Yu, Hao; Taleb, Tarik; Zhang, Jiawei; Wang, Honggang

Deterministic Latency Bounded Network Slice Deployment in IP-over-WDM based Metro-Aggregation Networks

Published in:
IEEE Transactions on Network Science and Engineering

DOI:
10.1109/TNSE.2021.3127718

Published: 01/01/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Yu, H., Taleb, T., Zhang, J., & Wang, H. (2022). Deterministic Latency Bounded Network Slice Deployment in IP-over-WDM based Metro-Aggregation Networks. *IEEE Transactions on Network Science and Engineering*, *9*(2), 596-607. https://doi.org/10.1109/TNSE.2021.3127718

Deterministic Latency Bounded Network Slice Deployment in IP-over-WDM based Metro-Aggregation Networks

Hao Yu, Tarik Taleb, *Senior Member, IEEE*, Jiawei Zhang, and Honggang Wang *Fellow, IEEE*

***Abstract*—As a critical enabler of the next generation networks, network slicing can dynamically and flexibly create virtual networks leveraging necessary network resources for services with different quality of service (QoS) requirements. To satisfy the strict QoS requirements of some vertical industries, e.g., industry automation, the Deterministic Networking (DetNet) concept has been recently proposed to investigate deterministic service provisioning with bounds on service latency, loss, and jitter. A critical issue for providing deterministic services is how to allocate the right amount of resources that ensures the QoS requirements within a network slice. In this paper, we first obtain the amount of network resources needed for a network slice, using the stochastic network calculus (SNC) method, to ensure the end-to-end latency requirement under specific traffic demands. We then study the network slice deployment problem in an IP-over-WDM (wavelength division multiplexing) metro aggregation network and propose a heuristic with three different objectives: minimizing traffic hops, minimizing lightpaths and minimizing wavelengths, which can help network service provider to optimize network deployment under different considerations. The simulation results show the comparison of the resource utilization under different strategies.**

***Index Terms*—5G, Beyond 5G, Network Slicing, Deterministic Networking, Stochastic Network Calculus, IP over WDM, and Network Optimization.**

I. Introduction

THE next generation networks are expected to provide a more flexible and scalable paradigm whereby network slices (NS) can be dynamically and virtually created to accommodate a wide variety of applications with different quality-of-service (QoS) requirements [1]. The service requirements of the fifth generation mobile networks (5G) necessitate the use of emerging technologies, such as software-defined networking (SDN) and network function virtualization (NFV), which enables the concept of Network Slicing (NS). NS is the key enabler towards the "network-as-a-service" paradigm [2], which partitions a physical network into multiple slices whereby each slice can be viewed as a chain of network functions, called Service Function Chain (SFC) [3] [4]. A SFC consists of a number of network functions of different types, instantiated and accessed following a specific order [5]. Inspired by the fundamental idea behind NFV, i.e. decoupling the software implementation of network functions from the hardware infrastructure [6], the virtualized network functions (VNFs) in a SFC can be deployed on general-purpose processors (GPPs) rather than proprietary hardware [7]. These VNFs can be easily installed/uninstalled in a server/virtual machine (VM), or migrate from one server to another in order to satisfy t
(

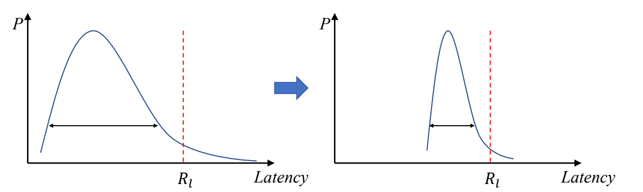

Fig. 1: The distribution of latency experienced by packets in a network slice.

Thanks to its flexibility and customization features, NS is supposed to support various vertical industries, such as industry automation, vehicle internet, which have highly strict QoS requirements. To this end, a new concept called Deterministic Networking (DetNet) [10] [11] is proposed by the IETF DetNet Working Group to study implementation of deterministic data paths for real-time applications (e.g., audio and video streaming, vehicle remote control) with extremely low packet loss rates, low packet delay variations (i.e., jitter), and bounded latencies. As shown in Fig.1, packets of a network slice would experience a certain latency with a probability P when passing through a switch, because of the statistic multiplexing feature of Ethernet networks. DetNet aims 1) to reduce the higher bound of latency to ensure latency distribution within the latency requirement R_l, in other words, avoid the long tail of latency distribution, and 2) to shorten the difference between the lower bound and the higher bound of latency, to ultimately reduce the jitter.

In this paper, we will investigate the resource provisioning problem so as to form a "deterministic latency-bounded network slice". For example, there are two network slices: one network slice consists of a service function chain $\{s_1, s_2, s_3\}$ whereby the end-to-end latency of 60% packets must be less than 50 ms, whiles another network slice runs another service

Hao Yu is with the Department of Communications and Networking, School of Electrical Engineering, Aalto University, Espoo 02150, Finland. E-mail: hao.yu@aalto.fi.

Tarik Taleb is with CWC, Oulu University, Oulu, Finland, and also with the Department of Computer and Information Security, Sejong University, Seoul 05006, South Korea. E-mail: tarik.taleb@oulu.fi.

Jiawei Zhang is with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, 100876, China. E-mail: zjw@bupt.edu.cn.

Honggang Wang is with the Department of Electrical and Computer Engineering, University of Massachusetts at Dartmouth, North Dartmouth, MA 02747 USA. E-mail: hwang1@umassd.edu.

function chain $\{s_2, s_3, s_4\}$ whereby the delay of 90% packets must be less than 10 ms. To create such two network slices, the network service provider (NSP) shall determine how to allocate network resources, including computation resources for VNFs along the service chain and bandwidth resources for routing the traffic between VNFs, in an adequate manner that ensures the deterministic latency requirements. The amount of resources required for such two slices are different, intuitively, different amount of resources yield different service latencies. Hereby, we divide this study into two parts which can be summarized as follows:

- What is the minimum amount of resources needed for a network slice given the traffic demand to satisfy the latency requirement?
- How to deploy network slices so that the overall network resource utilization is minimized?

To solve the first problem, we introduce the Stochastic Network Calculus (SNC) [23] method to estimate latency bound given the amount of resources allocated for a network slice under a given traffic arrival. Indeed, classical queuing theory analyzes a network of queues by providing an average performance estimation, which is not accurate enough. To solve this problem, Deterministic Network Calculus (DNC) [24] has been developed as a set of tools that can analyse network performance by providing a worst-case latency bound. In addition, in order to deal with SLA in such a form of latency bounds: e.g., 95% packets have less than 10 ms latency, SNC is proposed by introducing probability when calculating latency bounds, which is not too pessimistic for latency estimation. In turn, we can determine the amount of resources needed for a slice with SNC given the latency requirement of a network slice.

Regarding the second problem, we consider it in an IP-over-wavelength division multiplexing (WDM) based metro-aggregation network architecture. WDM-based optical underlying network provides not only huge network bandwidth capacity, but also lower switching latency compared with electric switching. Recently, China Telecom has proposed M-OTN (Mobile-Optical Transport Networks) to form 5G transport network architecture for hosting 5G services by flexible slicing in an optical-electric hybrid network. Although optical communication brings huge bandwidth capacities, we need also to balance the huge capacities and costs for fiber deployment. Thus, we integrate the VNF resource allocation with traffic grooming, which is the classical problem in IP-over-WDM networks, together to formulate the network slice deployment problem for minimizing the network cost. In this problem, VNF resource allocation can be modelled as a knapsack problem and traffic grooming is actually a routing and wavelength allocation (RWA) problem in IP-over-WDM networks. Based on this, we propose a deterministic latency bound based network slice deployment strategy with three different objectives: minimizing traffic hops, minimizing lightpaths and minimizing wavelengths (i.e., referred to as $minHop$, $minLP$ and $minWL$, respectively) to provide an efficient slice deployment solution according to NSP's different considerations on network deployment.

Our main contributions are summarized as follows:

- We first investigate the amount of resources needed for a network slice given the traffic demand and latency requirement. This information can help NSP to allocate enough quantity of network resources to a specific network slice without violating the latency requirement.
- We then propose an efficient network slice deployment strategy with the three objectives and compare the proposed strategy with two existing strategies.

Note that, the purpose of this paper is to minimize the overall network costs while ensuring the deterministic end-to-end latency of each network slice, rather than to minimize the end-to-end latency. We first derive the minimum amount of network resources that should be allocated to a network slice to meet the latency requirement with SNC. Then, we devise the strategy to deploy the network slices with the amount of network resources derived from the first step with the objective of minimizing the overall network cost.

