Slicing challenges for Operators

Luis M. Contreras
Telefónica, Spain

INTRODUCTION

An initial idea of network slices in standardization¹ was introduced in 2011 by the ITU-T (ITU-T, 2011) in the form of logically isolated network partitions (LINP), with a slice being considered as a unit of programmable resources such as network, computation and storage. Later, in 2016, the concept of network slicing has been formulated by the NGMN in (NGMN, 2015) as an enabler for (hopefully) achieving an efficient logical division of network resources (either physical or virtual) and service functions for different services on the same network infrastructure. The concrete settings of resources and service functions are selected and combined together with the aim of supporting a specific application or type of service.

The different types of services are in principle the ones under consideration for 5G services, as introduced in (ITU-R, 2015). These are the enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC) and ultra-Reliable and Low Latency Communications (uRLLC). The eMBB service type encompasses the challenge of providing an extremely high volume of delivered data, due to e.g., (ultra) high-definition video sharing. The mMTC kind of service focuses on applications where a large number of connected IoT devices, such as sensors, representing a large number of sessions collectively create a significant amount of data passing through the network. Moreover, these data are often associated with requirements like privacy, data ownership, etc. Finally, the URLLC type refers to services that require extremely low end-to-end latency, like Tactile Internet, Interactive Gaming, Automotive, or Industrial Automation. Figure 1 summarizes some characteristics of each type of service.

Figure 1. Characteristics of different kinds of 5G services

Combining services like eMBB, uRLLC, and mMTC on the same network, makes it quite difficult to define a common engineering pattern capable of supporting simultaneously all the requirements of each of them. An approach to manage such deployment needs is to segregate those services into specialized network partitions or slices, designed, and optimized for each of the types of services to be provided. Each slice is then tailored to such specific purpose, thereby avoiding unnecessary capabilities (for example, an instance of a mMTC slice may not require a TCP optimizer function, whereas an instance of an eMBB slice may need to invoke this capability). This form of flexible network resource consumption opens possibilities of business models for network operators. However, the associated challenges are not minor, as described along this chapter.

Figure 2 provides a high-level overview of the slicing concept. Through the instantiation of distinct network slices, the operator will be able to provide completely different services in a dynamic and isolated manner despite that all of them actually run over the same physical infrastructure. Thus, the operators can initiate a transition from the existing design choices that rely upon multi-service networks deployed over one architecture conceived to fit all kinds of services (Doverspike, Ramakrishnan, & Chase, 2010), towards the approach of getting logical networks defined per service.

Figure 2. Network slicing concept

Network slices are intended to behave like entirely independent networks. This implies that there should be mechanisms for properly isolating the slices (if required by the slice definition), thereby avoiding interference between one slice and the others. Furthermore, in some cases, there will be situations that will require that some of the slices to be interconnected to compose a service, raising the need for interworking (stitching) such independent logical constructs.

In order to be able of managing such complexity induced by the logical slice partitioning, the operator shall have technical means to take into account the specific requirements of each slice (for instance in terms of throughput, delay, and jitter), as well as suitable inter-slice information exchange mechanisms to avoid congestion issues and arbitrate resource needs, permanent or occasional.

It is clearly anticipated that the mechanisms for creating and operating the network slices will be supported by the softwarization trend enabled by the programmability of the network resources and the virtualization of the network functions, that is, leveraging Software Defined Networking (SDN) and Network Functions Virtualization (NFV) techniques. Apart from the elasticity provided by the virtualization of network functions when dynamically deploying network functions, flexible steering mechanisms (a.k.a., Service Function Chaining) are needed to shepherd the traffic flows according to the expected service behavior. In order to address these requirements, network programmability and its integration with the function virtualization described above, are considered as the baseline pieces to approach network slicing (Odini, 2017) playing an instrumental role in the control, management, and operation of future networks. Both SDN and NFV raise a number of challenges by themselves (Contreras, Doolan, Lønsethagen, & López, 2015). The author focuses here on the additional degree of complexity brought by the concept of network slicing that accompanies this technical transition initiated by the network operators.

VALUE PROVIDED BY NETWORK SLICING

Future adaptation demands of 5G networks, mainly in terms of performance and resources, cannot be achieved in a sustainable manner with existing static and inflexible approaches that are mostly based on manual interventions. The convergence of the telecom and IT industries will provide a common and very high capacity 5G infrastructure, complemented with dynamic service invocation capabilities.

5G is expected to enrich by design the customer-facing services (the so-called “vertical industries” or simply “verticals”) with new capabilities and quality features. Automation will revolutionize service provisioning, for the most prominent vertical services in the areas of Media and Entertainment, eHealth, Energy, Automotive, and Manufacturing-Factories of the future (5G-PPP, 2015).

To satisfy the expectations of end-user service quality of experience, those verticals need also to rely on wholesale infrastructure services. This includes certain requirements of connectivity and the chaining of specific service functions to support the service, all of that provided in the overarching form of Slice-as-a-Service (SlaaS) as containers of network, storage, and compute resources deployed in a multi-tenancy environment. These services, unlike recent past experiences, will rely upon a fully softwarized network architecture from core to edge that utilizes virtualized resources with the claim to orchestrate, trade, deploy, and manage services in a (hopefully) fast, agile, secure, and efficient way.

The 5G overall ecosystem can be perceived as a universe of multiple independent vertical ecosystems, each targeting specific service sectors. Then, the concept of slicing in 5G networks contains a business purpose, which is the support of vertical markets that can leverage (seemingly) dedicated networks to improve their production cycles (5G-PPP, 2016).

Network slicing, as a combination of network, computing and storage resource allocation, is intended to enable value creation for vertical segments that lack physical network infrastructures, or which can complement their own resources with the ones offered by the operators. The business objective is to improve and evolve the production capabilities of those industries. Network slicing can thus be seen as the evolution of current wholesale services, by improving the way network resources are allocated and consumed in an agile and ultimately efficient manner.

At the time of characterizing the relationship with the vertical customer, the GSMA has described (GSMA, 2017) a potential value chain for the support of network slicing by network providers, as shown in Figure 3.

Figure 3. Network slicing value chain (GSMA, 2017)

Three different kinds of value-added services are identified as a source of incremental value:

- Capability exposure: through this utilization model, the providers can offer Application Programming Interfaces (APIs) to the vertical business customers for granting the capability of managing their own slices. Such management actions can include e.g., dimensioning, configuration, etc.
- Integration at customer premises: complementary network segments, in some cases pertaining to the vertical business customer, become an integral part of the solution, requiring a truly convergent network including the integration in existing business processes as defined by the vertical customer.
- Hosting applications: the provider offers the capability of hosting virtualized network functions or applications, including the activation of the necessary monitoring information for those functions.

In addition to the value provided to external customers, the network slicing does also provide value to the operator for the implementation of its own or internal services and for allowing federating services on demand with other operators. For example, it seems feasible to dynamically create dedicated environments adapted to the characteristics of services that today have to co-exist. The referred services are those like content distribution (e.g., TV, VoD), communication services (e.g., VPNs with distinct purposes such as backhauling, site interconnection, or database transactions), cloud-like services (e.g., for processing capabilities), etc. All these internal services today require in most cases re-architecting the network for accommodating them, including the steering of the traffic among service nodes and the reconfiguration of the forwarding paths supported onto the transport nodes. The expectation with network slicing is to accommodate some specific services on dedicated slices, going in the direction of network adaptation instead of re-architecting it.

Environments Facilitating Network Slicing

Network providers are nowadays deploying cloud facilities, termed NFV infrastructures (NFVI) that will serve to host virtualized network functions (VNFs). An NFVI allows for deploying VNFs in a dynamic way that can be adapted to the specific needs of each requested service. In the case of Telefónica, those capabilities are based on UNICA (Analysys Mason, 2017), the NFVI architecture being globally deployed for all of its operations. The overarching architecture of UNICA is based on cloud concepts to allow large-scale deployment across multiple sites, but also covers provider's needs such as carrier grade performance, scalability, and operational capabilities.

