Yousaf, F.Z, et al. Network slicing with flexible mobility and QoS/QoE support for 5G networks, *in: 2017 IEEE International Conference on Communications. Workshops (ICC Workshops) [Proceedings],* 7 Pp. July 2017

DOI: https://doi.org/10.1109/ICCW.2017.7962821

Network Slicing with Flexible Mobility and QoS/QoE Support for 5G Networks

Faqir Zarrar Yousaf, Marco Gramaglia, Vasilis Friderikos, Borislava Gajic, Dirk von Hugo, Bessem Sayadi, Vincenzo Sciancalepore, Marcos Rates Crippa

***Abstract*—Network slicing is an emerging area of research, featuring a logical arrangement of resources to operate as individual networks, thus allowing for massively customizable service and tenant requirements. The focus of this paper is to present the design of a flexible 5G architecture for network slicing, building on SDN and NFV technologies as enablers. More specifically, we place the emphasis on techniques that provide efficient utilization of substrate resources for network slicing, ultimately optimizing network performance. The key areas of consideration in our architecture revolve around flexible service-tailored mobility, service-aware QoS/QoE control as well as network-wide orchestration.**

***Index Terms*—NFV; Orchestration; 5G networks; Mobility Management; QoS/QoE; MANO; Network Slicing**

I. Introduction

THE current trends in mobile networking show a growing need for flexibility. Driven by the new business paradigms such as 5G Verticals, the future 5G network should support very heterogeneous services on the same infrastructure. Services such as Internet of Things (IoT) and Vehicular Networking require from the mobile network very different Key Performance Indicators (KPIs): low latency, high capacity or service continuity. Supporting all these requirements on the same infrastructure entails a revolutionizing re-engineering of the network architecture that goes beyond the extensions of the current 3GPP-LTE one.

One of the most promising approaches to such re-engineering of the network is *Network Slicing* [1]. A traditional way to achieve a highly customized network is to deploy physical infrastructure for each service (or even one for each business). This approach clearly cannot be applied in a cost effective way and calls for technical solutions that allow for both efficient resource sharing and *multi-tenant* infrastructure utilization. Therefore, many of the research efforts in the definition of novel architectures are including network slicing in their proposals: 5G NORMA is one of them [2].

Network slicing has the capability of enabling (through the network architecture) future 5G networks that encompass the required scalability and flexibility characteristics, thus supporting diverse service scenarios and services. A network slice can be broadly defined as an end-to-end logically isolated network that includes 5G devices as well as access, transport, and core network functions. In the general case, these system functions can be also shared between different slices based on pre-defined policies and business criteria. The above requires abstraction of different physical infrastructures into a logical virtual network, in which virtual network functions (created by the decomposition of physical equipment into multiple, isolated instances) are operated.

More specifically, we next identify the main elements that must be provided by a mobile network architecture to support the network slicing paradigm:

- a **network slicing aware orchestration framework**, that performs both service and resource orchestration by also taking into account the (possibly conflicting) requirements introduced by different network slices,
- a **flexible network function control system**, that goes beyond the current definition of Element Management System (EMS) and Software-Defined Network (SDN) controller, offering thus a programmable unique control point for all the network functions belonging to a slice,
- a consistent **QoS/QoE management framework** that acts as trigger for network re-orchestration events and,
- **enhanced mobility management algorithms** that can take optimal decision for UE (User Equipment) mobility on a per-slice base.

The requirements and the research challenges related to the introduction of these elements were already analyzed and discussed in [2]. This paper focuses on the preliminary technical solutions needed to solve these challenges and finally establish the network slicing paradigm as one of the fundamental building blocks of future 5G networks as currently being targeted by the 5G NORMA project.

II. The 5G NORMA Architecture

Broadly speaking, a Virtual Network Function (VNF) should provide/support the same overall functionality as an equivalent blackbox based function (non virtual or bare metal). The key difference, though, is that a VNF could be deployed as a software instance capable of running on general purpose servers via virtualization technologies. To allow for such flexibility, running for example network functions at various network locations – even at the base station – we briefly detail below an architecture that is inline with the defined overall ETSI MANO ecosystem with the emphasis placed