The rest of the paper is organized as follows. Section II discusses the related work. Section III introduces a typical IP-over-WDM metro-aggregation network architecture. We formulate the target problem in Section IV and our solution is presented in Section V. Section VI shows the performance evaluation. Section VII concludes this paper.

II. Related Work

The authors in [12] proposed a multi-domain orchestration and management framework to address the challenges of network slicing when utilizing federated resources. Based on this, the authors in [13] introduced of a new architecture in compliance with the ETSI-NFV model and the 3GPP specifications. They then formulated a Mixed Integer Linear Programming (MILP) model and a greedy-based heuristic algorithm to solve the cross-domain network slicing deployment problem regardless the underlying topologies (i.e., both the VNF layer and the physical layer). This is while satisfying all constraints and specifications requested by the end-user or a given vertical's application by investigating the possible trade-offs between execution runtime and network slice deployment time. The results show that the proposed solution guarantees the required delay and bandwidth, ensures efficient usage of both the VNF nodes and the physical nodes, and reduces the service provider's Operating Expenditure (OPEX). In [14], the slice isolation requirement is mainly considered in WDM-OTN based metro aggregation networks for the radio access networks (RAN) slice deployment problem based on the 3-layer RAN architecture. This paper jointly solved the RAN function placement and RWA problems when deploying a RAN slice by proposing a heuristic to minimize 1) the number of active physical nodes (i.e, hosting RAN functions) and 2) the number of established wavelength channels under constraints of network capacity and latency requirements.

Zhang et al [15] presented a hierarchically two-phase solution for joint optimization of VNF chain placement and request scheduling problem. They formulated the VNF chain placement problem as a variant of variable-sized bin-packing problem and proposed a priority-driven weighted heuristic

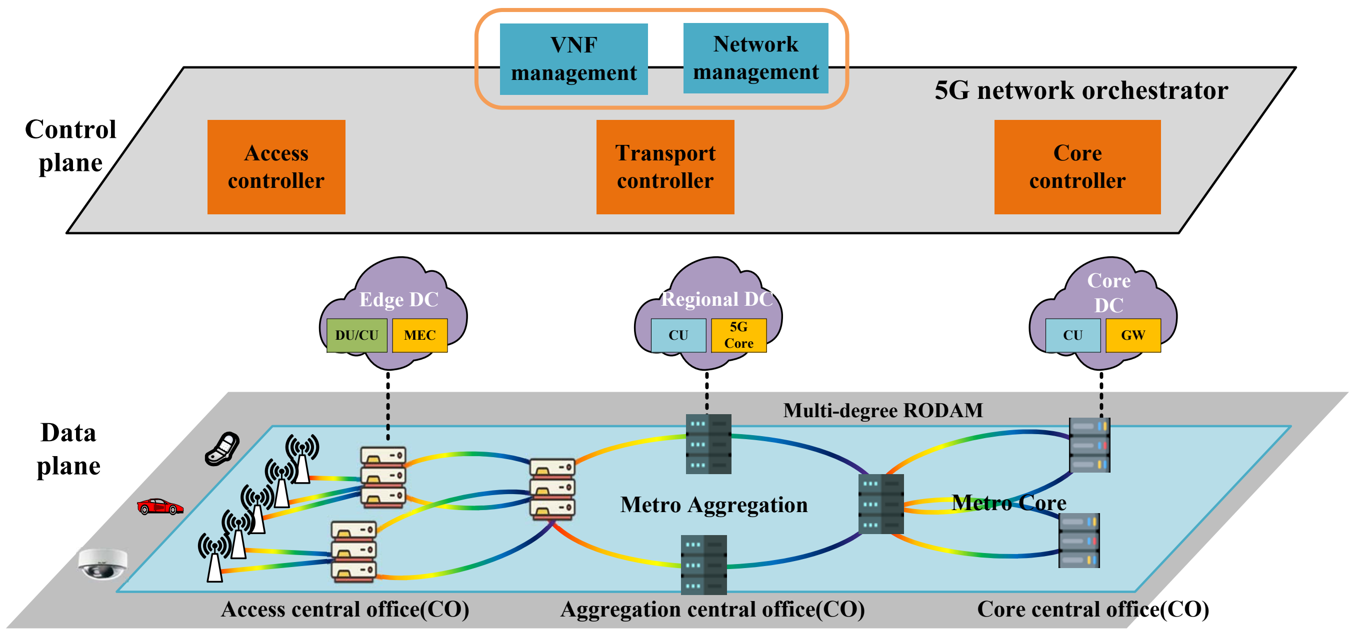

Fig. 2: 5G integrated metro aggregation network architecture.

for VNF chain placement to ensure a near-optimal solution. Then, to minimize the average response latency of each service instance, they also proposed another heuristic for request scheduling to optimize request scheduling cost. The authors in [16] investigated the problem of resource allocation, particularly the joint task of access control (AC) and VNF-forwarding graph embedding (FGE). Two variants of such a problem with respect to VNF splitting and multi-path routing were considered. By mathematically formulating the joint task as an ILP in case of unsplittable VNF and non-permissible multi-path routing, and as a MILP, when covering splittable VNF and multi-path routing scenarios, the authors employed successive convex approximation methods as well as heuristic approaches to find the solution of AC/FGE in a tractable manner. Through simulations, they demonstrated that such methods outperform earlier methods presented to address the AC/FGE problem.

Although the above mentioned research work attempt to solve the resource allocation and service scheduling problem from different perspectives, they did not consider slice deployment under a precise latency requirement and hence can not ensure an end-to-end deterministic service latency for network slices. To this end, the authors in [17] tried to investigate the relationship between the amount of resources allocated to a slice and the service latency bound. They applied stochastic network calculus to derive the latency bound for a given traffic distribution and amount of resources. This is also helpful for network service providers to improve network utilization without violating QoS constraints. The authors in [18] proposed a new DASH-based solution using the deterministic network calculus (DNC) theory in SDN-enabled networks to provide users with high video quality. The proposed solution leverages the benefits of SDN to provide the clients with the highest possible video quality whilst ensuring the delay bound.

With respect to these research work, in this paper, we focus on slice deployment with the objective of ensuring a deterministic latency bound in an IP-over-WDM metro aggregation network. We use SNC to determine the amount of resources to be allocated to a slice. We then solve the problem by jointly considering the VNF placement and RWA for network slices. The findings of this work shall help operators to efficiently deploy network slices so that ensures their respective DetNet requirements.

III. IP-over-WDM based Metro Aggregation Network Architecture

From the perspectives of flexible service provisioning, the 5G metro aggregation network architecture is conceived to be a deep convergence of cloud and network, supported by the NFV and SDN technologies. As shown in Fig.2, computation and bandwidth resources located within central offices (COs) can be managed federatively under the controller of different network domains with VNF and network management functions. The service quality of a network slice can be largely impacted by the allocated resources of connectivity (i.e., bandwidth) and computation.

For the underlying transport networks, to increase the network bandwidth capacity and alleviate electronic limitations on the device speed, the rise of optical fibre communications brings huge capacities to the Internet with the ability of multiplexing low rate signals into a higher rate transmission. WDM is a kind of multiplexing scheme that is analogous to frequency division multiplexing (FDM) in radio communications. In WDM, optical bandwidth is divided into several lower narrow-band channels, referred to as wavelengths. Signals are transformed into a compatible wavelength configuration from the client sides through transponders and are multiplexed using wavelength multiplexers. At the receiver sides, the composite

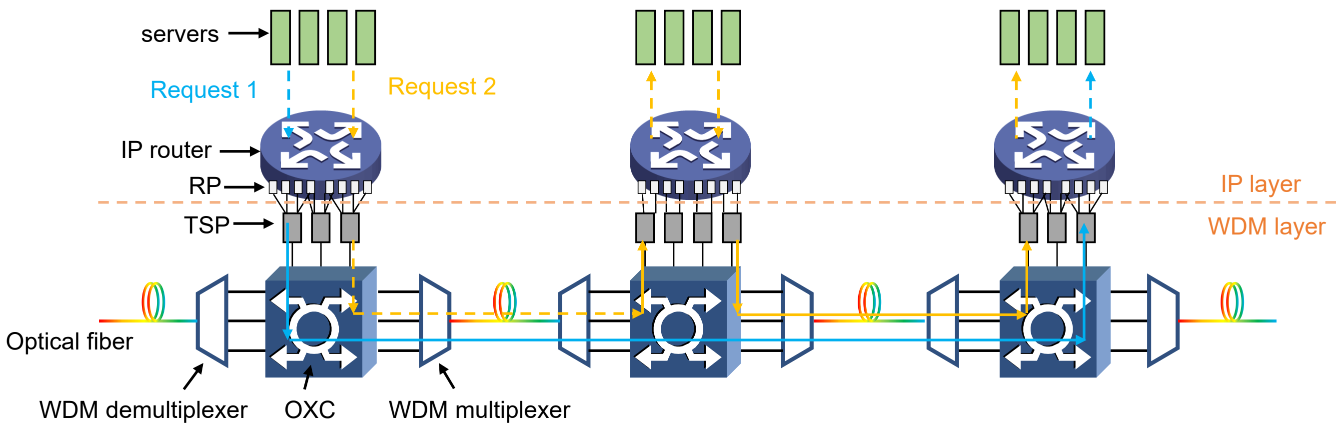

Fig. 3: IP over WDM network architecture with transparent mode (RP = router port, TSP = transponder, OXC = optical cross-connect).