In a generic manner, the NFVI infrastructure will be required to define several types of sites or points of presence along the operator's network to create the topology needed to satisfy the requirements of the services according to their needs in terms of latency, processing capacity, bandwidth, etc. The aim is to distribute workloads efficiently across the network.

New Business Proposition

Taking into account the similarities with the definition of wholesale service offerings today (2018), different situations can be considered when looking at SlaaS, leading to slices with different capabilities, specifically in terms of their management and control capabilities, and how much of them the network service provider hands over to the slice customer or tenant:

- Internal slices, understood as the partitions used for internal services of the provider, who remains in complete control of the corresponding slices and resources allocated to them.
- External slices, being those partitions hosting customer services, appearing to the customer as dedicated networks. In this case, a subsequent distinction applies:
 - Provider-managed slices, meaning that the provider keeps the full control and management of the slice. In other words, the customer can merely use the network resources of the provided slice, without any further capability of managing or controlling them.
 - Tenant²-managed slices, implying that the customer has full control of the resources and functions allocated. The tenant has access to a (limited) set of operations and/or configuration actions, and the provider just segregates the infrastructure necessary for that purpose. It is worth noting that the term “infrastructure” is used in the broadest sense that is, including all network functions that the customer can deploy and use on the slice.

It is clear that Tenant-managed slices have to be allocated one per customer, since the customer directly controls it. On the other hand, Provider-managed slices could accommodate different customers sharing the same kind of service in terms of service requirements (bandwidth, latency, number of handled sessions, etc.). Figure 4 presents these different options, illustrating that while the orchestration capability always resides in the provider side, the control of the slice or the service on top of such slice could reside in the provider and/or in the tenant sides depending on the kind of slice.

Figure 4. Types of considered slices and control responsibilities

Requirements for Accomplishing Business Objectives

A number of key requirements can be deemed critical for accomplishing the business objectives as previously described. This sub-section covers some of them.

SLA (Service Level Agreement) Management

The assurance of SLAs will become a key aspect of the provisioning of services on top of 5G networks. Acknowledging the relevance of this observation for operator's internal services, this is even more evident when considering vertical customers, who will base their production totally or partially on the negotiated network slice.

Any distortion of the negotiated SLA associated to the network slice can impact not only the technical behavior of the service offered to the end-user of the vertical but also its reputation or business leadership. Even more importantly, such distortion may have legal consequences of any kind due to the incapacity of honoring the contracted service expectations (e.g., in terms of latency, bandwidth, or availability) through the network slice based on the nature of the service. A concrete case could be the result of injured people in a factory or in an assisted driving service because of the misbehavior of the network slice supporting the corresponding service. SLAs then have consequences beyond what sales agreements with strict associated penalties imply.

The negotiated terms of an SLA define a number of service indicators to be satisfied. These service indicators will be translated into network indicators (to be understood in a broad sense, that is, including computing, networking and storage resources) that will be used as inputs for the configuration and the orchestration of resources to form the slice. The deployment of a vertical service will observe interdependencies in the different classes of resources to be provisioned and configured, their location, and their forecasted availability. Furthermore, depending on the class of slice provided (e.g., provider- or tenant-managed), the vertical customer could have direct control of those resources, introducing uncertainty with regards to the future behavior of the assets provided, which may raise some inconsistency issues.

Mechanisms for a timely analysis of the real network situation (including each particular slice) and quick reaction to fulfil an SLA are then required.

High Customization of the Slice

The vertical industries are different in nature, as described before. This implies that operators should deal with a large variety of needs and requirements to accommodate the vertical services, introducing an inevitable high customization at the time of provisioning network slices, tailored to each of the requested services.

The variety of services can be categorized according to different dimensions:

- **Major type of requested service:** The author has already discussed the major types of services to be supported by 5G networks, such as eMBB, uRLLC, and mMTC. Furthermore, some additional types could be considered, such as pure connectivity services (as an evolution of existing corporate communication services, for example), or even NFVI-as-a-Service services (as an evolution of the IaaS kind of services currently available in cloud environments). Finally, the same vertical customer can subscribe to services pertaining to more than one of these major types, and then require some interconnect of the provisioned slices to offer a smooth service experience for the vertical customer.
- **Particular type of requested service:** the major type of services described above can be seen as some kind of macroscopic description of the main characteristics of the service. Beyond such description, particular services will require specific conditions that eventually motivate a distinction at the time of provisioning the required slice. For instance, latency requirements of uRLLC services can differ

allowing for more relaxed conditions in some cases (Cominardi, Contreras, Bernardos, & Berberana, 2018).

- Vertical customer differentiation: obviously, a competitive industrial market will look for differentiation as a form of a competitive advantage compared to similar vertical competitors. Then, specific requirements coming from different customers pertaining to the same industrial sector at the time of requesting a network slice can also be expected.
- Location: it is also clear that different vertical customers will also have distinct geographical footprints, hence the need for extending the network slice coverage in a different and customer-specific way.
- Vertical service evolution: finally, the business of each vertical customer can evolve at a different pace, requiring upgrading and scaling resources at different speeds and with distinct constraints, including the need for invoking various kinds of slice instances over time.

As a consequence, the operator must endow itself with automation mechanisms as a means to simplify and foster the provisioning of the slices. These mechanisms will reside in the orchestration and management artifacts used to handle the resources (either physical or virtual) constituent of the slices. Programmability and virtualization (of functions and infrastructure) will be essential to reduce time to market, which essentially is the same as reducing the time to revenue. They will be essential to guarantee chances of capturing the market of 5G vertical services too, since conventional forms of service provision will not sufficiently scale to support this high customization (in terms of flexibility, speed, granularity, etc.).

Service Segregation

The critical importance of ensuring isolation among the slices provided and supported on top of the same infrastructure has already been mentioned.

The need for service segregation has been satisfied in the past by different means. A very basic option consisted in physically separating network infrastructures, each one being dedicated to enterprise, fixed residential, and mobile services, respectively. This design is clearly neither cost-efficient nor sustainable anymore.

Alternatively, logical separation has been also performed by means of the deployment of overlay solutions, e.g., in the form of VPNs. However, traditional tools and mechanisms for VPN provisioning are neither flexible nor agile, as previously pointed out in the high customization discussion.

Once again, it is required to evolve towards scenarios where dedicated resources (including network functions) can be automatically allocated according to the needs of the service, ensuring isolation, on top of the same infrastructure.

SURVEY ON SLICING INITIATIVES IN STANDARDIZATION

An important number of Standards Development Organizations (SDOs) and some other industrial associations are looking at the network slice concept from different angles and perspectives. From an operator's point of view, there is a risk of fragmenting the conceptual approach to network slices, since small differences can provoke incompatibilities among the different approaches. It is therefore necessary to reach consensus on common terms, definitions, rationale, ideas, and goals in order to properly normalize the network slice concept.

The NGMN Alliance provided a primary description of the network slice concept as mentioned in the introductory section. The NGMN view is that of a 5G slice as a composition of a collection of 5G

network functions and specific Radio Access Technology (RAT) settings that are combined for the specific use case or business model while leveraging NFV and SDN concepts. The network slice concept is organized in a layered manner (NGMN, 2016), differentiating the service instance layer, comprising the end-user of business services; the network slice instance (NSI) layer, as a set of functions forming a complete instantiated logical network; and the Resource layer, consisting of both physical and logical resources. In this layered view, the NSIs can be potentially shared among multiple service instances.

