Software Defined Network Service Chaining for OTT Service Providers in 5G Networks

Eftychia Datsika, Angelos Antonopoulos, Nizar Zorba, Christos Verikoukis

***Abstract*—The fifth generation (5G) wireless networks are expected to offer high capacity and accommodate numerous over-the-top (OTT) applications, relying on users' Internet connectivity, thus involving different stakeholders, i.e., network service providers (NSPs) and OTT service providers (OSPs). For the efficient management of OTT application flows, the implementation of service functions and their interconnection in service chains, namely the network service chaining (NSC), should consider the OSPs' performance goals and user management strategies. However, in current wireless network deployments, the NSPs have full control of NSC. Considering that user satisfaction from the offered services is common interest for both types of stakeholders, the OSPs need to participate in NSC and apply QoS and user prioritization policies to NSC resource management, which involves users connected in different network points, in a distributed manner. In this article, we describe 5G network management architectures and propose virtualization components that enable OSP-oriented NSC. We also outline the arising issues for OSPs in NSC and we introduce a distributed prioritization NSC management scheme for OTT application flows, based on matching theory. The evaluation results indicate the performance gains in OSPs' service levels that stem from the proposed scheme, demonstrating the benefits of introducing prioritization in NSC deployment.**

***Index Terms*—Over-the-top services, Network service chaining, SDN/NFV, 5G, Matching theory.**

I. Introduction

CURRENT research on wireless systems revolves around the accommodation on future networking demands of both end users and business enablers in the wireless market. The wireless traffic is expected to increase almost 10,000 times by 2030, comparing to 2010[1]. Along with the growth of circulating mobile data, the development of the fifth generation (5G) wireless networks is triggered by the appearance of novel Internet-based applications and business models, introducing multifaceted challenges in network operation.

Recent advances in wireless technology have created a plethora of over-the-top (OTT) services relying on broadband Internet service technologies, e.g., live video streaming, gaming, etc., offered by OTT service providers (OSPs). OTT services have different Quality of Service (QoS) requirements, depending on their data traffic type, e.g., video delivery requires high bandwidth, while VoIP needs low latency. These services encompass different user categories (e.g., free users or premium users paying for advanced QoS). The users access the OTT content through various devices, e.g., smartphones or tablets, and their Internet access is often based on cellular connectivity. In this case, the users are also customers of mobile network operators, which own the cellular network infrastructure and spectrum, acting as network service providers (NSPs). Therefore, a user may access different OTT applications, whereas the users of a specific OTT application may belong to different NSPs.

The co-existence of multiple network services and OTT applications' QoS demands stresses the need for dynamic network configuration in software level without modifying the network equipment. For this purpose, the network service chaining (NSC) can be employed, which allows the on-demand network services adjustment using service chains (SCs), i.e., sets of interconnected software-based service functions (SFs) [1]. The SFs determine the way packets of flows are treated while circulating through network elements. An SC defines the set and sequence of SFs related to a flow, namely the actions applied to a flow, e.g., it may refer to a policy with two SFs, one that enforces all HTTP traffic to pass through a firewall and another that applies content filtering. In brief, NSC is the process of flow classification, flow forwarding to appropriate SFs and instantiation of SCs that implement network services.

The development of modern OTT applications emphasizes the need for efficient NSC management in Radio Access Networks (RANs). SCs are configured specifically per data connection type, e.g., Internet connection or connection for multimedia messaging services in cellular networks. Nonetheless, as the variety of network services increases, SCs should be deployed in a fine-grained manner, e.g., per user type, creating sophisticated SF combinations. A flexible and cost-effective solution is the networking paradigm of *cloud computing* that allows the management of fully-fledged services in centralized data centers, forming cloud-based RANs (C-RANs) [2]. Offering a distributed alternative, *fog computing* is a variation of cloud computing that can improve the experienced QoS. It brings services closer to end users, allowing them to be hosted away from cloud data centers, at edge nodes. These nodes form a distributed networking structure that acts as intermediate management unit between cloud and users, implementing fog-based RANs (F-RANs).

The resource and NSC management is feasible through direct programmability of network services, using the Software

This work has been funded by the Research Projects AGAUR (2014-SGR-1551) and CellFive (TEC2014-60130-P).

E. Datsika is with IQUADRAT Informatica S. L., Barcelona, Spain. E-mail: edatsika@iquadrat.com.

A. Antonopoulos and C. Verikoukis are with Telecommunications Technological Center of Catalonia (CTTC/CERCA), Castelldefels, Spain. E-mail: {aantonopoulos, cveri}@cttc.es.

N. Zorba is with Qatar University, Doha, Qatar. E-mail: nizarz@qu.edu.qa.

[1] ITU-R, "Report M.2370-0, IMT traffic estimates for the years 2020 to 2030," http://www.itu.int/pub/R-REP-M.2370-2015, July 2015, Accessed on: 2017-06-21.

Defined Networking (SDN) and Network Function Virtualization (NFV), which allow the softwarization of network functions and virtualize the network infrastructure [3]. The SDN architecture offers to the stakeholders access to RAN software-defined controllers that implement functionalities of control, data and application plane [4]. NFV implements SFs as software programs running on servers [5]. Exploiting the SDN/NFV assets, the NSPs can deploy SCs over the RAN.