WDM signals are de-multiplexed into individual wavelength channels by received clients. Optical Cross Connects (OXC) are essential elements in the WDM network as they can provide switching ability of WDM channels between input and output ports, in addition to wavelength management and protection functionality. WDM, as the direct provider of capacity, can act as the logical channel to carry the IP payload, hence the current layered combination as IP over WDM [20]. In this combination, the IP layer is used for traffic aggregation, protection, and route calculation, whereas the WDM layer is responsible for providing routes, flow protection and network restoration. Note that in IP-over-wavelength routed networks, switching is performed based on the knowledge of the wavelengths and the ports of entry, which needs the coordination of packets and wavelength management function among different layers using the 5G network orchestrator. Moreover, the SDN technology needs to be extended to the optical domain to improve the controllability of the networks by flexibly configuring, modifying and managing the network based on different QoS requirements. The controllers in different domains are supposed to cooperate on the network configuration and management to provide customized network slices with deterministic QoS. Fig. 3 shows the components used in an IP over WDM network:

IP routers are the part for routing traffic flows from user devices (e.g., servers, cellular access points) which represent the client side of the optical terminals. They receive the aggregated traffic from edge routers and interface the optical side of the networks.

OXC routers, working in the translucent mode (combination of opaque and transparent mode [21]), consist of both electronic module and optical module offering both options for switching signals. If the signal just needs to passby the switching node, the wavelength routing can be applied only with optical module, no optical-electrical-optical (OEO) conversion happens, e.g., **request 1** shown in Fig. 3. However, if the signal needs to be processed in the intermediate switching node, electronic module comes in handy to provide traffic routing to the servers, e.g., **request 2**. The client side of the transponder (TSP) receives the signal from the IP router port (RP), and converts the client side signal into the required WDM wavelength signal.

We assume that the links that connect a server to a router in a node have capacities equal to the sum of the other ports of the router, so that there is no bottleneck in the local communication between a router and a server. In this paper, the latency contributors include computing latency in servers, electronic switching latency, optical switching latency and signal propagation latency in optical fibers. Because light travels approximately 1.5 times slower through optical fiber than in a vacuum, the latency is 5 μsec per kilometer. In the metro aggregation network context, optical switching and signal propagation latency contributor can be ignored, since it is negligible compared with other latency contributors (in order of milliseconds), thus, we only consider the computing and electronic switching latency in this paper.

IV. Deterministic Latency Bound based Slice Deployment Problem Formulation

A. Minimum Resource Capacity for a Network Slice Under Latency Requirement

A network slice request can be modelled as $r = \langle S, A, T, p \rangle$, where $S = \{s_1, s_2, ..., s_m\}$ is the m ordered VNFs forming service chain of this slice, $V = \{v_{m+1}, v_{m+2}, ..., v_n\}$ is the ordered virtual links between VNFs, A denotes traffic arrivals of this network slice, T is the end-to-end latency requirement, and p is the percentage of the traffic arrivals whose end-to-end latency must be less than T.

$C_i, i \in (1, n)$ denotes the required amount of packets which must be processed per second on a VNF s_i or transported per second on a link v_i so that the required latency can be satisfied. Let ω_i denote the amount of computation or bandwidth resources to process/transmit *one* packet on s_i or v_i. It is determined by the task type of the corresponding VNF s_i, different VNFs are featured with different ω_i. In this work, we assume that traffic arrivals to a network slice obey the Gaussian distribution, since various studies of real Internet traffic validated that the Internet traffic exhibits Gaussian characteristics [22]. That is, the amount of packet arrivals $A(t)$ during a time duration t follows a distribution $A(t) \sim N(\mu t, \sigma^2 t)$.

Based on the SNC theory, given $C = \{C_1, C_2, ..., C_n\}$, traffic arrivals A and the percentage p, an experienced end-to-end service latency $D_r(p, A, C_1, ..., C_n)$ of a network slice r can be derived. The derivation of latency function $D_r(p, A, C_1, ..., C_n)$ is not the focus of this paper and can be referred to [17]. In turn, given the latency requirement T of a network slice, we can obtain the $C = \{C_1, C_2, ..., C_n\}$ that should be allocated to this slice in order to satisfy the latency requirement $D_r(p, A, C_1, ..., C_n) \leq T$, which is formulated as follows:

Problem 1:

$$\underset{C_i}{min} \sum_{i=1}^{n} \omega_i C_i \tag{1}$$

$$s.t. \ D(p, A, C_1, ..., C_n) \leq T \tag{2}$$

We refer to the solution proposed in [17] to solve the Problem 1, and the minimum resource capacity in Problem 1 can be derived as follows[1]:

$$C_i = ln \frac{(\omega_i - \omega_1)\alpha + \omega_1 e^{\theta C_1}}{\omega_i} / \theta \tag{3}$$

All C_i can be expressed in terms of C_1. If we can find the minimum C_1 that makes $D(p, A, C_1) = T$, the Problem 1 is solved. We know that 99.7% traffic arrivals lie in the range of $[\mu - 3\sigma, \mu + 3\sigma]$ if we assume a Gaussian traffic distribution. Considering that the number of arrivals cannot be negative and the service rate should always exceed the mean arrival rate, the solution of C_1 locates in the range of $C_1 \in [\mu, \mu + 3\sigma]$. Thus, we can apply bisection search algorithm to find the minimum value of C_1 that can make $D(p, A, C_1) = T$ true, the probability of missing the optimal C_1 will be less than 0.03%.

B. Deterministic Latency Bound Based Slice Deployment

After determining the resource requirement of a single network slice, we investigate the slice deployment problem in IP-over-WDM metro aggregation networks, which can be formally stated as follows. Given i) a metro-aggregation network topology, represented by a graph $G(N, E)$, where N is the set of nodes and E is the set of optical fiber links, and ii) network slices with resource requirements C derived in Section IV.A, we need determine how to deploy slices over the physical infrastructure with three objectives: i) minimizing the total traffic hops, ii) minimizing the established lightpaths, and iii) minimizing the active wavelengths. In the next section, we propose the efficient algorithms to solve this problem for different objectives which are usually conflicting with each other.

[1]The definition of α and θ can also be referred in [17], we just need to know C_i is proportional to C_1 in this paper.

V. Deterministic Latency based Network Slice Deployment Strategy

In this section, we will introduce the deterministic latency based network slice deployment strategy. The objective of this strategy is to minimize the network resource utilization when deploying network slices with deterministic latency requirements. Basically, the network slice deployment can be divided into two parts: VNF placement and traffic routing between VNFs. In this paper, we focus on the traffic routing and wavelength allocation. The VNFs of a network slice are

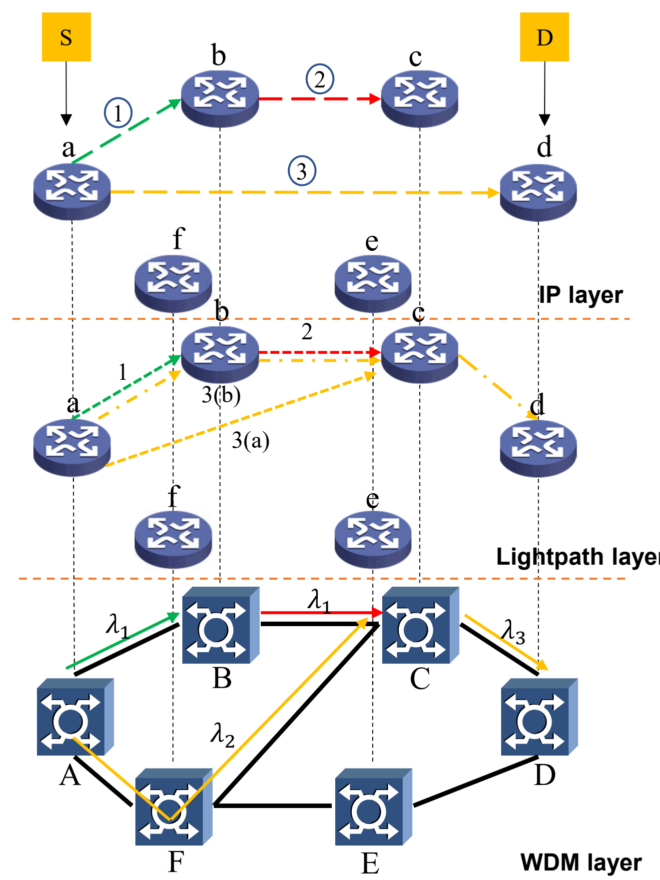

Fig. 4: Two possible request routing strategies.