The 3GPP differentiates (3GPP TS 23.501, 2017) among network slices and network slice instances. On one hand, a network slice represents a logical network providing specific network capabilities and network characteristics. On the other hand, a network slice instance is defined as a deployed network slice, that is, a specific set of network function instances and associated resources.

ETSI NFV maps some of the ideas of 3GPP onto its own architectural framework (ETSI GR NFV-EVE 012, 2017) concretely the network slice management artifacts defined as Communication Service Management Function (CSMF), which is responsible for translating communication service requirements into network slice requirements and the Network Slice Management Function (NSMF), which is responsible for the management (including lifecycle) of NSIs. These management functions are considered to be part of the OSS/BSS segment in ETSI NFV framework.

The BBF is also approaching network slicing (BBF, 2018) by augmenting the previous management functions by defining new and complementary ones, like Access Network Slice Management (ANSM), Core Network Slice Management (CNSM), and Transport Network Slice Management (TNSM). Each of them is intended to take care of the slice lifecycle management of each particular network slice sub-instance (i.e., access, core, or transport).

Also, ITU-T has defined (ITU-T, 2011) the concept of LINP composed of multiple virtual resources (i.e., abstraction of physical or logical resources), which is actually a realization of a network slice. The LINPs are isolated from other LINPs, having their own programmable control plane and data plane.

Last, the IETF is approaching the ideas of network slicing exploring management frameworks and architectural models (Geng et al., 2018).

CHALLENGES FOR NETWORK OPERATORS

The fundamental aspect of network slicing is that each slice will behave as if it were an independent network. Taking this into account, this section describes a list of key challenges from a provider's perspective.

Scalability

Scalability in the context of network slicing exhibits two dimensions: the scalability of the slice itself, and the overall scalability, as a function of the total number of slices.

The first dimension is related to resource allocation and accounting (including resources necessary for satisfying protection and availability), directly related with the kind of service requested and negotiated with the customers. Furthermore, the deployed slices can scale up or down over time, depending on several factors that include commercial success and optimization, e.g., following seasonal demand.

This scalability dimension requires orchestration mechanisms that dynamically add or remove assets to the slice in a consistent manner, based upon either computing or networking capabilities (or a combination thereof). Additionally, on the customer's side, these new resources have to be accounted as part of the provided solution. If the customer is responsible for the control and the management of the

provided resources, then those that have been introduced (to scale up) or removed (to scale down) have to be added or removed according to the decisions applied by the control and management functions of the customer. Charging should also be adapted dynamically according to the consumed capabilities at any given time.

The second scalability dimension refers to the global scalability that the operator can support in terms of quantity and types of slices to be orchestrated and managed. A too much fine-grained offering of slices can provoke an unmanageable number of artifacts to be orchestrated by the provider, making them unpractical. Some kind of aggregation or grouping will be needed for achieving tractability. Scalability can be much impacted by the number of external tenant-managed slices offered.

In order to partition network resources in a scalable manner, it is required to clearly define to what extent slice customers can be served or not by an existing slice. A proper application of different SLAs with the translation of service parameters into network ones (including compute needs) will be essential to understand to what extent a new demand can be accommodated with existing slices, from the service point of view. In addition, if the customer requires the responsibility of control and management capabilities, then the customer-specific “individualization” of the slice is a must.

Arbitration

In order to resolve conflicts and to ensure negotiated service levels, the provider needs to incorporate some arbitration mechanisms to allow an efficient usage of resources (including functions), preventing on the one hand resource over-dimensioning, and on the other hand service degradation or disruption. These mechanisms have to be in place not only among the different slices that are being deployed over the same infrastructure, but also within the individual slices themselves (like the three main classes described in the above categorization, as shown in Figure 4), since the relationship between a customer and a slice is not necessarily 1:1 (unlike the external, provider-managed slice case).

Arbitration will have to be applied not only to slice creation or customer activation, but also (and more importantly) when scaling and/or failure events happen, so resources are properly (re-)assigned according to the applicable SLAs. Such inter-slice arbitration may negatively influence the performance of other slices that share the same infrastructure, hence the critical importance of adequate arbitration solutions. Note that SLAs should enclose technical clauses which govern slice availability and the expected maximum duration before getting the slice running as expected. Such clauses need also to be taken into account.

The role of arbitration is to some extent equivalent to the role that existing QoS mechanisms play on current networks. QoS is primarily effective in situations of network congestion, when the availability of resources becomes compromised and their scarcity has to be managed to minimize any kind of impact on the service delivery. Similarly, arbitration needs to be in place when events in the network limit the availability of the resources that compose the different slices supported by the network, or their internal components.

This arbitration capability can collide with the requirement of slice isolation. That is, the arbitration of resources should maintain the principle of isolation among slices, to avoid any kind of degradation or interference from one slice to another, especially when such isolation is imposed by the associated SLA. The arbitration is foreseen as an internal capability of the operator, transparent to the customer that can influence the decisions and arbitration criteria only through the negotiated SLAs.

Finally, it can be expected that some prioritization could happen in the event of massive failure or outage. The criteria for establishing priorities could be diverse, ranging from the type of service to be ensured, the customers to be protected, the percentage of slice affected, the critical SLAs to be guaranteed, the

associated penalties, etc. Commercial, regulation and security aspects could motivate distinct options to be taken into consideration by each operator.

Slice Planning and Dimensioning

Over-dimensioning has often been the default design principle to deploy and operate networks for avoiding any kind of congestion. Through slicing, the location of the traffic sources and destinations becomes much less predictable, if predictable at all. This is especially relevant for the case of external tenant-managed slices, where the final decision of where to deploy traffic sources and destinations (as well as some intermediate service functions that could alter or modify the traffic profile) lies in the hands of the customer.

Two different time scales can be distinguished during these processes: microscopic and macroscopic planning and dimensioning. Microscopic planning and dimensioning refers to the process applied to each individual slice. On the contrary, macroscopic planning refers to the global process applied to the overall infrastructure, including resources and functions.

In the microscopic case, a primary source of information for the dimensioning process will be the SLA agreed with the customer, which specifies the expectations for the service to be provided. A number of network and compute KPIs will be derived from the service parameters expressed in the SLA. These will serve as inputs for determining and tweaking the required resources for honoring the requested service (as well as for identifying engineering parameters of the slice, like redundancy, etc.). This for sure is needed for data plane related resources, but also for control plane resources, including associated licenses or processing needs (e.g., for database dimensioning purposes). A certain level of over-allocation/overbooking can be expected in order to mitigate issues related to unexpected demands, failures, etc. Additionally, some excess could be derived from the resource quantization process (e.g., bandwidth granularity allocated in Gbps units, or licenses allocated in blocks of thousands of users).

The planning in the microscopic case will require some (traffic) forecast inputs from the slice tenant, either internal or external. The formalization of those inputs could vary, from being part of the negotiated SLA (renewed periodically) up to a different/specific administrative process. Adherence to the planned resources derives from a commercial obligation for the operator towards the customer. Then, an assessment of the allocated capabilities should be performed as well. The concept of in-operation network planning (Velasco, Castro, King, Gerstel, Casellas, & López, 2014) can be also assumed as valid for the network slicing case as a means to adjust any kind of resources to the observed demand evolution.

At a macroscopic level, the dimensioning process will take into consideration all the slice demands, the ones already in place plus the slices that are being instantiated. A punctual need coming from an aggregated demand could require borrowing resources from other slices (e.g., capabilities allocated for resource protection purposes) in order to satisfy the overall demand while the operator builds the necessary infrastructure for all the demands. Furthermore, overbooking is certainly an option when the probability of congestion is actually low or the scarcity of resources does not allow for an immediate upgrade of the slice in terms of resources.

