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

Mohammadreza Mosahebfard®, Member, IEEE, John S. Vardakas®, Senior Member, IEEE, and Christos Verikoukis®, Senior Member, IEEE

Abstract—Network Function Virtualization (NFV)-based 5G networks deliver specific services to end-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 Admission Ratio (AR) calculation 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 analysis, we not only take into account the occupancy distribution in both the network’s computational and communicational domains (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 proposed model is employed 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 simulations, 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 meet the 5G network expectations and provide the mentioned use cases to the end-users, the next generation of 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).

The deployment of Passive Optical Networks (PONs) at the fronthaul and backhaul network segments 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].

Data plane entities such as switches and routers are simplified by abstraction of their intelligence into one or more SDN controllers as control plane entities through SDN. This separation allows network programmability, which enables dynamic resource allocation and policy enforcement in accordance with the nature of the requested services. Furthermore, it allows intelligence to be moved from devices to a control plane that manages the overall devices. Apart from these, 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]. NFV, on the other hand, manages the computational resources within the data-center, including CPU cores, memory, and storage. Virtualized Network Functions (VNFs), which are installed on servers in data centers, replace dedicated hardware-based network functions. Moreover, NFV enables network operators to reduce capital and operational expenses as well as the delivery time of new services. 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. 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 solution to the challenge related to the effective management of a wide variety of services with different characteristics and requirements in a single physical network. It enables the execution of multiple logically isolated networks over a common shared physical

Manuscript received April 16, 2021; revised August 13, 2021; accepted September 9, 2021. Date of publication September 20, 2021; date of current version November 18, 2021. This work was supported in part by the research programs EU H2020 5GSTEP-FWD (722429), EU H2020 MARSAL (101017171), EU H2020 5G-SOLUTIONS (856691), H2020 EXPLOR (872897), ZE- ROTO6G (PID2020-119250GB-100), and 2017 SGR 891. The review of this article was coordinated by Prof. Jong-Moon Chung. (Corresponding author: Mohammadreza Mosahebfard.)

Mohammadreza Mosahebfard and John S. Vardakas are with the Iquadrat Informatica, 08006 Barcelona, Spain (e-mail: m.mosahebfard@iquadrat.com; j vardakas@iquadrat.com).

Christos Verikoukis is with the Telecommunications Technological Center of Catalonia (CTTC/CERCA), 08860 Castelldefels, Spain (e-mail: cveri@cttc.es).

Digital Object Identifier 10.1109/TVT.2021.313838

0018-9545 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
in [22] developed an infinite horizon Markov decision process model to
nature of service arrivals and departures. Furthermore, they have
on reliability-aware service placement considering the dynamic
and minimizing placement costs. This study focuses specifically
a variety of services to users using NFV, while maximizing AR
analyze a scenario in which a network operator aims to provide
in a cost-effective and seamless manner [19]. The authors of [20]
proposing higher capacity and more coverage for underserved areas in
forms into the common terrestrial cellular networks for achiev-
SDN, which combines cross-layer low and high altitude plat-
propose a novel hierarchical network architecture by employing
elastic optical networks (inter-data-center EONs). The objective
allocates the spectrum and computational resources is proposed
in order to handle the SFC requests efficiently in inter-datacenter
and in addition to heuristic algorithms, they employed relaxation
algorithm is proposed in [24] for VNF placement, which assigns
VNFs to flows. They have also studied flow routing problems in
SDN-based networks.
Lastly, our previous work in [25] aims to efficiently allocate
both optical and wireless resources in an SDN/NFV-based con-
verted optical-wireless network architecture, by determining the
slices of the network in a way that the specific delay and band-
width 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, mini-
um VNFs and vLinks capacities, etc.), which guarantee that
the resulting 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 do-
main of the network (servers and fiber links), but also consider
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 by deploying recur-
sive formulas and consequently employing the proposed model is
computationally efficient compared to optimization-based
approaches for performing network dimensioning.

A. Related Work
Resource management is a constant topic during the evol-
uution 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 opti-
mization frameworks. For example, authors in [17] consider an
NFV-based architecture in which the VNF requests are ad-
dressed by the network operator in order to maximize its own rev-
ue. This study investigated the joint resource allocation prob-
lem 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 to minimize 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 plat-
forms into the common terrestrial cellular networks for achiev-
ning higher capacity and more coverage for underserved areas in
a cost-effective and seamless manner [19]. The authors of [20]
analyze a scenario in which a network operator aims to provide
a variety of services to users using NFV, while maximizing AR
and minimizing placement costs. This study focuses specifically
on reliability-aware service placement considering the dynamic
nature of service arrivals and departures. Furthermore, they have
developed an infinite horizon Markov decision process model to
perform dynamic reliability-aware service placement, in which
the main and backup servers are both allocated simultaneously.
In [21], authors consider the coexistence of SDN control flows
and client flows together with NFV. In this work, a joint op-
timization for synchronizing the state of VNF instances and
scheduling the routes is presented in order to minimize the
rule occupation and VNF deployment costs. Beck et al. in [22]
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 backup candidates such as VNF
backups and link backups in order to protect network services
from failures. A resource dimensioning and routing problem in
NFV architectures is addressed in [23]. They proposed heuris-
tic algorithms for offline and online traffics with guaranteeing
uniform computational resources and bandwidth occupancy in
the network’s servers and links, respectively. They have also
proposed an algorithm to minimize the energy consumption at
the cost of SFC blocking. An energy-aware resource allocation
algorithm is proposed in [24] for VNF placement, which assigns
VNFs to flows. They have also studied flow routing problems in
SDN-based networks.