Faqir Zarrar Yousaf and Vincenzo Sciancalepore are with NEC Laboratories Europe, email:zarrar.yousaf|vincenzo.sciancalepore@neclab.eu. Marco Gramaglia (corresponding author) is with Universisdad Carlos III de Madrid and IMDEA Networks Institute, email: mgramagl@it.uc3m.es. Vasilis Friderikos is with King's College London, email: vasilis.friderikos@kcl.ac.uk. Borislava Gajic is with Nokia Bell Labs, Munich, email: borislava.gajic@nokia-bell-labs.com. Dirk von Hugo is with Telekom Innovation Laboratories, email: Dirk.von-Hugo@telekom.de. Bessem Sayadi is with Nokia Bell Labs, Paris, email: bessem.sayadi@nokia-bell-labs.com. Marcos Rates Crippa is with University of Kaiserslautern, email:crippa@eit.uni-kl.de

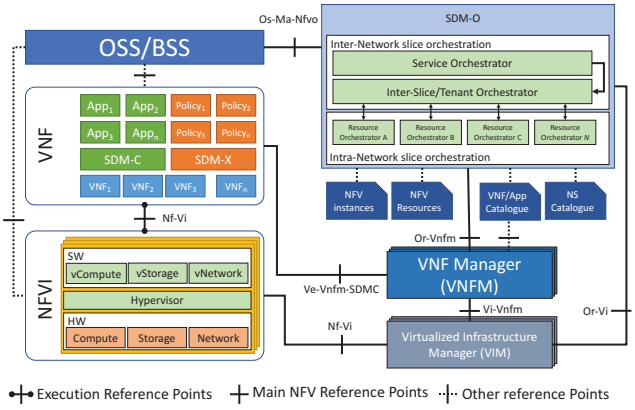


Fig. 1: The 5G NORMA MANO and Control Architecture.

on mobility and QoE [3] support. This architecture, built to natively support network slicing spanning several infrastructure domains, is composed by three main elements: the Software Defined Mobile Network Orchestrator (SDM-O), the Software Defined Mobile Network Controller (SDM-C), and the Software Defined Mobile Network Coordinator (SDM-X). Their role, broadly discussed in [2], is summarized next.

A. SDM-O and MANO framework

The overarching role of the SDM-O is to provide and maintain a suitable network function chaining so as to create an end-to-end service. The proposed orchestration capabilities are inline with (and provide extensions upon) the ETSI logical reference architecture for NFV MANagement and Orchestration (MANO) [4]. Several functional requirements were considered by 5G NORMA while designing the orchestration side as depicted in Figure 1. *i*) Different tenants should be able to perform both Management (i.e., VNF Manager, VNFM and Virtual Infrastructure Manager, VIM) and Orchestration (i.e., SDM-O) on a per slice basis if they want to. The MANO framework shall be flexible enough to allow this operation. *ii*) The SDM-O shall provide intelligent placement of VNFs together with optimized use of NFV resources, as VNFs can span across spatially distributed edge clouds (NFV Infrastructure, NFVI domains). This includes VNFs located close to the end user depending on the specific service request. *iii*) The orchestration framework shall tightly interact with the control elements (SDM-C and SDM-X, see Section II-B), reacting to QoS/QoE-based triggers.

The first requirement is achieved by splitting the Orchestration part in two submodules: the inter-slice and the intra-slice orchestration. The former module has a global view of the available resources, and optimally selects the resource quotas that are assigned to each slice (or tenant). This module also performs service orchestration as defined by ETSI. Then, the latter module performs resource orchestration on a per slice basis, directly acting on the VNFM and VIM modules.

The second requirement directly relates to the introduction of end-to-end network slicing in the architecture: VNFs will be orchestrated according to the target scenario. Therefore, the orchestration algorithms implemented in SDM-O shall encompass the efficient instantiation of virtualized functions in different network nodes for each service according to specific requirements and support multiple architecture instances simultaneously in the same physical network. Again, this is tackled by the intra-/inter-network slice orchestration split described previously.

The integration with SDM-C and SDM-X is paramount to achieving full QoE/QoS support in a network slice. These modules, described in Section II-B, are instantiated as further VNFs (see Fig. 1) and are the main triggers for QoE/QoS based re-orchestration as detailed in Section IV-B.

B. SDM-C / SDM-X

The Software Defined Mobile Network Controller (SDM-C) is in charge of controlling the network functions belonging to a slice and their associated resources using a Software Defined approach. There is an SDM-C instance per network slice, which retrieves network requirements through its northbound interface (connected to the MANO layer), and triggers the actions through its southbound interface (connected to both VNFs or Physical NFs), following the SDN paradigm. Such interfaces are used to fulfill slice QoE/QoS constraints. If QoE/QoS targets are not satisfied, the SDM-C instructs a re-orchestration. The advantages provided by the SDM-C can be summarized as follows:

Flexibility: Operators would be able to tailor the network to their needs by simply re-programming the controller.