A. Traffic grooming in IP-ove-WDM networks

For the sake of clarity, we divide the envisioned IP-over-WDM network into three layers: IP layer, lightpath layer and WDM layer. The traffic of network slices are generated in the IP layer and multiple IP traffic flows can be multiplexed into one wavelength channel which can be referred to as a lightpath. As shown in Fig.4, one connection in IP layer denotes one network slice. We assume the set of network slices to be deployed as $r_1(a, b), r_2(b, c), r_3(a, d)$. Specifically, $r_1(a, b)$ denotes that the source and destination nodes are a and b nodes on the physical topology. Initially, there are no established paths for hosting $r_1(a, b)$. A new lightpath needs to

TABLE I: Comparison of different routing options for $r_3(a, d)$

Option	Hops	Lightpaths	Wavelengths	NFV nodes
3(a)	2	2	3	a,c,d
3(b)	3	1	1	a,b,c,d

be established by routing along the shortest path $\{a, b\}^2$ on the physical topology and using wavelength λ_1 in link (a, b). The directed edge $\langle a, b\rangle$ is added in the ligthpath layer to represent a newly established lightpath. Since there is no intermediate node between the source and destination nodes of this network slice, a $single\ hop$ grooming is performed. If there is no restriction on VNF placement, the VNFs of $r_2(b, c)$ can be deployed onto nodes B and C with first-fit strategy if there are enough computing resources. Similarly, the network slice $r_2(b, c)$ is groomed by establishing a new ligthpath $\langle b, c\rangle$ using wavelength λ_1. The deployment of network slice $r_3(a, d)$ is slightly different. Since there is no direct link between physical nodes A and D, there are multiple candidate paths for $r_3(a, d)$. For example, as shown in Fig.4, the physical paths for network slice $r_3(a, d)$ could be either $\{a, b, c, d\}$ or $\{a, f, c, d\}$. For the network segment from A to C, there are two options that can be applied to the traffic routing as shown in Fig. 4:

$3(a)$: Establish a new ligthpath between node A and Node C by routing along the shortest path $\{a, f, c\}$ in the physical topology using wavelength λ_2 in each of the two links. A new lightpath edge will be added on the lightpath layer. Then, $single\ hop$ grooming is performed in this case without O-E-O conversion in the path.

$3(b)$: Reuse the existing lightpath $\langle a, b\rangle$, $\langle b, c\rangle$, if the residual capacity of both two lightpaths is enough for the traffic load of this network slice. Since the traffic is dropped at node B to IP router and then added to lightpath $\langle b, c\rangle$, even these two lightpaths use the same wavelength λ_1, $multi\ hop$ grooming is performed. An O-E-O conversion that happens at the node B which will cause a switching latency.

Although $3(a)$ grooms the traffic into one lightpath, which is a $single\ hop$ grooming (1 traffic hop), it will occupy a newly established wavelength λ_2 along the path $\{a, f, c\}$. On the contrary, $3(b)$ reuses the existing lightpaths and increases the wavelength resource efficiency, but it leads to a $multi\ hop$ grooming and an O-E-O conversion is required. A summary of these two options in terms of traffic hops, lightpaths, wavelengths and NFV nodes is given in Table. I. Obviously, the traffic hops are 2/3 if $3(a)/3(b)$ are applied respectively. The number of newly established lightpaths is 2 (i.e., $\langle a, c\rangle$, $\langle c, d\rangle$) in $3(a)$, while the number of newly established lightpaths is 1 (i.e., $\langle c, d\rangle$) in $3(b)$. As for wavelength, the number of newly activated wavelengths in $3(a)$ is 3 (i.e., λ_2 in (a, f) and (f, c), λ_3 in (c, d)), while $3(b)$ only activates one wavelength λ_3 in (c, d). Finally, there are three nodes (i.e., a, c, d) where VNFs of $r_3(a, d)$ can be deployed in $3(a)$, while the VNFs can be deployed on nodes a, b, c, d in $3(b)$. In general, given a network slice $r(s, d)$, the network operator has the following four different options to deploy $r(s, d)$.

[2] In this paper, we use () to denote a link in WDM layer, ⟨⟩ to denote a lightpath in lightpath layer, and {} to denote a path in WDM layer.

- $Option\ 1$: Search the lightpath layer to find an existing lightpath with sufficient spare capacity that connects directly s with d and then routes $r(s, d)$ on it (i.e., $single\ hop$ grooming). In this case, the VNFs are deployed into source and destination nodes of this lightpath.
- $Option\ 2$: Search the lightpath layer to find a logical path consisting of multiple lightpaths, each with sufficient spare capacity that connects s and d and then routes $r(s, d)$ on it ($multi\ hop$ grooming). In this case, the VNFs can be placed in the intermediate nodes along this logical path.
- $Option\ 3$: Establish a new lightpath directly between s and d and then route $r(s, d)$ on it ($single\ hop$ grooming).
- $Option\ 4$: Partly reuse the existing lightpath(s) with sufficient spare capacity and then establish new lightpath along the given path ($multi\ hop$ grooming).

B. Deterministic Latency based Slice Deployment

Although the required network resources of a network slice can be obtained with the solution in Section IV.A, the actual resource consumption of a slice relies on routing scheme as well, e.g., grooming the slice into a shorter path will result in less bandwidth consumption. Moreover, there also exist tradeoffs between the traffic hops, lightpaths and wavelengths by applying various routing schemes. $single\ hop$ grooming leads to less lightpaths. However, it causes more wavelengths consumption since the traffic flow occupies an exclusive wavelength channel along the path. $multi\ hop$ grooming can facilitate more bandwidth resource multiplexing, but will involve more lightpaths.

Based on the four options mentioned above, we introduce the proposed strategy, namely the Deterministic Latency based Slice Deployment (DLSD) strategy. The deployment strategy is designed with three different objectives: $minHop$, $minLP$, $minWL$ [26].

Before introducing our proposed solution, we need to modify option-4 with several objectives. As shown in Fig. 5, the solid lines represent existing lightpaths between nodes. For request $r_4(A, E)$ to be deployed, there are no single or multiple existing lightpath(s) connecting node A and node E. By applying option-4, different options can be chosen. If the objective is to minimize the traffic hops, a new lightpath $\langle B, E\rangle$ should be established directly (dashed line 1). If the objective is to minimize the number of lightpaths, either lightpath $\langle B, E\rangle$ or $\langle C, D\rangle$ can be applied. If the objective is to minimize the number of wavelengths, the lightpath $\langle C, D\rangle$ (

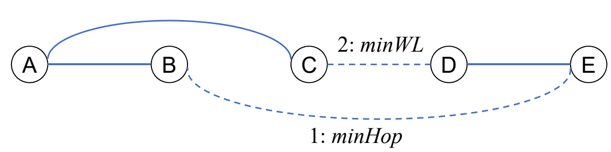

Fig. 5: Option-4 with different objectives.

Based on the above, the DLSD strategy with different objectives is described as follows:

Algorithm 1: Deterministic Latency based Slice Deployment (DLSD).

