolation in network slicing is a multi-faceted problem, articulated into three separate dimensions: (i) *performance*, ensuring that SLS is always met on each NS instance, regardless of workloads or faults from other running instances; (ii) *security,* ensuring that any type of intentional attach occurring in one NS instance have no impact on any other running instance; and (iii) *management*, ensuring that each slice instance can be operated as a separate network partition, with an independent lifecycle management. These three dimensions, originally discussed in GSMA PRD NG.127 [3], impose requirements on both NSP side and on NSC side, as elaborated below.

Performance dimension:

- The NSC shall be able to express the SLS for a service to be hosted in a NS. To that end, the NSC may use a NEST together with other deployment related info (e.g., on-prem execution of some NFs, asynchronous replication of applications and data across different regions for high availability support, etc.).
- The NSP shall be able to segregate resources of an NS instance from other NS instances. To that end, the NSP may leverage mechanisms providing means to split the infrastructure into a set of partitioned resources (resource chunks) and allocate partitioned resources into different NSs.

Note: This segregation impacts all resource domains, including radio resource domains (i.e., cell resources such as Radio Resources Control (RRC) connections, non-Guaranteed Bit Rate (GBR), Data Radio Bearers (DRBs), Physical Resource Blocks (PRBs), etc.), compute

resource domain (i.e., Infrastructure as a Service (IaaS) / Platform as a Service (PaaS) resources allowing for VNF/ Containerized Network Function (CNF) execution), and transport resource domain (i.e., WAN resources).

Security dimension:

- The NSC shall be able to express that wants its management data to be safe and separated from the rest of NSCs. This management data includes business related data (e.g., charging information, subscriber profile) and operation related data (e.g., performance assurance (PM)/fault supervision (FM)/provisioning (CM) data, log, trace, MDT, policy, analytics reports).
- The NSP shall provide means to guarantee the management data is securely stored, and accessible only for authorized NSC. Depending on the criticality of the data, the NSP may decide on different security management solutions for data protection, including integrity, confidentiality, and privacy protection solutions, even for the same NSC. This means that the NSP shall provide each NS instance with appropriate mechanisms preventing unauthorized entities to have read and write access to slice-specific management data and be able to record any of these attempts. Examples of these mechanisms can be found in the survey reported in [9].

Management dimension:

- The NSC shall be able to perceive the NS as a self-contained, dedicated network.
- The NSP shall provide means to allow for multi-tenancy support (controllability separation in the network), based on the definition of separate yet tailored management spaces for different NSCs. Each management space shall be provisioned with only the configuration and monitoring capabilities that the particular NSC needs to consume from their NS instance(s). The activation / de-activation of certain capabilities allows NSP to regulate how much control the NSC can take over allocated NS instance(s).
- The NSP shall have the ability to allow for the on-demand and controlled exposure of capabilities to the NSC, according to the settings of associated management space.

Note: The importance of capability exposure in slicing environments has been already highlighted in [10][11]. The NSP can leverage this feature to offer NSaaS in multiple forms, from *provider-managed slices* (i.e., the provider is in charge of the slice operation, while the customer can merely use the network resources of the provider slice, without any further capability of managing or controlling it) to *tenant-managed slices* (i.e., the customer takes full control of the slice, and the provider just segregates the infrastructure as required for that purpose), with some variants in between. By regulating the exposure, the NSP can define the visibility and the degree of control the NSC can take over the slice. Figure 3 shows the logic behind the capability exposure concept.

- The NSP shall be able to make capabilities available for consumption through APIs adhered to three main principles:
    - Open. Offered APIs need to leverage as much as possible on standard-based or de-facto APIs, following industry recommendations.
    - Global. Offered APIs should allow every NSC to have a uniform and consistent service experience across global footprint, with the effortless portability of

application across different operator platforms (relevant in mobility/roaming scenarios) and easy service replicability (in case there exists trans-national slices).

- User-friendly. Offered APIs need to be abstracted out of internal APIs, to hide telco complexity and make them easy to use (consume) by NSC, especially those with no telco expertise/background expertise.

- The NSP shall have mechanisms to manage interactions between NSC and NSP, including API registry & discovery, access control, and authorization.