B. Our Contribution
To position our contribution in detail, in this paper we consider
a converged optical-wireless 5G network, where the Remote Ra-
dio Heads (RRHs) are connected to a set of data-centers through
a network of PONs. In the mentioned 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 one or more SFCs. For this setup, we propose an analytical model for the AR calculation, which is then applied in order to determine 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 both communicational and computational levels 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 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 from [27], [28] as well as the Reduced Load Approximation (RLA) method from [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 simulations 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 determining the AR and the optimal 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.

II. SYSTEM MODEL

In this paper, we consider a converged optical-wireless network configuration to provide high speed connectivity in the access domain. Employing a PON-based fronthaul allows network operators to provide services that are capable of meeting the highly challenging operational framework of 5G, 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 as 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, ..., D) is equipped with C_d units of computational resources including CPU cores, RAM, and storage. Finally, link l (l = 1, ..., L) supports a total data rate of B_l. Moreover, the Baseband Unit (BBU) of each RRH in a specific PON is considered to be installed at the central office within that PON. The BBUs are optically connected to their corresponding RRH. We also take into account that the PON’s Optical Network Unit (ONU) is a part of each RRH.

In addition, the NFV Orchestrator (NFVO) and the SDN controller are responsible for managing 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. Next, the NFVO updates the SDN controller so that it interconnects the installed VNFs according to a 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 layer 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 responsible for implementing the TDM/WDM approaches in the network. It is also worth mentioning that the optical network elements, such as ROADMs and ONU's, do not impact the network performance, but they 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 $s$ ($s = 1, \ldots, S$)
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 $f_s$ ($f_s = 1, \ldots, F_s$). 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, which interconnects two consecutive VNFs of the chain. In addition, we assume that the service chain starts with 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 the vLinks. Moreover, each data center has an Internet Gateway (GW) through which the output data of the last VNF is transmitted to the Internet. $E$ identical Connecting Paths (CPs) in each service chain are also considered in order that they serve 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. All CPs are considered equal in each SFC. The reason is that a single CP should be able to steer the arriving 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 $V_{s,f} = \{v_i | i = 1, \ldots, N_{s,f}, v_i \neq v_j \iff i \neq j\}$. As it is clear in the definition of $V_{s,f}$, none of the VNF pairs are from the same type. Similarly, the set $P_{s,f} = \{p_i | i = 1, \ldots, N_{s,f}\}$ 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 $\lambda_{s,k}$. Consequently, the summation of independent Poisson processes will be a Poisson process. The total arrival rate of the resulting Poisson process for sub-service-class $k$ in slice $s$, which arrives at VNF $n_{s,f,n}$ 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 $k_s$ is exponentially distributed is shown by $\mu^{-1}_{s,k}$. Consequently, the offered traffic load of the sub-service-class $k$ 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 terms. In addition, VNF $n_{s,f,n}$ has two different coefficients named $\alpha_{s,f,n}$ (dimensionless) and $\beta_{s,f,n}$ (1/bps), which represent the consumed computational 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 $\beta_{s,f,n}$. Through (3) and (4), the aforementioned coefficients are used to calculate the capacities of vLinks and VNFs in each SFC, respectively. Likewise, we 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 $f_s$, the capacity of the first vLink of the chain is multiplied by all the rate coefficients of all prior VNFs to that vLink (i.e., $\alpha_{s,f,n}$, $n = 1, \ldots, n - 1$). This is logical as the added overhead to the traffic in each vLink, depends on all prior VNFs to that vLink. Consequently, the capacity of vLink $n_{s,f}$ is obtained as follows:

$$R_{s,f,n} = \left[ R_{s,f,1} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right],$$

$$n = 2, \ldots, N_{s,f}, \quad (3)$$

where $[x]$ denotes the least integer greater than or equal to $x$. Similarly, for calculating the required number of computational resources of the $n$th VNF of SFC $f_s$, and by employing (3), we multiply the consumed computational resources coefficient of the corresponding 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:

$$T_{s,f,n} = \left[ R_{s,f,1} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right],$$

$$n = 1, \ldots, N_{s,f} \quad (4)$$

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 ($R_{s,f}$) and all the prior VNFs to the $n$th vLink of the chain. Thus, $r_{s,f,n}^k$ can be calculated as follows:

$$r_{s,f,n}^k = \left[ \frac{R_{s}^n \prod_{q=1}^{n} \alpha_{s,f,q}}{R_{s,f,n}} \right] \cdot 100.$$