The planning procedures will also take into consideration the forecasted evolution of the existing slices plus the evolution of the (marketing and business) plans provided by the commercial and service units. The multi-tenancy approach facilitated by the slicing concept is fundamental to define rational investments for serving the expected demands, but also the programmability and virtualization techniques will be solicited to make a better use of the available resources by reconfiguring services in a flexible manner.

Proper planning, dimensioning, and enforcement are needed to make the transition towards this new form of service sustainable, starting with an appropriate data collection on resource usage, especially the virtual ones which need abstract models for usage reporting.

Multi-domain

Multi-domain can be interpreted in different ways, since the notion of “domain” can reflect different concepts. For slice provisioning, two meanings are considered: technological domain, taking into account the applicability of slicing to different technologies, and administrative domain, where more than one provider is involved in the provisioning of network slice(s) to a customer.

Slice-based services will necessarily require the integration of different technology domains. The complete end-to-end nature of slices involves distinct computing environments and transport technologies (data switching, optical, etc.), and linking them will require a consistent orchestration approach. For those services making use of radio access technologies, the slice concept has also to be extended to the Radio Access Network (RAN) and the Core Network (CN), with their own slicing specificities (Kaloxylos et al., 2018).

The deployment of slices in the networking technological domains will be necessarily different since the forms of traffic segregation are specific to each of the technologies. Furthermore, a combination of logical (e.g., VLAN tag-based in Ethernet systems, label-based in MPLS, etc.) and physical (e.g., lambda-based in optical, slot-based in Flexible Ethernet, etc.) separation methods can be expected. Complementary computing capabilities have to be added with their own form of separation (e.g., hypervisor-based or container-based). The form of combining and integrating them for a single slice will require the combined action of multiple controllers, including an overarching system that maintains a global, systemic view of all the resources, including those that are available.

The interaction with different technologies is clearly fundamental for slice provisioning. Notwithstanding, vertical customers could require no restriction for the slice requested in terms of coverage, service capability, resource constraints, geographical footprint, etc., avoiding any potential limitation of the network provider with whom they have a commercial relationship as their privileged provider. This leads to the necessity of enabling multi-domain slicing, which implies functional and commercial interfaces to be normalized for the sake of massive adoption.

Both business and technical implications can be deemed necessary for such multi-operator slice provisioning context. From the business side, the following implications can be listed:

- Coordination models: there is a need for business coordination to define how multiple stakeholders interact for the provisioning of a multi-domain slice in order to trade low-level resources and elementary services combined and orchestrated to deploy slices end-to-end.
- Inter-provider SLAs: each provider in each administrative domain should have its own SLA assessment capabilities internally, including interfaces with SLA aggregation components that automate the multi-domain aggregation process.
- Pricing schemes: bilateral negotiation can be expected as a regular mechanism to establish pricing agreements. Even for simple pricing formulas, the values of the parameters under consideration should be dynamically adapted, e.g., according to demand or resource/service availability.
- Service specification and customer facing advertisement: each provider may consider not only its own capabilities in its domain but also slice offerings and capabilities available in neighboring domains. Therefore, catalogue synchronization is required to be performed across domains.

From a technical standpoint, implications include:

- Multi-domain orchestrator: from the provider-to-provider's viewpoint, only certain entities within each domain should interact with each other for handling the inter-domain activities in order to keep the slice provisioning consistent end-to-end. Those entities can be identified as multi-domain orchestrators. This kind of orchestrator will be in charge of abstracting the underlying infrastructure in its domain before it announces (to neighboring providers) what capabilities and functions the operator can provide.
- Slice decomposition: The vertical customer will request a slice to a provider, the latter becoming the origin provider for that customer. The origin provider should then incorporate sufficient logic for decomposing the slices across the different domains.
- Discovery of domains: although manual configuration can be used, automatic procedures are desirable for speeding up service provisioning in the network softwareized era.
- Common abstraction models: a common understanding of the description of the resources (i.e., network, compute and storage) and the capabilities per domain is needed.
- Standard interfaces, protocols, and APIs: these will be required for remote control and management of functions and slices in other domains.

Orchestration and Control of the Slices

The request for network slices that is sent by customers will require the orchestration of resources in order to address the said request. SDN and NFV techniques can be used by an operator to orchestrate slices (Ordonez-Lucena, Ameigeiras, López, Ramos-Munoz, Lorca, & Folgueira, 2017). Such orchestration needs to have full control and visibility of the nodes, topology, functions, and capabilities (such as bandwidth or compute power) to make decisions. Example situations can be found in (Bryskin et al., 2018).

End-to-end slices are a service management issue. The enablers of management and orchestration are the usage of open and standard interfaces for interacting with the nodes and systems, as well as the definition of normalized models for service and devices. An overview of network management and orchestration can be found in (Contreras et al., 2018).

The result of the orchestration process is the allocation of the resources, as well as the management of their lifecycle, including service assurance and fulfillment. Then, the constituent blocks of the slice are controlled by the operator.

Different customers can require distinct levels of control for the resources they have requested to the provider. Extreme cases can be customers that do not require any capability of control and management of the allocated assets (just pure communication service), and on the other end, customers requiring full control of their assets. A gradual level of control can be found in between.

Then, the operator should provide configuration and administration capabilities to the customers according to the levels of control that they request. These capabilities could come by simply exposing some interfaces for that required control actions (e.g., APIs), up to granting direct access to the resources (e.g., IP address to access the element console). The more abstracted way, the less invasive for the operator.

3GPP (3GPP TR 28.801, 2017) defines a number of management functions needed to manage network slices to support communication services. These functions are:

- Communication Service Management Function (CSMF), which is responsible for translating the communication service-related requirements into network slice-related requirements.
- Network Slice Management Function (NSMF), which is responsible for the management and the orchestration of an instance of a network slice.

- Network Slice Subnet Management Function (NSSMF), which performs the same task as the NSMF, but at a sub-instance level.

Both NSMF and NSSMF can be considered as functionally similar. ETSI NFV (ETSI GR NFV-EVE 012, 2017) has mapped these management functions with ETSI's NFV orchestration framework. The resulting mapping locates this functionality as part of the broader OSS/BSS components.

Regarding the programmable control of the slices, two levels of SDN control can be considered for a given slice. On one hand, there is the tenant's controller that permits to configure the involved functions. On the other hand, there is the infrastructure controller that is used to program the underlying infrastructure resources to provide end-to-end connectivity. These two levels of control facilitate interplay of actions at both service function and connectivity infrastructure levels, enabling flexibility on the slice instantiation and provisioning, assuming cross-layer coordination.

In the NFV framework, such a double control level is proposed in (ETSI GS NFV-EVE 005, 2015) that describes the usage of SDN in NFV environments. Figure 5 shows this approach. The infrastructure controller relates to the Transport SDN controller, while the tenant controller is equivalent to the service controller, as proposed in (Contreras, Bernardos, López, Boucadair, & Iovanna, 2015).

Figure 5. Service and connectivity control

Slice Operation

The operation of slices, once instantiated, share many aspects with conventional networks. However, some new mechanisms and artifacts will be needed. This section considers monitoring and maintenance as relevant aspects of the operation of a network, identifying how the application of maintenance and operational procedures related to a slice can raise new requirements.

Monitoring and performance information is essential for a healthy operation of a network. Monitoring data are usually associated to specific resources, either network ones (e.g., packet errors) or compute ones (e.g., CPU load). This can be complemented with indicators of the service functions that compose the service (e.g., number of active users of the service). Some mechanisms have to be defined in order to properly display and abstract the information for each slice tenant (or user). To this respect, external slices

have a higher degree of complexity since the information to be exposed, and the constraints to access it, have to be defined or even better, agreed between the provider and the customers.