Programmability: It allows third parties to acquire network resources on-demand satisfying their individual Service Level Agreement (SLA) while enhancing the user perceived QoE with customized network resources.

Unified control: Adopting a logically centralized control unifies heterogeneous network platforms and provides a simplified operation of the wireless network. With SDM-C, network operators only need to control a set of central entities (namely, the controllers) that control the entire network, which possibly includes heterogeneous radio technologies.

New services: New services can be easily introduced by directly modifying the network behavior by means of applications running on the SDM-C northbound interface. This would considerably save time in developing, debugging, and deploying new network functions.

While SDM-C provides isolation between slices resources/functions, shared components need a dedicated controller to fully exploit the multiplexing gain. This controller is called SDM-X. Specifically, elements such as transmission points, radio resources, transport and fronthaul capacity are often realized as shared resources. Once they are collected in a common pool, an interaction between SDM-X and SDM-C is in place to dynamically use the shared physical resources during operational flows. Physical resources are intended as radio and transmission over other media, processing within areas of computing resources, and storage for user/data plane and control plane information. While resource pooling for storage and processing power may be less demanding due to theoretically large resource pools, the scarcity of radio resources in many cases requires an advanced resource management solution [5]. The SDM-X takes decisions based on the policies provided by SDM-O how to fulfill the demands

of several partially competing network slices simultaneously. Slice performance demands might be identified in terms of: *i*) throughput, which requires a management of the radio resources, *ii*) latency, which requires a management of the placement of the network functions and usable storage entities, *iii*) management of processing / compute and storage resources in the neighborhood of access nodes, which may also impact latency and error recovery performance, *iv*) reliability and resilience, which is also greatly influenced by proper mechanisms for dynamic sharing of all three kinds of resources.

III. Mobility Management per Slice

Within a Software Defined Mobile network the capability of flexibly supporting mobility of users and their terminals, as well as sessions and flows and even (virtual and physical) network entities is predominantly seen as a challenge in terms of changing performance (e.g. throughput) and user perceived QoS/QoE. We next detail specific technical requirements and solutions that allow to reach adaptive and flexible mobility management solutions.

A. Binding Mobility Management to a Network Slice

Given the variety of services and their corresponding mobility management requirements that are envisioned within 5G networks, a more flexible mechanism for selection of mobility management schemes is necessary. Envisaged 5G services and slices exhibit different demands for mobility support in terms of e.g., terminal speed, session continuation requirements, and of stability of the endpoint address. The mobility management scheme needs to be selected flexibly according to the context of the service or a network slice.

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

The binding functionality includes three blocks as depicted in Figure 2: *i*) a Binding Policy Management, *ii*) a Binding Function and, *iii*) a Network Slice Selection and Configuration system.

The binding between the mobility management scheme and the network slice configuration needs to be done based on

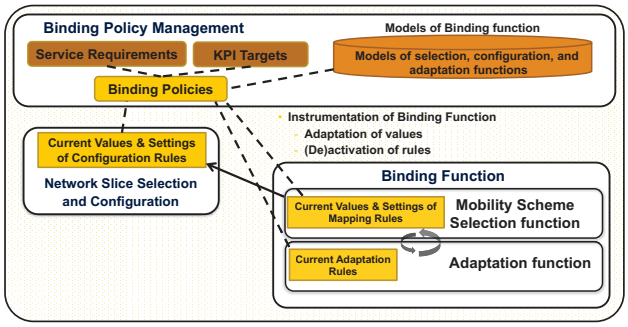


Fig. 2: Binding functionality – building blocks.

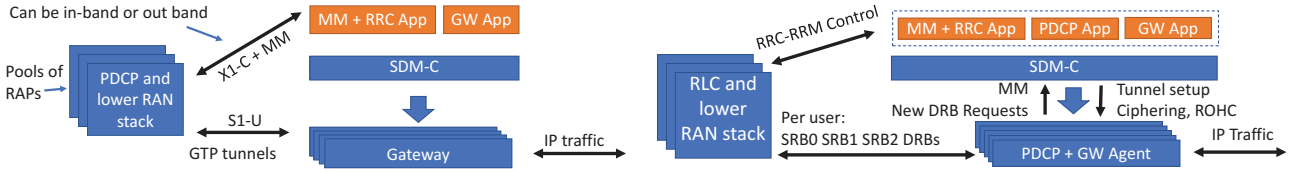
predetermined policies which are maintained in the Binding Policy Management, see Figure 2. The Binding Policy Management function translates service requirements, operator targets, and KPIs to policies that have to be enforced on the network slice. Based on this, the Binding Function executes the actual selection of the mobility management scheme according to the predetermined rules, as well as the adaptation of such selection if needed. The Binding Function includes:

- A Mobility Scheme Selection function that translates mobility requirements to mobility mechanisms/schemes based on policy and mapping rules as provided by the Binding Policy Management function as well as context information available from the network.
- An Adaptation function that is applied during the runtime, which verifies the actuality of the mapping between current mobility requirements of a slice and currently deployed mobility management schemes and performs according modifications, i.e. re-mapping between mobility requirements and mobility management schemes.