Similarly, 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 ($R_{s,f}^k$) and all the prior VNFs in the corresponding chain. Therefore, $t_{s,f,n}^k$ is obtained with the following formula:

$$t_{s,f,n}^k = \left[ \frac{R_{s}^k \beta_{s,f,n} \prod_{q=1}^{n-1} \alpha_{s,f,q}}{T_{s,f,n}} \right] \cdot 100. \quad (6)$$
We define the binary variable \( h^{d}_{s,f,n} \) to check if the VNF \( n_{s,f} \) is placed in server \( d \) or not:

\[
h^{d}_{s,f,n} = \begin{cases} 1 & \text{if VNF } n_{s,f} \text{ is initiated in server } d \\ 0 & \text{otherwise} \end{cases}
\]  \( (7) \)

Set \( \Delta_d \) that shows all the VNFs placed in server \( d \) is also defined as follows:

\[
\Delta_d = \{ (s, f, n) \mid h^{d}_{s,f,n} = 1 \ \forall s, f, n \}. \tag{8}
\]

On the other hand, for the communicational part, another binary variable is defined to check if vLink \( n_{s,f} \) is passing through link \( l \) or not:

\[
j^{l}_{s,f,n} = \begin{cases} 1 & \text{if vLink } n_{s,f} \text{ is placed in link } l \\ 0 & \text{otherwise} \end{cases}
\]  \( (9) \)

Similar to the set \( \Delta_d \), the set \( \Gamma_l \) indicates all the vLinks utilizing physical link \( l \) to steer their traffic and is defined as follows:

\[
\Gamma_l = \{ (s, f, n) \mid j^{l}_{s,f,n} = 1 \ \forall s, f, n \}. \tag{10}
\]

We define set \( H_{s,f} = \{ d \mid h^{d}_{s,f,n} = 1 \ \forall n \in \{1, \ldots, N_{s,f}\} \} \) representing all the servers hosting the VNFs of SFC \( f_s \), while set \( J_{s,f} = \{ l \mid j^{l}_{s,f,n} = 1 \ \forall n \in \{1, \ldots, N_{s,f}\} \} \) 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:

\[
\sum_{d=1}^{D} h^{d}_{s,f,n} = 1 \ \forall (s, f, n). \tag{11}
\]

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:

\[
1 \leq \sum_{l=1}^{L} j^{l}_{s,f,n} \leq L \ \forall (s, f, n). \tag{12}
\]

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:

\[
\sum_{s=1}^{S} \sum_{f=1}^{F_s} \sum_{n=1}^{N_{s,f}} h^{d}_{s,f,n} t_{s,f,n} \leq C_d \ \forall d = 1, \ldots, D. \tag{13}
\]

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:

\[
\sum_{s=1}^{S} \sum_{f=1}^{F_s} \sum_{n=1}^{N_{s,f}} j^{l}_{s,f,n} r_{s,f,n} \leq B_l \ \forall l = 1, \ldots, L. \tag{14}
\]

### 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 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 the \( K_s \) sub-service-classes of corresponding service type. 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 in 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 \( f_s \) 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 section IV-A. For this purpose, in Section III-A, we will firstly determine the occupancy distribution of the VNFs by considering the aforementioned Markov chain. In the next steps, this occupancy distribution will be considered for determining various parameters that are required in order to calculate the SFC establishment AR. These values, which are 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, are calculated 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]. This method was considered for the determination of bandwidth occupancy distribution in a single communication link that supports multiple service classes. It considers the summation of all steady-state equations of all supported sub-service-classes providing the following formula of the occupancy distribution in VNF $n_{s,f}$:

$$
q_{s,f,n}(x) = \frac{1}{x} \sum_{i=1}^{K_{s,f}} A_{s,k} t_{s,f,n}^i q_{s,f,n}(x - t_{s,f,n}^i),
$$

$$x = 1, \ldots, T_{s,f,n},
$$

$$Q_{s,f,n} = \sum_{i=0}^{T_{s,f,n}} q_{s,f,n}'(x),
$$

$$q_{s,f,n}(x) = \frac{1}{Q_{s,f,n}} q_{s,f,n}'(x),
$$

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:

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

(16)

where $y_{s,f,n}^k(x + t_{s,f,n}^i)$ is the average number of active requests of sub-service-class $k$ in state $(x + t_{s,f,n}^i)$. 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 $x + t_{s,f,n}^k$:

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

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 $x$ is obtained as follows:

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

(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. $M_{s,f}$ is the release rate of the SFC $f_s$ and is equal to the sum of the rates from state $t_{s,f,n}^k$, $k = 1, \ldots, K_s$, 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:

$$M_{s,f,n} = \sum_{i=1}^{K_s} \mu_{s,i} t_{s,f,n}^i y_{s,f,n}^k(t_{s,f,n}^i) \hat{q}_{s,f,n}(t_{s,f,n}^i)
$$

$$= \sum_{i=1}^{K_s} \mu_{s,i} t_{s,f,n}^i y_{s,f,n}^k(t_{s,f,n}^i) \frac{q_{s,f,n}(t_{s,f,n}^i)}{1 - q_{s,f,n}(0)},
$$

(19)

where $\hat{q}_{s,f,n}(x)$ is the conditional probability that $x$ resources of VNF are occupied, given that the VNF resources are still occupied in the server. On the other hand, as all VNFSs and vLinks of a specific SFC are occupied and become idle at the same time, only the first VNF of the service chain can be considered 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 $f_s$:

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

(20)
D. Interarrival Time for SFC Creation Request

In addition to the mean service rate of SFC, the interarrival time for SFC creation requests need to be calculated. An SFC is established through the arrival of a user request that may belong to any of the sub-service-classes of the corresponding slice. Thus, the total arrival rate of all sub-service-classes arriving at SFC \( f_s \), which is represented by \( \lambda_{s,f} \), is given as:

\[
\lambda_{s,f} = \sum_{i=1}^{K_s} \lambda_{s,f,i}.  
\]  

In order to calculate the mean interarrival rate of the SFC creation requests, the schematic diagram presented in Fig. 3 is considered. 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, \( 1/M_{s,f} \), by the total interarrival time, \( 1/\lambda_{s,f} \), the mean number of users that arrived within the SFC service time can be found. By increasing the latter number by one, the next user arrival is counted, which actually is by itself an SFC creation request. Therefore, the interarrival time of the SFC creation request is stated as follows:

\[
\frac{1}{\Lambda_{s,f}} = \left( \frac{1}{\frac{1}{\lambda_{s,f}} - \frac{1}{M_{s,f}}} + 1 \right) \frac{1}{\lambda_{s,f}},  
\]

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 for establishing 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:

\[
AR_{s,f,k} = AR_{s,f}^{\text{AR}^\text{Comp.}} AR_{s,f,k}^{\text{AR}^\text{user}},  
\]

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

A. SFC Establishment AR

In this Section, the SFC establishment AR, the first term in (23), is determined by utilizing the parameters calculated in Sections III-C and III-D. To this end, two different 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 cores, RAM, and memory).

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

\[
AR_{s,f}^{\text{AR}^\text{Comp.}} = (1 - P_{s,f}^{\text{Comp.}}) (1 - P_{s,f}),  
\]

where \( P_{s,f}^{\text{Comp.}} \) represents the blocking probability of realizing SFC’s vLinks on top of fiber links and \( P_{s,f}^{\text{Comp.}} \) is the SFC blocking probability of SFC’s on top of servers. These two values are calculated by considering the analysis presented in Sections IV-A1 and IV-A2, respectively.

1) Blocking Probability in the Communicational Domain:

\( P_{s,f}^{\text{Comp.}} \) is considered to be 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 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). This method calculates the probability of blocking based on an approximate estimate of the reduction in the offered traffic load in each link of the route (the vLink in our analysis) because 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 \( \Gamma_l \) 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, \( M_{s,f}^{\text{AR}^\text{Comp.}} \) and \( \Lambda_{s,f} \), are considered as the arrival rate and service time of each vLink in link \( l \). By employing (1) from [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:

\[
P_{s,f,n}^d = E \left[ B_t, \sum_{(s,f,n) \in \Gamma_1} (M_{s,f,n} A_{s,f,n}) \prod_{i \in J_{s,f} \setminus \{l\}} (1 - P_{s,f,n}^l) \right],
\]

(25)

where

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

is the Erlang loss formula in which \( C \) is the capacity of the link and \( \rho \) 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:

\[
P_{s,f}^{\text{Comp.}} = 1 - \prod_{n=1}^{N_{s,f}} (1 - P_{s,f,n}^l) \forall l \in J_{s,f}.
\]

(26)

2) Blocking Probability in the Computational Domain: \( P_{s,f,n}^d \) is considered to be the probability that the \( d \)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 \( d \), which hosts VNFs of set \( \Delta_d \), the occupancy distribution inside the server, \( q_d(x) \), 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 in set \( \Delta_d \) 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]:

\[
P_{s,f,n}^d = \sum_{x=C_d-T_s, f_n+1}^{C_d} q_d(x).
\]

(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:

\[
P_{s,f}^{\text{Comp.}} = 1 - \prod_{n=1}^{N_{s,f}} (1 - P_{s,f,n}^d) \forall d \in H_{s,f}.
\]

(28)

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

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 traffic. In other words, provided that one VNF or vLink of the service chain is full, the request will be blocked. As 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

\[
\text{Algorithm 1: AR Calculation’s Steps.}
\]

