

Modelling the Admission Ratio in NFV-based Converged Optical-Wireless 5G Networks

Mohammadreza Mosahebfard, *Member, IEEE*, John Vardakas, *Senior Member, IEEE*,
and Christos Verikoukis, *Senior Member, IEEE*

Abstract—Network Function Virtualization (NFV)-based 5G networks deliver specific services to a set of users through the creation of Service Function Chains (SFCs), which are composed of a number of Virtualized Network Functions (VNFs) interconnected via a set of virtual links (vLinks). VNFs consume computational resources of the network's servers, while vLinks utilize the communicational resources of the network. The efficient utilization of network resources remains a challenge in NFV-based networks. To this end, in this paper, we propose an analytical framework for the calculation of the Admission Ratio (AR) in NFV-based converged optical-wireless 5G networks. The proposed methodology employs a network slicing architecture in which different network slices form end-to-end logically isolated networks and each slice delivers a specific service type to the users through its SFC(s). In the proposed analytical model, we not only take into account the occupancy distribution in both computational and communicational domains of the network resources (servers and fiber links), but we also consider the SFC establishment AR by taking into account different sub-service-classes belonging to different slices. The accuracy of the model is evaluated through the comparison of analytical and simulation results and was found satisfactory. Furthermore, the network dimensioning is performed by employing the proposed model for the determination of the optimal (minimum) capacity for all SFCs elements (VNFs and vLinks) in a way that users belonging to a specific slice experience a predefined value of AR as minimum. Additionally, our calculations deploy recursive formulas, which have a low computational complexity, as opposed to time-consuming simulation approaches without requiring the application of complex optimization algorithms.

Index Terms—Analytical model, admission ratio, network slicing, network function virtualization, resource management.

I. INTRODUCTION

5G networks on one hand are expected to provide high data rate, ultra-low latency, high user mobility, ultra-reliable, and ultra-dense communications. On the other hand, they should be able to provide a variety of services to the end-users, each with distinctive characteristics and requirements, such as autonomous driving, augmented and virtual reality, tactile Internet, and smart city to name a few [1], [2]. In order to achieve the 5G network expectations and provide the mentioned use cases to the end-users, the next generation mobile networks are expected to be more agile, flexible, scalable and software configurable by utilizing a set of emerging technologies such as Software Defined Networking (SDN), Network Function Virtualization (NFV), and Network Slicing (NS).

M. Mosahebfard and J. Vardakas are with Iquadrat Informatica, Barcelona, Spain (e-mail: m.mosahebfard@iquadrat.com; jvardakas@iquadrat.com).

C. Verikoukis is with the Telecommunications Technological Center of Catalonia (CTTC/CERCA), Castelldefels, Spain (e-mail: cverik@cttc.es).

The deployment of Passive Optical Networks (PONs) at the fronthaul and backhaul is an efficient solution for the provision of a reliable and fast connection to the end-users of the 5G network [3], [4], [5]. Additionally, the employment of the converged optical-wireless infrastructure is essential to meet the overall 5G network requirements. Eventually, this convergence will lead to reductions in capital and operational expenditure, as well as increased flexibility and scalability for mobile network operators [6], [7].

On one hand, SDN is a novel paradigm simplifying the data plane entities such as switches and routers and abstracts their intelligence using one or more SDN controllers as control plane entities. This separation enables network programmability, which allows to dynamically allocate resources and enforce policies based on the nature of required services. The separation also allows intelligence to be moved from devices to a control plane that manages the overall devices. In other words, SDN is in charge of communicational resources management [8], where the communication between the control plane and the data plane entities is ensured by a southbound communication protocol, such as the OpenFlow [9], [10]. On the other hand, NFV is responsible for managing the computational resources residing inside data-centers including CPU, memory, and storage. It moves legacy physical network functions from dedicated hardware towards software applications named Virtualized Network Functions (VNFs), which are installed on top of servers in data-centers. The exploitation of NFV results in a significant reduction in the delivery time of new services as well as the network's capital and operational expenditures. That is, in an NFV based network, a number of VNFs are chained in a predefined order to form a specific Service Function Chain (SFC) and consequently deliver a specific service. In order to implement such services, one critical task is to perform the SFC placement in the underlying physical network bound to diverse resource and service requirements [11], [12].

In addition to SDN and NFV, NS is an appropriate answer to the challenge related to the effective management of a wide variety of services with distinct characteristics and requirements in a single physical network. The NS technique focuses on the execution of a number of logically isolated networks on top of a common shared physical infrastructure [13]. It dedicates different slices to different service types such as the three 3GPP service types: i) enhanced Mobile Broadband (eMBB) ii) massive Machine-Type Communications (mMTC) iii) Ultra-Reliable and Low Latency Communication (URLLC).

Each slice provides service to a number of sub-service-

classes, which are services with similar Quality of Service (QoS) requirements. For instance, industrial automation, autonomous driving, and remote medical assistance are the sub-service-classes of the URLLC service type, and will be provided to the users by the URLLC slice [14], [15], [16]. One of the main issues in such NFV-based 5G networks is the joint management of both computational and communicational resources, while considering a number of SFCs providing connectivity to the end-users of different slices. Moreover, the Admission Ratio (AR), which determines the percentage of successful connections, can be considered as a practical criteria for the evaluation of such NFV-based converged optical-wireless networks in which users with different QoS requirements belonging to different slices are serviced. To this end, in this paper we propose an analytical model for the calculation of the AR, by taking into account the statistical features of the traffic inside the servers, fiber links, and the SFC elements, virtual Links (vLinks) and VNFs. Finally, we apply the proposed analytical model to perform network dimensioning that is determining the optimal capacity values for vLinks and VNFs in each SFC in order to guarantee a predefined AR for users arriving at a specific slice.

A. Related Work

Resource management is a permanent topic during the evolution of 5G, and a key challenge for future 5G and beyond networks, which has been widely investigated specifically in virtualized environments. This problem has been studied mainly through the development of simulation environments and optimization frameworks. For example, authors in [17] consider an NFV-based architecture in which the VNF requests are served by the network operator who tries to maximize its own revenue. In this regard, they have investigated the joint resource allocation problems of admission control and VNF forwarding graph embedding and in addition to heuristic algorithms, they employed relaxation and SCA methods. In [18], a static network planning that jointly allocates the spectrum and computational resources is proposed in order to handle the SFC requests efficiently in inter-datacenter elastic optical networks (inter-data-center EONs). The objective of this work is minimizing the total number of deployed VNFs and spectrum resources of inter-data-center EONs. Qiu et al. propose a novel hierarchical network architecture by employing SDN, which combines cross-layer low and high altitude platforms into the common terrestrial cellular networks for achieving higher capacity and more coverage for underserved areas in a cost-effective and seamless manner [19]. In [20], authors analyze the situation in which a network operator aims to offer a variety of services to users using NFV, while trying to maximize the AR and minimize the placement cost. Their research focuses specifically on reliability-aware service placement considering the dynamic nature of service arrivals and departures. They have also employed an infinite horizon Markov decision process model for performing dynamic reliability-aware service placement, where they consider the simultaneous allocation of the main and backup servers. Authors in [21] consider the coexistence of SDN control flows

and client flows together with NFV. In this work, a joint optimization for synchronizing the state of VNF instances and scheduling the routes is proposed attempting to minimize the rule occupation cost and VNF deployment cost. In [22], authors have formulated an optimization problem with the objective of minimizing the number of SFC requests that get rejected by considering the infrastructure failure possibility. They have also studied different resource backups such as VNF backups and link backups in order for protecting network services from failures. Eramo et al. in [23] consider the resource dimensioning and routing problem in NFV architectures. They proposed heuristic algorithms for offline and online traffics with guaranteeing uniform computational resources and bandwidth occupancy in the servers and the link, respectively. They have also proposed an algorithm for minimizing the energy consumption resulting in high decrease in energy consumption but in the cost of SFC blocking. An energy-aware resource allocation algorithm is proposed in [24] for VNF placement, assigning VNFs to flows, and flow routing problems in SDN-based networks.

Finally, our previous work in [25] targets to efficiently allocate both optical and wireless resources in an SDN/NFV-based converged optical-wireless network architecture, by determining the slices of the network in a way that the specific delay and bandwidth requirements of the multiple services are met. However, this approach is simulation-oriented, which significantly limits the flexibility when applied to different network configurations, and thus it cannot be used for network dimensioning purposes.

To the best of our knowledge, this is the first work that proposes a mathematical framework for the calculation of the AR in NFV-based 5G networks, and for the determination of crucial network parameters (such as offered traffic load, VNFs and vLinks capacities, etc.), which guarantee that the resulted AR is above a predefined threshold. In the proposed analytical model, we not only take into account the occupancy distribution in both computational and communicational domains of the network (servers and fiber links), but also consider as the SFC establishment AR by taking into account different sub-service-classes belonging to different slices. Each network slice provides a particular service-type to the users assigned to it. Thereafter, as another novelty of this work compared with the state-of-the-art, we employ the proposed model in order to determine the optimal (minimum) capacity for all SFCs' elements (VNFs and vLinks) in a way that users belonging to a specific slice experience a predefined value of AR as minimum. Moreover, our calculations are performed with deploying recursive formulas and consequently employing the proposed model is computationally efficient compared to optimization-based approaches for performing network dimensioning.