Finally, a Slice Selection and Configuration function is responsible for selection of the right templates for slice instantiation (e.g. with right VNFs type selection and right composition and configuration of different VNFs). In essence, for a given service requirements and network context the binding functionality gives two important outputs: suitable mobility management mechanism/scheme and the fitting template for slice instantiation/configuration. E.g., for slices that require low latency communication and limited mobility support the MM scheme might imply deploying local mobility anchors, whereas for slices with highly mobile users and session continuity requirements the central anchors and tunneling capabilities might be envisioned by the MM scheme. The corresponding slice templates will include the exact selection, placement and configuration of mobility anchors and other network functions needed for realization of selected MM scheme.

B. Flexible mobility management with SDM-C

In the current mobile network architecture, the Mobility Management (MM) functionality is a process that involves different physical entities in the network (i.e., eNB, MME, gateways,...). As discussed in Section II, the current *softwarization* trend will trigger the transition from a *network of entities* to a *network of functions*. To that end, an essential functionality as MM must also be transformed and made aware of the novel QoS/QoE, Orchestration and control mechanisms.



(a) PDCP-RRC split

(b) RLC-PDCP split

Fig. 3: Flexible SDM-C MM application options.

Following the design and architectural principles defined in Section II, the mobility management as a whole can be managed as a SDM-C application. Taking advantage of a unified QoS/QoE framework like the one described in Section IV, the management of user mobility is a thorough process that involves network function control and orchestration to achieve an optimized functionality on a per slice basis.

By exploiting these characteristics the network flexibility is increased: the adaptation of the network slice capacity according to the instantaneous traffic demands and required KPIs entails the re-configuration and re-orchestration of the network at many levels. Therefore, besides the selection of the most appropriate MM algorithm or the parameters that may influence the MM algorithm behavior, the MM shall be able to control different network configurations seamlessly.

One of the key technologies for the enhancement of the flexibility is RAN as a Service (RANaaS) [6]. This capability, envisioned as one of the future pillars of 5G networks [2], allows to split the current monolithic RAN stack into atomic functions that may be orchestrated in different ways, exploiting either the multiplexing gain of baseband processing centralization or the flexible resource utilization of decentralization of edge computing. In this very heterogeneous context, an enhanced MM shall *i*) jointly optimize RAN and Core network functions by leveraging on the centralized network control capabilities of SDM-C and, *ii*) steer user flows across different network functions according to the RANaaS functional split implemented in the network. The former functionality is implemented within an SDM-C application, while the latter is provided by a set of *plugins* installed on the Southbound Interface of the controller. We next sketch the overall ideas of a software defined MM algorithm that can cope with the changing environment of a RANaaS-enabled network. According to the selected functional split, different optimization options and network control challenges arise.

PDCP-RRC: this functional split is a pure c-/d-plane split, as PDCP is the highest layer that deals with user data, handling GTP traffic towards the gateways (and the Internet). Using the SDM-C approach, the functionality currently carried out by NAS, MME and RRC can be centralized and implemented as an SDM-C application. That is, a pool of virtualized Radio Access Points (RAP), implementing the RAN stack up to the PDCP, may be controlled by a centralized MM application that can take optimal handover decision according to the load of RAPs. On the other hand, the SDM-C southbound interface needs to interact with both the RAPs and the gateways (that may be joined in a single entity) by managing directly NAS, RRC and GTP session requests from the d-layer network function. This split is depicted in Figure 3a.

RLC-PDCP: this split involves managing directly data radio bearers between a pool of RAPs that implement the RAN stack up to the RLC and a centralized entity that needs to perform several functions in addition to the former split (i.e., *PDCP-RRC*) including e.g., (de)ciphering. The centralization of these previously distributed network functions allows for enhanced routing optimization, multipath or radio bearer based mobility. On the other hand, the southbound interface shall be able to manage, among other information, data radio bearers and their mapping to the RLC channels. Figure 3b describes this case.

These two examples just serve as illustration of how enhanced MM algorithms can take optimal decision depending on the *cloudification* of the network and the requirements that have to be fulfilled.