Input: network slice requests $\boldsymbol{R}$, physical topology $G_P(V_P, E_P)$
Output: (VNFs placement, RWA) scheme

1 Initialize $G_L(V_L, E_L)$ by $V_L \leftarrow V_P, E_L \leftarrow \emptyset$
2 $\boldsymbol{S} \leftarrow PreProcess(\boldsymbol{R})$ **while** $\mathbf{S} \neq \emptyset$ **do**
3   for $r \in S$
4   **if** *objective* = $minHop$ **then**
5   *Path* $\leftarrow Search(G_L, r)$ $\#Option$ 1
6   **if** Path $= NULL$ **then**
7       $l \leftarrow rwa(G_p, r)$ $\#Option$ 3
8   **if** $l = NULL$ **then**
9   Apply Option-4 with $minHop$
10         $S \leftarrow S - \{r\}$
11   **else**
12         $G_L \leftarrow grooming(G_L, l, r)$
13         $S \leftarrow S - \{r\}$
14   **else**
15       $G_L \leftarrow grooming(G_L, Path, r)$
16       $S \leftarrow S - \{r\}$
17   **else if** *objective* = $minLP/minWL$ **then**
18   *Path* $\leftarrow Search(G_L, r)$ $\#Option$ 1&2
19   **if** $Path = NULL$ **then**
20   Option-4 with $minLP/minWL$
21       $S \leftarrow S - \{r\}$
22   **else**
23       $G_L \leftarrow grooming(G_L, Path, r)$
24       $S \leftarrow S - \{r\}$

$minHop$ – Minimize the total number of traffic hops: For a given network slice, $single\ hop$ grooming option-1 is selected first. If it does not succeed, then $single\ hop$ grooming option-3 is tried. If option-3 fails, then option-4 is chosen with the objective of minimizing traffic hops. Here, we define the traffic hops of a network slice as the number of hops between source and destination nodes. For example, if there is no O-E-O conversion in this network slice, the traffic hops of this network slice is one.

$minLP$ – Minimize the number of established lightpaths: For a given network slice, option-1 is selected first. If is does not succeed, then option-2 is tried. If both options fail, option 4 with the objective of minimizing the number of lightpaths can be applied.

$minWL$ – Minimize the number of wavelengths used for the establishment of lightpaths: For a given network slice, option-1 is tried first. If it fails, then option-2 is tried. If both options fail, a new lightpath is going to be established. Option 3 and option 4 with the objective of minimizing the number of wavelengths are compared, and the option with fewer wavelengths will be chosen.

The complete algorithm pseudocode is shown in **Algorithm 1**. Initially, a graph $G_L(V_L, E_L)$ of lightpath layer is created to store the existing lightpath information (line 1) and $G_P(V_G, E_G)$ represents the physical network topology (i.e., WDM layer). Then, network slices are processed by $PreProcess()$ to sort the requests in terms of service latency requirements (line 2). Network slices with strict latency requirements are handled with high priority. If the algorithm objective is $minHop$, we will search if there are existing lightpaths in G_L that can host the requested network slice r with $Search()$ (line 5). Otherwise, a new lightpath is going to be established with $rwa(G_P, r)$ performing the required RWA on G_P (line 7). In $rwa(G_P, r)$, we employ shortest path algorithm and the well-known First-Fit policy for wavelength assignment. The lightpath l, returned by $rwa()$, is then used for grooming r, then deploying the requested network slice r with $grooming(G_L, l, r)$ (line 12). If there is no available wavelength for establishing a new lightpath for r, option-4 is then performed with $minHop$ for a multi-hop routing scheme (line 10). If there is an existing lightpath for accommodating network slice request r, $grooming(G_L, Path, r)$ is carried out to multiplex the traffic into existing lightpath G_L (line 15). If the objective is $minLP/minWL$, we first search for existing lightpath(s) in G_L (line 18). If none is available, option-4 (multi-hop grooming) is performed which tries to reuse lightpaths and avoids establishing a new lightpath (line 20).

It shall be noted that in the cases of $minHop/minLP/minWL$, the VNF placement will be performed after the path is selected. For example, if the path selected for a network slice is a $single - hop$ path, the VNFs of the network slice r will be placed in the source and destination nodes of the selected path. If the path selected is a $multi - hop$ path, the VNFs will be also placed on the intermediate node(s) besides the source and destination nodes of the selected path. In order to prevent congestion at specific nodes (that would ultimately impact the network slice deployment), VNFs are placed onto nodes with the maximum residual resources along the path, with the objective of balancing the usage of computing resources among the physical nodes. Specifically, after selecting a path, the resource utilization of each node along the path is checked, and VNFs are placed onto the physical node(s) with the least computation load.

C. Complexity Analysis

Now we give a brief complexity analysis of the proposed algorithm.

Proof: The $search()$ procedure adopted is performed on the lightpath layer to find the shortest lightpath(s) with adequate capacity based on Breadth First Search (BFS). Thus, the time complexity of $search()$ is $O(|V_L| + |E_L|)$ i.e. $O(n + |E_L|)$, due to $|V_L| = |V_P|$. $rwa()$ also employs BFS for finding the shortest path in WDM layer, then the time complexity of $rwa()$ is $O(|V_P| + |E_P|)$ i.e. $O(m+n)$. For the option-4, the worst-case result is to perform a $rwa()$ on WDM layer, so the time complexity is also $O(m + n)$. The overall time complexity of proposed algorithm is as follow. When the worst case happens, i.e. no existing lightpath(s) is available and no new lightpath can be established, both $search()$, $rwa()$ and option-4 contribute to the running time, then an order of

$n+m+|E_L|$ time is incurred for a network slice. Therefore, we can obtain the worst-case time complexity of proposed algorithm as $O(|R|(n+m+|E_L|))$. ■

VI. Performance Evaluation

Hereunto, we evaluate the performance of our proposed network slice deployment strategy.

Simulation Setup: In the simulations, we consider a topology containing 30 physical nodes as in [27]. The computing resource capacity of each physical node is set to 2000 units. One resource unit refers to the ability to handle one packet of 64 Bytes at 10k packet per second (kpps). According to [28], one CPU core can handle packets of 64 Bytes at 1.5 Mpps, which equals 150 units of resources, thus 2000 units of resources indicate 14 CPU cores. The number of wavelengths of each physical link between adjacent nodes are set to be infinite, and the bandwidth capacity of a single wavelength is set to 1 Gbps. We assume that each network slice traverses a service function chain consisting of at most three VNFs [29] and two links which take different amounts of resources ω_i to process/transmit one packet. Assuming that the packet size is 64 Bytes, if the mean packet arrival rate is 200 kpps, thus the mean data rate is 100 Mbps.

Introduction of Compared Algorithms: We will give a brief description about compared algorithms: K-shortest path first fit (KSPFF) and C-Hop [30].

- KSPFF first calculates K shortest paths in WDM layer. Then, among these K paths, it will search the first available wavelength along the candidate paths until a shortest path k with available wavelength is found. The VNFs of network slice are distributed on the nodes along the selected path and the traffic is routed using the first available wavelength.
- The authors in [30] also proposed a new traffic grooming strategy named C-Hop. The basic idea of C-Hop strategy is to groom traffic with a fixed number of hops h on the logical topology. It initially sets the $h=1$ and groom the requests on the path with $h=1$ traffic hop. After removing the requests that are groomed with $h=1$ traffic hop successfully from the request set, it grooms the requests in the next iteration with $h+1$ until all the requests are groomed successfully.

A. Performance of Resource Allocation for a Network Slice under Latency Requirement

We consider network slices running service function chains of three VNFs with traffic arrivals of 200/400 kpps whereby the SLAs are specified as the combination of latency L and probability p (i.e., 50 ms/70%, 80 ms/70%, 50 ms/90%, 80 ms/90%). The amount of resources needed to process one packet on each VNF are set as $\omega_1=1, \omega_2=1, \omega_3=2$, the amount of resources needed to transmit one packet is set as $\omega_4=1$. Obviously, the required capacity increases along with the increment of the mean rate of packet arrivals, as shown in Table.II. Higher packet arrivals lead to much more resource consumption for a network slice. Actually, SNC is supposed to provide the worst case latency bound. It over-estimates the resources to reduce the probability of congestion. Effectively, the network slices experience a much lower E2E latency when the calculated resources are indeed allocated to them.

TABLE II: Resource Capacity Needed for a Network Slice.