B. Our contribution

To position our contribution in detail, in this paper we consider a converged optical-wireless 5G network, where the Remote Radio Heads (RRHs) are connected to a set of data-centers through a network of PONs. In the mentioned

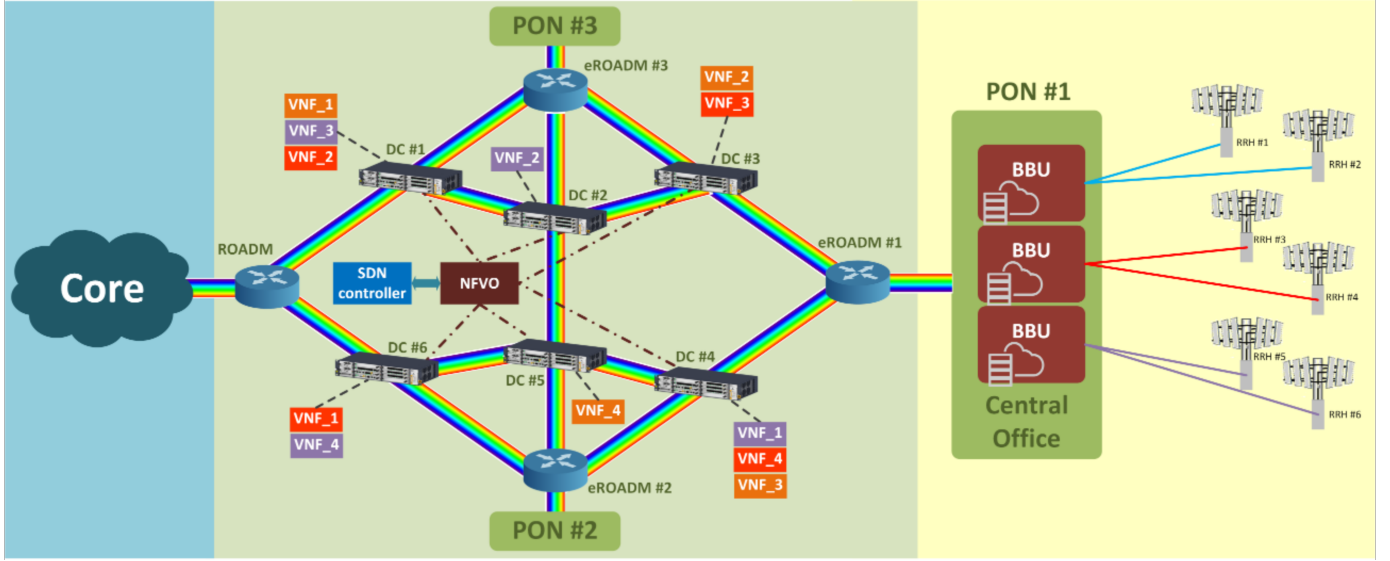


Fig. 1. Network architecture.

architecture, VNFs run on top of commodity servers, while the vLinks are initiated on the top of fiber links. The network is divided into multiple slices, where each slice is able to provide connectivity to end-users with different arrival and service rates requesting for different sub-service-classes. Furthermore, each slice can provide multiple SFCs. For this setup, we propose an analytical model for the AR calculation which is then applied in order to formulate a mathematical model for the determination of the optimal capacities for the VNFs and vLinks that guarantee a predefined value of AR. Our contributions are summarized as follows:

- An analytical model for the determination of the AR is proposed and evaluated. The proposed model determines the AR as the probability that a user connection is accepted for service when enough bandwidth resources at the communicational level and computational resources at the data-centers are available.
- The proposed model determines the occupancy distribution of the first VNF by employing the recursive formula in [26], which merely considers a single communication link supporting multiple service-classes. However, in our approach we take into account a network of data-centers interconnected through multiple fiber links. We furthermore calculate the service-rate and the interarrival time for SFC creation requests. We then employ the obtained occupancy distribution of the first VNF of the chain in order to calculate the AR inside the corresponding SFC. It is also worth to mention that the aforementioned occupancy distribution is extracted by employing the one-dimensional Markov chain, which significantly reduces the computational complexity of the proposed analytical framework.
- Additionally, by utilizing the obtained service-rate and the interarrival time for SFC creation requests, we calculate the SFC establishment AR. To this end, we employ the multirate Kaufman-Roberts formula presented

in [27], [28] and the Reduced Load Approximation (RLA) method ([29]) in order to determine the blocking probabilities in the computational and communicational infrastructures that are then used for the determination of the AR.

- The accuracy of the analytical model is validated through simulation and is found to be quite satisfactory. Furthermore, the proposed analytical framework is computationally efficient since it is based on recursive formulas.
- Moreover, the main contribution of the work is the application of the proposed mathematical framework for the effective determination of the optimal values of the system parameters (e.g. VNFs and vLinks capacities) that are required so that the overall AR is above a predefined threshold. It should be highlighted that the determination of the optimal values of the network resources is achieved by considering our proposed analytical model; this methodology has a very low computational complexity, as opposed to time-consuming simulation approaches.

C. Organization of the paper

The remainder of the paper is organized as follows. In Section II, we present the main contribution of this work by providing the system model's physical infrastructure and network slicing layers' description in Section II-A and II-B, respectively, while the SFC model and its placement constraints are provided in Section II-C and II-D, respectively. Section III describes the proposed Markov chain model for the SFC. The analysis for the determination of the AR and of the optimum network parameters are presented in Section IV and Section V, respectively. Section VI is our evaluation scenario, where we evaluate the effectiveness of the proposed analysis. Finally, in Section VII, conclusions and potential direction of future research are drawn. We also summarize the notations used throughout the paper in Table I.

TABLE I
NOTATION OF THE SYSTEM MODEL

Symbo	Description	Symbo	Description
$A_{s,k}$	Offered traffic load of sub-service-class k_s in slice s	$t_{s,f,n}^k$	Percentage of consumed computational resources by sub-service-class k_s in VNF $n_{s,f}$.
B_l	Capacity of fiber link l	V	Number of VNFs types in the network
C_d	Number of computational units in server d	$V_{s,f}$	VNFs of SFC f_s
D	Number of servers	<i>Greek symbols</i>	
E	Number of PONs	$\alpha_{s,f,n}$	Communicational coefficient of VNF $n_{s,f}$
f_s	f th SFCs in slice s	$\beta_{s,f,n}$	Computational coefficient of VNF $n_{s,f}$
F_s	Number of SFCs in slice s	Δ_d	Set of all VNFs deployed in server d
$h_{s,f,n}^d$	Binary variable to check if the VNF $n_{s,f}$ is placed in server d or not	Γ_l	Set of all vLinks utilizing physical link l to steer their traffic
$j_{s,f,n}^l$	Binary variable to check if vLink $n_{s,f}$ is utilizing link l or not	$\lambda_{s,f,k}$	Mean arrival rate for sub-service-class k in SFC f_s
K_s	Number of sub-service-classes in slice s	$\lambda_{s,k}$	Mean arrival rate for sub-service-class k_s
L	Number of fiber links	$\lambda_{s,k,e}$	Mean arrival rate arriving at PON e for sub-service-class k_s
$n_{s,f}$	n th VNF/vLink of SFC f_s	<i>Subscripts</i>	
$N_{s,f}$	Number of VNFs/vLinks in SFC f_s	d	Server number from the D server set
$P_{s,f}$	vLinks of SFC f_s	e	PON number from the E PON set
$R_{s,f,n}$	Occupied capacity by vLink $n_{s,f}$	f	SFC number from the F_s SFCs set
$r_{s,f,n}^k$	Consumed capacity by sub-service-class k_s in vLink $n_{s,f}$.	k	Sub-service-class number from the K_s sub-service-classes set
R_s^k	Required data rate for sub-service-class k_s	l	Link number from the L server set
S	Number of slices	n	VNF/vLink number from the $N_{s,f}$ VNF/vLink set
$T_{s,f,n}$	Number of occupied computational resources by VNF $n_{s,f}$	s	Slice number from the S slice set.

II. SYSTEM MODEL

In this paper, we consider a converged optical-wireless network configuration to provide high speed connectivity in the access domain. By employing a PON-based fronthaul, network operators are able to efficiently support services that can meet the highly challenging 5G operational framework, especially in two critical Next Generation Fronthaul Interface (NGFI) application scenarios: the ultra dense scenario, where thousands of users located in limited space city-landscape, and the hotspot scenario, where a dense population is located within very confined areas, such as arenas or stadiums [3], [30].

A. Physical infrastructure

As depicted in Fig. 1, we consider a network of PONs interconnected through D servers (residing in D different data-centers), L fiber links interconnecting the servers, and a number of Reconfigurable Optical Add-Drop Multiplexers (ROADMs) acting like optical switches. E PONs are connected to the servers-fibers network through the edge ROADMs (eROADMs). The eROADMs are responsible for steering the end-users' traffic into the inner part of the architecture. We assume that commodity server d ($d = 1, \dots, D$) is equipped with C_d units of computational resources including CPU cores, RAM, and storage. Finally, link l ($l = 1, \dots, L$) supports a total data rate of B_l . Moreover, each Baseband Unit

(BBU) of a specific RRH in a specific PON is considered to be installed in the central office residing in that PON. Each of these BBUs is optically connected with its corresponding RRH. We also take into account that the PON's Optical Network Unit (ONU) is a part of each RRH.

Furthermore, the NFV Orchestrator (NFVO) and the SDN controller as the resource management blocks are in charge of the management of the computational and communicational resources, respectively. That is, when the first user belonging to a particular slice requests for a specific sub-service-class, the NFVO decides where to install the VNFs for the corresponding SFC to that slice on the top of servers in data-centers. As the next step, the NFVO updates the SDN controller so that it interconnects the installed VNFs in the predefined order by assigning a specific amount of bandwidth to each vLink of the service chain. The characteristics of SDN do not affect the AR model and analyses, but it is considered as the communicational resources management entity in our network. Moreover, from the physical point of view, since two or more vLinks can be instantiated on a single link, they need to be differentiated from each other. To this end, either a Time Division Multiplexing (TDM) approach, or a Wavelength Division Multiplexing (WDM) approach, where the total capacity on each link is equal to the number of wavelengths multiplied by the capacity of each wavelength, can be deployed. In the proposed system model, ROADMs are the in-charge components for implementing the TDM/WDM

approaches in the network. It is also worth mentioning that the optical network elements, such as ROADMs and ONUs, do not impact the network performance, but they are a part of the network equipment and are necessary for such optical-wireless 5G networks to function.

B. Network slicing layer

At the network slicing layer, we consider that the network is divided into S distinctive slices, where slice s ($s = 1, \dots, S$) provides service to only one service type that supports K_s sub-service-classes. The sub-service-class connection requests have different bandwidth requirements, as well as different arrival and service procedures. In addition, there are F_s SFCs in slice s , where the f th SFC of slice s is represented by f_s ($f_s = 1, \dots, F_s$). The SFC f_s is made of $N_{s,f}$ VNFs, and $N_{s,f}$ vLinks. A vLink is formed of at least one physical link and interconnects two consecutive VNFs of the chain. To this end, we assume that the service chain starts from the first vLink connecting the corresponding eROADM to the first VNF of the chain and continues with an ordered set of VNFs interconnected through the rest of vLinks. Moreover, an Internet Gateway (GW) is present at each data-center and the output data of the last VNF is transmitted to the Internet through the GW. In addition, we consider E identical Connecting Paths (CPs) in each service chain in order that they play the role of the first vLink of that SFC, where their role is steering the arrived traffic at E different PONs to the first VNF of the corresponding service chain. In each SFC, the capacities of all CPs, which play the role of the first vLink, are considered to be equal as a single CP should be capable of steering the arrived traffic to the first VNF of the SFC in the case that all connection requests for a specific service chain arrive at a single PON.

Furthermore, there are V different types of VNFs in the network, (e.g. firewall, packet data network gateway, and load balancer). We represent the VNFs forming the SFC f_s as $V_{s,f} = \{v_i | i = 1, \dots, N_{s,f}, v_i \neq v_j \Leftrightarrow i \neq j\}$. As it is clear in the definition of $V_{s,f}$, none of the VNF pairs are from the same type. Similarly, the set $P_{s,f} = \{p_i | i = 1, \dots, N_{s,f}\}$ shows the vLinks of the f -th chain of slice s .