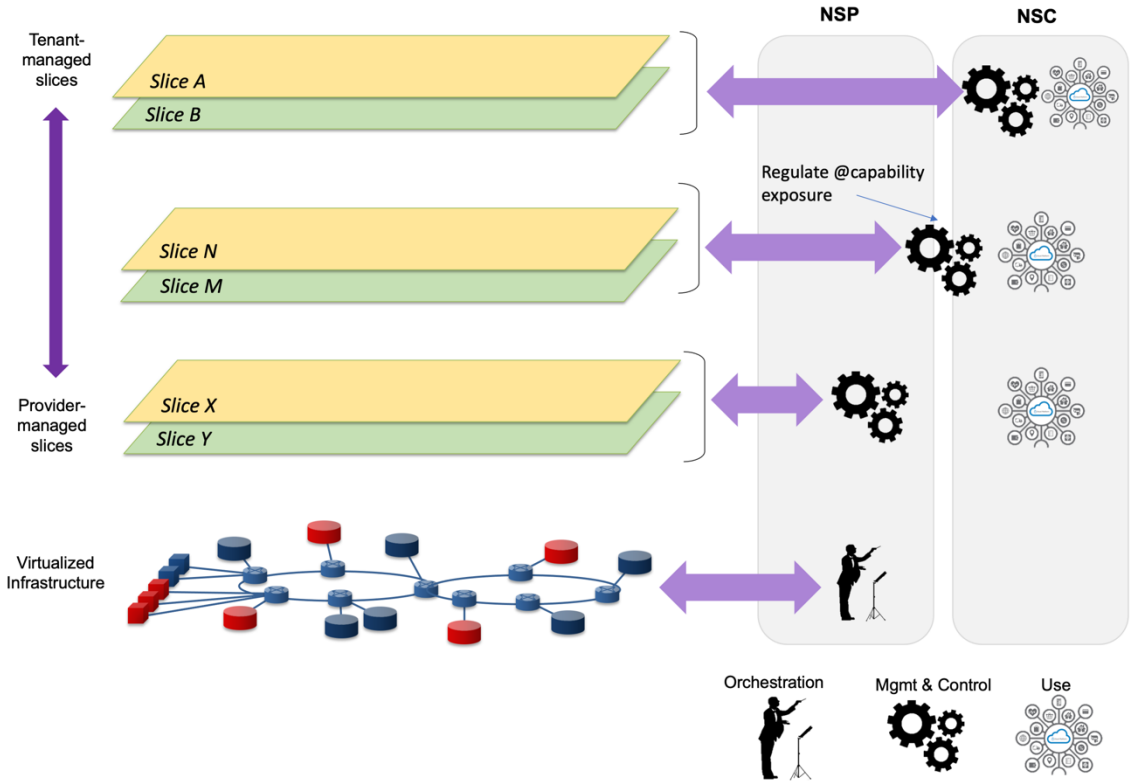

Figure 3 Capability exposure in network slicing environments.

3.1.8 NSaaS enabling exposure

3.1.8.1 Requirements

The E2E network slicing and its architecture shall fulfil the following requirements related to NSaaS enabling exposure:

1. Based on NSP's policy, operational capabilities exposed to an NSC in NSaaS should include provisioning, performance management and fault management.
2. NSP shall be able to make the requested capabilities available for consumption through a service platform.
3. NSP shall support mechanisms to police the interaction with individual NSCs, in relation to consumption of exposed capabilities.

3.1.8.2 Description

One of the key benefits of NSaaS is the ability of individual NSCs to retain control of their allocated slices, specially in what concerns (i) the usage and monitoring of hosted services, and (ii) the configuration and management of subscriber devices. To that end, the NSP shall be able to make network slicing related capabilities (including infrastructure, network and OAM related capabilities) available for consumption to these NSCs.

The scope of this clause is to identify the enablers for network slicing capability exposure in NSaaS offering.

An NSP offers a service-tailored connectivity pipe to one (or more) Application Server (AS) hosted by the Data Network (DN). This AS can be associated with NSP services (e.g. communication services) or 3rd party services. Devices subscribed to one service can establish communication with the AS through the corresponding network slice, which will provide an enhanced connectivity profile in terms of functionality, performance and/or security.