At the time of creating a slice, a number of resources are allocated to a given customer. As a consequence, the monitoring information associated to the allocated resources has to be extracted in order to ensure proper operation of the slice. All or some of them will be presented to the customer in order to provide the necessary information. It is important to preserve privacy as the information related to other slices must not be leaked into other slices. The monitoring and performance-related information would also serve as a reference for assessing the compliance of what has been delivered with what has been agreed in the SLA.

Then, the monitoring information has to be properly processed to be provided to the customer (possibly according to a specific format and semantic). Only the information strictly associated to the customer's resources has to be exposed, which implies some filtering on the global indicators. Furthermore, slices can be supported on top of virtual resources. This means that the physical resources can change over time while the virtual ones allocated for the slice will appear as unchanged, even though they may have been, for example, migrated from one virtual machine to another. This implies that the monitoring information could be originated from different sources, thereby requiring dynamic aggregation and association during the slice lifetime.

All the received information has to be processed and correlated in the same manner as in today's legacy networks. This implies the replication of operational procedures per slice, hence raising a scalability issue. Two approaches could be followed: (1) initial processing of all the indicators and further personalization per slice; or (2) initial separation per slice of the corresponding indicators for further individual processing and correlation. The first approach seems to be more scalable and practical, but requires more coordination and integration from the OSS/BSS point of view.

Slice Marketplace

The interaction with the vertical customers is a critical feature for understanding the needs of the service to be provided. From that interaction, the provider will obtain the necessary information for creating (or reusing) a slice and mapping the customer service to the allocated resources.

The level of detail in the request sent by the customer will impact the functionality required on the provider's side to address such request. The customer requests could be expressed in terms of service or resources. The former will imply that the request identifies the characteristics of the service to be delivered, without any further details about what is needed in terms of resources to be allocated (also known as intent-based requests). In contrast, the latter will imply that the request details the resources identified as needed by the customer. Clearly, the provider should be able to translate the service semantics into resource semantics for the actual allocation of the slice resources, based upon a computation logic that can take into account the outcomes of the possible negotiation between the customer and the provider as input data, but also the network planning policies, the status of the network, its resources, the location of the end-user, whether the end-user is in motion or not, etc.

Different kinds of customers can have different levels of know-how and skills, thereby determining what approach to follow. From the provider's perspective, service semantic requests can allow to reach a broader market, the one formed by vertical customers neither specialized nor skilled in the knowledge of what communication services are needed for their specific business. Considering the type of slices shown in Figure 4, most probably the external, customer-managed slices will be requested through the resource semantics approach, while the external, provider-managed slices will be requested through service semantics. The support of the different slice classes will require a consistent set of abstractions either at service or resource levels to allow the aforementioned semantics to be expressed consistently. As a consequence, proper abstractions and templates have to be defined to ensure the provisioning of a service

portfolio providing a consistent view of the network and its resources, as well as their integration with the internal network management and orchestration systems.

Security

In any shared infrastructure, security is a key element to guarantee proper operation to each user. Slice customers must be appropriately authenticated, their rights enforced by authorization mechanisms, and the operations they perform accounted for, so that further auditing can be applied in case of any problem. This becomes crucial when considering the external, customer-managed slices, since the possibility of altering the behavior or status of the allocated functions and resources increases.

Each vertical customer will offer services to his/her end-users thanks to the slices that have been deployed (similar to the current MVNO service offerings). This means that the vertical customer has also the responsibility of enforcing security measures in order to protect the service (and the end-users, indirectly), which indirectly affects the security exposure of the allocated resources. The security measures of every vertical customer will be multiple and diverse, thereby requiring the provider to harden the allocated resources in order to prevent whatever issue (e.g., attacks generated from within a slice managed by a tenant). Even if the provider can influence the security implemented by the vertical customer, the control will not be total. Generic guidelines and best (current) practices should be defined and updated according to new threats and security problems as they are met.

Another key issue is the privacy of customer data, as well as the privacy of the end-users' data making use of the service offered by the vertical customer. All this information has to be properly stored and encrypted (whenever applicable) in order to prevent any exposure of such data to the provider in general, or to other customers who use the provider's infrastructure. This is essential when some of the functions and resources can be shared between different customers in the same slice or across slices.

Beyond this, measures have to be in place to proactively detect and mitigate active security attacks, thereby avoiding that a security breach that affects one slice does not propagate into the infrastructure and other slices.

Slice Aging

The same dynamicity for the allocation of slices to tenants, leveraging SDN and NFV techniques, applies as well to the slices' lifecycle. One of the promises of automation is the possibility of invoking and deploying services faster than today, thereby overcoming the currently weak service agility. This facility in the creation of services opens new business opportunities since targeted services, scoping events and situations of short duration, can be made available easily. The duration of the existence of a slice can be certainly variable, so differentiating between short versus long is relative: "long" slices are those created as a semi-permanent service. Hourly, daily, or even seasonal slices can be considered as short-aged ones. Furthermore, the frequency at which a given slice is requested can be another timing parameter to be taken into consideration. Depending on the frequency of slice requests, some resources can be freed or should be kept as booked until a new forthcoming service request shows up (that is, the resources may remain unavailable for any other slice request, except for temporary needs, for example).

This dynamic situation (motivated by allocating and freeing resources), implies that whatever the decision or the action made by the operator with respect to using some resources, the operator should consider not only the resource view and status at the time of the specific request but also over a broader timeline, since any decision at the moment of the request can negatively influence future decisions.

Slices based on calendaring considerations (e.g., day/night operation) will need some guarantees, thus motivating a certain level of resource booking. The matching of the resource availability with the time duration request will introduce more constraints as far as resource allocation for slices is concerned.

Whatever the duration of a slice, the creation and operation of a slice will require a non-negligible number of administrative and technical registers. Administrative notifications, data and billing records, systems configurations, etc., are proportional to the number of slices.

A mix of long- and short-lived slices co-existing on top of the same infrastructure should be expected. This will impact providers in several manners, from resource planning (slice demand forecast including traffic and resources to be consumed) to security (data preservation per tenant).

Slice Isolation

This main requirement for isolation constitutes the essential feature any network service provider has to support in order to deliver slices to its customers. The degree of isolation achieved will be critical in determining the ability of a certain provider to address the different classes of slices discussed above, and how they can be requested and used by its customers.

The isolation has to be applied at different levels, such as control plane and data plane isolation, as well as resource and function isolation. Even the sharing of constituent elements within each of these planes can be allowed, the allocated capabilities have to be segregated in order to avoid any kind of misbehavior induced by any other customer in the system.

Data plane isolation can be achieved by several means, and with different degrees. The encapsulation of data in different tunnels, one per vertical customer, can be a primary measure for achieving such isolation in a shared environment. More extreme situations, like strict allocation of assets (as enabled e.g., by the concept of calendar slots in Flexible Ethernet (OIF, 2017)) will allow the exclusive allocation of a specific amount of resources to slices. It will also be possible to find several options with a higher or lower level of isolation.

For the control plane, the separation can be achieved even by supporting such separation in the actual control plane capabilities or alternatively by replicating control plane capabilities dedicated to specific customers per slice. The first approach requires the control plane element to implement such isolation mechanisms, e.g., via the creation of different virtual spaces per customer. The second approach naturally grants isolation since the different replicas act as independent control elements.

A pairing among the form of managing the data plane and the control plane has to be defined. For instance, sharing the data plane capabilities while dedicating control plane ones can be problematic. Multiple replicas of control elements (i.e., hard isolation at the control plane level) acting on the same data plane elements that isolate traffic by means of encapsulation in an overlay model context (i.e., soft isolation at the data plane level) can lead to inconsistencies, hence jeopardizing the isolation. As for resource isolation, the implications reside in the level of partitioning that the provider is willing to or can implement. Resource isolation can apply to compute nodes, ranging from the dedication of specific compute resources like the bare-metal approach versus the sharing of computing capabilities by means of hypervisors. It can also apply to transport resources, like the allocation of specific lambdas in optical nodes to specific slices versus the accommodation of traffic of different customers carried by the same lambda.