In 5G networks, stakeholders like NSPs and OSPs may coexist, whose common interests are the provision of high quality services and the users' satisfaction. As the OTT applications' performance is intertwined with network connectivity service levels that affect the overall user experience, the OTT QoS is both NSPs' and OSPs' concern. However, with the current network architectures, the NSPs have total control of NSC, thus the OSPs cannot supervise the OTT applications' key performance indicators (KPIs), e.g., grade of service, or fully manage their users, as the Internet connections are controlled by the NSPs. Even in cases that users with high priority should be accommodated first by the SFs, the OSPs are not able to apply their user prioritization policies. Therefore, 5G network architectures should allow the OSPs to intervene in the NSC customization in two ways: i) select the SFs that should be implemented and ii) indicate the resources required for the SCs' implementation. As multiple OTT applications might access the same network concurrently, centralized optimization methods that require the aggregation of all flow information become impractical, due to the high overhead of control data transmissions and poor adaptability to dynamic network conditions. Hence, distributed self-organizing approaches should be employed for the OSP-oriented NSC deployment.

Even though wireless network virtualization facilitates OTT applications' management through the exposure of network resources, several issues may arise for the OSPs regarding their participation in NSC. Motivated by the lack of literature that studies the NSC deployment from the OSPs' viewpoint, in this article, our aim is threefold:

(i) We describe network management architectures and propose virtualization components that enable dynamic SC configuration by OSPs, exploiting SDN/NFV.

(ii) Considering the characteristics of the OTT applications and the needs of the OSPs as industry verticals in 5G wireless networks, we investigate the challenges that arise in NSC management process.

(iii) We study the realization of flexible OTT-oriented NSC and propose a matching-theoretic NSC management algorithm for OTT application flows that enables the OSPs to make decisions regarding the user prioritization policy and dynamically select suitable resources. The performance evaluation demonstrates that the OTT applications' service levels are improved when the OSPs declare their preferences over the resource assignment.

II. OSPs' Requirements for the Deployment of Network Service Chains

Numerous OTT applications already exist, (Skype, WhatsApp, etc.), which rely on Internet connectivity, often provided by the users' NSPs. Hence, the OTT flows circulate over different NSPs' network infrastructure. Furthermore, the OSPs face the dynamic nature of their services, as business decisions may require new SFs in order to capture the users' demands for new features. In this context, we next describe the OSPs' requirements regarding the NSC deployment.

A. OSPs' access to NSPs' networks

The OSPs should interact with the NSPs for the orchestration of NSC, monitoring their users' status (connection quality, location, subscription details, etc.) and combining this information for the construction of OTT flow profiles. Allowing the OSPs' intervention in NSC implies their access in NSPs' network resources, which can be financially advantageous for both parties. As their revenues seem to be correlated, achieving high OTT QoS can be also to the NSPs' best interest, if the OSP-NSP cooperation is balanced, e.g., through negotiation of agreements that regulate the degree of OSPs' intervention and sharing of gains [6].

The OSP-NSP interaction requires that the NSPs expose their service capabilities through properly designed Application Programming Interfaces (APIs), as described in the Service Capability Exposure Function concept of 3rd Generation Partnership Project[2]. As multiple NSPs co-exist, the OSPs may be associated with multiple network slices, with different characteristics and services. Therefore, the OSPs should modify the SCs according to network service capabilities and available resources of the involved network slices.

Enabling the OSPs to develop SCs might entail preferential management of certain flows over the Internet. If OSPs apply flow prioritization in NSC, the NSPs' resources may not be shared fairly among OTT flows, creating concerns about the network neutrality. Although prioritization policies are necessary in certain cases, e.g., for gaming applications with low latency requirements, NSPs' resources should be accessed in an impartial manner, without monopolizing their utilization by some OSPs only. Thus, OSP-oriented NSC should balance flow prioritization and fair access to NSPs' networks.

B. Adaptation to OTT service market dynamics

The OSPs need to deploy services dynamically over static networks and compose business strategies, customizing the SCs. As the OSPs' revenue is highly dependent on the timely development of high quality services, the flexibility in constructing SCs in real-time is of crucial importance for the acceleration of OTT services' time-to-market.

The analysis of OTT application traffic is useful for the customization of SCs as a response to OTT service market dynamics, which may require the development of novel services or the addition of new features in OTT services. This update process might induce the addition of SFs in OSPs' SCs, e.g., for the enrichment of an online gaming application, an

[2] 3rd Generation Partnership Project, "Technical Specification Group Services and System Aspects; Architecture Enhancements for Service Capability Exposure (3GPP TR 23.708 version 13.0.0 Release 13)," https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=869, June 2015, Accessed on: 2017-06-21.

SF that adapts graphics rendering to the capabilities of users' devices might be added. Essentially, dynamic NSC implies end-to-end network intelligence and adaptive network service composition [1]. These features can be achieved by examining the OTT users' behavior and the market trends using the flow information of properly designed SFs. A typical example of such SFs is the deep packet inspection (DPI) that provides the OSPs with network analytics needed to define user policies.

III. Network management for OTT service delivery

The OTT application users are attached to RANs composed of heterogeneous network nodes (small cells, Wi-Fi access points, etc.), owned and managed by different NSPs, and are served by data centers with different capabilities and architectural design. Hence, the 5G network design should facilitate NSC for OSPs', offering two fundamental functionalities [7]:

(i) Holistic network view: OSPs should be aware of NSPs' resources and networking capabilities in order to create the appropriate SCs using virtualization techniques.

(ii) Support for network slicing for different OSPs: NSPs should be able to allocate resources to multiple OSPs, creating proper network slices.