It is also assumed that end-users requesting for the k th sub-service-class of the supported service type in slice s arrive at the coverage area of a RRH in PON e , which is connected to the network through the eROADM e . The arrival procedure of the connection requests follows a Poisson process with mean arrival rate $\lambda_{s,k,e}$. Consequently, as the summation of independent Poisson processes will be a Poisson process, the total arrival rate of the resulting Poisson process for sub-service-class k_s in slice s arriving to VNF $n_{s,f,1}$ through E CPs will be:

$$\lambda_{s,k} = E \lambda_{s,k,e}. \quad (1)$$

On the other hand, the mean service time of sub-service-class k_s exponentially distributed is shown by $\mu_{s,k}^{-1}$. Consequently, the offered traffic load of the sub-service-class k_s in slice s is calculated by the following formula:

$$A_{s,k} = \frac{\lambda_{s,k}}{\mu_{s,k}}. \quad (2)$$

C. SFC features

The initial data rate that is requested by sub-service-class k_s is represented by R_s^k (bps), and is determined based on the corresponding Service Level Agreement (SLA) terms. In addition, VNF $n_{s,f}$ has two different coefficients named $\alpha_{s,f,n}$ (dimensionless) and $\beta_{s,f,n}$ (1/bps), which are the consumed communicational resources and computational resources coefficients, respectively. The total amount of added overhead to the input flow of the VNF is determined by multiplying the total input traffic of the VNF and $\alpha_{s,f,n}$. On the other hand, the total amount of consumed computational resources in VNF $n_{s,f}$ is calculated by multiplying the total input traffic of VNF $n_{s,f}$ by $\beta_{s,f,n}$. The aforementioned coefficients are used in the calculations of the capacities of vLinks and VNFs in each SFC presented through (3) and (4), respectively. We also calculate the usage percentage of the resources in each VNF and vLink for each arrival traffic through (5) and (6), respectively.

In order to determine the vLinks' capacities, it is assumed that the supported data rate of the first vLink of the chain is predefined and shown by $R_{s,f,1}$ and this vLink can handle the traffic of a group of users. In order to determine the supported data rate on the n th vLink of SFC f_s , we calculate the capacity of the first vLink of the chain by the multiplication of all rate coefficients of all prior VNFs to that vLink ($\alpha_{s,f,n}$, $n = 1, \dots, n-1$). This is logical as the added overhead to the traffic in each vLink, depends on all prior VNFs to that vLink. Consequently, the supported capacity for vLink $n_{s,f}$ is obtained as follows:

$$R_{s,f,n} = \left[R_{s,f,1} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right], \quad n = 2, \dots, N_{s,f}, \quad (3)$$

where $\lceil x \rceil$ denotes the least integer greater than or equal to x . Similarly, for calculating the number of computational resources allocated to the n th VNF of SFC f_s and by employing (3), we multiply the consumed computational resources coefficient of the VNF, $\beta_{s,f,n}$, by the supported data rate of the prior vLink to the corresponding VNF. Therefore, the number of consumed computational resources by VNF $n_{s,f}$ is represented by $T_{s,f,n}$ and is obtained by employing the following equation:

$$T_{s,f,n} = \left[R_{s,f,1} \beta_{s,f,n} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right], \quad n = 1, \dots, N_{s,f}. \quad (4)$$

Additionally, we need to calculate the usage percentage of the resources of SFC f_s by each input flow. This amount can be calculated in a similar way that we have obtained the capacity of vLinks and VNFs using (3) and (4). The percentage of the consumed capacity of vLink $n_{s,f}$ by k -th sub-service-class, $r_{s,f,n}^k$, directly depends on the demanded rate by the input traffic stream (R_s^k) and all the prior VNFs to the n th vLink of the chain. Thus, $r_{s,f,n}^k$ can be calculated as follows:

$$r_{s,f,n}^k = \left[\left(R_s^k \prod_{q=1}^n \alpha_{s,f,q} \right) * \frac{100}{R_{s,f,n}} \right]. \quad (5)$$

The percentage of the consumed computational resources of VNF $n_{s,f}$ by k -th sub-service-class, $t_{s,f,n}^k$, directly depends on the sub-service-class required rate (R_s^k) and all the prior VNFs in the corresponding chain. Therefore, $t_{s,f,n}^k$ is obtained by utilizing the following formula:

$$t_{s,f,n}^k = \left\lceil \left(R_s^k \beta_{s,f,n} \prod_{q=1}^{n-1} \alpha_{s,f,q} \right) * \frac{100}{T_{s,f,n}} \right\rceil. \quad (6)$$

We define the binary variable $h_{s,f,n}^d$ to check if the VNF $n_{s,f}$ is placed in server d or not:

$$h_{s,f,n}^d = \begin{cases} 1 & \text{if VNF } n_{s,f} \text{ is initiated in server } d \\ \text{otherwise} & \end{cases}. \quad (7)$$

Set Δ_d that shows all the VNFs placed in server d is also defined as follows:

$$\Delta_d = \{(s, f, n) \mid h_{s,f,n}^d = 1 \ \forall s, f, n\}. \quad (8)$$

On the other hand, for the communicational part, another binary variable is defined as follows to check if vLink $n_{s,f}$ is utilizing link l or not:

$$j_{s,f,n}^l = \begin{cases} 1 & \text{if vLink } n_{s,f} \text{ is placed in link } l \\ \text{otherwise} & \end{cases}. \quad (9)$$

Similar to the set Δ_d , the set Γ_l indicates all the vLinks utilizing physical link l to steer their traffic and is defined as follows:

$$\Gamma_l = \{(s, f, n) \mid j_{s,f,n}^l = 1 \ \forall s, f, n\}. \quad (10)$$

We define set $H_{s,f} = \{d \mid h_{s,f,n}^d = 1 \ \forall n \in \{1, \dots, N_{s,f}\}\}$ representing all the servers hosting the VNFs of SFC f_s , while set $J_{s,f} = \{l \mid j_{s,f,n}^l = 1 \ \forall n \in \{1, \dots, N_{s,f}\}\}$ shows all the links hosting the vLinks of service chain f_s .

D. SFC placement constraints

We have identified four constraints that should be considered in order to place a specific SFC's elements on the top of the network physical infrastructure. The first constraint is defined to guarantee that each VNF is deployed in only one server and cannot split. It can be written as the following equation:

$$\sum_{d=1}^D h_{s,f,n}^d = 1 \ \forall (s, f, n). \quad (11)$$

The second constraint is introduced for assuring that each vLink, which connects two consecutive VNFs, passes through at least one and at most all the network links. It is stated as follows:

$$1 \leq \sum_{l=1}^L j_{s,f,n}^l \leq L \ \forall (s, f, n). \quad (12)$$

Next constraint states that the total consumed computational resources in the whole network should be less than the available computational resources in all servers. More specifically, the amount of the occupied resources by VNFs on each server

should not exceed its capacity. It is expressed as the following inequality:

$$\sum_{s=1}^S \sum_{f=1}^{F_s} \sum_{n=1}^{N_{s,f}} h_{s,f,n}^d t_{s,f,n} \leq C_d \ \forall d = 1, \dots, D. \quad (13)$$

The last constraint is on the network's communication resources, where the total utilized capacity on the network links should be less than the summation of the available capacity of all physical links. In other words, the total occupied resources by vLinks passing through a specific link should not exceed the total capacity of that link. This constraint is stated as follows:

$$\sum_{s=1}^S \sum_{f=1}^{F_s} \sum_{n=1}^{N_{s,f}} j_{s,f,n}^l r_{s,f,n} \leq B_l \ \forall l = 1, \dots, L. \quad (14)$$

III. MARKOV CHAIN MODEL FOR SFC

As we mentioned in Section II-B, a network slice provides service to the users requesting for a specific service type by creating the corresponding SFC. Users belonging a slice are serviced by the established SFC, which is formed of a number of VNFs and vLinks placed on the top of physical infrastructure. When the first arriving user requests for a specific service, the NFVO initiates the first version of the corresponding SFC. In order to determine the mean service rate of SFC and the interarrival time for SFC creation requests we propose a Markov chain model for SFC.

The first SFC establishment process starts in the NFVO by arriving the first user asking for any of K_s sub-service-classes of corresponding service type in slice s . The requested service chain is created if and only if there are enough resources in both computational and communicational domains to deploy all the $N_{s,f}$ VNFs and $N_{s,f}$ vLinks of the SFC on the top of the physical infrastructure. On the other hand, the occupied resources of an existing SFC are released at the same time, right after the departure of the last active user in that chain. More specifically, by considering (3) and (4), it is concluded that all VNFs and vLinks of a specific SFC are occupied and become idle at the same time. The reason is that, each request consumes the same percentage of the resources on all VNFs, and vLinks of an SFC. Hence, in order to calculate the service time and the arrival rate (or equivalently the mean service rate and the interarrival time) for a specific SFC, it is enough to take into account one of the VNFs or vLink for the analyses. In this way, the obtained parameters will be the same for all elements of the service chain.

To this end, we construct a one-dimensional Markov chain for VNF $n_{s,f}$ inside SFC f_s , where the number of occupied resources in this VNF is considered to be the state of the system. Fig. 2 represents the constructed one-dimensional Markov chain for n th VNF of chain f in slice s , which supports $K_{s,f}$ different sub-service-classes. We employ the mentioned Markov chain for extracting the service chain's parameters, which will be utilized to calculate the SFC creation AR in the corresponding slice. For this purpose, in Section III-A, we will firstly determine the occupancy distribution of the VNFs by considering the aforementioned Markov chain. In

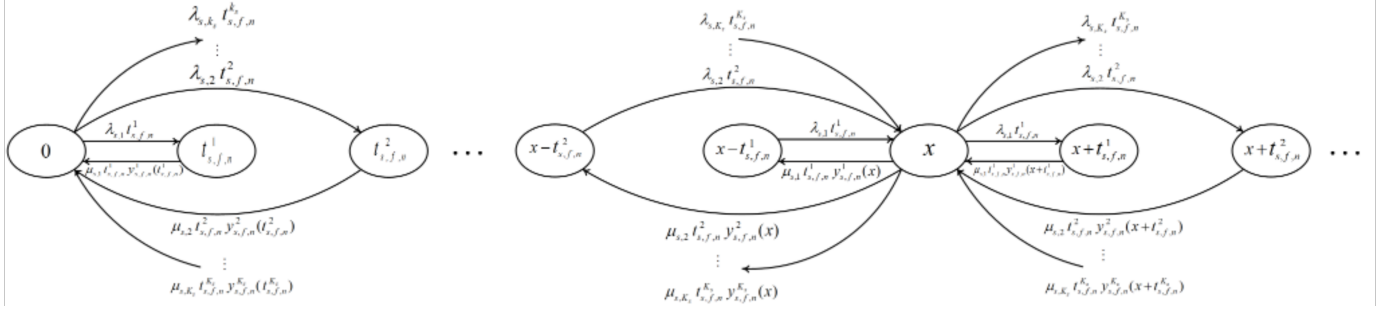


Fig. 2. One-dimensional Markov chain for the VNF $n_{s,f}$, which is the n th VNF of SFC f_s .

the next steps, this occupancy distribution will be considered for determining various parameters that are required in order to calculate the slice AR, e. g. the mean number of in-service users in the SFC, the mean service rate of the SFC, and the mean interarrival time for SFC creation request in Sections III-A, III-B, and III-D, respectively.