Arrival (kpps)		SLA		Service Capacity (kpps)			
μ	σ	T (ms)	p (%)	C_1	C_2	C_3	C_4
200	25	50	70	231.24	228.5	225.31	231.24
200	25	50	90	250.66	246.78	243.33	249.7
200	25	80	70	226.26	223.58	220.45	226.26
400	55	80	90	551.25	547.22	542.6	551.25

B. Performance of Network Slice Deployment with different objectives

Next, we evaluate the performance of the proposed strategy with three objectives. We set the mean packet arrival rate to 400 kpps, the latency requirement to 80 ms and $p=90\%$

1) Number of Hops: As shown in Fig.6, although $minHop$ targets on the minimum number of traffic hops among the three objectives, $KSPFF$ results in least traffic hops, since it always performs $single\ hop$ grooming by directly establishing a lightpath connecting source and destination nodes. For the other strategies, as the number of network slices increases, the traffic grooming of network slices gradually changes from $single\ hop$ grooming to $multi\ hop$ grooming. When there are 300 network slices in the networks, the VNFs of all network slices can be deployed directly on the source and destination nodes of the selected path, thus resulting in 300 traffic hops, since the network resources are relatively adequate and each network slice experiences only one traffic hop. As the number of network slices increases, more lightpaths have been established in the networks. Thus, $minLP$ will have more opportunity to perform $multi\ hop$ grooming with two or more existing lightpaths than $singlehop$ grooming, which will ultimately lead to more traffic hops than $minHop$ case. As for $minWL$, reducing the establishment of wavelengths is the main consideration: it will reuse the existing lightpaths as much as possible as $minLP$ performs, hence the traffic hops resulted by $minLP$ and $minWL$ are similar. Note that, $C-Hop$ leads to more traffic hops than $minHop$, since it always tries to reuse the existing lightpaths with less traffic hops, while $minHop$ establishes a new $single\ hop$ lightpath when it cannot find a existing lightpath with single hop. Thus, $minHop$ results in less traffic hops.

2) Number of Lightpaths: As we can see in Fig.7, when the number of network slices is 300, each network slice occupies only one lightpath, since option-3 or option-1 are the main choices for each network slice at this time. As the number of network slices increase, $minLP/minWL$ try to multiplex more network slices into the existing lightpaths; option-2 or option-4 is then more likely to be selected in $minLP/minWL$ when deploying the network slices. Specifically, the difference between $minHop$ and $minLP$ is that $minHop$ does not select option-2 when there is no existing direct lightpath. It establishes a new lightpath instead. On the other hand, $minLP$ adopts option-2 to avoid new lightpaths. Therefore, the number of established lightpaths of $minHop$ are higher than that of $minLP/minWL$. Note that, despite

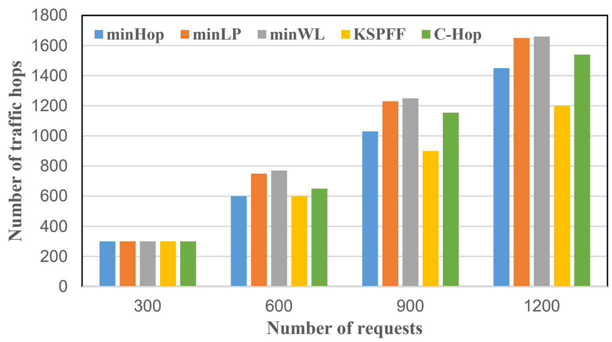

Fig. 6: The number of traffic hops vs the number of network slice requests.

$minLP$ aims at minimizing the lightpath and $C-Hop$ aims at minimizing the traffic hops, the number of lightpaths established in both cased are similar. The reason is that both cases try to groom the traffic into existing lightpaths, but with different ways, thus they have the same performance in terms of established lightpaths.

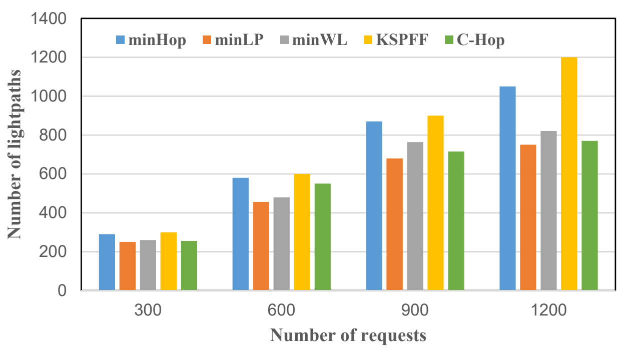

Fig. 7: The number of lightpaths vs the number of network slice requests.

3) Number of Wavelengths: The number of wavelengths used in the networks mean the sum of activated wavelengths in each physical link of the network topology, and a lightpath may contain multiple wavelengths. The number of wavelengths for a network slice usually depend on the number of links along the selected path. From the results shown in Fig.8, for $minWL/minLP$, the number of used wavelengths are much less than that of $minHop$, as both $minWL/minLP$ take advantage of lightpath/wavelength multiplexing. Furthermore, $minWL$ consumes less wavelengths than $minLP$ for the reason that it will check the number of wavelengths that need to be activated and select the lightpath with less links in option-4.

4) Impact of SLA on the resource utilization: Then, we evaluate the impact of SLA on the resource usage. We set $p = 70\%$ and $p = 90\%$, and the E2E service latency requirements vary from 20 ms to 500 ms. The number of wavelengths and lightpaths are then compared among the different latency requirements and probability conditions. We evaluate $minWL$ in terms of the number of wavelengths and $minLP$ in terms of the number of lightpaths, separately. From Section VI.A, we know that stricter latency requirements and larger p result in

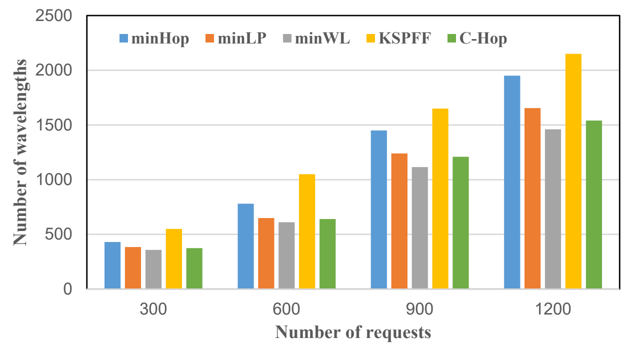

Fig. 8: The number of wavelengths vs the number of network slice requests.

more resource consumption for a network slice. For example, if the latency requirement changes from 20 ms to 50 ms, the resource consumption will increase by 1.5 % on average. If p increases from 70% to 90%, the resource consumption increases by 8.5 %. Therefore, as in Fig.9, the results show that the number of both wavelengths and lightpaths decrease with more moderate E2E service latency requirements and lower p values.

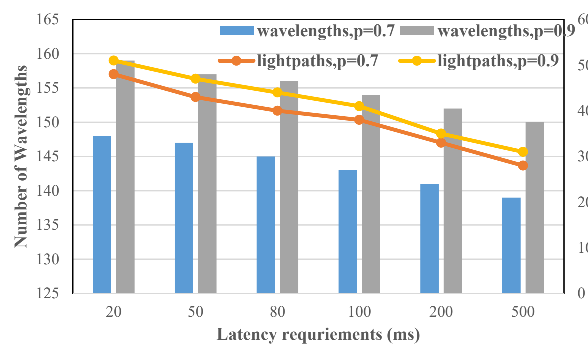

Fig. 9: Wavelength and lightpaths vs latency requirement T and p.

C. Performance of Network Slice Deployment under Dynamic Network Slice Arrivals

We also evaluate the performance of proposed network slice deployment scheme over the dynamic slice arrivals. The arrival rates of network slices follow by the distributions shown in Fig. 10. We set the total time units to be 50 and the number of slice arrivals are counted every 1 time unit. Furthermore, to emulate dynamic load environments, we also set a lifetime for each network slice. The lifetime of each network slice follows the exponential distribution with an average of 10 time units. During the lifetime, the network slices should be allocated enough resources to ensure that the latency requirements are met. Then a network slice will release the occupied resources when its lifetime is expired. Also, for each network slice, the packet arrival distribution is set as $\mu = 200$ and $\rho = 25$. The maximum tolerated latency for each network slice is set to be 50 ms and $p = 70\%$. In this case, we perform the DLSD strategy to deploy the network slices in sequential order based on two kinds of network slice arrivals, then compare the results of the objective $minHop$, $minLP$ and $minWL$ with two benchmark strategies.

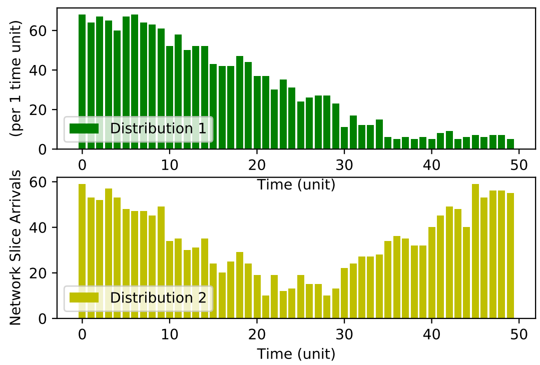

Fig. 10: Two SFC arrival distributions.