A. *OSP-friendly network management architectures*

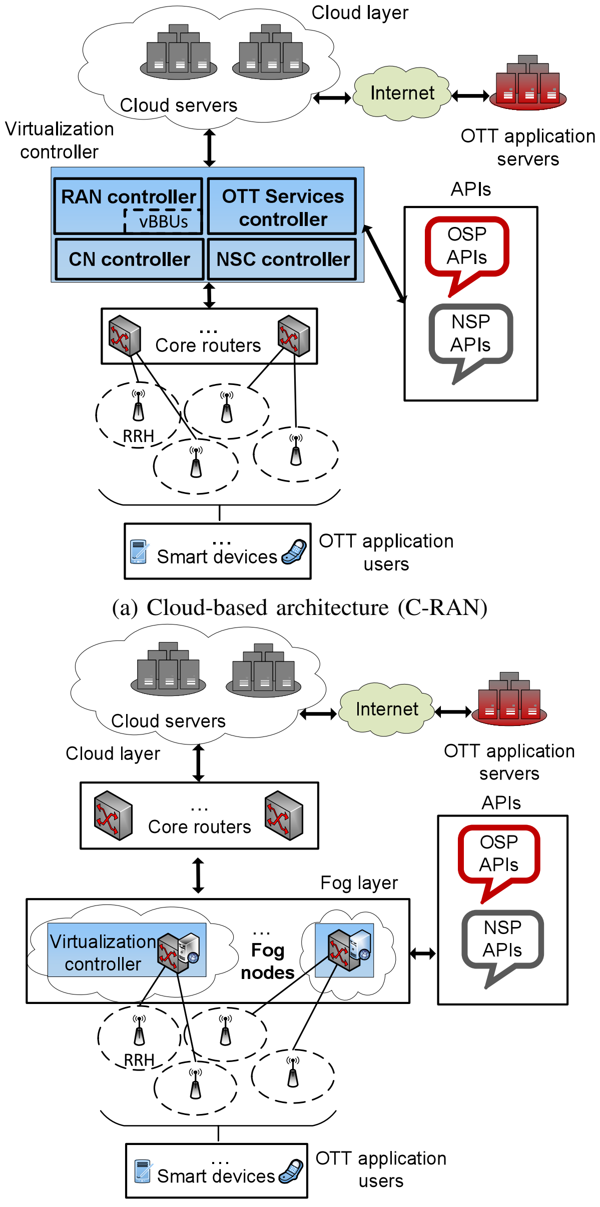
For SC development, information of OTT users often scattered in different RAN connection points should be aggregated. A well-known RAN-wide management technology is cloud computing networking [2]. A C-RAN consists of three main structural elements: i) several access points (APs) with Remote Radio Head units (RRHs), ii) a virtual Base Band Unit (vBBU) pool connected with the APs that performs baseband operations, and iii) core routers that connect the RRHs with the cloud (Fig. 1(a)).

Even though the centralized approach is useful for NSC, locating the RAN management unit away from users leads to high latency and overhead. Alternatively, fog computing places cloud services close to network edge [2]. An F-RAN is a distributed system which controls a set of RRHs through fog nodes (FNs), i.e., network devices as local servers with storage and computing capabilities (Fig. 1(b)). FNs bring the network management operations closer to end users, are connected with switches or routers in RAN edge and communicate with the vBBU pool. Each FN is a small data center implementing SFs for users connected to the APs it manages. The cloud data center acts as a global administration point.

B. *NSC virtualization frameworks for OSPs*

In the aforementioned architectures, SFs are applied in different network nodes with different order and features. This procedure requires the programmability of network functions and the abstraction of RAN resources for network slicing. For this purpose, the NFV and SDN frameworks can be incorporated in the network management architectures and provide virtualization components [4] (Fig. 2).

The NFV is a key technology for the customization of SCs and provides the essentials for virtualized management and organization (MANO). It enables the instantiation of virtual network functions (VNFs), manages the NFV infrastructure (NFVI) resource requests and offers complete services by combining the VNFs. The VNF Managers initialize and configure VNF instances and their interconnections with the NFVI that contains virtual computing resources (CPU, memory, etc.), storage resources and virtual machines. The Virtualized Infrastructure Manager (VIM) controls the underlying resources of NFVI, allocating them appropriately to VNF instances.



(a) Cloud-based architecture (C-RAN)

(b) Fog-based architecture (F-RAN)

Fig. 1: Network management architectures for OSPs

For SFs' management, SDN offers the capability of VNF orchestration using various components. In the considered architectures, a virtualization controller consists of four types of controllers: 1) the RAN controller that orchestrates the RRHs, allocates the spectrum resource blocks (RBs) and performs flow scheduling at each RRH, ii) the Core Network (CN)

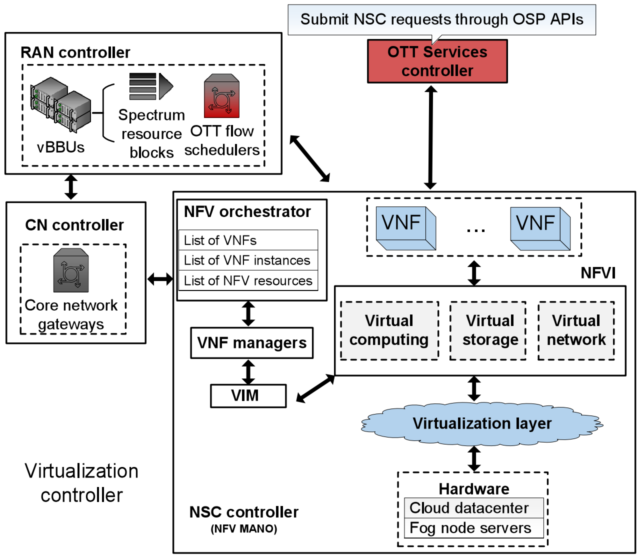


Fig. 2: SDN/NFV framework for OSPs' NSC

controller that manages the CN-related gateways, iii) the NSC controller that stores the information for NSC deployment and coordinates the VNFs, and iv) the OTT Services controller that is used by the OSPs for OTT service surveillance and submission of NSC requests. The OTT Services controller communicates with the VNFs and provides an overview of the implemented SFs (VNF instances), allowing the OSPs to assess the NSC performance and decide upon the SCs, by submitting NSC requests using the OSP APIs.