Function isolation will basically consist in the instantiation of separated service function instances per vertical customer versus the sharing of a given service function instance to be used by multiple verticals. An intuitive example could be a firewall. Implications like the number of compute capabilities associated to each option, the connectivity with the rest of the service functions per each customer, the number of licenses to be consumed, etc., have to be taken into consideration.

PROSPECTIVE APPROACHES

This section presents some prospective approaches covering some of the aforementioned aspects.

Slicing Across Multiple Administrative Domains: the 5G-Exchange Project

Even in the existing networks, the majority of the services involve more than one administrative domain, from basic Internet access to services provided by Over-The-Top players, up to inter-domain VPN services deployed for multinational companies. This multi-administrative interaction requires increasing computing capabilities that providers start to offer for internal and external services, and needs solutions for interworking between partners in a standard manner, e.g., for trading and operational procedures as well as resource allocation among providers.

The 5G-Exchange (5GEx) project³ has developed an architectural framework for multi-domain orchestration among providers that enables the overarching service category of SlaaS. The framework is an extension of the ETSI NFV model for orchestration purposes across multiple administrative domains. Initial steps towards the specification of this model have been reported in (ETSI GR NFV-IFA 028, 2018).

Three are the services that can be provided by 5GEx under the umbrella of SlaaS: (1) the NFVI-as-a-Service (NFVIaaS), where one provider makes a specific amount of infrastructure resources available to another provider; (2) the VNF-as-a-Service (VNFaaS), where one provider makes a number of VNFs available to another provider; and, (3) the enriched connectivity service, where assured quality forwarding paths for medium-to-large traffic aggregates and value-added forwarding paths at the flow level are provided. In all these cases, providers also offer the necessary capabilities for the configuration and the administration of the services provided in the remote domain as part of the service.

The starting point for 5GEx has been the idea of enabling a logical exchange (as an analogy with the physical exchange environments that exist nowadays e.g., for peering purposes) for a global and automated orchestration of multi-domain 5G services. The objective is to enable mechanisms for trading resources (including service functions) together with the necessary orchestration, management, and control capabilities between partnering providers, and which are not necessarily directly and physically interconnected.

Figure 6 shows the high-level architecture of 5GEx. The providers that are involved in the delivery of a service each operate a different administrative domain. The interaction among providers is done through multi-domain orchestrators (MdOs). The MdO is a functional extension of the MANO NFVO used for the orchestration of services across multiple administrative domains.

Figure 6. 5GEx reference architectural framework

Each MdO enables the offering of resources of its own domain, and at the same time coordinates the resource and service orchestration in other domains, via the interaction with other MdOs. The domain orchestrators in each of the administrative domains are responsible for performing virtualization, service orchestration, and/or resource orchestration within their domain, and the corresponding tasks are coordinated by the MdO. These domain orchestrators use the abstractions exposed by the underlying resource domains according to a variety of technologies, both for network and compute resources.

There are three main interworking interfaces identified in the 5GEx architecture framework. The interface I1, which is focused on Customer-to-Business (C2B) interactions, exposes service specification APIs allowing business customers to specify their requirements for a service to the MdO. The interface I2, which is scoped for Business-to-Business (B2B) relationships, allows the interactions among MdOs of distinct providers to request and orchestrate resources across administrative domains. Lastly, the interface I3 permits the orchestration of resources within the same administrative domain through the interaction between each particular MdO and the internal domain orchestrators.

Figure 7 presents the detailed 5GEx functional architecture where a number of components are identified for enabling the multi-domain SaaS provision. In this case, all the providers are assumed to support the same components and modules, although in Figure 7, the complete view is only shown for the provider on the left for the sake of simplicity.

Figure 7. Functional architecture of 5GEx

Hereafter are briefly described some of the most relevant components of Figure 7. A full description can be found in (5GEx, 2016) (5GEx, 2017).

- The *Inter-Provider NFVO* is the NFVO that implements multi-provider service decomposition, and is responsible for performing the end-to-end network service orchestration. The Network Service Orchestrator (NSO) and Resource Orchestrator (RO) capabilities are contained here.
- The *Topology Distribution* module exchanges topology information with its peer MdOs.
- The *Resource Repository* keeps an abstracted view of the resources at the disposal of each domain reachable by the MdO.
- The *Topology Abstraction* module performs topology abstraction by elaborating the information stored in the Resource Repository and Topology Distribution modules.
- The *SLA Manager* is responsible for reporting the performance of its own partial service graph (that is, its piece of the multi-domain service).
- The *Policy Database* that contains policy information (for instance, topology information to be advertised to other providers).
- The *Resource Monitoring* module dynamically instantiates monitoring probes on the resources of each domain involved in the implementation of a given service.
- The *Service Catalogue* is in charge of exposing the available services to customers and to other MdOs operated by other providers.
- The *MD-PCE* (Multi-Domain Path Computation Element) is devoted to compute traffic-engineered paths and to set up the connections between domains.

From an interface perspective, the functional split is related to service management (-S functionality), VNF lifecycle management (-F), catalogues (-C), resource topology (-RT), resource control (-RC), and monitoring (-Mon). Candidate solutions have been proposed for the implementation of these interfaces in (5GEx, 2016) (5GEx, 2017).

The left MdO is the entry point for a service request coming from the customer, through the I1 interface. Using I1-C, I1-S and I1-F, the vertical customer will be able to request VNF instantiation and

configuration, apart from expressing how these VNFs should be chained by means of a service forwarding graph.

The MdO will decompose the service by the NFVO of provider A. The NFVO will make use of resources offered by other providers if provider A cannot satisfy the service needs itself. The availability of resources from other parties is collected via I2-RT, and the availability of services offered by such parties is obtained through I2-C.

The MdO from provider A will make use of I2-S and I2-RC for requesting and controlling the necessary resources and services. This MdO will also make use of I3 interface for governing its own resources accordingly, in a similar manner.

In order to accomplish the negotiated SLA between the parties (both the customer and the entry provider, and the providers participating to the delivery of the end-to-end service), monitoring capabilities are deployed and monitoring information is collected, using I1-Mon, I2-Mon, and I3-Mon for the respective capabilities.

As a reference of the different roles in the exchange, note that provider B (the one in the middle) participates to the delivery of the end-to-end service only by providing data plane connectivity between providers A and C.

Preparing the Network to be Consumed in the Form of Slices: the 5G-TRANSFORMER Project

The delivery of services to vertical customers should be facilitated, so that they do not have to worry about the actual (technological) implementation of a service. The 5G-TRANSFORMER project⁴ explores how to progress the current state-of-the-art developing solid and efficient solutions for orchestration and management of the vertical services.

The architectural proposition of the 5G-TRANSFORMER is built upon three major components: (i) Vertical Slicer (VS), (ii) Service Orchestrator (SO), and (iii) Mobile Transport and computing Platform (MTP), as represented in Figure 8. These three modules aim to allow any vertical industry to obtain an end-to-end slice tailored to its needs.

The vertical slicer is the common entry point for all the vertical customers into the system. Services for vertical customers are offered through a high-level interface that is designed to allow the verticals to focus on the service logic and requirements, without the need to worry about how they are eventually deployed at the resource level that is, providing service semantics for the service request. The VS offers a catalogue of vertical service blueprints, based on which the service requests are invoked by the vertical customer. After the appropriate translation of service requirements into slice-related requirements, a decision is made at the vertical slicer level on whether the service is included in an already existing slice or if a new one needs to be created. The network slice manager is the core component of the VS, and it is in charge of the lifecycle management of network slice instances.