The fact that makes network slicing an E2E concept is that the device-to-AS connectivity pipe involves all the technical domains within the NSP's managed network, including the RAN, CN and TN (see Figure 1). The TN domain is in charge of providing infrastructure level connectivity between the gNB (the entry point to the network for the device) and the DN (where the AS's are hosted). In this data path, different segments can be outlined: i) fronthaul segment (RU/DU connectivity, O-RAN interface), ii) midhaul segment (DU/CU-UP connectivity, 3GPP F1 interface), iii) backhaul interface (CU-UP/UPF connectivity, 3GPP N3 interface; UPF/PSA-UPF connectivity; 3GPP N9 connectivity) and DN segment (PSA-UPF/AS connectivity, 3GPP N6 interface). It is worth noting to clarify that 3GPP slice concept leaves out DN segment. As a result:

- for those cases where PSA-UPF and AS are co-located, NSPs provide 3GPP network slices;
- otherwise, NSPs provide 3GPP network slice and DN segment together.

In this latter situation, the DN info (e.g., Key Performance Indicators (KPIs) and alarms) should be aggregated with 3GPP info, in order to provide E2E view. This aggregation is enforced in the E2E service management block (see Figure 1), before being made available for consumption to NSC. In order to get the DN info, it is assumed that the DN is within the NSP's administrative domain and/or is cloud network which provides the network capability exposure to be monitored and/or notify some events in its own network section so as that NSP can manage. Other scenarios (e.g. the DN belonging to a 3rd party local network) is out of the scope of this analysis.

NSP shall have controllability and auditability means when exposing (network slicing related) infrastructure (e.g. transport), network and OAM capabilities to NSCs, typically B2B and B2B2C customers such as hyperscalers, application developers and industry specific enterprise customers (e.g., the verticals). Controllability means that the NSP can regulate the particular set of resources each NSC is allowed to access and under which conditions, leveraging Role Based Access Control (RBAC). Auditability means that every (request-response / subscribe-notify) message exchanged between the NSP and an NSC needs to be logged with accurate timestamps (for traceability) and support non-repudiation (for SLA

verification). For these two means, the NSP can rely in API gateway solution. An API gateway provides all the functionalities that are needed to policy the interaction between the NSP and the NSCs, in relation to API invocation. These functionalities include API publication & discovery, access control (authentication & authorization), accounting and logging, among others. However, further details of API gateways are out of scope of this document.

3GPP has defined requirements related to network slice lifecycle management in 3GPP TS 22.261 [12]. In relation to the enablers for exposure in NSaaS offering, the following points are applied:

Based on NSP's policy, operational capabilities exposed to an NSC in NSaaS should include:

Provisioning

- **Provisioning** [13]: also referred to as lifecycle management, it encompasses the set of activities that allow establishing and maintaining the desired state for a Network Slice Instance (NSI) throughout its lifecycle, from commissioning (i.e., NSI creation) to de-commissioning (i.e., NSI deletion). The establishment of the NSI's desired state referred to as configuration management in the traditional Operation Support System (OSS) jargon, requires injecting appropriate configuration parameters into the NFs which are in scope of the NSI [2]. These parameters make sure the NSI is ready to serve as expected. The configuration management also cover to configure the information which associate a UE to use a network slice, for instance, to subscribe a dedicated network slice for particular service. On the other hand, *the maintenance* consists in continuously evaluating the difference between desired and actual state in light of an optimization policy, which may also change from time to time, and attempt to modify the state of NSI resources accordingly. An NSC can be granted with permission to i) set a non-default optimization policy, or bring a new one, according to their specific needs; ii) proactively trigger modification actions, either at network function layer (e.g., inject configuration parameters of one or more slice's constituent NFs), at virtualized resource layer (e.g., scaling out the NSI capacity), or both of them at the same time.
- **Provisioning data reporting** [13]: it refers to the collection of operations that allow the NSC to subscribe to certain events related to NSI state, so it can be notified upon accordingly. Examples of these events can include changes on NSI status (e.g., instantiated, activated, removed), NSI capacity (e.g., scale in, out), on NSI configuration (e.g., parameter value change on one of the NSI's constituent NFs).