C. Further considerations regarding inter-slice Mobility

Design considerations on the amount and type of parameters to be configured within a slice-specific MM application (MM-App) as described in Section III-B include the issue whether the MM is invoked on a dedicated per-slice function or across multiple slices: E.g. a simple only on-demand MM would be associated only to a low/no-mobility slice (e.g., for IoT, Fixed/Home-Net slice etc.) whereas an MM-App supporting a range of terminal speeds and seamless session continuity might apply for multiple sessions/UEs within same regional context (within train/bus, on highway, etc.) but belonging to different slices (e.g. verticals' automotive slices and an enhanced Mobile Broadband slice). Such an MM-App would then be reused across different slices.

Further design criterion could be the availability and usage of Layer sensitive information across layers (e.g. to proactively invoke handover on MAC or IP layer based on signal strength at the physical layer). Such a feature as well as the capability to adjust to variable service demands (e.g. in terms of QoS/QoE) allowing for higher granularity in terms of mobility support of course would depend on the specifics of the involved access technologies. Another parameter guiding the MM-App design is the potential differentiation for a hierarchical mobility treatment (e.g. local and global mobility). Finally, the degree of flexibility which a chosen MM approach supports (e.g. whether a feature may be changed on the move according to changing environments) is restricted to certain variables (e.g. no multi-link or access heterogeneity is possible in case of single-interface devices). Based on such parameters considered in a mobility support design, the correspondingly required effort in terms of process complexity and amount as well as frequency of necessary signaling messages can be estimated and used to decide on the most appropriate MM module to be applied to a specific network slice.

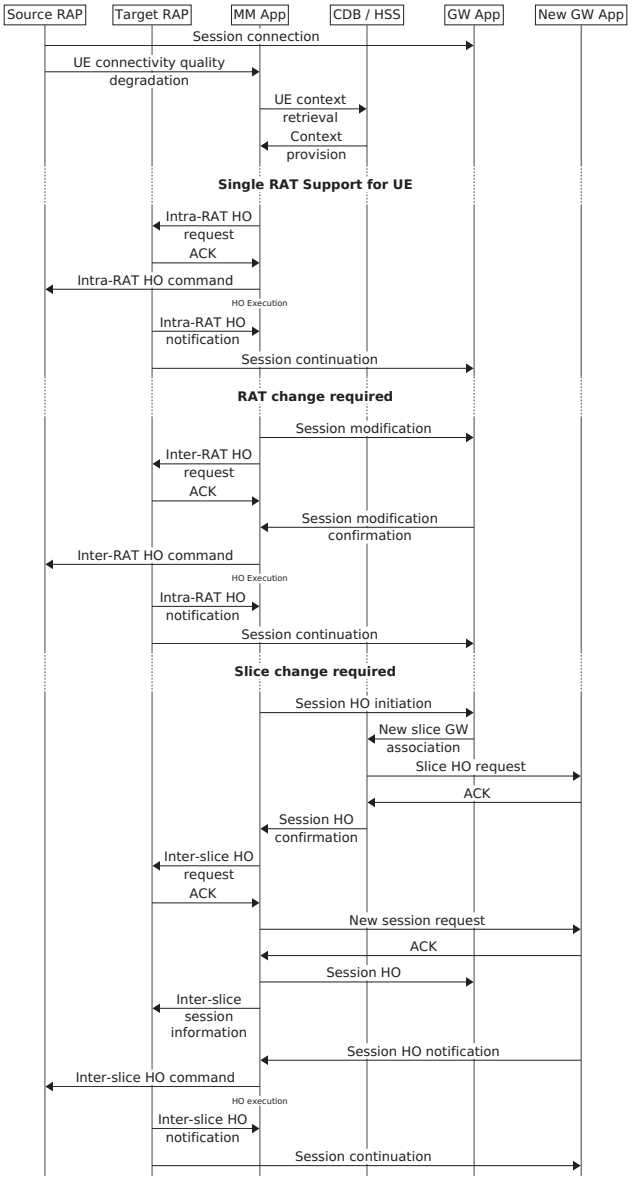


Fig. 4: Handover procedure as exemplary MM process.

Also, the number of network entities included in the required MM processes and the related messages to be exchanged may determine the effort spent by a specific MM application. Entities involved cover UE and RAP nodes and beside MM-App also core network entities as a GW-App (i.e. an SDM-C application for configuring the Gateway between 5G NORMA architecture and the outside world) and customer/subscription data base (CDB/HSS) for storage of subscription policy information. Figure 4 shows an exemplary Message Sequence Chart (MSC) with increasing amount of entities included in more complex Handover (HO) processes for visualization of Intra-RAT HO (between Source and Target RAP of same technology), Inter-RAT HO (where RAPs operate at different technologies and a corresponding modification of the GW-App may be required), and Inter-Slice HO where not only the Target RAP but also the New GW-App belong to a different network slice.