A. Occupancy distribution of VNF

In this part, we derive the occupancy distribution in VNF $n_{s,f}$ with total capacity of $T_{s,f,n}$, which handles the traffic of $K_{s,f}$ different sub-service-classes, each consuming $t_{s,f,n}^k$ of the total computational resources of the VNF. This occupancy distribution is obtained from the Markov chain of Fig. 2, by following the method in [26] that was considered for the determination bandwidth occupancy distribution in a single communication link that supports multiple service classes. This method considers the summation of all steady-state equations of all supported sub-service-classes, which provides the following recursive formula of the occupancy distribution in VNF $n_{s,f}$:

$$\left\{ \begin{array}{l} q'_{s,f,n}(x) = \frac{1}{x} \sum_{i=1}^{K_{s,f}} A_{s,k} t_{s,f,n}^i q'_{s,f,n}(x - t_{s,f,n}^i), \\ x = 1, \dots, T_{s,f,n}, \\ Q'_{s,f,n} = \sum_{i=0}^{T_{s,f,n}} q'_{s,f,n}(i), \\ q_{s,f,n}(x) = \frac{1}{Q'_{s,f,n}} q'_{s,f,n}(x), \end{array} \right. \quad (15)$$

where $q'_{s,f,n}(0) = 1$, $q'_{s,f,n}(x) = 0$ for $x < 0$, and $q'_{s,f,n}(x)$ and $q_{s,f,n}(x)$ are the unnormalized and normalized occupancy distribution of the VNF $n_{s,f}$, respectively.

B. Average number of in-service users in SFC

In this part, we employ the extracted occupancy distribution in Section III-A in order to determine the average number of in-service users in SFC. This number is a crucial parameter, which is required to obtain the mean service rate of the SFC discussed in Section III-C. In order to calculate this number for all sub-service-classes, a statistical equilibrium equation between the total request stream incoming to state x and the

total request stream outgoing from state x should be taken into account. This equation is written as follows:

$$\sum_{i=1}^{K_s} \lambda_{s,i} t_{s,f,n}^i q_{s,f,n}(x) = \sum_{i=1}^{K_s} \mu_{s,i} t_{s,f,n}^i y_{s,f,n}^i(x + t_{s,f,n}^i) q_{s,f,n}(x + t_{s,f,n}^i), \quad (16)$$

where $y_{s,f,n}^k(x + t_{s,f,n}^k)$ is the average number of active requests of sub-service-class k in state $(x + t_{s,f,n}^k)$. Equation (16) is satisfied if the local balance equations for the streams of every single sub-service-class is fulfilled. The local balance equation for sub-service-class k in SFC is derived from the Markov chain of Fig. 2 by considering the incoming and outgoing request streams for sub-service-class k between states x and $x + t_{s,f,n}^k$ and it is shown as follows:

$$\lambda_{s,f,k} t_{s,f,n}^k q_{s,f,n}(x) = \mu_{s,k} t_{s,f,n}^k y_{s,f,n}^k(x + t_{s,f,n}^k) q_{s,f,n}(x + t_{s,f,n}^k). \quad (17)$$

Thus, (16) is derived by summing up the corresponding equations of (17) for all sub-service-classes. By substituting x for $(x - t_{s,f,n}^k)$ in (17), the average number of in-service users of sub-service-class k in state x will be obtained by the following formula:

$$y_{s,f,n}^k(x) = A_{s,k} \frac{q_{s,f,n}(x - t_{s,f,n}^k)}{q_{s,f,n}(x)}. \quad (18)$$

C. Mean service rate of SFC

By using the occupancy distribution of VNF and the mean number of in-service users in SFC obtained in Sections III-A and III-B, respectively, we are able to calculate the mean service rate of SFC. The termination rate of the SFC is determined based on the rate that the number of occupied resources become zero. This happens when the last user departs from the system, who might be using any of the supported sub-service-classes. The rate $M_{s,f}$ is the release rate of the SFC f_s and is equal to the sum of the rates from state $t_{s,f,n}^k$, $k = 1, \dots, K_s$ leading to state 0 for all sub-service-classes, given that the system is in state $t_{s,f,n}^k$. This can be written as follows:

$$\begin{aligned} M_{s,f,n} &= \sum_{i=1}^{K_s} \mu_{s,i} t_{s,f,n}^i y_{s,f,n}^i(t_{s,f,n}^i) q_{s,f,n}(t_{s,f,n}^i) \\ &= \sum_{i=1}^{K_s} \mu_{s,i} t_{s,f,n}^i y_{s,f,n}^i(t_{s,f,n}^i) \frac{q_{s,f,n}(t_{s,f,n}^i)}{1 - q_{s,f,n}(0)}, \end{aligned} \quad (19)$$

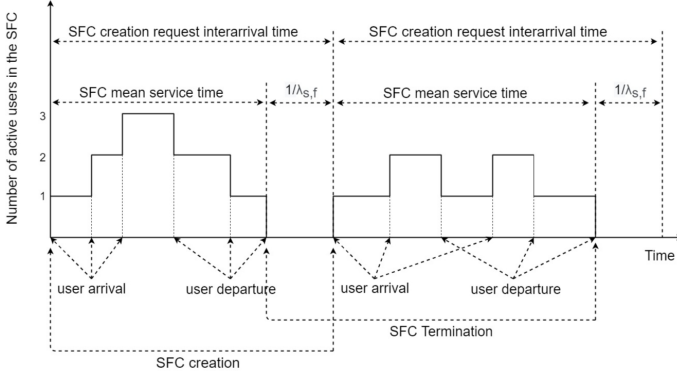


Fig. 3. Schematic diagram showing the interarrival time of SFC establishment requests and its service time.

where $\hat{q}_{f,n}(x)$ is the conditional probability that $x\%$ of the VNF capacity is occupied, given that the VNF resources are still occupied in the server. On the other hand, as we have mentioned in the beginning of the current section, all VNFs and vLinks of a specific SFC are occupied and become idle at the same time, we can consider only the first VNF of the service chain for calculating the mean service time of the SFC. To this end, we substitute (2) and (18) in (19) and we have the following formula for the mean service rate of SFC f_s :

$$M_{s,f} = \frac{q_{s,f,1}(0)}{1 - q_{s,f,1}(0)} \sum_{i=1}^{K_s} \lambda_{s,f,i} t_{s,f,1}^i. \quad (20)$$

D. Interarrival time for SFC creation request

After calculating the mean service rate of SFC, the interarrival time for SFC creation request needs to be calculated. An SFC is created through the arrival of a user's request that may belong to any sub-service-class. Thus, the total arrival rate of all sub-service-classes arriving at SFC f_s is required, which is represented by $\lambda_{s,f}$ and is calculated by the following formula:

$$\lambda_{s,f} = \sum_{i=1}^{K_s} \lambda_{s,f,i}. \quad (21)$$

In order to calculate the mean interarrival rate of the SFC creation requests, we consider the schematic diagram presented in Fig. 3. The time between two consecutive arrivals of SFC creation requests is equal to the sum of the mean SFC service time plus a delay until the next arrival occurs. Therefore, by dividing the mean service time of the SFC by the number $1/\lambda_{s,f}$, which is the total interarrival time, we can find the mean number of users that arrived within the SFC service time. By increasing the latter number by 1, we count the next user arrival, which actually is by itself a SFC creation request. Thus, interarrival time of the SFC creation request is stated as a multiple of mean requests' interarrival time as follows:

$$\frac{1}{\Lambda_{s,f}} = \left(\left\lfloor \frac{1}{\frac{M_{s,f}}{\lambda_{s,f}}} \right\rfloor + 1 \right) \frac{1}{\lambda_{s,f}}, \quad (22)$$

where $\Lambda_{s,f}$ is the SFC's arrival rate and $\lfloor x \rfloor$ denotes the smallest integer not exceeding x .

IV. AR CALCULATIONS

Having determined the SFC parameters, we now proceed on the determination of the AR. There are two cases that an arriving request from sub-service-class k_s is accepted:

- i) When there is not any active SFC in the corresponding slice and there are enough resources in both computational and communicational domains to deploy all VNFs and vLinks of the first SFC in that slice.
- ii) When there are enough resources in the existing SFC for handling the new arriving traffic flow. In other words, all the VNFs/vLinks of the existing service chain have enough capacity to process/steer the traffic of the arrival request.

By considering the two cases discussed above, the total AR in SFC f_s for sub-service-class k_s can be obtained by applying the following formula:

$$AR_{s,f,k} = AR_{s,f}^{sfc} AR_{s,f,k}^{user}, \quad (23)$$

where the terms $AR_{s,f}^{sfc}$ and $AR_{s,f,k}^{user}$ are the AR of the SFC creation and the AR of the user inside the SFC, respectively. The terms, $AR_{s,f}^{sfc}$ and $AR_{s,f,k}^{user}$, will be calculated in the Sections IV-A and IV-B, respectively.

A. SFC establishment AR

In this part, the AR of SFC establishment is determined by utilizing the parameters calculated in Sections III-C and III-D. In order to calculate the AR of the SFC creation request, two cases are considered, where an SFC creation request is not admitted:

- i) There are not enough capacity in at least one of the selected fiber links to launch the vLink(s) of the chain.
- ii) One or more VNFs cannot be hosted by the selected server(s) due to the lack of computational resources (CPU, RAM, and memory) in the server(s).

Consequently, the following formula is proposed to calculate the AR of the creation request of the service chain f_s :

$$AR_{s,f}^{sfc} = (1 - P_{s,f}^{Com.}) (1 - P_{s,f}^{Comp.}), \quad (24)$$

where $P_{s,f}^{Com.}$ represents the blocking probability in the communicational domain of the SFC and $P_{s,f}^{Comp.}$ is the SFC blocking probability in the computational domain, which are calculated by considering the analysis presented in Sections IV-A1 and IV-A2, respectively.