Performance management

- **Performance Measurement job control** [14]: it refers to the operations that allow the NSC to manage (create, delete, list) measurement jobs applicable to an

NSI. When configuring a performance measurement job, the NSC can specify i) the performance measurements to be collected, either raw data or analytics data; and ii) the consumption pattern, either using streaming method or the file method. In a nutshell, the performance data control can be tagged as “proactive performance management”.

- In the streaming method, the performance data is sent (from the NSP to the NSC) when ready. The volume of the data sent with this method is expected to be small, and the granularity period of the data streams needs to be configurable and is expected to be short.
- In the file method, the data is accumulated for a time duration before it is sent to the NSC; the data will be delivered as a file.

- **Performance threshold monitoring** [14]: it refers to the collection of operations that allow the NSC to subscribe to certain events related to NSI performance, so it can be notified upon accordingly. Examples of these events are threshold crossing. In a nutshell, the performance data report can be tagged as “passive performance management”.

Fault management

- **Fault supervision data control** [15]: it refers to the operations that allow the NSC to manage (create, delete, list) alarms applicable to an NSI. When configuring an alarm, the NSC can specify i) the fault(s) associated to the alarm; and ii) optionally to be informed on suggestions for remediation actions. In a nutshell, the fault data control can be tagged as “proactive fault management”.
- **Fault supervision data report** [15]: it refers to the collection of operations that allow the NSC to subscribe to certain events related to NSI malfunctioning, so it can be notified upon accordingly. Examples of these events are node malfunctioning or parameter value mismatching. In a nutshell, the fault data report can be tagged as “passive fault management”.

NSP shall be able to make the requested capabilities available for consumption through a service platform.

Note1: The NSP service platform should offer self-management channels to NSC, via REST APIs. The implementation details of this platform are out of the scope of the present document. For those cases where the NSP role is played by a telco operator, the NSP platform can be mapped to the Operator Platform (OP) concept that GSMA defines.

Note2: Some NSCs may not have a deep network slicing understanding. Therefore, it is needed to offer these NSCs user-friendly APIs, hiding unneeded telco complexity, with the sole focus on NSC operational needs. How to define these user-friendly APIs out of SDO-defined network slicing APIs is out of scope of the present document.

NSP shall support mechanisms to police the interaction with individual NSCs, in relation to consumption of exposed capabilities. In particular:

- NSP shall provide mechanisms to onboard NSCs which can consume the exposed capabilities and manage (create, update, delete) their profiles. Onboarding is one-time registration process that enables NSC to access the exposed capabilities, as mentioned in clause 4.9 in TS 23.222 [16].

Editor's Note: The exact functionality associated with boarding of NSCs is FFS,

- NSP shall provide mechanisms to publish the exposure information of NSP's capabilities to be used by NSCs, in order to discover and subsequently access the NSP's capabilities.
- NSP shall provide mechanisms to control the access to the NSP's capability for each NSC, based on NSP's policy or agreement with NSC. Access control relies upon authentication and authorization mechanisms. For those B2C/B2B2C use cases, it may be also including user consent.
- NSP shall provide mechanisms for logging the interactions with individual NSCs, for traceability and SLA verifications purposes.

3.1.9 Support for differentiated handling of Traffic Categories

Please see GSMA PRD NG.141 [24] for the definition of Traffic Categories.

4 Pathways for a phased rollout of NSaaS

Network slicing is an E2E solution, covering all the technology domain spanning from UE to the data network. The provisioning and operation of the different slices, and the lifecycle management of hosted services, is done with an OSS composed of a number of management domains, including vertical management domains (e.g., RAN, CN, TN management domains) and horizontal management domains (e.g., MANO, E2E network & service management domains) [3]. However, the reality proves that the maturity level of network slicing varies across all these technology and management domains. In fact, the degree of penetration of slicing features in the different technology domains is not the same. For example, while the core network has incorporated network slicing support since the first 5G release (3GPP Release 15), the transport network does not support any native network slicing feature yet, and first solutions have only recently been integrated into the radio access network. The main reason why the maturity level varies across technology domains (and their corresponding management domains) is mainly due to the existing lack of coordination among SDOs in the standardization arena, with a high number of participating SDOs. In the current landscape, each SDO addresses a portion of the E2E problem, developing slicing specifications for this portion under assumptions that do not necessarily match the assumptions made by other standard bodies, which typically address other portions. A clear example of this mismatching can be observed on the priorities that different SDOs set in relation to which slicing features need to be worked out in each release. In fact, these priorities are quite different across SDOs, both in time and scope.