1: Input: Network graph, predefined paths, \( \lambda_{s,k,c}, \mu_{s,k} \), 
2: \( r_{s,f,1}^k = R_{s,f,1}^k, E, C_d, B_t, \beta_{s,f,n}, \alpha_{s,f,n}, R_{s,f,1}, S \), 
3: \( F_s, N_{s,f}, V, h^d_{s,f,n}, h^l_{s,f,n}, h^d_{s,f,n}, h^l_{s,f,n}, \Delta_d, \Gamma_l \)
4: Determine \( \lambda_{s,k} \) and \( A_{s,k} \) using (1) and (2).
5: for \( s = 1, \ldots, S \) do
6: for \( n = 1, \ldots, N_{s,f} \) do
7: if \( n = 1 \) then
8: \( R_{s,f,1} \) is predefined.
9: else
10: Calculate \( R_{s,f,n} \) with (3).
11: end if
12: Calculate \( T_{s,f,n} \) employing (4).
13: end for
14: for \( k = 1, \ldots, K_s \) do
15: \( \alpha_{s,f,n} \) do
16: if \( n = 1 \) then
17: Calculate \( r_{s,f,1}^k \) with (6).
18: else
19: Calculate \( r_{s,f,n}^k \) by using (6).
20: end if
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 \( s = 1, \ldots, S \& f_s = 1 \) do
27: Calculate \( M_{s,f} \) and \( A_{s,f}^{-1} \) with (20) and (22).
28: end for
29: for \( l = 1, \ldots, L \) do
30: for all vLinks \( n_{s,f} \) where \( (s, f, n) \in \Gamma_l \) do
31: Employ (25) to calculate \( P_{s,f,n}^d \).
32: end for
33: end for
34: for each server \( d = 1, \ldots, D \) do
35: Calculate \( q_d(x) \) using (15).
36: for all VNFs \( n_{s,f} \) where \( (s, f, n) \in \Delta_d \) do
37: Calculate \( P_{s,f,n}^d \) with (27).
38: end for
39: end for
40: for each slice \( s = 1, \ldots, S \) do
41: Calculate \( P_{s,f}^{\text{Comp.}} \) employing (28).
42: end for
43: \( \text{Calculate } P_{s,f}^{\text{Comp.}} \) with (29).
44: end for
45: \( \text{Calculate } AR_{s,f} \) using (24) and (30).
46: \( \text{Calculate } AR_{s,f} \) with (23).
of all the blocking states, which is defined as follows:

\[ P_{s,f,k}^{\text{user}} = P_{s,f,1}^{\text{VNF}} = \sum_{x=T_{s,f,1} - t_{s,f,1}^k}^{T_{s,f,1}} q_{s,f,1}(x), \]  

(29)

where \( q_{s,f,1}(x) \) is the occupancy distribution of the first VNF in SFC \( f \) 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:

\[ A_{s,f,k}^{\text{user}} = 1 - P_{s,f,k}^{\text{user}}. \]  

(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. Initially, the necessary parameters are inserted to initiate the algorithm and calculate \( \lambda_{s,k} \) and \( A_{s,k} \) using (1) and (2). Then, a loop with a length of \( S \) is started. In each iteration of this loop, in Lines 4-11, there is an inner loop with the length of \( N_{s,f} \) (equal to number of VNFs and vLinks in SFC \( f_s \)). 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 \( T_f \) 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 \( P_{s,f,n}^{\text{VNF}} \), the probability that \( n \)th VNF of SFC \( f_s \) cannot be initiated in server \( d \) due to the lack of computational resources (\( P_{s,f,n}^{\text{user}} \)) 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 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 of each VNF and vLink for each SFC in the network in order to achieve 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 guarantee the AR requirements in each slice. Specifically, this procedure is realized by reversing the AR calculation process, 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:

\[ R_{s,f,n} = R_{s,f,1} \prod_{i=1}^{n-1} \alpha_{s,f,i}, \]

(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:

\[ T_{s,f,n} = \left[ R_{s,f,1} \beta_{s,f,n} \prod_{i=1}^{n-1} \alpha_{s,f,i} \right], \]

(32)

The input parameters for performing the network dimensioning analysis are the predefined AR values 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, the accuracy of the proposed mathematical framework is evaluated, 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 remains 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 Fig. 5, respectively, where the data-centers are represented with green and ROADMs with red color. In both setups, the capacity of each data-center is set to 50 units of computational resources \( C_d = 50 \) and the capability of each fiber link is equal to 40 Gbps \( B_l = 40 \times 10^9 \). In the small-scale network, 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 the two PONs. On the other hand, in the large-scale network, nine data-centers are considered, as in the case of the evaluation scenario of [32], interconnected through eight ROADMs and 72 fiber links, where the eROADMs (nodes 1, 3, 15 and 17) are connected to the 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 ARs for the three slices, we assume that all SFCs in small-scale network are formed of three VNFs and three vLinks ($\forall s = 1,\ldots, 3$, $N_s = 3$), while this number for the other setup is set to four ($\forall s = 1,\ldots, 3$, $N_s = 4$). Moreover, the parameters for all deployed VNFs are set to $\alpha_{s,f,n} = 2 \times 10^{-9}$ (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 $\mu_{1,1} = \mu_{2,1} = \mu_{3,1} = 0.5$ minutes and $R_1 = 100$, $R_2 = 90$, and $R_3 = 80$ Mbps. For the small-scale setup, in each SFC, by employing (3) and (4), the capacities of vLinks 1 to 3, are equal to 4, 4.2, and 4.4 all in Gbps, respectively, while the capacities of VNFs 1 to 3 will be 8, 9, and 9 computational units, 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, while the capacities of VNFs 1 to 4, are equal to 10, 11, 12, and 12 computational units, respectively.