1) *Blocking probability in the communicational domain:* $P_{s,f}^l$ is the probability that vLink $n_{s,f}$ cannot be deployed on link l due to the lack of resources. It can be determined by considering the fact that each vLink utilizes one or more than one fiber links to interconnect two consecutive VNFs of the chain. To this end, we consider the RLA method [29], which has been developed in order to determine the blocking probabilities in a network that supports multiple source-destination routes of multiple links, by considering the offered traffic load for each route. More specifically, the RLA method considers the occupancy distribution of the resources in each link of the route, which is provided by (15), then it determines the blocking probability based on the approximate

determination of the reduction of the offered traffic load in each link of the route (the vLink in our analysis) due to the fact that there is a probability that all the resources are occupied in the rest of the links of the route under study.

Hence, to apply the RLA method in our approach, we consider the set Γ_l presenting the vLinks, which use physical link l as the medium to steer their traffic. Moreover, the characteristics of the offered traffic load of the SFCs, $M_{s,f}^{-1}$ and $\Lambda_{s,f}$, are considered as the arrival rate and service time of each vLink in link l . By employing (1) of [29], the probability that vLink $n_{s,f}$ cannot be deployed on top of all physical links of set $J_{s,f}$ is calculated through the following formula:

$$P_{s,f,n}^l = E \left[B_l, \sum_{(s,f,n) \in \Gamma_l} (M_{s,f,n} \Lambda_{s,f,n}) \prod_{i \in J_{s,f} - \{l\}} (1 - P_{s,f,n}^i) \right], \quad (25)$$

where

$$E[C; \rho] = \frac{\rho^C / C!}{\sum_{n=0}^C \rho^n / n!}$$

is the Erlang loss formula in which C is the capacity of the link and ρ is the offered traffic load of the vLinks that are going to be initiated on that link. By having all the blocking probabilities of creating vLink $n_{s,f}$ on all its host links, the blocking probability in the communicational domain of the SFC f_s is calculated through the following formula:

$$P_{s,f}^{Com.} = 1 - \prod_{n=1}^{N_{s,f}} (1 - P_{s,f,n}^l) \quad \forall l \in J_{s,f}. \quad (26)$$

2) *Blocking probability in the computational domain:* $P_{s,f,n}^d$ is the probability that n th VNF of SFC f_s cannot be initiated in server d due to the lack of computational resources. The arrival and service requests for the computational resources follow the same procedures as the corresponding communicational procedures, which allow the utilization of a traffic loss model for the determination of the targeted blocking probability. By using the recursive formula of (15), this time for each server d hosting VNFs of set Δ_d , the occupancy distribution inside server d ($q_d(x)$) is determined. It should be noted that for employing (15) for server d , the corresponding values of the capacity, arrival rate, and service time of VNFs of set Δ_d should be taken into account.

In the next step and by having $q_d(x)$, the value of $P_{s,f,n}^d$ can be obtained by employing the multirate Kaufman-Roberts formula presented in [27], [28]:

$$P_{s,f,n}^d = \sum_{x=C_d - T_{s,f,n} + 1}^{C_d} q_d(x). \quad (27)$$

Finally, the probability that one or more VNFs of the SFC f_s cannot be deployed in one or more servers of set $H_{s,f}$ is calculated through the following formula:

$$P_{s,f}^{Comp.} = 1 - \prod_{n=1}^{N_{s,f}} (1 - P_{s,f,n}^d) \quad \forall d \in H_{s,f}. \quad (28)$$

By substituting (26) and (28) in (24) the SFC establishment AR is obtained. Next, we calculate the AR inside the SFC.

B AR inside the SFC

A connection request in a specific slice is granted if there is at least an active service chain with enough space in all of its VNFs and vLinks to handle the new asked traffic. In other words, provided that one VNF or vLink of the service chain is full, the request will be blocked. As it was discussed in Section III, all the VNFs and vLinks of an SFC, get full and become idle at the same time. Therefore, the user blocking probability inside an SFC can be stated based on the probability that the first VNF of that chain gets full. The probability that the first VNF ($n_{s,f,1}$) does not have enough capacity to handle the new traffic flow of the sub-service-class k_s is determined by summing up the probabilities of all the blocking states and is defines as follows:

$$P_{s,f,k}^{user} = P_{s,f,1,k}^{VNF} = \sum_{x=T_{s,f,1} - t_{s,f,1}^k + 1}^{T_{s,f,1}} q_{s,f,1}(x), \quad (29)$$

where $q_{s,f,1}(x)$ is the occupancy distribution of the first VNF in SFC f_s obtained from (15), $t_{s,f,1}^k$ is the usage percentage of the user's flow in the first VNF, and it is assumed that $T_{s,f,1} = 100$ as $t_{s,f,1}^k$ is stated in percentage. Hence, the probability that a user's request gets accepted inside the existing service chain is calculated by the following formula:

$$A_{s,f,k}^{user} = 1 - P_{s,f,k}^{user}. \quad (30)$$

By substituting (24) and (30) in (23), the total AR for k th sub-service-class in slice s is obtained.

Algorithm 1 represents the steps of AR calculation based on the proposed model. At the beginning, we input the necessary parameters to initiate the algorithm and determine $\lambda_{s,k}$ and $A_{s,k}$ using (1) and (2). Then, we have a loop with a length of S . In each iteration of this loop, we have an inner loop with the length of $N_{s,f}$ (equal to number of VNFs and vLinks in SFC f_s) in Lines 4-11. In each iteration of this inner loop, the number of occupied computational resources by VNFs and the amount of occupied communicational resources by vLinks of each SFC are determined. Moreover, we have another inner loop with the length of K_s (equal to the number of sub-service-classes in slice s) in Lines 12-21. In each iteration of this inner loop, the percentage of consumed computational and communicational resources by each sub-service-classes flow are obtained. Next, in the occupancy distributions of all VNFs and vLinks are determined in Lines 23-25.

Afterwards, in Lines 26-28, for each slice, we calculate the mean service rate of the first SFC and the interarrival time for SFC creation request. Then, the probability that vLink $n_{s,f}$ cannot be deployed on top of all physical links of set $J_{s,f}$ is calculated for all Γ_l presenting the vLinks, which use physical link l as the medium to steer their traffic (Lines 30-32). In Line 34, the blocking probability in the communicational domain of the SFC f_s is determined using (26). Similar to $P_{s,f,n}^l$, the probability that n th VNF of SFC f_s cannot be initiated in server d due to the lack of computational resources ($P_{s,f,n}^d$) is calculated in Lines 35-40. Then in Line 41, we determine the probability that one or more VNFs of the SFC f_s cannot be deployed in one or more servers of set $H_{s,f}$. In the Next step, the probability that the first VNF ($n_{s,f,1}$) does not have enough

Algorithm 1 AR calculation's steps.

```

1: Input: Network graph, predefined paths,  $\lambda_{s,k,e}$ ,  $\mu_{s,k}$ ,
 $r_{s,f,1}^k = R_s^k$ ,  $E$ ,  $C_d$ ,  $B_l$ ,  $\beta_{s,f,n}$ ,  $\alpha_{s,f,n}$ ,  $R_{s,f,1}$ ,  $S$ ,  $F_s$ ,  $N_{s,f}$ ,
 $V$ ,  $h_{s,f,n}^d$ ,  $j_{s,f,n}^l$ ,  $h_{s,f,n}^d$ ,  $j_{s,f,n}^l$ ,  $\Delta_d$ ,  $\Gamma_l$ .
2: Determine  $\lambda_{s,k}$  and  $A_{s,k}$  using (1) and (2).
3: for  $s = 1, \dots, S$  do
4:   for  $n = 1, \dots, N_{s,f}$  do
5:     if  $n = 1$  then
6:        $R_{s,f,1}$  is predefined.
7:     else
8:       Calculate  $R_{s,f,n}$  with (3).
9:     end if
10:    Calculate  $T_{s,f,n}$  employing (4).
11:  end for
12:  for  $k = 1, \dots, K_s$  do
13:    for  $n = 1, \dots, N_{s,f}$  do
14:      if  $n = 1$  then
15:         $r_{s,f,1}^k = R_s^k$ .
16:      else
17:        Calculate  $r_{s,f,n}^k$  with (5).
18:      end if
19:      Calculate  $t_{s,f,n}^k$  by using (6).
20:    end for
21:  end for
22: end for
23: for all VNFs & vLinks do
24:   Use (15) to calculate  $q_{s,f,n}(x)$ .
25: end for
26: for  $s = 1, \dots, S$  &  $f_s = 1$  do
27:   Calculate  $M_{s,f}$  and  $\Lambda_{s,f}^{-1}$  with (20) and (22).
28: end for
29: for  $l = 1, \dots, L$  do
30:   for all vLinks  $n_{s,f}$  where  $(s, f, n) \in \Gamma_l$  do
31:     Employ (25) to calculate  $P_{s,f,n}^l$ .
32:   end for
33: end for
34: Calculate  $P_{s,f}^{Com.}$  employing (26).
35: for each server  $d=1, \dots, D$  do
36:   Calculate  $q_d(x)$  using (15).
37:   for all VNFs  $n_{s,f}$  where  $(s, f, n) \in \Delta_d$  do
38:     Calculate  $P_{s,f,n}^d$  with (27).
39:   end for
40: end for
41: Calculate  $P_{s,f}^{Comp.}$  employing (28).
42: for slice  $s = 1, \dots, S$  do
43:   Calculate  $P_{s,f,k}^{user}$  with (29).
44: end for
45: Determine  $AR_{s,f}^{sfc}$  and  $A_{s,f,k}^{user}$  using (24) and (30).
46: Calculate  $AR_{s,f,k}$  with (23).

```

capacity to handle the new traffic flow of the sub-service-class k_s is calculated (Lines 42-44). Finally, after determining the user and SFC AR in Line 45, the total AR in SFC f_s for sub-service-class k_s can be obtained.

V. NETWORK DIMENSIONING

One of the applications of the proposed model is determining the minimum capacity for each VNF and vLink in each SFC in the network for achieving a predefined amount of AR. In this Section, we employ the proposed model and the aforementioned analysis for the determination of the optimal network parameters to satisfy the AR requirements in each slice. More specifically, this procedure is realized by reversing the AR calculations, so that the optimal (minimum) capacity for the first VNF of the SFC is obtained, in a way that the AR remains above a predefined threshold. Afterwards, we calculate the minimum required capacities for the vLinks of SFC through the following formula:

$$R_{s,f,n} = \left[R_{s,f,1} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right], \quad n = 2, \dots, N_{s,f}, \quad (31)$$

where

$$R_{s,f,1} = \frac{T_{s,f,1}}{\beta_{s,f,1}}.$$

Similarly, the minimum required capacities for the rest of VNFs of the service chain is obtained by applying the following equation:

$$T_{s,f,n} = \left[R_{s,f,1} \beta_{s,f,n} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right], \quad n = 2, \dots, N_{s,f}. \quad (32)$$

The input parameters for performing the network dimensioning analysis are predefined values of AR for each slice, the number of created SFCs in each slice, and the required data rates for sub-service-classes in each slice.