The above-mentioned issue, together with the inherent integration issues existing in carrier networks (e.g., multi-vendor solutions, brownfield facilities), outlines the main challenges that operators need to work out to enable the full promise of network slicing, which is NSaaS. However, solving these issues towards this ultimate goal may require years. In the meantime, operators need to look for workarounds to start commercializing and monetizing network slicing. In this regard, the only realistic option for operators is to introduce slicing capabilities progressively, starting with early-stage network slicing and to use a vision of what is desired in the long term to guide progress and focus; in other words, operators shall outline a phased NSaaS rollout. For this activity, it is needed to identify all solutions enabling network slicing and position them into a multi-faceted common space, by means of a technology radar. The technology radar allows modelling this common space, consisting of:

- **Multiple dimensions**: identify the criteria based on which solutions can be profiled. A given managed domain (RAN, CN, TN, OSS) can have one or more dimensions.
- **Multiple rings**: represent different timelines, associated to the General Availability (GA) of solutions.

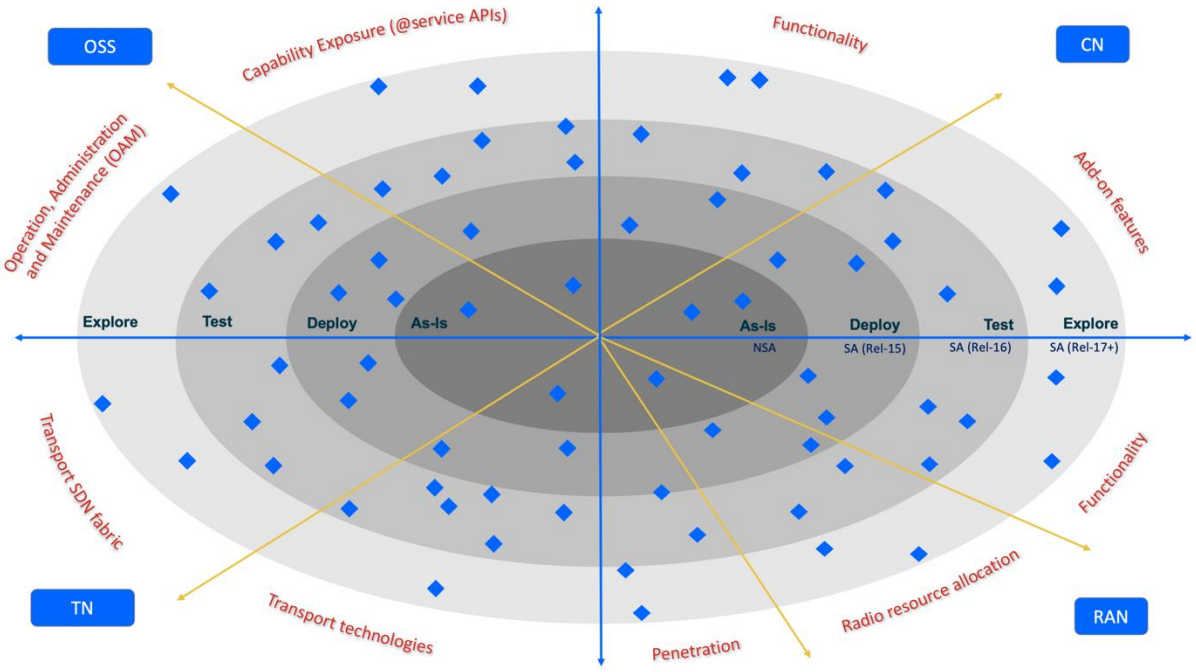

Figure 4 Technology radar toward NSaaS.