IV. Open issues in NSC deployment for OSPs

The SCs are ordered sequences of SFs combined in order to process application flows, which are forwarded to the APs . For the OSPs', the NSC should be performed according to the flows' QoS requirements, the user information and the OSPs' preferences regarding the KPIs. Although the current network management architectures and virtualization frameworks are useful for NSC, several issues arise for the OSPs, which are outlined in this section (Fig. 3).

A. Assessing the OTT users requirements

The OTT applications access RANs via the users' Internet connections. OTT flows have different network and application related user characteristics, which influence the experienced QoS. Thus, it is important for OSPs to evaluate the NSC requirements and coordinate appropriately the SCs related to a heterogeneous set of users over multiple network slices.

As flows are related to different users, the OSPs should obtain information regarding users' location and downlink channel conditions. Even if the OSPs' KPIs characterize the OTT applications, the users' specific context may affect the SC construction, e.g., if the users of a video streaming application experience poor downlink channel conditions, it might be unreasonable to use an SF for video optimization. Moreover, the OSPs might serve users associated with different network slices, i.e., APs and data centers with different capabilities possibly owned by different NSPs [1]. The NSPs' resources made available to users may impose the bounds in NSC efficiency, e.g., a small FN at the network edge may not support advanced flow processing for all users.

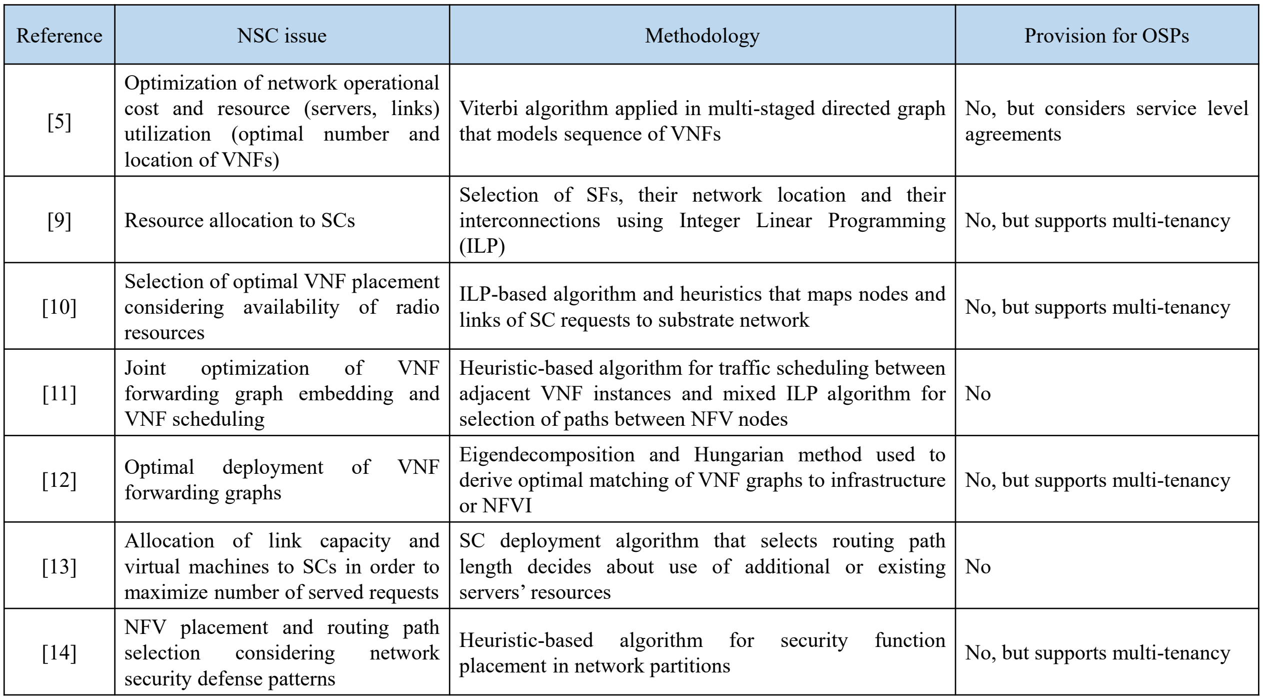
B. Mapping the OSPs' SCs to the NSPs' resources

In SDN/NFV enabled networks, the composition of SCs implies the selection of network services that will be implemented as VNF instances, the orchestration of VNFs on a server or cluster of servers, the establishment of proper traffic routing paths among the VNF instances and the allocation of resources to VNF instances (SC embedding problem). The resources are computational and storage resources of hardware and virtual machines or infrastructure and spectrum elements corresponding to different virtual networks, managed by NSPs' data centers.