VI. EVALUATION AND DISCUSSION

In this Section, we evaluate the accuracy of the proposed mathematical framework, by comparing the analytical results, with the corresponding results from a custom-made system-level simulator, by considering two evaluation scenarios, a small-scale and a large-scale network. Moreover, this Section showcases the effectiveness of our proposed mathematical model in determining the optimal network parameter values (i.e. offered traffic load, VNF size, etc.) that are required so that the total AR is above a predefined threshold.

A. Evaluation of the model's accuracy

In order to evaluate the accuracy of the proposed model, we consider two different simulation setups, a small-scale and a large-scale network depicted in Fig. 4 and 5, respectively, data-centers are represented with green and ROADMs with red. In both setups, the capacity of each data-center is set to 50 units of computational resources ($C_d = 50$) and the capacity of each fiber link is set to 40 Gbps ($B_l = 40 \times 10^9$). In the small-scale network case there are four data-centers, as in the case of the evaluation scenario of [31], interconnected through four ROADMs and 26 fiber links, where the eROADMs (nodes 3 and 6) are connected to two PONs. On the other hand, in the large-scale network we consider nine data-centers, as in the case of the evaluation scenario of [32], interconnected through eight ROADMs and 72 fiber links, where the eROADMs

TABLE II
THE PREDEFINED PATHS AND HOST SERVERS CONSIDERED IN THE EVALUATIONS OF THE PROPOSED MODEL (SMALL-SCALE NETWORK).

	Slice 1 SFC 1, SFC 2	Slice 2 SFC 1, SFC 2	Slice 3 SFC 1, SFC 2
Host servers for VNFs	{2,5,4}, {4,1,5}	{4,2,1}, {1,5,2}	{5,2,4} {1,4,2}
Links for CPs - SFC 1	{5}, {6}	{7,20}, {12}	{6,22,23}, {14,25,15}
Links for CPs - SFC 2	{12,2,16}, {13,11,26,20}	{14,25,17,20}, {7,21,15}	{13}, {5,4,15}
Links for vLink 2	{4,18,23}, {9,21,15}	{8,5}, {1,13}	{24,12,2,16}, {2,17,20}
Links for vLink 3	{11,26,20}, {1,13}	{4,15}, {11,25,16}	{3,6}, {8,5}

TABLE III
THE PREDEFINED PATHS AND HOST SERVERS CONSIDERED IN THE EVALUATIONS OF THE PROPOSED MODEL (LARGE-SCALE NETWORK).

	Slice 1 SFC 1, SFC 2	Slice 2 SFC 1, SFC 2	Slice 3 SFC 1, SFC 2
Host servers for VNFs	{2,6,11,14}, {16,9,4,7}	{14,7,4,2}, {6,9,12,16}	{12,7,9,1} {16,9,6,14}
Links for CP 1	{6}, {9,67,68}	{26,61,71,59}, {6,17,24}	{14,29,48}, {60,66}
Links for CP 2	{7}, {45,69,59,55,65}	{23,40,57}, {23}	{23,40,53,49}, {7,1,26,61,66}
Links for CP 3	{50,54,38,21,8}, {66}	{66,56,58}, {28,4,6,10,23}	{50}, {71,67}
Links for CP 4	{64,5,6}, {67}	{59}, {47,42,25}	{67,55,49}, {67}
Links for vLink 2	{17,24}, {56,38}	{70,64,28}, {19,35}	{62,28}, {63,50,33,36}
Links for vLink 3	{40,46}, {32,16}	{4,14}, {32,48}	{31,36}, {21,24}
Links for vLink 4	{69,59}, {3,26}	{18,8}, {54,65}	{41,46}, {40,57}

(nodes 1,3,15 and 17) are connected to four PONs. We evaluate the accuracy of the proposed analytical model through different examples, where the analytical results are compared with corresponding simulation results. To this end, we assume that there are four different types of VNFs in the network ($V = 4$). Furthermore, in order to perform a fair comparison between the AR of the three slices, we assume that all SFCs in small-scale network are formed of three VNFs and three vLinks ($\forall s = 1, \dots, 3, N_s = 3$), while this number for another setup is set to four ($\forall s = 1, \dots, 3, N_s = 4$). Moreover, the parameters for all deployed VNFs are $\alpha_{s,f,n} = 2 \times 10^{-9}$ (1/bps) and $\beta_{s,f,n} = 1.05$. For each slice, the arrival requests are from the same sub-service-class class ($K_s = 1, \forall s$) and their traffic characteristics at each of the PONs are $\mu_{1,1}^{-1} = \mu_{2,1}^{-1} = \mu_{3,1}^{-1} = 0.5$ minutes and $R_1^1 = 100$, $R_2^1 = 90$, and $R_3^1 = 80$ Mbps. For the small-scale setup, in each SFC, by employing (3) and (4), we will have 4, 4.2, and 4.4 all in Gbps as the capacities of the vLink 1 to vLink 3, respectively, and 8, 9, and 9 computational resources units as the capacities of the VNF 1 to VNF 3, respectively. Similarly, for the large-scale network setup, the capacity of vLinks 1 to 4 are 5, 5.5, 5.5, and 6 all in Gbps, respectively, and 10, 11, 12, and 12 computational resources units as the capacities of the VNF 1 to VNF 4, respectively. We consider two evaluation scenarios for each of the aforementioned networks. The first scenario refers to the case of a single SFC per slice, while the second scenario refers to the case where there are two SFCs in each network slice. Based on the SFC deployment constraints presented in Section II-D, we consider a set of predefined host nodes to deploy the VNFs for each SFC, and subsequently, a number of predefined paths for interconnecting the initiated VNFs in both small-scale and large-scale networks, which are

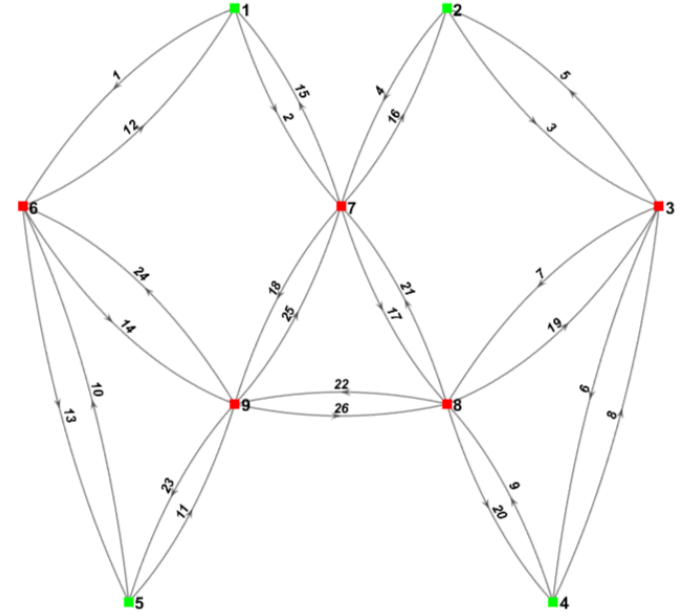


Fig. 4. Considered small-scale network graph for the simulations.

presented in Tables II and III, respectively. The simulation results for the AR of requests per slice are presented as mean values from seven runs with confidence interval of 95%. For each simulation run we assume 1,000,000 originating requests, while a stabilization period that corresponds to the first 100,000 requests is considered in order for the system to reach the steady state.

The analytical and simulations results of the AR for the single SFC per slice case are presented in Tables IV and VI,

TABLE IV
THE ANALYTICAL VERSUS SIMULATIONS RESULTS FOR THE AR IN SMALL-SCALE NETWORK CASE - ONE SFC PER SLICE: $F_s = 1$.

Arrival rate	Slice 1		Slice 2		Slice 3	
	Analysi	Simulations	Analysi	Simulation	Analysi	Simulations
50	0.998	0.9977 ± 0.0001	0.999	0.9998 ± 0.0000	1.000	1.0000 ± 0.0000
55	0.994	0.9923 ± 0.0002	0.999	0.9991 ± 0.0001	1.000	0.9999 ± 0.0000
60	0.985	0.9806 ± 0.0004	0.996	0.9965 ± 0.0002	0.999	0.9996 ± 0.0001
65	0.969	0.9612 ± 0.0007	0.990	0.9901 ± 0.0002	0.999	0.9985 ± 0.0001
70	0.946	0.9348 ± 0.0009	0.978	0.9777 ± 0.0005	0.996	0.9951 ± 0.0002
75	0.918	0.9038 ± 0.0013	0.960	0.9590 ± 0.0006	0.991	0.9881 ± 0.0005
80	0.887	0.8699 ± 0.0014	0.936	0.9346 ± 0.0010	0.981	0.9759 ± 0.0008

TABLE V
THE ANALYTICAL VERSUS SIMULATIONS RESULTS FOR THE AR IN SMALL-SCALE NETWORK CASE - TWO SFCs PER SLICE: $F_s = 2$.