Figure 4 illustrates an example of how a technology radar could look like [17]. As seen, it captures solutions impacting all operator managed domains into four rings:

- As-is ring represents solutions that are available in today's carrier networks. These solutions are typically associated with technologies that operators have high confidence in, with low risk and recommended to be available across the entire service footprint. In terms of 5G roll-out strategy, this corresponds to 5G NSA (Non-Standalone).
- Deploy ring covers the slicing solutions that can be applied in early 5G SA (Standalone) networks, based on 3GPP Release 15 standards. Some operators have already started to activate their SA networks, while some others expect to get them operationally ready within next year. With this timing in mind, we can say that this ring captures proven slicing solutions that operators may integrate in the short-term.
- Test ring: captures slicing solutions that are much more focused on satisfying requirements from uRLLC (Ultra-Reliable and Low Latency Communications) and mIoT services. Associated with brand new Rel-16 features, these solutions have great potential but are unproven in production networks, hence it is worth operators investing in prototyping efforts in order to evaluate their performance and impact. This evaluation is typically done with commercial trials, either bilateral or multi-vendor, and different Proof of Concepts (PoCs). The upgrade towards Rel-16 is expected within the next 2–3 years; this means that test ring represents slicing solutions that might be available in the medium term.
- Explore ring includes slicing solutions that are foreseen in the long run, starting in the next 4–5 years. These solutions, tied to features from 3GPP Rel-17 onwards, promise to provide great potential, though their impact and commercial availability is still far from crystal clear. The role of the operator is to keep track of their evolution through exploratory activities such as the ones done in research and innovation projects, e.g., HEXA-X [18].

In addition, the solutions are profiled into different dimensions, as follows:

- CN domain dimensions: functionality and add-on features. The *functionality* dimension deals with the discussion on how to use CN functions for the construction of different NSs, scouting different deployment options for these slices depending on the isolation and business requirements of hosted services. On the other hand, the *add-on features* dimension refers to the set of value-added solutions that complement and extend baseline slice functionality. The network operator can optionally make use of these solutions to either (i) provision enriched services to the customer, i.e., new revenue streams; or (ii) streamline internal network operation, i.e., OPEX savings.
- RAN domain dimensions: functionality, radio resource allocation and penetration. The *functionality* dimension provides a deep dive on the applicability of open RAN principles on NR protocol stack functions to design and configure RAN slices, going from monolithic solutions towards more flexible, service-tailored composition patterns. The *radio resource allocation* dimension discusses the availability of solutions to segregate and dispatch cell radio resources to competing RAN slices, so that their targeted KPIs are met. Finally, the *penetration* dimension refers to the penetration of RAN slicing technology within the operator's footprint.
- TN domain dimensions: transport technologies, and transport Software Defined Networking (SDN) fabric. The *transport technologies* dimension captures the protocol encapsulation and data plane solutions across the different network segments, allowing for the realization of soft and hard slicing, with some solutions in between. On the other hand, the *transport SDN fabric* dimension refers to the control and management plane aspects, discussing the application of programmability and automation through SDN.
- OSS domain dimensions: OAM, and capability exposure. On the one hand, the *OAM* dimension refers to the set of activities related to NS life cycle management (model-driven fulfilment and data-driven assurance operations). On the other hand, the *capability exposure* dimension touches on the need for NSPs to make NS capabilities available for consumption to NSCs, through open, global and user-friendly service APIs.

With this baseline template for the technology radar, the operator is in position to define their own roll-out strategy for NSaaS. To that end, the following steps are needed.

First, the operator should fill the radar out with network slicing solutions. The position of each solution in the radar can be done according to three different criteria: (i) the technology maturity of the solution, which is related to the readiness of the corresponding standards; (ii) the roadmap of commercial products, which specifies when the features associated with the solution will be available; and (iii) the relevance for the customers, which determines the prioritization of the solution over others.

Upon completion of the first task, the operator should then define its own go-to-market strategy [19], based on deciding which solutions will be made available, when, for which customers, and under which business models.

Annex A Collaboration with external organizations

A.1 Gap Analysis

Technical requirements are addressed in Section 3 of this PRD. In order to achieve the requirements, identification of gaps between them and existing specification in the external organization is of importance. The following sub-sections state the observed gaps for outreach between GSMA and the external organizations.