IV. Quality of Service/Experience Support

In addition to mobility, flexible QoS and/or QoE support depending on the requirements of different slices should be supported. This function, especially important for supporting tactile/haptic Internet applications requiring ultra-low delays and increased reliability levels, is considered as one of the main pillars of future 5G Networks. Orchestration, Network Control and Monitoring must be QoS/QoE aware, allowing a per-slice configuration. As explained in Section II, all these operations must be performed separately in each network slice.

A. QoS multipath support and effect on VNF orchestration

The SDM-O enables routing (for spatially distributed virtual/physical network functions), chaining as well as hosting (efficient use of edge cloud resources); in addition to baseline functionalities [7] both physical and virtual network elements need to be supported. Supporting multipath routing gives the flexibility for innovative VNF chaining configurations depending on the service characteristics. Figure 5 shows the effect of VNF hosting location on the overall path diversity and available path selection in the network.

As shown in the figure there are three alternative shortest paths between the sources and the destinations but the VNF allocation in Case I allows only to explore one of the shortest paths. On the contrary, the VNF allocation shown in Case II allows for all three shortest paths to be utilized by the SDM-C and depending on overall aggregate traffic levels and individual QoS per service request the availability on selecting across various shortest paths can increase network efficiency. Shortest paths can be readily available using e.g., OSPFv3, which is an open standard protocol, using Dijkstra's algorithm to find shortest paths to each and every destination node and/or network. At the same time it allows the support of multiple routes to same destination having equal cost for to make the routing process faster and balancing the load equally on various paths; extensions can also be utilized for more advanced routing decisions.

Therefore, in order to allow a rich routing environment that is not adversely affected by the location of VNFs in the network, special attention should be paid on proposed optimization algorithms to allow full use of available paths in the network. An illustrative example on how VNF location can limit the degree of multipath support is shown in Figure 5.

B. QoS/QoE-aware re-orchestration imperatives

After deployment and activation of network slices, the challenge is to ensure the QoS/QoE compliance of the services that are being supported by the network slices. This compliance is being managed and enforced by the SDM-O with the help of SDM-X and SDM-C functional elements. Each time the SDM-O detects service degradation it needs to derive appropriate orchestration actions in order to address the causes of degradation and maintain the service QoS/QoE within prescribed limits.It is important for the SDM-O to derive the most appropriate orchestration action in order to minimize service disruption that may occur during the application of the management action and to stabilize the service with minimum disruption time. This challenge is made more difficult owing to

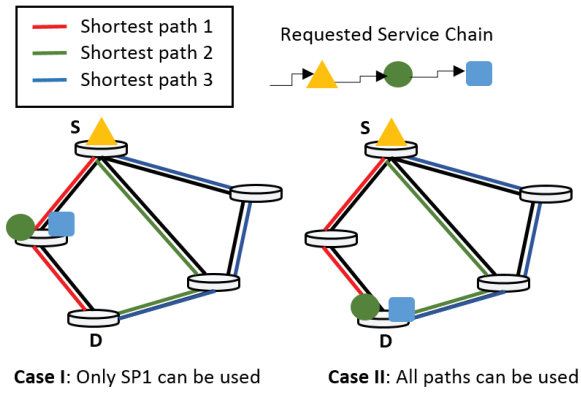


Fig. 5: Effect of vNF orchestration on the location and selection of available (shortest) paths in the network.

the network slice topology and configuration. For example, a single network slice may support one or more services, where each service has different QoS/QoE requirements. In addition, two or more network slices with different QoS/QoE requirements may share common resources, such as some VNFs as Common Network Function (CNF), so that management of each slice has to consider and coordinate with all other slices accessing the same resource. Moreover, a single network slice may traverse over multiple administrative domains (i.e., the NFVIs) making the implementation of QoS/QoE MANO very complex.

In view of the above complexities and challenges, effective methods needs to be devised and architectural extensions proposed to that the SDM-O is able to determine the exact cause of the QoS/QoE degradation in order to accurately derive and implement appropriate management actions. A functional overview of the QoS/QoE-aware slice management and orchestration system, compliant with the MANO architecture presented in Figure 1, is depicted in Figure 6.