We consider two evaluation scenarios for each of the aforementioned network setups. The first scenario refers to the case with 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 presented in Table II and Table III, respectively. The simulation results of the AR for requests per slice are presented as mean values from seven runs with confidence interval of 95%. In each simulation run we assume 1,000,000 originating requests, and for the system to reach steady state we consider the first 100,000 requests as the stabilization period.

The analytical and simulations results of the AR for the case with single SFC per slice are presented in Table IV and Table VI, while the results for the case in which there are two SFCs per slice are presented in Table V and Table 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, $E$, is calculated by the following formula:

$$E = \frac{v_a - v_s}{v_s} \times 100\%,$$

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, the maximum related error is:

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

which shows the high accuracy of the proposed model. As can be seen in Table IV and Table VI ($F_s = 1$) as well as Table V and Table 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, for the second case, where there are two SFCs in each slice, the users experience higher AR values as 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 optimal capacities.
of VNFs and vLinks within SFCs so that a predefined AR threshold can be achieved. We consider different values for the required data rate in each slice: \( R_1 = 150 \), \( R_2 = 100 \), and \( R_3 = 50 \) all in Mbps. Fig. 6 depicts the minimum required computational resources for the first 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 these values for the cases with two and three SFCs per slice, respectively.

By comparing the three cases of Figs. 6, it can be concluded that the higher the number of SFCs is, the smaller size of VNF needs to be. 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. Furthermore, by comparing the three cases in Fig. 6 for a specific arrival rate in a particular slice, it can be concluded 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 VNF of the two chains residing in eMBB slice. Finally, it can be concluded that a slice with less number of service chains will have a greater number of users affected by an SFC failure, because more users are assigned to the SFC than in the case of a slice with more 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 and consequently the total amount of consumed resources by the SFC (all VNFs and vLinks) to guarantee a desired AR value. This helps to calculate the size of SFCs precisely in order to guarantee a predefined AR value.

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. It illustrates the analytical the AR values for the three 3GPP service types versus the user arrival rates. In all cases, the AR drops below the value of 1 for a different value of the user arrival rate. As anticipated, the eMBB service has the highest threshold value, followed by URLLC and mMTC, as it is the result of their varying resource needs. Specifically, the maximum offered arrival rate that guarantees a 100% AR in eMBB slice is 18 connection requests per minute, while this value for URLLC and mMTC is 65 and 550, respectively.

### Table VI

| Arrival rate | Slice 1 Analysis | Slice 1 Simulations | Slice 2 Analysis | Slice 2 Simulations | Slice 3 Analysis | Slice 3 Simulations |
|--------------|------------------|---------------------|------------------|---------------------|------------------|---------------------|
| 70           | 0.9967 ± 0.0004  | 0.9965 ± 0.0001     | 1.0000           | 1.0000 ± 0.0000     |
| 75           | 0.9913 ± 0.0007  | 0.9855 ± 0.0003     | 0.9999           | 0.9999 ± 0.0000     |
| 80           | 0.9813 ± 0.0011  | 0.9657 ± 0.0004     | 0.9977           | 0.9977 ± 0.0001     |
| 85           | 0.9661 ± 0.0011  | 0.9878 ± 0.0005     | 0.9990           | 0.9989 ± 0.0002     |
| 90           | 0.9462 ± 0.0013  | 0.9797 ± 0.0007     | 0.9972           | 0.9970 ± 0.0005     |
| 95           | 0.9222 ± 0.0014  | 0.9652 ± 0.0009     | 0.9931           | 0.9931 ± 0.0008     |
| 100          | 0.8954 ± 0.0018  | 0.9462 ± 0.0013     | 0.9861           | 0.9861 ± 0.0009     |