A.1.1 Co-operation of application layer and transport layer

Regarding transport network, IETF is on the way to specify the IETF Network Slice framework in their Internet-Draft “draft-ietf-teas-ietf-network-slices” and "draft-contreras-teas-3gpp-ietf-slice-mapping". As mentioned above, harmonization with 3GPP domain is an important factor. GSMA needs to encourage IETF to work on this standardization with harmonization perspective.

A.1.2 SLA assurance in E2E

3GPP has documented management system in 3GPP domain such as architecture framework and management services in TS 28.533 [20] as well as the 5G network resource model (NRM) in TS 28.541 [21]. For SLA assurance in E2E, it is required to clarify and align specifications between 3GPP and other organizations (i.e., IETF and O-RAN).

Regarding transport network, IETF is on the way to specify the IETF Network Slice framework in their Internet-Draft “draft-ietf-teas-ietf-network-slices” and the IETF Network Slice service Yang model in their Internet-Draft draft-ietf-teas-ietf-network-slice-nbi-yang. The model can be used by an IETF Network Slice customer (for example TN NSSMF for 3GPP slices) to manage IETF Network Slice from an IETF Network Slice Controller. As mentioned above, harmonization with 3GPP domain is an important factor. GSMA needs to encourage IETF to work on this standardization with harmonization perspective.

A.1.3 Deterministic and operator-controlled UE behavior

Identified gaps regarding the requirements on deterministic and operator-controlled UE behaviour are as the followings:

- 3GPP specifications need to be updated to improve the UE behavior control to address the issues identified in 3.1.3 (including Standardized Traffic Categories support)
- GSMA lacks the identification and definition of standardized traffic categories.

The following text indicates steps for the requirements on this.

- when the UE is registered with the network, it receives the necessary control information by HPLMN and if applicable and allowed by the VPLMN. 3GPP to define the best suitable technical option.

- (Example solution should for Standardized Traffic Categories support) the URSPs TD definition is augmented in 3GPP TS 24.526 [22] as a "Connection Capability" codepoint is associated with each Standardized Traffic Category. Applications that require specific connectivity defined for a specific Standardized Traffic Category request connectivity from the UE by indicating the related required Connection Capability.

Editor's note: 3GPP could be tasked to find the best technical option to support standardized traffic categories once GSMA identifies them.

A.1.4 Multi-aspect resource optimization

Identified gaps regarding the requirements on MARO are as the followings:

- 3GPP has documented the use case description as summarized in 3GPP Release 16 TR 28.861 [23]; 3GPP also has input data for the different 3GPP domains, but details like resource parameters used for configuration are not described for this use case.
- 3GPP has not agreed to include this use case into normative specification. Hence, lacking is a normative specification of the management services and procedures related to the provision of MARO. Note that the specific solution requirements are not in scope of GSMA to specify, but GSMA should identify the requirement to optimize simultaneously across more than one dimension at a time if there are areas missing so far and encourage 3GPP to work towards solutions address this problem space.

Also, the input that would come from non-3GPP domains may need to be evaluated, possibly the candidate inputs expanded.

The following text indicates example steps for the requirements on this.

The following is an example of possible steps; this should not be seen as a solution description. The provider of MARO takes network status inputs from multiple domains and derives the policies for network configuration and optimization. This may be specific to a single network slice, or it may be applicable to multiple network slices.

The provider of MARO consumes PM, FM and CM services across different domains. The information retrieved may consists of network resource-related data (both physical and virtual resources), which may be characterized as, but not restricted to:

- The information from one or multiple network slices, slice subnets and/or infrastructure domains (physical and virtual)
- For each domain/subnet/slice, information on allocated resources and their utilization, event history (e.g., recent overload situations), history of actions taken (e.g., MLB decisions, scaling of resources)

The provider of MARO then:

- Analyses the collected data for resource optimization across different infrastructure domains for either a single network slice or multiple network slices
- Correlates the data across network slices or slice subnets. Note that although from network planning some initial mapping on resource allocation across domains is known, correlation of performance in the different domains is still necessary. For example, even given CU/DU split with known resource allocation in each domain the relative performance in each domain (CU-virtual resource utilization, vs. DU physical load) must be correlated.
- Calculates/generates network resource provisioning policy improvements. Non-exhaustive list of examples of policy improvements may be:
    - Adjusting the policy on scaling of virtual resources based on the observation that RAN MLB (on physical resources) was performing well but the virtual resources were not adequately provisioned.
    - Change the characteristics of transport links based on observed transport network performance bottleneck.
    - Change in the network slice or subnet resource provisioning for inter-slice resource optimization

Exposes obtained policy improvements as a service toward other network functions.