The vertical slicer is the component of the system that is aware of the business needs of the vertical, its SLA requirements, and how they are satisfied by mapping them with existing slices. In case that one customer is served with more than one slice, a prioritization policy may be enforced between slices delivered to such customer, based upon considerations about the agreed SLA.

The VS sends request towards the SO to create or update the NFV-based network services (NFV-NS) that implement the slices. In turn, such NFV-NS will be created or updated through a Network Service Descriptor (NSD), which is a service forwarding graph composed of a set of service functions (e.g., in the

form of VNFs) chained with each other, and the corresponding fine-grained instantiation parameters (e.g., deployment flavor) that are sent to the SO.

Figure 8. 5G-TRANSFORMER system architecture

The SO is responsible for providing the end-to-end orchestration of the NFV-NS across multiple administrative domains by interacting with the SOs of other administrative domains, besides interacting with the local MTPs. The SO interacts with SOs of other federated administrative domains in a peer-to-peer fashion to make decisions about the end-to-end (de)composition of virtual services.

Even if a service is deployed across several administrative domains, a vertical customer still uses a single VS to access the system, the one of the primary provider. The federation is opaque to the vertical customer, which has no notion of the multi-domain provisioning resources, since the SO hides this information.

The SO embeds the network service orchestrator (NFV-NSO) and the resource orchestrator (NFVO-RO) with functionalities equivalent to those of a regular NFV orchestrator. Since the slices handled at the VS will in general serve complex end-to-end services, the corresponding network service will be generally a composition of nested NFV-NSs. The lifecycle management of this complex NFV-NS is the role of the NFV-NSO.

Eventually, a resource-related request is generated towards the underlying NFVO-RO to assign virtual resources towards the deployment of the (constituent) NFV-NS. The NFVO-RO functionality of the SO handles resources coming from the local MTP (physical or virtual) and from the SOs of other administrative domains (in an abstracted manner).

The MTP is responsible for the orchestration of resources and the instantiation of VNFs over the infrastructure under its control, as well as for the management of the underlying physical transport network, and computing and storage infrastructures. In general, there will be multiple network/resource segments inside an MTP (e.g., data centers, mobile network, or wide area network).

Regarding the way the customer requests services, the vertical slicer offers Vertical Service Blueprints (VSB) and Descriptors (VSD), enabling the modeling of vertical services by using parameters adapted to the information and capabilities of the vertical (i.e., model services without detailing the infrastructure resources required to support such service). 5G-TRANSFORMER uses an extended version of ETSI NSDs (ETSI GS NFV-IFA-014, 2018) as a notation for network slice templates.

To instantiate a vertical service, a vertical first selects a blueprint, complements the blueprint with the missing information to prepare a VSD. The vertical slicer maps this with an NSD, which describes the network slice for this vertical service. The VSBs are then parameterized versions of VSDs. Details on VSBs, VSDs, and gaps compared to ETSI NSDs can be found in (5G-TRANSFORMER, 2018).

Lightweight Slicing Methods for Limited environments: the NECOS Project

The multiplicity of services with a variety of properties and characteristics will imply the need for deploying functions and using resources spread across the network. These environments can present episodes of resource scarcity either temporary (e.g., for unexpected events) or almost permanently (e.g., because of limitations on space and energy towards the very edge of the network). Developing forms of creating and operating slices in a lightweight mode is essential to make the infrastructure compatible with these operational situations.

The NECOS project⁵ aims to realize an integrated platform encompassing cloud and network management, service orchestration, and distributed resource monitoring to support slicing across cloud capabilities. The target is to propose a lightweight approach for automating the process of optimal cloud slicing configuration, identified as Lightweight Slice Defined Cloud (LSDC). The purpose of LSDC is to automate the process of cloud configuration by creating cloud-based slices across all the available resources in a set of federated data centers. The objective is also to provide a uniform management of the currently separated computing, connectivity and storage resources. The main characteristics of LSDC are:

- Slice-as-a-Service as a service model to be offered to customers. A slice as a grouping of physical or virtual (network, compute, storage) resources can act as a sub-cloud, sub-network and can accommodate service components, seemingly independent of other slices. The management software can dynamically map service components on top of the already available cloud platform features and functions, creating slices on-demand, and (re)configuring them as appropriate to provide the end-to-end service. The slice management takes over the control of all the service components, virtualized network functions and system programmability functions assigned to the slice.
- Configuration of slices across physical resources spanning multiple cloud infrastructures (from large centralized data centers to the computing capabilities located at the edge), in order to better accommodate the various service demands. Adaptation and reconfiguration are done at the slice level, rather than the whole cloud. This configuration is achieved by using specially-designed software that can describe and manage the various aspects of slices within the cloud environment. The actions performed in each aspect that comprises the cloud environment – from the networking between virtual machines, to the SLAs of the hosted applications – are to be managed via software. This reduces the complexity related to configuring and operating the infrastructure, which in turn eases the management of the cloud infrastructure. Such infrastructure tends to expand at large scales, and is commonly composed of thousands of servers and network elements, which support tens of thousands of virtual machines, virtual networks, and applications.
- Lightweight and uniform management and virtualization systems as the target design, with small footprint components, and which can be deployed over a large number of small servers and cloud systems both at the core and the edges of the network. These lightweight elements enable the integration of core data centers and mobile edge into cloud networks.

As service providers offer more applications on the cloud and as consumers' demand increases at an exponential rate, eventually even the largest cloud infrastructure will run out of computing, networking, and storage resources. While cloud infrastructures can be over-provisioned to handle spikes on demand, this is hugely expensive and not scalable.

Figure 9. NECOS platform

Figure 9 depicts the platform design that represents the LSCD concept and which exposes interfaces for resource allocation and service deployment purposes. It includes as components the Cloud Manager, the Network Manager, and the Control Element for VMs, and the Service Orchestration.

This view is complemented by Figure 10 which depicts the NECOS native integration of cloud computing and advanced networking enabling cloud networking slicing capabilities in multi-domain scenarios.

Figure 10. NECOS high-level architecture

CONCLUSION AND FUTURE DIRECTIONS

Network slicing is a promising approach for the provisioning of network services, and it aims to deal with the diversified demand of 5G services with very stringent requirements, and which ambitions to support new wholesale offerings. While leveraging recent but well-established technology substrates such as NFV and SDN seems a straightforward path to achieve network slicing, there are still several important challenges that have to be addressed to make the concept operationally feasible, viable and deployable for providers.

In this chapter, the author has discussed a number of issues and challenges that have to be taken into consideration for ensuring practical operability of the slicing solutions. Some other points can be equally important, but have not been addressed for the sake of conciseness. These are aspects like the interface between the customer and the provider, details on the interfaces for federating infrastructures, regulation, roaming, charging and billing, etc.

Network slicing is currently more an objective than a reality since the implications of creating a network (the slice) within a network (either physical or virtual) affects all the dimensions of current service delivery procedures. Control, management, and data planes are affected by these changes, thereby making their evolution intricate and inter-related.

The final target can be similar to what is depicted in Figure 11: an infrastructure operated as a network that can host some other networks within, these networks being run and operated independently, with the necessary isolation, and probably in conjunction with other resources and functions offered by alternative providers.

Figure 11. End-to-end slicing of the operator's infrastructure

Coordinated orchestration of network and compute capabilities is necessary but not sufficient. The exposure of control and management capabilities to the vertical customers through standard interfaces will be the next step in order to cover the business expectations generated by the broad concept of network slicing. Supporting dynamic means to capture slice requirements will accelerate and ease the slice provision (including foster recursive slice instantiation when multiple domains and providers are to be solicited).