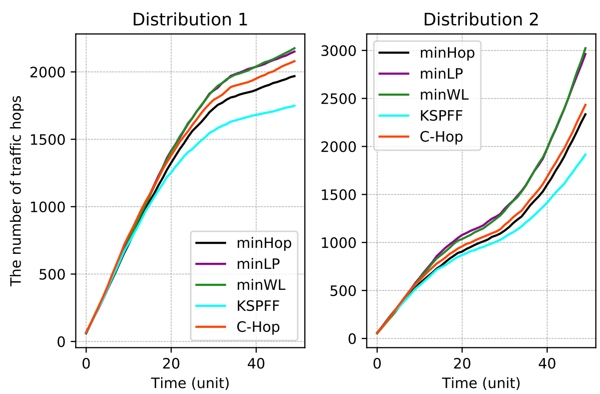

Fig. 11: The number of traffic hops for two slice arrival distributions over time.

As shown in Fig. 11, the number of traffic hops of different strategies over time are compared. Note that, the number of traffic hops mean the sum of traffic hops of accumulated network slices deployed in the networks over time. The traffic hops count when a network slice is deployed until its lifetime is expired. Since the slice arrivals under distribution 1 are decreasing over time, the slope of curve in Fig. 11(a) also decreases over time. The traffic hops of $KSPFF$ are minimal as it directly establishes or reuses one lightpath between source and destination nodes of network slices. If a new lightpath is established, one traffic hop is counted, and the number of traffic hops will not increase if the network slice reuses an existing lightpath in the networks. Hence, $KSPFF$ has the highest number of wavelengths established in the networks, as shown in Fig. 12. $minLP/minWL$ result in most traffic hops in the networks, as they try to reuse as much lightpaths or wavelengths as possible regardless traffic hops, as shown in Fig. 12, they establish least wavelengths compared with other strategies. $minHop$ and $C-Hop$ achieve a balance between traffic hops and the established wavelengths in the networks. And the traffic hops of $minHop$ are slightly higher than $C-Hop$, since $C-Hop$ will directly establish a new lightpath between source and destination if there is no existing available lightpath, whereas $minHop$ will partly establish new lightpaths in option-4 with $minHop$ (i.e., partly reusing the existing lightpaths). Considering the slice arrivals under distribution 2 will increase in the latter part, thus the slope of curve in Fig. 11(b) increases sharply after 30 time units. The performance of each strategies are similar with the case with distribution 1 due to the reason discussed above.

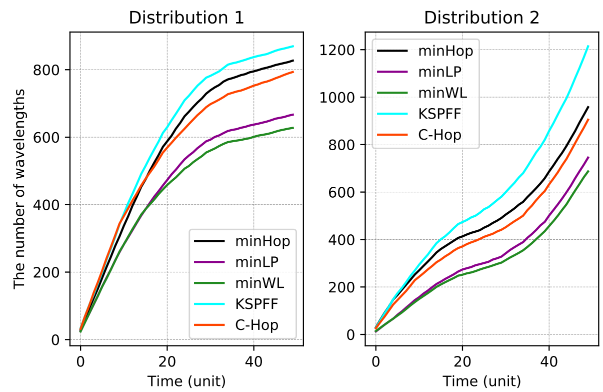

Fig. 12: The number of wavelength for two slice arrival distributions over time.

VII. Conclusion

In this paper, we investigated the deterministic latency bound based network slice deployment problem. First, We introduced an SNC-based method to determine the relationships between traffic arrivals, QoS requirements, and the amount of required resources. We apply SNC to derive the amount of resources needed for a network slice given a specific traffic distribution and E2E service latency requirement. Then, with the amount of required resources, we propose an effective algorithm to deploy network slices into IP-over-WDM metro aggregation networks to minimize network bandwidth resources utilization. The performance of the proposed algorithm with three objectives is evaluated, and encouraging results were obtained. Finally, the results indicate that the proposed strategy will help network service providers improve resource efficiency according to different optimization objectives while ensuring the QoS requirements. In the future, we will further study the deterministic traffic flow scheduling including queuing configuration and time slot scheduling, and provide a more accurate timing analysis to evaluate the proposed scheduling strategy. Then, we combine the deterministic traffic flow scheduling with network slicing to enable the networks with deterministic performance guarantee for the time critical applications in 5G and beyond 5G.

Acknowledgment

This work was partially supported by the European Union's Horizon 2020 Research and Innovation Program through the MonB5G Project under Grant No. 871780. It was also supported in part by the Academy of Finland 6Genesis project under Grant No. 318927 and IDEA-MILL with grant number 33593.

REFERENCES

[1] 5G Americas, “Global organizations forge new frontier for 5G,” Jul. 2016, White Paper.
[2] 5G Americas, “Network slicing for 5G networks & services,” Nov. 2016, White Paper.
[3] H. Hantouti, N. Benamar, and T. Taleb, “Service Function Chaining in 5G & Beyond Networks: Challenges and Open Research Issues,” in IEEE Network Magazine, Vol. 34, No. 4, Aug. 2020, pp. 320 - 327.
[4] H. Hantouti, N. Benamar, T. Taleb, and A. Laghrissi, “Traffic Steering for Service Function Chaining,” in IEEE COMST, Vol. 21, No. 1, Feb. 2019, pp. 487-507.
[5] A.M. Medhat, T. Taleb, A. Elmangoush, G. Carella, S. Covaci, and T. Magedanz, “Service Function Chaining in Next Generation Networks: State of the Art and Research Challenges,” in IEEE Communications Magazine. Vol. 55, No. 2, Feb. 2017, pp. 216-223.
[6] T. Taleb, A. Ksentini, M. Chen, and R. Jantti, “Coping with Emerging Mobile Social Media Applications through Dynamic Service Function Chaining,” in IEEE Trans. on Wireless Communications, Vol. 15, No. 4, Apr. 2016. pp. 2859 – 2871.
[7] R. A. Addad, D.L.C. Dutra, M. Bagaa, T. Taleb and H. Flinck, “Towards studying Service Function Chain Migration Patterns in 5G Networks and beyond,” in IEEE Globecom’19, Hawaii, USA, Dec. 2019.
[8] R. A. Addad, T. Taleb, H. Flinck, M. Bagaa and D.L.C. Dutra, “Network Slice Mobility in Next Generation Mobile Systems: Challenges and Potential Solutions,” in IEEE Network Magazine, Vol. 34, No. 1, Jan. 2020, pp. 84 – 93.
[9] R. A. Addad, D.L.C. Dutra, M. Bagaa, T. Taleb, and H. Flinck, “Fast Service Migration in 5G Trends and Scenarios,” in IEEE Network Magazine, Vol. 34, No. 2, Mar. 2020, pp. 92 - 98
[10] Finn, N., Thubert, P., Varga, B., and J. Farkas, ”Deterministic Networking Architecture,” RFC 8655, DOI 10.17487/RFC8655, October 2019, ¡https://www.rfc-editor.org/info/rfc8655¿.
[11] Grossman, E., Ed, ”Deterministic Networking Use Cases”, RFC 8578, DOI 10.17487/RFC8578, May 2019, ¡https://www.rfc-editor.org/info/rfc8578¿.
[12] T. Taleb, I. Afolabi, K. Samdanis and F. Z. Yousaf, “On Multi-domain Network Slicing Orchestration Architecture & Federated Resource Control,” in IEEE Network Magazine, Vol. 33, No. 5, pp. 242 - 252, Sep. 2019.
[13] R. A. Addad, M. Bagaa, T. Taleb, D. L. C. Dutra and H. Flinck, “Optimization Model for Cross-Domain Network Slices in 5G Networks,” in IEEE Transactions on Mobile Computing, vol. 19, no. 5, pp. 1156-1169, 1 May 2020.
[14] H. Yu, F. Musumeci, J. Zhang, M. Tornatore and Y. Ji, “Isolation-Aware 5G RAN Slice Mapping Over WDM Metro-Aggregation Networks,” in Journal of Lightwave Technology, vol. 38, no. 6, pp. 1125-1137, 15 March15, 2020.
[15] Q. Zhang, Y. Xiao, F. Liu, J. C. S. Lui, J. Guo and T. Wang, “Joint Optimization of Chain Placement and Request Scheduling for Network Function Virtualization,” 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, pp. 731-741, 2017.
[16] M. A. Tahmasbi Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi and B. H. Khalaj, “vSPACE: VNF Simultaneous Placement, Admission Control and Embedding,” in IEEE Journal on Selected Areas in Communications, vol. 36, no. 3, pp. 542-557, March 2018.
[17] Q. Xu, J. Wang and K. Wu, “Learning-Based Dynamic Resource Provisioning for Network Slicing with Ensured End-to-End Performance Bound,” in IEEE Transactions on Network Science and Engineering, vol. 7, no. 1, pp. 28-41, 1 Jan-March 2020.
[18] O. El Marai, J. Prados-Garzon, M. Bagaa and T. Taleb, “Ensuring High QoE for DASH-Based Clients Using Deterministic Network Calculus in SDN Networks,” in IEEE GLOBECOM’19, Waikoloa, HI, USA, 2019, pp. 1-6.
[19] V. Paxson and S. Floyd, “Wide area traffic: The failure of poisson modeling,” IEEE/ACM Trans. Netw., vol. 3, no. 3, pp. 226–244, Jun. 1995.
[20] S. Dixit, “IP over WDM: building the next-generation optical internet,” John Wiley & Sons, 2004.
[21] A. Tzanakaki, I. Zacharopoulos, and I. Tomkos, “Broadband building blocks [optical networks],” IEEE Circuits and Devices Magazine, vol. 20, no. 2, pp. 32-37, 2004.
[22] R. Van De Meent, M. Mandjes, and A. Pras, “Gaussian traffic everywhere?” in Proc. IEEE Int. Conf. Commun., 2006, pp. 573–578.
[23] M. Fidler, “An end-to-end probabilistic network calculus with moment generating functions,” in Proc. 14th IEEE Int. Workshop Quality Service, pp. 261–270, 2006.
[24] Y. Jiang, “Network calculus and queueing theory: Two sides of one coin,” in Proc. 4th Int. ICST Conf. Perform. Eval. Methodologies Tools, Art. no. 37, 2009.
[25] J.Y. Le Boudec and P. Thiran, “Network Calculus: A Theory of Deterministic Queuing Systems for the Internet,” Berlin, Heidelberg: Springer-Verlag, 2001.
[26] H. Zhu, H. Zang, K. Zhu. and B. Mukherjee, “A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks,” IEEE/ACM Transactions on Networking, vol. 11, no. 2, pp. 285-299, April 2003.
[27] S. Orlowski, R. Wessaly, A. Tomaszewski, and R. Wess, “Sndlib 1.0 survivable network design library,” in Networks, pp. 276–286, 2010.
[28] Z. Xu, F. Liu, T. Wang, and H. Xu, “Demystifying the energy efficiency of Network Function Virtualization,” in Quality of Service (IWQoS), 2016 IEEE/ACM 24th International Symposium on. IEEE, pp. 1–10, 2016.
[29] Q. Zhang, Y. Xiao, F. Liu, J. C. Lui, J. Guo, and T. Wang, “Joint optimization of chain placement and request scheduling for network function virtualization,” in Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst., pp. 731–741, 2017.
[30] A. Bhattacharya, H. Kumbhakar, M. Chatterjee and D. Saha, “Controlling Traffic Hops - A New Approach for Traffic Grooming in WDM Optical Networks,” 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-5.