Mapping the SCs to resources is a complex process that matches various types of physical resources with the VNF instances related to the SFs. For instance, virtualized C-RAN resources, e.g., RRHs and fiber links, are assigned to different tenants using auction mechanisms [8]. Hybrid NFV-based networks that incorporate network functions provided by dedicated physical hardware and virtualized instances may also exist [9]. Spectrum resource blocks can be allocated to APs associated with the VNF instances that serve the users connected to them, e.g., using VNFs that schedule the wireless resources of network slices in RAN nodes [10].

The SC embedding problem is multi-fold, since decisions for NFV placement affect resource allocation and vice versa [11]. As NSPs manage the network resources and are concerned about overall VNF operational cost, VNF instances can be deployed in a way that VNF host selection cost, traffic forwarding cost and energy consumption are minimized [5]. The VNF orchestration might imply a trade-off between the latency induced by the VNF placement and chaining in servers and switches and the efficiency of resource utilization [12]. Suitable VNF locations and flow routing paths can be also defined according to capacity constraints of virtual machines that host the VNFs and of the links among them, optimizing the amount of these resources that are allocated to SFs [13].

Existing works arrange the VNFs aiming to optimize NSP-related aspects. Nonetheless, OSPs should participate in NFV orchestration by expressing their preferences over SFs. Users of the same OSP may be connected to different APs and served by different data centers, e.g., in F-RANs. Different FNs may serve users, thus different VNFs are required in each FN, considering the OSPS' flow management policies. Moreover, the OSPs usually have their own policies regarding service differentiation, which provides the rules that deem the flows to be of higher of lower importance. The flows have different characteristics and different user priorities exist. This prioritization should be depicted in NSC, not only during NFV placement, but also in resources allocated to SCs, as VNF instances related to higher priority flows should be arranged first. For example, in an F-RAN accessed by flows with multiple priorities, the SC embedding involves not only the prioritized arrangement of VNF instances in FNs, but also the prioritization of spectrum allocation to users connected to APs.



Reference	NSC issue	Methodology	Provision for OSPs
[5]	Optimization of network operational cost and resource (servers, links) utilization (optimal number and location of VNFs)	Viterbi algorithm applied in multi-staged directed graph that models sequence of VNFs	No, but considers service level agreements
[9]	Resource allocation to SCs	Selection of SFs, their network location and their interconnections using Integer Linear Programming (ILP)	No, but supports multi-tenancy
[10]	Selection of optimal VNF placement considering availability of radio resources	ILP-based algorithm and heuristics that maps nodes and links of SC requests to substrate network	No, but supports multi-tenancy
[11]	Joint optimization of VNF forwarding graph embedding and VNF scheduling	Heuristic-based algorithm for traffic scheduling between adjacent VNF instances and mixed ILP algorithm for selection of paths between NFV nodes	No
[12]	Optimal deployment of VNF forwarding graphs	Eigendecomposition and Hungarian method used to derive optimal matching of VNF graphs to infrastructure or NFVI	No, but supports multi-tenancy
[13]	Allocation of link capacity and virtual machines to SCs in order to maximize number of served requests	SC deployment algorithm that selects routing path length decides about use of additional or existing servers' resources	No
[14]	NFV placement and routing path selection considering network security defense patterns	Heuristic-based algorithm for security function placement in network partitions	No, but supports multi-tenancy

Fig. 3: NSC issues and existing solutions

C. Constructing secure SCs

An important aspect of the NSC procedure is the deployment of safe SFs considering different security standards. Selecting the location and order of SFs based solely on the SC performance estimation does not always lead to secure NSC. Particularly in RANs accessed by various OTT applications, the instantiation of SCs in a secure manner is not trivial, as security constraints of different OSPs have to be imposed on NSC. A recent work proposes the use of network security patterns in order to capture the network security constraints in a C-RAN [14]. Still, the OSPs need to devise their own security policies that may change according to their users' demands or OTT application characteristics. These policies should be "translated" to SFs organized jointly for all OSPs accessing a RAN, in a way that all OTT services' security needs are met.

V. Flexible NSC for OSPs

The implementation of SCs should match the OTT applications' particularities and OSPs' policies. Considering the need for distributed control of prioritization in NSC over networks accessed by multiple OSPs, we propose a NSC management algorithm that allows the OSPs to define their policies and declare their preferences over SFs and resources in a distributed manner, based on matching theory. Moreover, we examine the effects of flow prioritization in NSC by assessing the performance of the proposed algorithm.

A. OTT flow prioritization using matching theory

OTT users may be connected to different network points and prioritization has to be applied in various abstraction levels, i.e., VNF instantiation and RAN resource allocation. For the composition of SCs, the OSPs need information related to i) the availability of network resources (e.g., spectrum RBs) and SFs provided by the NSPs (e.g., DPI), and ii) the OTT application flows' characteristics. These characteristics include required data rates, user subscription status, content type, etc., which are known to the OSPs. Flows' characteristics referring to the OTT users' cellular connections, i.e., parameters related to downlink channel conditions (e.g., supported modulation and coding schemes), are provided by the NSPs', along with the information about network resources and functions, and can be accessed through the OSP APIs.

Introducing the concept of matching theory in NSC enforces the role of OSPs in OTT flow management. The NSC management employs the notion of matching game, which models the interactions between NSPs and OSPs [15]. The OSPs create ranked lists of preferences over virtual resources according to flows' requirements. Each list item is a virtual resource request (VR), i.e., a combination of parameters related to each flow i $(id_i, AP_i, RB_i, priority_i, chain_i)$, where AP_i is the AP of the user with flow i, RB_i is the number of spectrum RBs needed for the flow's QoS demand, in terms of data rate, latency or other metric, when the user is connected in AP_i, $priority_i$ is the flow's priority level, and $chain_i$ is an ordered set of SFs, which declares a chain of VNFs required for the specific flow's processing. Multiple VRs may be related to the same flow but at most one will be finally accommodated.