### Table VII

| Arrival rate | Slice 1 Analysis | Slice 1 Simulations | Slice 2 Analysis | Slice 2 Simulations | Slice 3 Analysis | Slice 3 Simulations |
|--------------|------------------|---------------------|------------------|---------------------|------------------|---------------------|
| 140          | 0.9998 ± 0.0001  | 0.9999 ± 0.0000     | 1.0000           | 1.0000 ± 0.0000     |
| 150          | 0.9990 ± 0.0004  | 0.9999 ± 0.0000     | 1.0000           | 1.0000 ± 0.0000     |
| 160          | 0.9960 ± 0.0009  | 0.9997 ± 0.0000     | 1.0000           | 1.0000 ± 0.0000     |
| 170          | 0.9833 ± 0.0014  | 0.9987 ± 0.0002     | 1.0000           | 1.0000 ± 0.0000     |
| 180          | 0.9652 ± 0.0018  | 0.9950 ± 0.0008     | 0.9999           | 0.9999 ± 0.0000     |
| 190          | 0.9403 ± 0.0023  | 0.9867 ± 0.0018     | 0.9995           | 0.9994 ± 0.0001     |
| 200          | 0.9119 ± 0.0025  | 0.9722 ± 0.0030     | 0.9980           | 0.9977 ± 0.0004     |
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, in a sliced network, the minimum required capacity for the VNFs and vLinks of the service chain(s) are calculated in order to keep the AR higher than a predefined threshold. It was also concluded that when the number of SFCs is higher, the less optimal the usage of the infrastructure resources will be. 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. Next, we will investigate the case that there are various 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 consumption is minimized, while the AR is remained above an agreed amount.

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 Commun. Surv. Tuts.*, vol. 22, no. 2, pp. 905–929, Apr.–Jun. 2020.
[2] M. Lashgari, C. Natalino, L. M. Contreras, L. Wosinska, and P. Monti, “Cost benefits of centralizing service processing in 5G network infrastructures,” in Proc. Asia Commun. Photonics Conf., 2019, pp. 1–3.

[3] G. Kalfas et al., “Next generation fiber-wireless fronthaul for 5G mmWave networks,” IEEE Commun. Mag., vol. 57, no. 3, pp. 138–144, Mar. 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 et al., “QoS-aware resource management for converged fiber wireless 5G fronthaul networks,” in Proc. IEEE Global Commun. Conf., 2018, pp. 1–5.

[6] S. Pattaranoon et al., “5G mm wave networks leveraging enhanced fiber-wireless convergence for high-density environments: The 5G-PHOS approach,” in Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast., 2018, pp. 1–5.

[7] A. M. Abdalla, J. Rodriguez, I. Effergani, and A. Teixeira, “Optical and Wireless Convergence for 5G Networks,” Hoboken, NJ, USA: Wiley, Inc., 2019.

[8] E. Datsika et al., “SDN-enabled resource management for converged fi-wi 5G fronthaul,” IEEE J. Sel. Areas Commun., vol. 39, no. 9, pp. 2772-2788, Sep. 2021.

[9] O. Sadio, I. Ngom, and C. Lishou, “Design and prototyping of a software defined vehicular communication,” IEEE Trans. Veh. Technol., vol. 69, no. 2, pp. 842–850, Jan. 2020.

[10] H. Hantouti, N. Benamar, T. Taleb, and A. Laghrissi, “Traffic steering for service function chaining,” IEEE Commun. Surv. Tuts, vol. 21, pp. 487–507, Jan.–Mar. 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 Trans. Veh. Technol., vol. 69, no. 2, pp. 1706–1718, Feb. 2020.

[12] X. Cheng, Y. Wu, G. Min, and A. Y. Zomaya, “Network function virtualization in dynamic networks: A stochastic perspective,” IEEE J. Sel. Areas Commun., vol. 36, no. 10, pp. 2218–2232, Oct. 2018.

[13] D. Chandramouli, R. Liebhart, and J. Pirskanen, “5G for the Connected Altitude Platforms”, in Proc. ACM SIGCOMM Workshop Hot Top, Middleboxes: Network Function Virtualization, 2015, pp. 55–60.

[14] T. Guo, and A. Suárez, “Enabling 5G-RAN slicing with EDF slice scheduling,” IEEE Trans. Veh. Technol., vol. 68, no. 3, pp. 2865–2877, Mar. 2019.

[15] J. Tang, B. Shim, and T. Q. S. Quek, “Service multiplexing and revenue maximization in sliced C-RAN incorporated with ULLC and multicast eMBB,” IEEE J. Sel. Areas Commun., vol. 37, no. 4, pp. 881–895, Apr. 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 Trans. Veh. Technol., vol. 69, no. 2, pp. 2079–2091, Feb. 2020.

[17] M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahnoodi, and B. H. Khalaj, “vSPACE: VNF simultaneous placement, admission control and embedding,” IEEE J. Sel. Areas Commun., 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 Commun. Lett., vol. 20, no. 8, pp. 1547–1550, Aug. 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 Commun., vol. 26, no. 6, pp. 140–148, 2019.

[20] M. Karimzadeh-Farshafaran, V. Shah-Mansouri and D. Niyato, “A dynamic reliability-aware service placement for network function virtualization (NFV) in multi-tenant 5G networks,” IEEE J. Sel. Areas Commun., vol. 38, no. 2, pp. 318–333, Feb. 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,” in Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Ubiquitous Comput. Commun., Big Data Cloud Comput., Social Comput Netw., Sustainable Comput. Commun., 2018, pp. 9–14.