Arrival rate	Slice 1		Slice 2		Slice 3	
	Analysi	Simulations	Analysi	Simulation	Analysi	Simulations
100	0.999	1.0000 ± 0.0000	0.999	1.0000 ± 0	0.999	1.0000 ± 0
110	0.999	0.9993 ± 0.0001	0.999	1.0000 ± 0.0000	0.999	1.0000 ± 0
120	0.997	0.9961 ± 0.0002	0.999	0.9999 ± 0.0005	0.999	1.0000 ± 0
130	0.990	0.9855 ± 0.0006	0.998	0.9989 ± 0.0002	0.999	1.0000 ± 0.0000
140	0.974	0.9659 ± 0.0008	0.994	0.9950 ± 0.0002	0.999	0.9997 ± 0.0000
150	0.949	0.9358 ± 0.0017	0.984	0.9841 ± 0.0003	0.999	0.9985 ± 0.0002
160	0.917	0.9006 ± 0.0019	0.965	0.9647 ± 0.0009	0.996	0.9937 ± 0.0006

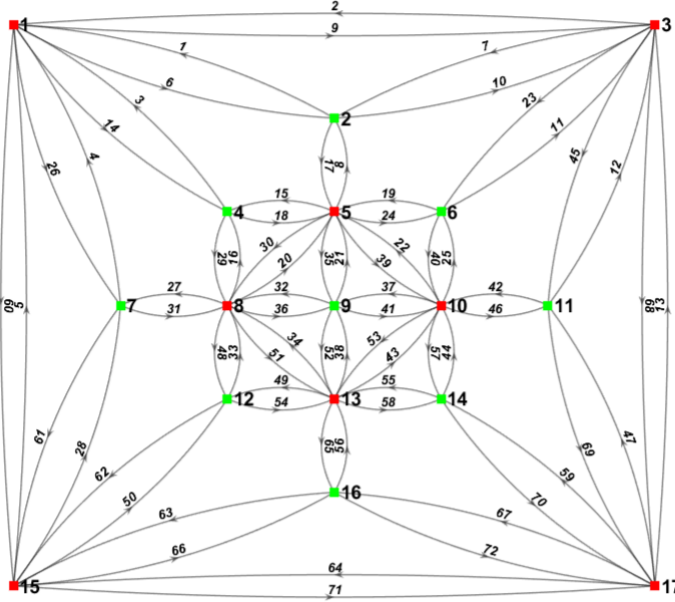


Fig. 5. Considered large-scale network graph for the simulations.

while the results for the case in which we consider two SFCs per slice are presented in Tables V and VII. The analytical results for AR are obtained from (23), which requires the use of (24) and (30). As the comparison of the analytical and the corresponding simulation results of Tables IV-VII reveal, the accuracy of the proposed analytical models is completely satisfactory. To be more specific, we calculate the relative error between analytical and simulation results, in order to quantify the accuracy of the proposed analysis. The relative error is

calculated by the following formula:

$$E = \left| \frac{v_a - v_s}{v_s} \right| * 100 \%,$$

where v_s is the value obtained from the simulations, while v_a is the value extracted from the analytical model. By considering Table IV and the results related to arrival rate of 80 connection requests per minute for slice 1, which is the case with the largest gap between the analytical and simulations values, we calculate the maximum related error:

$$E = \left| \frac{0.8699 - 0.8872}{0.8699} \right| * 100 \% = 1.99\%,$$

which shows the high accuracy of our proposed model. As can be seen in Tables IV and VI ($F_s = 1$) as well as V and VII ($F_s = 2$), by increasing the data rate the AR decreases as the number of in-service request in each slice increases, by having this in mind that we do not change the mean service time of the arrived requests. Furthermore, we notice that the AR for slice 3 is higher compared with the other two slices as its requested data rate is smaller. Moreover, it is obvious that for the second case, where there are two SFCs in each slice, the users experience higher values of the AR the load values are doubled.

B. Network dimensioning results

In this Section, we employ the proposed mathematical model, which calculates the optimal network parameters while the AR is above a predefined threshold in a significantly shorter computational time compared to simulations. To this end, we consider the aforementioned set of values of the network parameters in Section VI-A in order to determine the

TABLE VI
THE ANALYTICAL VERSUS SIMULATIONS RESULTS FOR THE AR IN LARGE-SCALE NETWORK CASE - ONE SFC PER SLICE: $F_s = 1$.

Arrival rate	Slice 1		Slice 2		Slice 3	
	Analysi	Simulations	Analysi	Simulation	Analysi	Simulations
70	0.996	0.9966 ± 0.0004	0.999	0.9995 ± 0.0001	1.000	1.0000 ± 0.0000
75	0.991	0.9913 ± 0.0007	0.998	0.9985 ± 0.0003	0.999	0.9999 ± 0.0000
80	0.981	0.9814 ± 0.0011	0.995	0.9957 ± 0.0004	0.999	0.9997 ± 0.0001
85	0.966	0.9664 ± 0.0011	0.989	0.9898 ± 0.0005	0.999	0.9989 ± 0.0002
90	0.945	0.9462 ± 0.0013	0.979	0.9800 ± 0.0007	0.997	0.9970 ± 0.0005
95	0.921	0.9222 ± 0.0014	0.965	0.9652 ± 0.0009	0.993	0.9931 ± 0.0008
100	0.895	0.8954 ± 0.0018	0.946	0.9464 ± 0.0013	0.986	0.9861 ± 0.0009

TABLE VII
THE ANALYTICAL VERSUS SIMULATIONS RESULTS FOR THE AR IN LARGE-SCALE NETWORK CASE - TWO SFCs PER SLICE: $F_s = 2$.

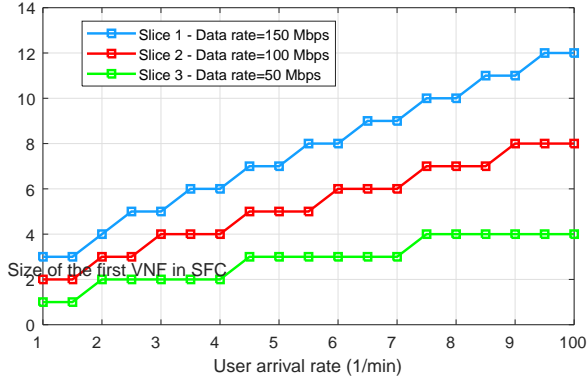
Arrival rate	Slice 1		Slice 2		Slice 3	
	Analysi	Simulations	Analysi	Simulation	Analysi	Simulations
140	0.999	0.9997 ± 0.0001	0.999	1.0000 ± 0.0000	1.000	1.0000 ± 0
150	0.999	0.9984 ± 0.0004	0.999	1.0000 ± 0.0000	1.000	1.0000 ± 0
160	0.996	0.9938 ± 0.0009	0.999	0.9997 ± 0.0000	1.000	1.0000 ± 0
170	0.988	0.9933 ± 0.0014	0.999	0.9987 ± 0.0002	1.000	1.0000 ± 0.0000
180	0.973	0.9652 ± 0.0018	0.996	0.9950 ± 0.0008	0.999	0.9999 ± 0.0000
190	0.951	0.9403 ± 0.0023	0.988	0.9867 ± 0.0018	0.999	0.9994 ± 0.0001
200	0.924	0.9119 ± 0.0025	0.975	0.9722 ± 0.0030	0.998	0.9977 ± 0.0004

optimal capacities of VNFs and vLinks within SFCs so that a predefined AR threshold can be achieved. We just consider different values for the required data rate in each slice, which are given as $R_1^1 = 150$, $R_2^1 = 100$, and $R_3^1 = 50$ all in Mbps. Fig. 6 depicts the minimum number of required computational resources for the starting VNF of the SFC versus the arrival rate for achieving the ARs equal to or higher than 99.99% for three different cases: Fig. 6(a) provides the first VNF's computational capacity values for the case where one SFC per slice is considered, while Fig. 6(b) and Fig. 6(c) illustrate this values for the cases of two and three SFCs per slice, respectively.

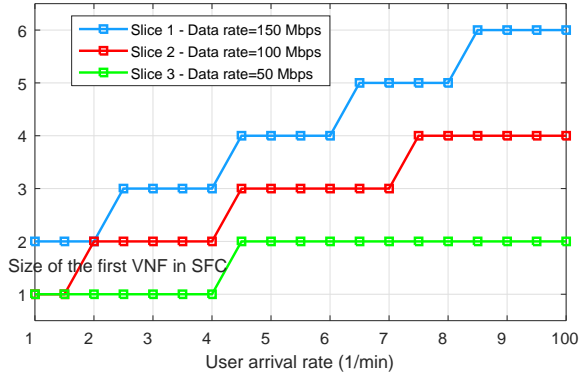
By comparing Figs. 6(a), 6(b), and 6(c), it can be concluded that the higher the number of SFCs is, the smaller size of VNF needs to be deployed in the servers. It can be also seen that by increasing the amount of arrival data rates, bigger VNFs and vLinks in terms of capacity are needed in the SFC for keeping the AR above a predefined value and this is the reason of ascending trend of the plots in Fig. 6. Furthermore, by comparing the three cases in Fig. 6 for a specific arrival rate in a particular slice, it can be seen that the number of occupied resources by the first VNF of the SFC and consequently the total number of occupied resources by the SFC is different in each case. As an example, for the arrival rate of 30 connection requests per minute in eMBB slice, in the case with one SFC per slice the first VNF of the chain utilizes 5 computational resources while in the case with two SFCs per slice, the first VNF of each SFC in eMBB slice occupies 3 computational resources, which results in the total number of 6 computational resources occupied by the first VNFs of the two chains residing in eMBB slice. Finally, it can be concluded that for a slice with less number of SFCs more number of users will be affected in

result of an SFC failure as more number of users are assigned to that SFC compared with the case that the slice has more number of in service chains. In other words, in the case with higher number of SFCs per slice, the users are divided between the existing SFCs and consequently, less number of users are affected in the result of possible SFC failures. To sum up, Fig. 6 aims at showing the amount of consumed computational resources by the first VNF of the chain and consequently the total amount of consumed resources by the SFC (all VNFs and vLinks) in order to guarantee a desired value of AR. This helps to calculate the size of SFCs precisely in order to guarantee a predefined value of AR.

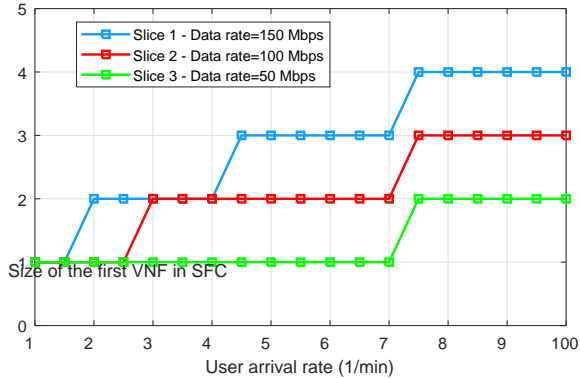
Apart from this, the proposed analysis may be used in order to determine the maximum values of the offered traffic load of each service type in different slices while guaranteeing a 100% AR. To this end, we consider the previous evaluation scenario in Section VI-A and assume that there are three slices each providing one of the 3GPP service types (eMBB, mMTC, and URLLC). Fig. 7 illustrates the maximum supported data rate for different slices each providing one of the 3GPP service types while guaranteeing AR value of 1. Fig. 7(a) illustrates the analytical results for the AR of the three service types versus the user arrival rate. In all cases, the AR drops below the value of 1 for a different threshold value of the user arrival rate. As it was anticipated, the eMBB service has the highest threshold value, followed by URLLC and mMTC, as it is a result of the resource requirements of each service type. More specifically, the maximum offered arrival rate that guarantees the AR of one in eMBB slice is 18 connection requests per minute, while this value for URLLC and mMTC is 65 and 550 connection requests per minute, respectively.