The VRs are added to the OSPs' preference lists according to flows' priority levels. The priorities are determined by each OSP, according to the performance goals, e.g., maximize the number of flows that achieve the QoS demands. Therefore, the QoS metrics of each flow and KPIs of each OSP that affect

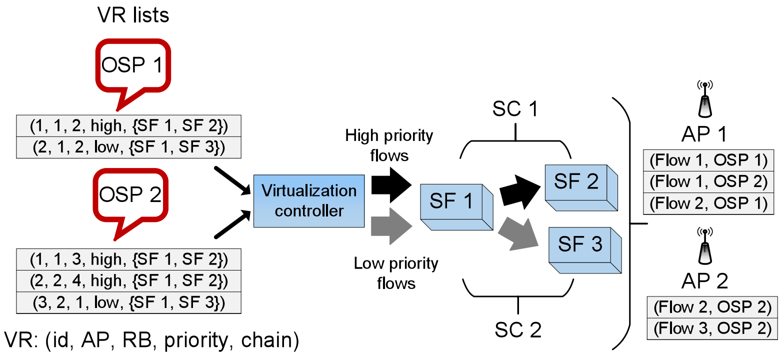

VR lists
OSP 1
(1, 1, 2, high, {SF 1, SF 2})
(2, 1, 2, low, {SF 1, SF 3})
OSP 2
(1, 1, 3, high, {SF 1, SF 2})
(2, 2, 4, high, {SF 1, SF 2})
(3, 2, 1, low, {SF 1, SF 3})
VR: (id, AP, RB, priority, chain)
Virtualization controller
High priority flows
Low priority flows
SC 1
SF 1
SF 2
SF 3
SC 2
AP 1
(Flow 1, OSP 1)
(Flow 1, OSP 2)
(Flow 2, OSP 1)
AP 2
(Flow 2, OSP 2)
(Flow 3, OSP 2)

Fig. 4: NSC example using the matching-theoretic OTT flow prioritization algorithm

the construction of the preference list might be different. Each flow may be associated with different SFs, thus different SCs may be added in a VR, e.g., an OSP can apply an SF, such as DPI only for premium users.

When new flows arrive, the OSPs decide about the necessary SFs and RBs and create preference lists, placing VRs for high priority flows first (Fig. 4). The VRs of flows with the same priority are ordered by ascending number of RBs, e.g., for OSP 2, the flow $i = 1$ needs three RBs and is placed before the flow $i = 2$. Subsequently, the matching process begins using the various components of the virtualization controller (Fig. 2). The VRs are submitted through the two OSPs' APIs to the OTT Services controller and the NSC controller initiates the matching process, handling by priority the requests. The RAN controller allocates the RBs, in a way that in both APs, the high priority flows receive RBs first. The VNF manager organizes the VNF instances and their interconnections in the data center connected to the two APs. The requests for each VNF are aggregated and suitable resources (CPU, memory) in NFVI are assigned by the VIM according to flows' priorities. For all flows, one VNF instance for each of the SFs (SF 1, SF 2, SF 3) is required. The high priority flows access the SF 1 and SF 2 instances, whereas the low priority flows pass through the SF 1 and SF 3 instances, thus two SCs are created, namely SC 1: $\{$SF 1, SF 2$\}$ and SC 2: $\{$SF 1, SF 3$\}$. Once RBs and VNF instances are organized, SCs can be implemented.

Finally, a stable matching between flows and resources is reached, including allocations acceptable by both OSPs and NSPs. Each acceptable matching is individually rational and is not blocked by a flow-resource combination, making it the most preferable match for the flow [15].

B. Matching-theoretic NSC performance

We assess the performance of the OTT application flow prioritization matching algorithm (MAFP), against a best-effort approach without prioritization (BE) and a fair allocation (FA) scheme that splits radio resources evenly among all OSPs in each AP. Considering the importance of user satisfaction, the OTT service levels are evaluated in terms of grade of flow accommodation (GFA), namely the percentage of flows that are not served with the requested QoS out of the flows of all OSPs. We also study the resource utilization levels, estimating the number of NFV instances required to serve the users. Two scenarios with different users' minimum data rate demands are tested: i) a file downloading (FD) scenario with minimum acceptable data rate equal to 64 kbps, and ii) a video streaming (VS) scenario with data rate equal to 128 kbps. In the presented results, a 95 % confidence interval is considered.