[22] M. T. Beck, J. Botero, and K. Samelin, “Resilient allocation of service function chains,” in Proc. Conf. Netw. Function Virtual. Softw. Defined Netw., 2016, pp. 128–133.

[23] V. Eramo, A. Tosti, and E. Mucci, “Server resource dimensioning and routing of service function chain in NFV network architectures,” J. Electron. Comput. Eng., vol. 2016, pp. 1–12, Apr. 2016.

[24] M. M. Taji, S. Salsano, L. Chiaraviglio, M. Shojaifar, and B. Akbari, “Joint energy efficient and QoS-aware path allocation and VNF placement for service function chaining,” in Proc. IEEE Trans. Netw. Service Manag., vol. 16, no. 1, pp. 374–388, 2018.

[25] M. Mosahefbadi, J. Vardakas, K. Ramantas, and C. Verikoukis, “SDN/NFV-based network resource management for converged optical-wireless network architectures,” in Proc. IEEE 21st Int. Conf. Transparent Opt. Netw., 2019, pp. 1–4.

[26] I. D. Moscholios, M. D. Logothetis, and G. K. Kokkinakis, “Connection-dependent threshold model: A generalization of the erlang multiple rate loss model,” Int. J. Perform. Eval., vol. 48, no. pp. 177–200, 2002.

[27] J. Kaufman, “Blocking in a shared resource environment,” IEEE Trans. Commun., vol. 29, no. 10, pp. 1474–1481, Oct. 1981.

[28] J. Roberts, “A service system with heterogeneous user requirements,” in Proc. Perform. Data Commun. Syst. Appl., G. Pupolle, Ed., 1981, pp. 423–431.

[29] S. P. Chung and K. W. Ross, “Reduced load approximations for multivariate loss networks,” IEEE Trans. Commun., vol. 41, no. 8, pp. 1222–1231, Aug. 1993.

[30] R. P. Leal, F. José da Silva Velez, L. M. Campos, and A. G. Armada, “TeenUp5G: A multidisciplinary approach to research and training on new RAN techniques for 5G ultra-dense mobile networks,” in Proc. 12th IEEE Int. Symp. Commun., Syst., Netw. Digit. Signal Process., 2020, pp. 1–6.

[31] L. A. Freitas et al., “Slicing and allocation of transformable resources for the deployment of multiple virtualized infrastructure managers (VIMs),” in Proc. 4th IEEE Conf. Netw. Softwarization Workshops, 2018, pp. 424–432.

[32] A. Gushchin, A. Walid, and A. Tang, “Scalable routing in SDN-enabled networks with consolidated middleboxes,” in Proc. ACM SIGCOMM Workshop Hot Top, Middleboxes: Network Function Virtualization, 2015, pp. 55–60.

Mohammadreza (Reza) Mosahefbadi (Member, IEEE) received the B.Sc. degree in electronics engineering from SUTECH, Shiraz, Iran, in 2013, and the M.Sc. degree in communications engineering from TMU, Tehran, Iran, in 2016. In 2018, Reza joined the R&D Department of iQuadrat Informatica, Barcelona, Spain, as a Marie Curie Ph.D. Fellow with the framework of the 5GSTEP-FWD H2020 European project. He is currently the Ph.D. candidate with Signal Theory and Communications Department, UPC (BarcelonaTech). His research interests mainly include resource management and power reduction in converged optical-wireless NFV-based 5G networks with different slices instances, each providing a specific service type to the end-users.

John S. Vardakas (Senior Member, IEEE) received the Dipl.-Eng. in electrical computer engineering from the Democritus University of Thrace, Greece, in 2004 and the Ph.D. degree from the Electrical Computer Engineering Department, University of Patras, Greece, in 2012. He has authored more than 40 journal articles and 70 conference articles, while he has participated in more than 15 competitive research programs. His research interests include teletraffic engineering, performance analysis and simulation of communication networks and smart grids. He is a member of the Technical Chamber of Greece (TEE).

Christos Verikoukis (Senior Member, IEEE) received the Ph.D. degree from UPC, Barcelona, Spain, in 2000. He is currently the Research Director (R4) with the Telecommunications Technological Centre of Catalonia (CTTC/CERCA), Castelldefels, Spain, and an Adjunct Professor with the University of Barcelona. He has authored more than 143 journal articles, 220 conference articles, three books, 14 book chapters, and five patents. He has participated in more than 30 competitive research projects, while he has supervised 19 Ph.D. students and 10 Postdoctoral researchers. He is the IEEE ComSoc EMEA Director and a Member-at-Large of IEEE ComSoc GITC. He was the recipient of the the Best Paper Award at the IEEE ICC 2011 and 2020, the IEEE GLOBECOM 2014 and 2015, and the EuCNC 2016, and the EURASIP 2013 Best Paper Award of the Journal on Advances in Signal Processing.