(a)



(b)

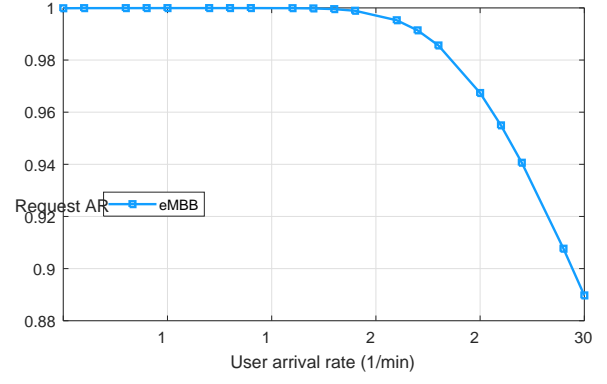


(c)

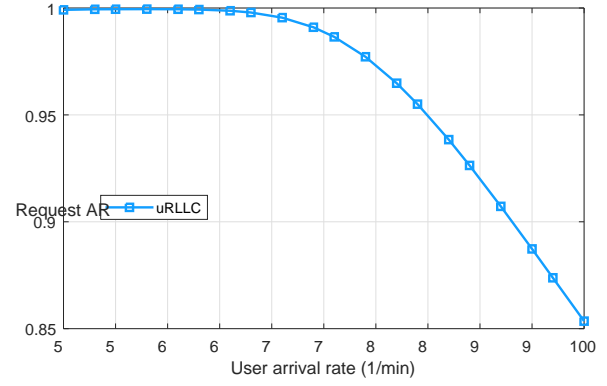
Fig. 6. The required number of computational resources for the first VNF of the SFC versus the arrival rate to guarantee that the ARs is equal to or more than 99.99% for three cases: (a) one SFC per slice, $F_s = 1$, (b) two service chains in slice, $F_s = 2$, and (c) three SFCs in slice, $F_s = 3$.

VII. CONCLUSION

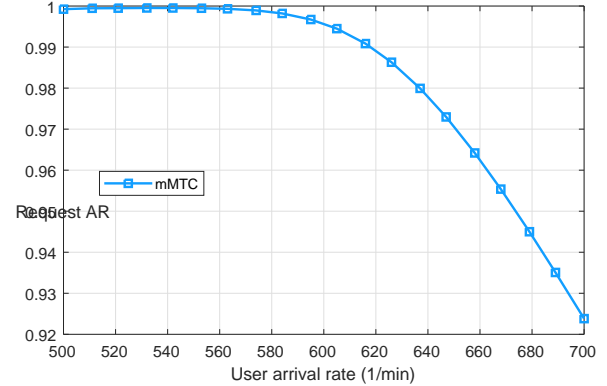
In this paper, we proposed an analytical model for the AR in NFV-based converged optical-wireless networks by taking into account different network slices each offering a number of sub-service-classes of a specific service type. The accuracy of the proposed model was confirmed by comparing the analytical with the simulation results for different cases. Furthermore, we employ this mathematical model to perform the network dimensioning in order to guarantee a predefined value of AR. More specifically, we determined the minimum required



(a)



(b)



(c)

Fig. 7. AR versus the arrival rate for the three 3GPP service types: (a) eMBB, (b) uRLLC, and (c) mMTC.

capacity for the VNFs and vLinks of the service chain(s) of the slice to keep the AR higher than a predefined threshold. We also concluded that when the number of SFCs is higher, the less optimum the usage of the infrastructure resources will be more efficient. On the other hand, by employing more number of SFCs in slice, less number of users will be affected by SFC failures, which are due to the possible operational failures in data-centers or communicational devices in the physical parts of the network. In our future work, we use the results gained from network dimensioning section as an input to guarantee the AR. In the next step, we will investigate the case that

we have different copies of the requested VNFs in different network's servers. In this case, the network's resource manager entity tries to select the VNFs and set the connecting paths in a way that the whole network energy consumptions is minimized, while the AR is remained above an agreed amount.

ACKNOWLEDGEMENT

This work was funded by the EU H2020 research program 5GSTEP FWD (Grant Agreement No. 722429).

REFERENCES

- [1] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models", *IEEE Communications Surveys and Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.
- [2] M. Lashgari, C. Natalino, L. M. Contreras, L. Wosinska, and P. Monti, "Cost benefits of centralizing service processing in 5G network infrastructures", *Asia Communications and Photonics Conference (ACP)*, pp. 1–3, 2019.
- [3] G. Kalfas, C. Vagionas, A. Antonopoulos, E. Kartsakli, A. Mesodiakaki, S. Papaioannou, P. Maniotis, J. S. Vardakas, C. Verikoukis and N. Pleros, "Next generation fiber-wireless fronthaul for 5G mmWave networks", *IEEE Communications Magazine*, vol. 57, no. 3, pp. 138–144, 2019.
- [4] A. D. Hossain, and A. R. Hossain, "A distributed control framework for TDM-PON based 5G mobile fronthaul", *IEEE Access*, vol. 7, pp. 162102–162114, 2019.
- [5] E. Datsika, E. Kartsakli, J. S. Vardakas, A. Antonopoulos, G. Kalfas, P. Maniotis, C. Vagionas, N. Pleros, and C. Verikoukis, "QoS-aware resource management for converged fiber wireless 5G fronthaul networks", *IEEE Global Communications (GLOBECOM)*, pp. 9–13, 2018.
- [6] S. Papaioannou, G. Kalfas, C. Vagionas, et al., "5G mm Wave networks leveraging enhanced fiber-wireless convergence for high-density environments: the 5G-PHOS approach", *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–5, 2018.
- [7] A. M. Abdalla, J. Rodriguez, I. Elfergani, and A. Teixeira, "Optical and wireless convergence for 5G networks", *John Wiley & Sons*, 2019.
- [8] E. Datsika, J. Vardakas, K. Ramantas, P. V. Mekikis, I. T. Monroy, L. A. Neto, and C. Verikoukis, "SDN-enabled resource management for converged Fi-Wi 5G fronthaul", *IEEE Journal on Selected Areas in Communications*, 2021.
- [9] O. Sadio, I. Ngom, and C. Lishou, "Design and prototyping of a software defined vehicular communication", *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 842–850, 2020.
- [10] H. Hantouti, N. Benamar, T. Taleb, and A. Laghrissi, "Traffic steering for service function chaining", *IEEE Communications Surveys Tutorials*, vol. 21, pp. 487–507, 2019.
- [11] Y. Yu, X. Bu, K. Yang, H. K. Nguyen, and Z. Han, "Network function virtualization resource allocation based on joint benders decomposition and ADMM", *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1706–1718, 2020.
- [12] X. Cheng, Y. Wu, G. Min, and A. Y. Zomaya, "Network function virtualization in dynamic networks: a stochastic perspective", *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 2218–2232, 2018.
- [13] D. Chandramouli, R. Liebhart, and J. Pirskanen, "5G for the connected world", *John Wiley & Sons*, 2019.
- [14] T. Guo, and A. Suárez, "Enabling 5G RAN slicing with EDF slice scheduling", *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2865–2877, 2019.
- [15] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB", *IEEE Journal on Selected Areas in Communications*, vol. 37, pp. 881–895, 2019.
- [16] H. Chien, Y. Lin, C. Lai, and C. Wang, "End-to-End slicing with optimized communication and computational resource allocation in multi-tenant 5G systems", *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2079–2091, 2020.
- [17] M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi, and B. H. Khalaj, "vSPACE: VNF simultaneous placement, admission control and embedding", *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 542–557, Mar. 2018.
- [18] W. Fang, M. Zeng, X. Liu, W. Lu, and Z. Zhu, "Joint spectrum and IT resource allocation for efficient VNF service chaining in inter-data-center elastic optical networks", *IEEE Communication Letters*, vol. 20, pp. 1539–1542, 2016.
- [19] J. Qiu, D. Grace, G. Ding, M. D. Zakaria, and Q. Wu, "Air-ground heterogeneous networks for 5G and beyond via integrating high and low altitude platforms", *IEEE Wireless Communications*, vol. 26, no. 6, pp. 140–148, 2019.
- [20] M. Karimzadeh-Farshbafan, V. Shah-Mansouri and D. Niyato, "A dynamic reliability-aware service placement for network function virtualization (NFV)", *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 2, pp. 318–333, 2020.
- [21] J. Shi, J. Wang, H. Huang, L. Shen, J. Zhang, and H. Xu, "Joint optimization of stateful VNF placement and routing scheduling in software-defined networks", *IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCLOUD/SocialCom/SustainCom)*, pp. 9–14, 2018.
- [22] M. T. Beck, J. Botero, and K. Samelin, "Resilient allocation of service function chains", *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pp. 128–133, 2016.
- [23] V. Eramo, A. Tosti, and E. Miucci, "Server resource dimensioning and routing of service function chain in NFV network architectures", *Journal of Electrical and Computer Engineering*, vol. 2016, pp. 1–12, Apr. 2016.
- [24] M. M. Tajiki, S. Salsano, L. Chiaraviglio, M. Shojafar, and B. Akbari, "Joint energy efficient and QoS-aware path allocation and VNF placement for service function chaining", *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 374–388, 2018.
- [25] M. Mosahebfard, J. Vardakas, K. Ramantas, and C. Verikoukis, "SDN/NFV-based network resource management for converged optical-wireless network architectures", *IEEE International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4, 2019.
- [26] I. D. Moscholios, M. D. Logothetis, and G. K. Kokkinakis, "Connection-dependent threshold model: a generalization of the Erlang multiple rate loss model", *Performance Evaluation*, vol. 48, no. 1–4, pp. 177–200, 2002.
- [27] J. Kaufman, "Blocking in a shared resource environment", *IEEE Transactions on communications*, vol. 29, no. 10, pp. 1474–1481, 1981.
- [28] J. Roberts, "A service system with heterogeneous user requirements", G. Pujolle (Ed.), *Performance of Data Communications systems and their applications*, pp. 423–431, 1981.
- [29] S. P. Chung, and K. W. Ross, "Reduced load approximations for multirate loss networks", *IEEE Transactions on Communications*, vol. 41, no. 8, pp. 1222–1231, 1993.
- [30] R. P. Leal, F. José da Silva Velez, L. M. Campos, and A. García Armada, "TeamUp5G: a multidisciplinary approach to training and research on new RAN techniques for 5G ultra-dense mobile networks", *12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, pp. 1–6, 2020.
- [31] L. A. Freitas, V. G. Braga, S. L. Correa, L. Mamatasz, C. E. Rothenberg, S. Clayman, K. V. Cardoso, "Slicing and allocation of transformable resources for the deployment of multiple virtualized infrastructure managers (VIMs)", *4th IEEE Conference on Network Software and Workshops (NetSoft)*, pp. 424–432, 2018.
- [32] A. Gushchin, A. Walid, and A. Tang, "Scalable routing in SDN-enabled networks with consolidated middleboxes", *ACM SIGCOMM Workshop on Hot Topics in Middleboxes and Network Function Virtualization*, pp. 55–60, 2015.