ACKNOWLEDGMENT

This work has been partly funded by the European Commission through the H2020 projects 5GE-Ex (Grant Agreement no. 671636) and 5G-TRANSFORMER (Grant Agreement no. 761536), and EUB project NECOS (Grant Agreement no. 777067). This information reflects the consortia views, but neither the consortia nor the European Commission are liable for any use that may be done of the information contained therein.

REFERENCES

- 3GPP TS 23.501 (2017). *System Architecture for the 5G System*.
- 3GPP TR 28.801(2017). *Study on management and orchestration of network slicing for next generation network (Release 15)*.
- 5G-PPP (2015). White papers on vertical industries. Retrieved from <https://5g-ppp.eu/white-papers/>
- 5G-PPP (2016). *5G empowering verticals*. Retrieved from https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf
- 5G-TRANSFORMER (2018). *Definition of vertical service descriptors and SO NBI*. Deliverable 3.1.

5GEx (2016). *5GEx Initial System Requirements and Architecture*. Deliverable 2.1.

5GEx (2017). *5GEx Final System Requirements and Architecture*. Deliverable 2.2.

Analysys Manson (2017). *Telefónica's UNICA architecture strategy for network virtualization*. Retrieved from <http://www.analysysmason.com/telefonica-UNICA-architecture-strategy-for-network-virtualisation-report>

BBF (2018). *End-to-End Network Slicing*. SD-406 (work in progress).

Boucadair, M., Lévis, P., Griffin, D., Wang, N., Howarth, M., Pavlou, G., Mykoniati, E., Georgatsos, P., Quoitin, B., Rodríguez Sánchez, J., & García-Osma, M.L. (2007). A Framework for End-to-End Service Differentiation: Network Planes and Parallel Internets. *IEEE Communications Magazine*.

Bryskin, I., Liu, X., Guichard, J., Lee, Y., Contreras, L.M., Ceccarelli, D., & Tantsura, J. (2018). *Use Cases for SF Aware Topology Models*, draft-bryskin-teas-use-cases-sf-aware-topo-model-03 (work in progress).

Cominardi, L., Contreras, L.M., Bernardos, C.J., & Berberana, I. (2018). Understanding QoS applicability in 5G transport networks, in *Proc. of the 13th IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Valencia, Spain.

Contreras, L.M., Bernardos, C.J., López, D.R., Boucadair, M., & Iovanna, P. (2015). *Cooperating Layered Architecture for SDN*, draft-contreras-sdnrg-layered-sdn-04.

Contreras, L.M., Doolan, P., Lønsethagen, H., & López, D.R (2015). Operation, organization and business challenges for network providers in the context of SDN and NFV. *Elsevier Computer Networks*, Vol. 92, pp. 211-217.

Contreras, L.M., López, V., Vilalta, R., Casellas, R., Muñoz, R., Jiang, W., ... Toka, L. (2018). Network management and orchestration. In P. Marsch, Ö. Bulakci, O. Queseth, & M. Boldi (Eds.), *5G System Design: Architectural and Functional Considerations and Long Term Research*, John Wiley & Sons.

Doverspike, R.D., Ramakrishnan, K.K., & Chase, C. (2010). "Structural Overview of ISP Networks. In C.R. Kalmanek, S. Misra, Y.R. Yang (Eds.), *Guide to Reliable Internet Services and Applications*, Springer-Verlag, 2010.

ETSI GR NFV-IFA 028 (2018). *Report on architecture options to support multiple administrative domains*. V3.1.1.

ETSI GR NFV-EVE 012 (2017). *Report on Network Slicing Support with ETSI NFV Architecture Framework*. V3.1.1.

ETSI GS NFV-IFA-014 (2018). *Management and Orchestration; Network Service Templates Specification*. V2.4.1.

ETSI GS NFV-EVE 005 (2015). *Report on SDN Usage in NFV Architectural Framework*. V1.1.1.

Galis, A., Abramowicz, H., Brunner, M., Raz, D., Chemouil, P., & Pras, A. (2009). Management and Service-aware Networking Architectures (MANA) for Future Internet - Position Paper: System Functions, Capabilities and Requirements. *IEEE 2009 Fourth International Conference on Communications and Networking in China –ChinaCom*.

Galis, A., Denazis, S., Brou, C., & Klein, C. (ed) (2004). *Programmable Networks for IP Service Deployment*. Artech House.

Geng, L., Qiang, L., Ordonez, J., Adamuz-Hinojosa, O., Ameigeiras, P., Lopez, D., Contreras, L.M. (2018). *COMS architecture*, draft-geng-coms-architecture-02 (work in progress).

GSMA (2017). *An introduction to network slicing*. Retrieved from <https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf>

ITU-R (2015). *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*. Recommendation M.2083-0. Retrieved from https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

ITU-T (2011). *Framework of network virtualization for future networks*. Recommendation Y.3011.

ITU-T (2012). *Framework of Network Virtualization for Future Networks, Next Generation Network –future Networks*. Recommendation Y.3011.

Kaloxyllos, A., Mannweiler, C., Zimmermann, G., Di Girolamo, M., Marsch, P., Belschner, J., ... Nikaein, N. (2018). *Network Slicing*. In P. Marsch, Ö. Bulakci, O. Queseth, & M. Boldi (Eds.), *5G System Design: Architectural and Functional Considerations and Long Term Research*, John Wiley & Sons, 2018.

NGMN (2015). *5G White Paper*. Retrieved from https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

NGMN (2016). *Description of Network Slicing Concept*.

Odini, M.-P. (2017). *SDN and NFV Evolution Towards 5G*. IEEE Softwarization. Retrieved from <https://sdn.ieee.org/newsletter/september-2017/sdn-and-nfv-evolution-towards-5g>

OIF (2017). *Flexible Ethernet Implementation agreement*. OIF-FLEXE-01.1.

Ordonez-Lucena, J., Ameigeiras, P., López, D., Ramos-Munoz, J.J., Lorca, J., & Folgueira, J. (2017). Network slicing for 5G with SDN/NFV: concepts, architectures, and challenges. *IEEE Communications Magazine*.

Velasco, L., Castro, A., King, D., Gerstel, O., Casellas, R., & López, V. (2014). In-Operation Network Planning. *IEEE Communications Magazine*.

PRE-PRINT VERSION

KEY TERMS AND DEFINITIONS

5G: fifth generation of mobile systems conceived to support the needs of an hyper-connected society demanding very high data rate access, requiring a wider coverage, and offering an increasing number of almost permanently connected devices.

Network slice: logical division of network resources (either physical or virtual) and service functions for different services sharing the same network infrastructure.

Network Functions Virtualization: networking approach that envisages the instantiation of service functions on commodity hardware, breaking the traditional monolithic approach of functional software and hardware.

Network Functions Virtualization Infrastructure: set of compute, storage, and networking assets hosting and connecting the virtualized service functions as enabled by the NFV architecture.

Orchestration: coordination of systems and actions that permit the instantiation of functions and/or the control and allocation of resources in order to compose a network service.

Software Defined Networking: networking approach that proposes the decoupling of both the control and data planes in network equipment and logically centralizing that control while leaving the network element to forward traffic, and enforcing policies according to instructions received from a controller.

X-as-a-Service: on-demand instantiation of service X by a given provider, as requested by a customer of such provider.

ENDNOTES

¹ Some research works introduced seminal ideas on network slicing even before, like (Boucadair et al. 2007), (Galis, Denazis, Brou, & Klein, 2004), (Galis et al., 2009).

² Tenant and customer are used indifferently through the chapter as the consumer of the network slice, as opposed to end-user, which is referred as the user served by the tenant.

³ <http://www.5gex.eu/>

⁴ <http://5g-transformer.eu/>

⁵ <http://www.h2020-necos.eu/>

PRE-PRINT VERSION