Hao Yu received the B.S. and Ph.D degree in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015 and 2020. He was also a Joint-Supervised Ph.D. Student with the Politecnico di Milano, Milano, Italy. He is currently a Postdoctoral Researcher with the Department of Communications and Networks, Aalto University, Espoo, Finland. His research interests include network virtualization, software defined networking, time sensitive network, deterministic networks.

Tarik Taleb received the B.E. degree (with distinction) in information engineering and the M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a Professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. He is the founder and the Director of the MOSA!C Lab, Espoo, Finland. He is a part-time Professor with the Center of Wireless Communications, University of Oulu, Oulu, Finland. He was an Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a laboratory fully funded by KDDI until 2009. He was a Senior Researcher and a 3GPP Standards Expert with NEC Europe Ltd., Heidelberg, Germany. He was then leading the NEC Europe Labs Team, involved with research and development projects on carrier cloud platforms, an important vision of 5G systems. From 2005 to 2006, he was a Research Fellow with the Intelligent Cosmos Research Institute, Sendai. He has also been directly engaged in the development and standardization of the Evolved Packet System as a member of the 3GPP System Architecture Working Group. His current research interests include architectural enhancements to mobile core networks (particularly 3GPP's), network softwarization and slicing, mobile cloud networking, network function virtualization, software defined networking, mobile multimedia streaming, intervehicular communications, and social media networking.

Prof. Taleb was a recipient of the 2017 IEEE ComSoc Communications Software Technical Achievement Award in 2017 for his outstanding contributions to network softwarization and the Best Paper Awards at prestigious IEEE-flagged conferences for some of his research work. He was a corecipient of the 2017 IEEE Communications Society Fred W. Ellersick Prize in 2017, the 2009 IEEE ComSoc Asia–Pacific Best Young Researcher Award in 2009, the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2008, the 2007 Funai Foundation Science Promotion Award in 2007, the 2006 IEEE Computer Society Japan Chapter Young Author Award in 2006, the Niwa Yasujirou Memorial Award in 2005, and the Young Researcher's Encouragement Award from the Japan Chapter of the IEEE Vehicular Technology Society in 2003. He is a member of the IEEE Communications Society Standardization Program Development Board. He is/was on the Editorial Board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE Wireless Communications Magazine, the IEEE JOURNAL ON INTERNET OF THINGS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, and a number of Wiley journals.

Jiawei Zhang received the Ph.D. degree from the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications (BUPT), China. He currently is an associate Professor with BUPT. Dr. Zhang has authored and co-authored more than 30 OFC/ECOC papers and top journal papers in optical communication and networks. His research interests include the collaboration of optical networks with IP, wireless and cloud/edge, currently with an emphasis on the advanced technologies for providing deterministic connections for future network applications. He served on the Technical Program Committees for the IEEE DRCN 2018-2020, IEEE ICNC 2017-2018, ACP2020, and for the Workshop on Cloud Computing Systems, Networks and Applications at the IEEE GLOBECOM 2014-2016, ICC 2015-2016, and INFOCOM 2017-2018 conferences. He also severed as a Guest Editor of the special issue on Resilience in future 5G Photonic Networks of Photonic Network Communications journal (Springer).

Honggang Wang received the Ph.D. degree in computer engineering from the University of Nebraska-Lincoln, Lincoln, NE, USA, in 2009. Before he joined the University of Massachusetts Amherst (UMass) Dartmouth, North Dartmouth, MA, USA, in 2009, he was with the Bell Labs Lucent Technologies China, Shanghai, from 2001 to 2004 as a Member of Technical Staff. He is currently a Tenured Associate Professor with the UMass Dartmouth and is an affiliated Faculty Member of Advanced Telecommunications Engineering Laboratory, University of Nebraska-Lincoln. He is also the Faculty Member of Biomedical Engineering and Biotechnology Ph.D. Program (BMEBT) with the UMass Dartmouth. He has authored or coauthored more than papers in his research areas. His research interests include Internet of Things, wireless health, body area networks (BAN), cyber and multimedia security, mobile multimedia and cloud, wireless networks and cyber-physical system, and BIG DATA in mHealth.,Dr. Wang was an invited participant by National Academic Engineering (NAE) for 2017 German–American Frontiers of Engineering Symposium, as one of about 50 outstanding Engineers (ages 30–45) from the U.S. Companies, Universities, and Government Labs. He was the General Chair/Co-Chair and Technical Program Committee (TPC) Chair/Co-Chairs for several IEEE/ACM conferences. He is also the Steering Committee Co-Chair of IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) and TPC Co-Chair of IEEE CHASE 2016, which is a leading international conference in the field of connected health. He was the TPC Member for IEEE INFOCOM 2013–2015, International Conference on Distributed Computing Systems 2015, IEEE International Conference on Data Mining 2018, ACM Multimedia 2017, and the Area Chair for IEEE International Conference on Multimedia and Expo 2016 and 2017. He has also been the Editor in Chief (EiC) for the IEEE Internet of Things Journal (SCI impact factor: 9.5), an Associate Editors for the IEEE Transactions on Big Data, IEEE Transactions on Multimedia, an Editor for the IEEE Transactions on Vehicular Technology, Associate Editor for the IEEE Network Magazine. He is currently the Vice Chair of IEEE eHealth Technical Committee and the Chair of IEEE Comsoc Multimedia Communications Technical committee (IEEE MMTC). He was the recipient of the IEEE Multimedia Communications Technical Committee (MMTC) Outstanding Leadership Award (2015) and IEEE HEALTHCOM 2015 Outstanding Service Award. His research is supported by National Science Foundation, Department of Telecommunications, President office, and UMass Healey Grant (total over 2 M dollars). He is an IEEE distinguished Lecturer. He is the "Scholar of The Year" (in 2016, only one per year at UMass Dartmouth).