TABLE I: Simulation settings

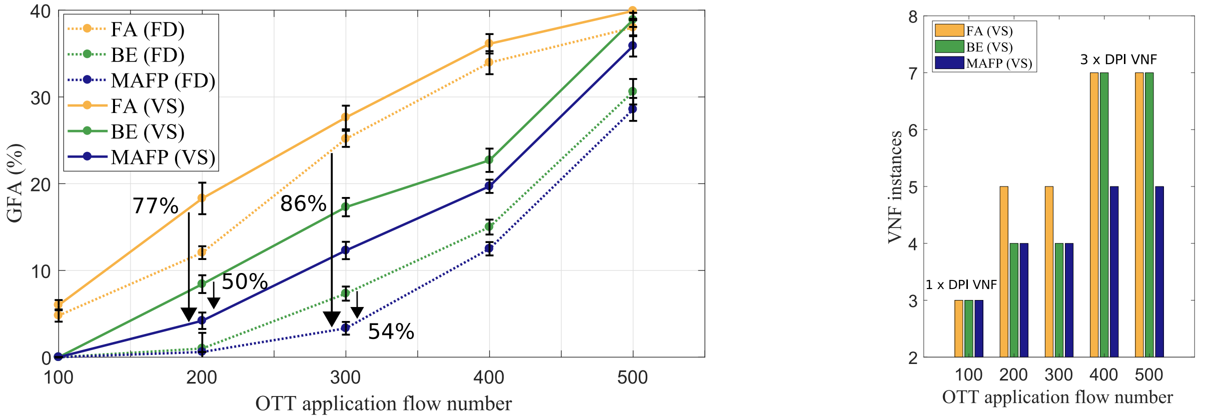
Setting	Value
Network	F-RAN
APs	8
RBs per AP	50
Minimum required data rate	64 (FD), 128 (VS) kbps
AP range	200 m
AP transmission power	33 dBm [3]
FN capabilities	10 CPU cores, 500 GB memory
Resources per VNF type (for 100 flow requests/sec) [9]	Routing, Firewall: 1 CPU core, 10 MB memory (each), DPI: 1 CPU core, 500 MB memory
OTT application flows' number	{100, 200, 300, 400, 500}
Full DPI Service SC	{DPI SF}

Sampled DPI Service SC	90% of flows:
	{Routing SF, Firewall SF}
	10% of flows:
	{Routing SF, DPI SF}

The FN implements two SC types, one that offers a Full DPI service where all flows access the DPI SF, and one for a Sampled DPI service, where a portion of flows access the DPI SF, as regulated by OSPs (Table I). In our simulations, two OSPs serve users of either high or low priority. Half of the users belong to one OSP, whereas 60% of each OSP's users have high priority. High priority flows access the Full DPI service and low priority flows use Sampled DPI service.

In Fig. 5 (a), the GFA performance of all schemes is depicted. The increase of flows degrades the performance of both schemes, as fewer flows are served using the available RBs at each AP. Nonetheless, the MAFP achieves lower GFA than BE and FA, and enables the accommodation of more flows, reaching a reduction of 54%-86% (FD scenario), and 50%-77% (VS scenario), for 300 and 200 users, respectively. As

[3] G. Miao, N. Himayat, Y. Li and D. Bormann, ”Energy Efficient Design in Wireless OFDMA,” IEEE Int. Conf. on Commun., 2008, pp. 3307-3312.



(a) Grade of flow accommodation (%)

(b) VNF instances for VS scenario

Fig. 5: Performance with and without flow prioritization

flows are sorted by number of required RBs and priority, more flows are served and the high priority flows are guaranteed to receive resources first. The performance gain is lower when higher data rate is required, as more RBs are needed.

Focusing on the VS scenario, Fig. 5 (b) shows the VNF instances for the implementation of Full and Sampled DPI services. The GFA levels influence the NSP's resource utilization, as different number of VFNs must be instantiated. Three VNF instances including one DPI VNF instance for 100 flows are needed for all schemes. In contrast, for 400 or more flows, five or more VNF instances are used (three DPI VNF instances). As the number of flows increases, more VNF instances are required. With BE and FA, more VNF instances are needed for more than 400 flows, as more flows of low priority are served and require Routing and Firewall VNFs.

Overall, prioritization in NSC affects the resource allocation in the APs and the VFN instantiation. It improves the OTT service levels, as the OSPs declare their preferences. In reality though, the NSPs may supervise the resource utilization, ensuring the application of cooperation terms and fairness among OSPs. Still, using matching theory, the OSPs can express their preferences over resources and SFs. Therefore, the available resources are no longer allocated in a best effort manner but the OTT application flows' requirements are considered in the NSC. Last, the OSPs can request the prioritization of certain flows according to the required QoS and their KPIs and flows with higher priority are guaranteed to receive resources first.

VI. Conclusions

In this article, we have described virtualization components for 5G network architectures that serve OTT application users and the challenges that arise in NSC deployment for OSPs. The network elements and their resource availability are different from one NSP to another. Furthermore, content and user types, QoS levels or KPIs change with the evolution of OTT applications and OSPs' business decisions. Thus, the successful deployment of OTT applications requires flexible adaptation of network services, according to OSPs' flow management and prioritization strategies. Considering this context, we have presented a matching-theoretic OTT flow prioritization algorithm for NSC, which improves OTT applications' service levels, achieving more efficient resource management. The performance evaluation results can provide valuable insights for OSPs in the 5G wireless market.

We should note that the NSC deployment creates various practical challenges for both NSPs and OSPs. First of all, the NSC deployment implies the exposure of the NSPs' network resources and SFs. As the resources provided to OSPs are affected by both the network capabilities (feasibility of exposure) and the NSPs' business goals (expected profit from exposure), the decision about the network exposure levels requires the joint consideration of network–related and financial parameters, which is an open research issue of NSC resource management. Moreover, the increasingly complex OSPs' requirements should be efficiently mapped in SFs, stressing the need for sophisticated and self-organizing solutions that customize the SCs properly. Still, this mapping process should not compromise the security of SCs. Ensuring that the security levels requested by OSPs and NSPs are maintained in NSC deployment can be an issue with high technical complexity. To this end, we believe that our study has shed some light on the OSPs' requirements and can motivate further investigation of NSC for OTT applications.

References

[1] F. Paganelli, M. Ulema, and B. Martini, "Context-aware Service Composition and Delivery in NGSONs over SDN," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 97–105, Aug. 2014.