The policy improvement service exposed can be either:

- Directly applied by the provider of MARO by utilizing the configuration management service of the related network function
- Consumed by the provider of domain-specific control which:
    - takes into account all the existing policy improvement requests related to that particular domain/resource
    - analyses potential conflicts among policy improvements
    - resolves identified conflicts among policy improvements and derives the actual policy improvement to be applied within that specific domain
    - applies the resulting policy improvement through the configuration management of involved network function

A.1.5 Feasibility check and resource reservation

3GPP Release 17 specifications provide solutions to support feasibility check and resource reservation capabilities in network slicing scenarios.

TS 28.531 [13] specifies the use cases and the procedures for such capabilities. Specifically, it reports on the ability of Network Slice Management Service Provider (NSMS_P) to determine whether the service requirements captured in a ServiceProfile can be fulfilled, upon Network Slice Management Service Consumer (NSMS_C) request; if so, the NSMS_P is able to reserve resources for their allocation to the corresponding network slice. These resources can be secured/guaranteed for a certain validity period, which is configurable by the NSMS_C.

TS 28.541 [21] specifies the stage 2 and 3 solutions for these capabilities, with the definition of FeasibilityCheckAndReservationJob IOC. This IOC allows the NSMS_C:

- To decouple feasibility check from resource reservation, for those cases where NSMS_C only wants to get informed on whether a network slicing request is feasible or not, but does not want to get resources secured
- To obtain the progress information of feasibility check work on NSMS_P side, following asynchronous communication patterns in request-response messages.
- To get informed on the result of feasibility check work, upon completion on NSMS_P side. If unfeasible, the NSMS_C can be (i) informed of the reason why the service requirements captured in ServiceProfile cannot be fulfilled; and (ii) be recommended with alternative service requirements that can be fulfilled instead.

As an additional note, it is worth mentioning that the same solutions summarized above for network slices also are applied for network slice subnets.

A.1.6 Automatic provisioning

No specific gaps are identified in this version.

A.1.7 Isolation

No specific gaps are identified in this version.

A.1.8 NSaaS enabling exposure

Regarding the management service exposure, as typified by lifecycle management, to external entity beyond 3GPP management system, 3GPP has defined requirements in 3GPP TS 22.261. However, 3GPP has not defined normative specification on the management service exposure, while 3GPP has an ongoing study and is drafting 3GPP TR 28.824 where some possible solutions are proposed, but there are neither conclusions nor recommendations in this TR.

On the other hand, regarding operation capability which may be covered by network exposure, if NSC requests to create and use a dedicated network slice for the NSC, there is no interface to configure the information about the created network slice which associate a UE.

3GPP has considered to integrate transport network as 3GPP-external domain in management system, for example, 3GPP TS 28.533 in an architecture framework and management services perspective and 3GPP TS 28.541 in 5G network resource model perspective. However, the description on the integration of virtualized data network in order to aggregate DN info (e.g., KPIs and alarms) with 3GPP info cannot be found in the existing specification.

Annex B Document Management

B.1 Document History

Version	Date	Brief Description of Change	Approval Authority	Editor / Company
1.0		New PRD first version	NG, ISAG	Masaharu Hattori / KDDI
2.0	20/11/2022	CR1002	NG,ISAG	Masaharu Hattori / KDDI
3.0	23/06/2023	CR1003, CR1004, CR1005	NG, ISAG	Masaharu Hattori / KDDI
4.0	23/06/2024	CR1006	NG, ISAG	Masaharu Hattori / KDDI

B.2 Other Information

Type	Description
Document Owner	NG-NRG
Editor / Company	Masaharu Hattori / KDDI

It is our intention to provide a quality product for your use. If you find any errors or omissions, please contact us with your comments. You may notify us at prd@gsma.com

Your comments or suggestions & questions are always welcome.