Before we describe the key architectural extensions, it should be noted that the run-time slice QoS/QoE orchestration has more diverse and stringent requirements than those of individual service instances. This is due to the fact that a single network slice may support multiple service instances and thus the algorithmic impact may be felt across them, with the potential of affecting thousands of users and tens of thousands of service instances in the case of a mobile network slice. The first major requirement of QoS/QoE-aware slice MANO system is hence a pre-emptive (i.e., predictive) QoS/QoE management system. This contrasts the current reactive approach. Then, the QoS/QoE management system should be able to accurately determine and quantify the performance bottlenecks which would enable the SDM-O to derive a more suitable MANO action in order to ensure long-term sustainability and survivability of the network slice. In perspective of these main requirements, the two essential components, proposed in Figure 6, that deal with the aforementioned requirements are:

Service Quality Monitoring (SQMon) Each slice instance has a SQMon component, which is responsible for the *QoS*

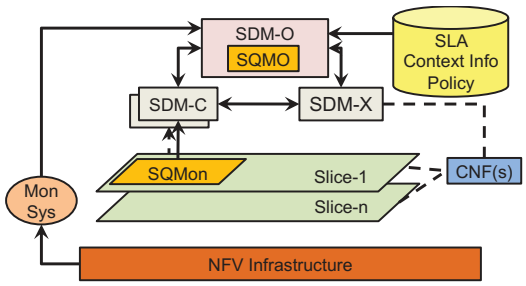


Fig. 6: A QoS/QoE-network slice aware MANO architecture.

monitoring and *QoE mapping* of the service instance(s) supported by the respective slice. It can also monitor the resource utilization by individual VNF(s) that make up the slice instance. The SQMon is a multi-parameteric and configurable entity that can be configured and tuned to reflect each service functional and performance peculiarities and based on them is able to monitor the service's QoS and map it to QoE. The SQMon will periodically keep the SDM-C updated of the network slice QoS/QoE status, which will determine any QoS/QoE degradation event. The SDM-C will provide the SDM-O with the network slice QoS/QoE performance reports.

Slice QoS/QoE MANO (SQMO) It is a part of the SDM-O and the main component where the QoS/QoE-aware management and orchestration of the network slice and/or its resources takes place. The SQMO can also be the part of the inter-slice/tenant orchestrator (See Figure 1) that will derive appropriate MANO decisions on orchestration actions (e.g., scale in/out/up/down, migrate, update/upgrade). The unique feature of the SQMO is that in addition to the slice QoS/QoE reports from the SDM-O, it relies on a variety of other inputs, such as SLA, user/device/service context-information, physical/virtual resource utilization of the infrastructure, etc., thereby enabling it to make an accurate determination of the reasons for QoS/QOE degradation. This informational variety allows the SDM-O to make accurate MANO decisions depending on the triggers.

As shown in Figure 6, the SDM-O is able to receive the infrastructure resource utilization reports through a monitoring system, which is a prerequisite to any MANO system. The monitoring system should be able to monitor the resource utilization by a slice and has the granularity to monitor each individual resource unit assigned to individual VNFs making up the slice. The SDM-O is also able to receive performance and resource related data from the CNF that is being shared between slices. With the availability of such variety of information, the SQMO is not only able to accurately calculate the amount/type of additional resources required to maintain the slice QoS/QoE, but is also able to derive the appropriate MANO action to ensure service continuity within the service quality bounds. Figure 7 shows an outline of a generally more complex QoS/QoE-aware orchestration logic that enables the SQMO to determine the best MANO action in recognition of specific events. In Figure 6 it is assumed that the SQMon is able to monitor and recognize QoS/QoE service degradation and the SQMO is able to *detect* the reason for service degradation and thus is able to determine the most suitable MANO action that the SDM-O has to to execute. The

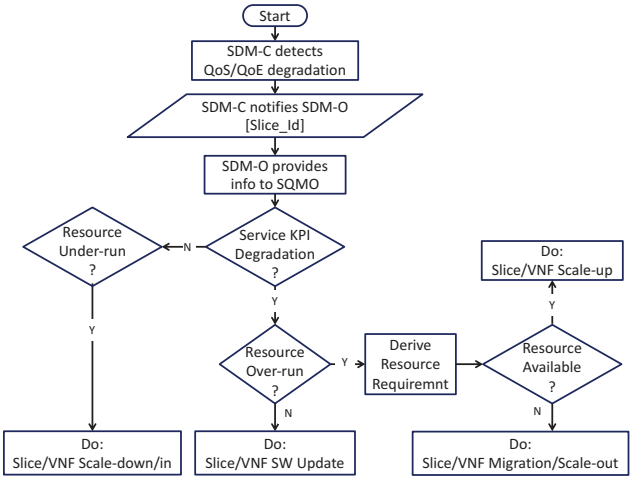


Fig. 7: Outline of a QoS/QoE-aware network slice management and orchestration procedure.