[2] C. X Mavromoustakis, G. Mastorakis, and C. Dobre, "Advances in Mobile Cloud Computing and Big Data in the 5G Era," *Studies in Big Data*, vol. 22, 2016.

[3] A. M. Medhat, T. Taleb, A. Elmangoush, G. A. Carella, S. Covaci, and T. Magedanz, "Service Function Chaining in Next Generation Networks: State of the Art and Research Challenges," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 216–223, Feb. 2017.

[4] V.-G. Nguyen, T.-X. Do, and Y. Kim, "SDN and Virtualization-based LTE Mobile Network Architectures: A Comprehensive Survey," *Wireless Personal Commun., Springer*, vol. 86, no. 3, pp. 1401–1438, Feb. 2016.

[5] F. Bari, S. R. Chowdhury, R. Ahmed, R. Boutaba, and O. C. M. B. Duarte, "Orchestrating Virtualized Network Functions," *IEEE Trans. on Network and Service Management*, vol. 13, no. 4, pp. 725–739, Dec. 2016.

[6] A. Antonopoulos, E. Kartsakli, C. Perillo, and C. Verikoukis, “Shedding Light on the Internet: Stakeholders and Network Neutrality,” *IEEE Commun. Mag.*, vol. PP, no. 99, pp. 2–9, 2017.

[7] G. Tseliou, K. Samdanis, F. Adelantado, X. C. Pérez, and C. Verikoukis, “A Capacity Broker Architecture and Framework for Multi-tenant support in LTE-A networks,” in *IEEE Int. Conf.on Commun.*, 2016, pp. 1–6.

[8] S. Gu, Z. Li, C. Wu, and H. Zhang, “Virtualized Resource Sharing in Cloud Radio Access Networks Through Truthful Mechanisms,” *IEEE Trans. on Communications*, vol. 65, no. 3, pp. 1105–1118, March 2017.

[9] H. Moens and F. De Turck, “Customizable Function Chains: Managing Service Chain Variability in Hybrid NFV Networks,” *IEEE Trans. on Network and Service Management*, vol. 13, no. 4, pp. 711–724, Dec. 2016.

[10] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, “Scheduling Wireless Virtual Networks Functions,” *IEEE Trans. on Network and Service Management*, vol. 13, no. 2, pp. 240–252, Apr. 2016.

[11] L. Wang, Z. Lu, X. Wen, R. Knopp, and R. Gupta, “Joint Optimization of Service Function Chaining and Resource Allocation in Network Function Virtualization,” *IEEE Access*, vol. 4, pp. 8084–8094, Nov. 2016.

[12] M. Mechtri, C. Ghribi, and D. Zeghlache, “A Scalable Algorithm for the Placement of Service Function Chains,” *IEEE Trans. on Network and Service Management*, vol. 13, no. 3, pp. 533–546, Sept. 2016.

[13] T. W. Kuo, B. H. Liou, K. C. J. Lin, and M. J. Tsai, “Deploying Chains of Virtual Network Functions: On the Relation Between Link and Server Usage,” in *IEEE INFOCOM 2016-35th Annual IEEE Int. Conf. on Computer Commun.*, 2016, pp. 1–9.

[14] A. S. Sendi, Y. Jarraya, M. Pourzandi, and M. Cheriet, “Efficient Provisioning of Security Service Function Chaining Using Network Security Defense Patterns,” *IEEE Trans. on Services Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[15] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, “Matching Theory for Future Wireless Networks: Fundamentals and Applications,” *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.

Eftychia Datsika received her B.Sc. and M.Sc. degree in Computer Science from the Computer Science Department, University of Ioannina, Greece in 2010 and 2012, respectively. She is currently a Researcher in IQUADRAT Informatica S.L., Barcelona, Spain. Her research interests lie in the area of resource management in Long Term Evolution Advanced networks, software defined networking, network service chaining and matching theory.

Angelos Antonopoulos received the Ph.D. degree from the Technical University of Catalonia (UPC), in 2012. He is currently a Researcher with CTTC/CERCA. He has authored over 70 peer-reviewed publications on various topics, including energy efficient network planning, 5G wireless networks, cooperative communications and network economics. He received the Best Paper Award at IEEE GLOBECOM 2014 and EuCNC 2016, the Best Demo Award at IEEE CAMAD 2014 and the First Prize at IEEE ComSoc Student Competition.

Nizar Zorba received the B.Sc. degree in electrical engineering from Jordan University of Science and Technology, Irbid, Jordan, in 2002, the M.Sc. degree in data communications and the MBA degree from the University of Zaragoza, Zaragoza, Spain, in 2004 and 2005, respectively, and the Ph.D. degree in signal processing for communications from Universitat Politecnica de Catalunya, Barcelona, Spain, in 2007. He led and participated in over 25 research projects; and authored five patents, two books, seven book chapters, and over 100 peer-reviewed journals and international conferences. His research interests are quality of service/experience, energy efficiency, and resource optimization.

Christos Verikoukis (Ph.D., UPC, 2000) is a Fellow researcher at CTTC/CERCA and an adjunct associate professor at UB. He is a co-author of 4 books, 18 chapters, 2 patents, 108 journal papers and over 180 conference papers. He participated in more than 30 competitive projects and served as the principal investigator of national projects. He supervised 15 Ph.D. students and 5 postdoctoral researchers. He received a best paper award in IEEE ICC 2011, IEEE GLOBECOM 2014 & 2015, EUCNC/EURACON2016 and the EURASIP 2013 Best Paper Award for the Journal on Advances in Signal Processing. He is currently Chair of the IEEE ComSoc CSIM TC.