SQMO can also determine the required resources, based on SLA and context information, to maintain the service quality.

It should be noted that there are situations where the service degradation is not due to resource paucity but it may be due to software related issues. Figure 7 depicts how the SQMO is able to derive the most suitable MANO action by accurately detecting specific issues. For example, when the SQMO detects that there is QoS/QoE degradation despite enough resources then it can deduce the problem to be with the slice (or VNF) software itself and hence would trigger to attempt a software upgrade. In case of resource over-run, the SQMO will first try to determine if the required resources are available in the existing infrastructure host or not. If available the SQMO will indicate to the SDM-O to perform a slice scale-up operation or else do a migration to a suitable host within the slice infrastructure.

V. Open Ended Issues and the Road Ahead

Throughout this paper, we have described different solutions for achieving flexible mobility management and QoE/QoS support in a novel network architecture capable of supporting the network slicing paradigm. However, there are still open points in the algorithm definitions that will surely drive the research process. We briefly discuss them next.

A. Network slicing control granularity

As already eluded before, network slicing offers an effective way to support different use cases having diverse requirements, and exploit the benefits of a common network infrastructure. It enables operators to establish different deployments, architectural flavors, and performance levels for each use case or service group and run all network implementations in parallel. The current discussions in the industry have been focused on the overall concept and requirements of network slicing. However, the granularity of controlling and managing the slices (per flow, per service type, per device type, etc.) remains an open issue. Typically, the Core Network can be disaggregated to elementary functions (authentication, session management, etc.) which are deployed as VNFs. These VNFs could be in essence chained via a tunnel-based implementation which follows the user as it moves. Of course, this assumes a technology feasibility in terms of virtualization performance. In this scenario, the intelligence is more shifted towards the SDM-O. Following a different approach where VNFs like, e.g. the gateways are deployed as SDN-Apps (GW-App), then the tunnel management is handled by the SDM-C. In this scenario, the intelligence is more in the SDM-C. Therefore, what is the best granularity of the VNFs is an important challenge that will derive the way the network slices should be orchestrated and controlled to balance between flexibility and cost efficiency of the network management.

B. Interplay between SDM-C and SDM-X

The salient assumption of the SDM-C and SDM-X elements as presented in previous sections is that they can be considered as two different logical entities/elements with somehow distinguished roles. However, an interesting interplay between those two elements arises when trying to optimize dedicated and shared resources with an overlapping subset between them (e.g., frequency bands a subset of which is allowed to be shared). In this case a carefully design is required because of the *coupling* between these two elements due to the shared resources. Hence, even though these two elements can be considered logically as independent there can be instances the roles of which may blur depending on the type of shared resources and the algorithmic framework on network optimization that is implemented.

VI. Conclusions

Undeniably, network slicing is emerging as an important building-block of future 5G networks. In this paper, we focused on proposed extensions for NFV orchestration that provide network slice tailored support for mobility and QoS/QoE whilst ensuring efficient utilization of the substrate network resources. However, significant open-ended questions remain in this scope area and to this end, future research directions have also been briefly outlined.

Acknowledgment

This research work has been performed in the framework of H2020-ICT-2014-2 project 5G NORMA. The authors would like to acknowledge the contributions of their colleagues, although the views expressed are those of the authors and do not necessarily represent the project.

References

[1] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, B. Sayadi, *Mobile network architecture evolution toward 5G,* IEEE Communications Magazine, 2016.

[2] M. Gramaglia, I. Digon, V. Friderikos, D. von Hugo, C. Mannweiler, M. A. Puente, K. Samdanis, B. Sayadi, *Flexible connectivity and QoE/QoS management for 5G Networks: The 5G NORMA view,* 5GArch workshop, 2016.

[3] P. Le Callet, S. Möller, A. Perkis, *Definitions of Quality of Experience,* European Network on Quality of Experience in Multimedia Systems and Services, March 2013

[4] ETSI GS NFV-MAN, *Network Functions Virtualisation (NFV); Management and Orchestration*, 2014.

[5] V. Sciancalepore, K. Samdanis, X. Costa, D. Bega, M. Gramaglia, A. Banchs, *Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization,* in IEEE INFOCOM 2017.

[6] iJOIN, *Final definition of iJOIN architecture*, 2015.

[7] ETSI GS NFV-MAN, *Network Functions Virtualisation (NFV); Architectural framework*, 2014.