Traffic-Weighted Availability-Guaranteed Network Slice Composition with VNF Replications

Riti Gour[†], Varin Sikand[*], Joseph Chang[*], Zhouxiang Wu[*], Genya Ishigaki[ϕ], and Jason P. Jue[*]
[†] Division of Mathematics, Computing, and Statistics
Simmons University, Boston Massachusetts 02115, USA
[*]Department of Computer Science
The University of Texas at Dallas, Richardson, Texas 75080, USA
[ϕ]Department of Computer Science,
San Jose State University, San Jose, California 95192, USA
Email: riti.gour@simmons.edu, {vss180000, jxc210027, zhouxiang.wu, and jjue}@utdallas.edu, genya.ishigaki@sjsu.edu

***Abstract*—In this work, we consider the network slice composition problem for Service Function Chains (SFCs), which addresses the issue of allocating bandwidth and VNF resources in a way that guarantees the availability of the SFC while minimizing cost. For the purpose of satisfying the availability requirement of the SFC, we adapt a traffic-weighted availability model which ensures that the long-term fraction of traffic supported by the slice topology remains above a desired threshold. We propose a method for composing a single or multi-path slice topology and for properly dimensioning VNF replicas and bandwidth on the slice paths. Through simulations, we show that our proposed algorithm can reduce the total cost of establishment compared to a dedicated protection approach in 5G networks.**

***Index Terms*—5G networks, network slicing, service function chains, availability.**

I. Introduction

The emergence of Network Function Virtualization (NFV) has introduced the possibility of supporting high bandwidth and low latency services by allowing network operators to manage and expand their network capabilities on demand using virtual, software-based applications [1]. NFV has also led to a reduction in the overall cost and an increase in resource utilization as seen from the perspective of a network operator. Virtual Network Functions (VNFs) in an NFV environment may be chained together to form a Service Function Chain (SFC). Different SFCs for distinct applications can be supported over the same physical infrastructure where computing resources are provided to support VNFs and bandwidth resources are allocated for the logical links in the SFC.

Recent telecommunication networks such as 5G networks are utilizing the concept of network slicing to support multiple isolated virtual networks over the physical infrastructure, as shown in Fig. 1. In network slicing, logically isolated networking and computing resources are tailored according to application service requirements and are allocated on a common physical infrastructure [2]. Availability is an important QoS requirement for network slices and is often specified in the service level agreement between the network operator and the slice operator. *Availability* is defined as the fraction of time for which the network slice is up and providing resources for the services running on the slice [3].

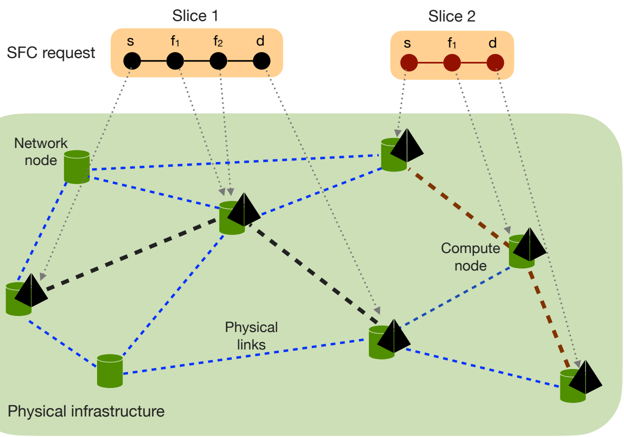

Fig. 1. Network slice on the physical infrastructure.

A traditional approach for guaranteeing availability is to provision extra resources as a protection mechanism, resulting in additional cost for network operators. *Traffic-weighted availability* is another metric that accounts for the long-term fraction of traffic that is supported by the slice [4]. The main advantage of traffic-weighted availability stems from the fact that it helps in reducing the amount of resources that are allocated for satisfying the availability requirement of a service. In contrast to traditional availability, traffic-weighted availability minimizes the number of redundant resources allocated to a slice and allocates just enough resources to satisfy the required demands.

In this work, we study a problem of network slice composition which requires determining the mapping of VNFs and bandwidth flows of an SFC on slice nodes and links respectively, and then the mapping of slice nodes and links on the physical infrastructure. In the mapping process, proper dimensioning of computing and bandwidth resources is required such that the availability requirement of the SFC is satisfied and the cost of establishment for the network operator is minimized. An important decision to make while designing a slice topology is whether to consider a single-path topology or to resort to a multi-path topology with the motivation to prevent the slice from reserving unnecessary resources. If the availability of the VNFs and physical resources on a single path can satisfy the availability requirement of the SFC, then a single-path slice topology with the appropriate bandwidth allocation is sufficient to support the SFC. However, if the availability of a single path is less than the required availability, then a multi-path slice topology must be utilized. Once the topology is constructed, we utilize the concept of traffic-weighted availability to further reduce the amount of bandwidth that is allocated on each path of the slice.

In our previous work [4], we proposed a slice composition problem and constructed a multi-path slice topology by allocating appropriate amount of bandwidth on each path to satisfy the availability requirement. In this work, we present a slice composition problem to meet the availability requirement of the SFC while considering multiple VNF replications over each path in the slice topology, and then allocating sufficient bandwidth to meet the availability requirement. Unlike our previous work, this work focuses on how VNF replications on each path of the slice affects the bandwidth allocation to guarantee the availability requirement.

The rest of the paper is organized as follows. In Section II we present the related background literature. In Section III we present the problem statement and availability analysis in detail. In Section IV we discuss our algorithm for slice composition and provisioning of bandwidth and VNF resources for the slice. In Section V we discuss the effectiveness of our approach through simulations. We conclude the paper in Section VI.

II. Background Literature

Several recent works have discussed the SFC mapping problem, which includes VNF placement and flow allocation [5]–[8], while not addressing the SFC availability aspect. In [9], the authors study an availability-aware SFC mapping problem, but their work only considers VNF availability, while the physical network and computing components are considered to be always available. In [10] the authors consider the availability requirement of the SFC, and a joint path-VNF backup method is proposed where the backup path is shared between two working paths. In [11] the authors study a resilient network slice embedding problem in which the VNFs of an SFC are instantiated on appropriate PMs to minimize the number of affected SFCs during a physical machine failure. The authors in [12] propose the concept of parallelized SFC in which they split large data flows into multiple small sub-flows, and a hybrid placement algorithm is designed which aims to map SFC on datacenter networks while meeting the availability requirement of the SFC. The difference between our work and the above works is in the aspect of using network protection for each path when the availability cannot be met. In our case, although a slice topology can utilize multiple paths, the bandwidth resources that are allocated on the backup are just enough to meet the availability requirement.

III. System Model

A. Problem Statement

We are given a physical infrastructure denoted by a graph $G = (V, E)$, where $V = V_{net} \cup V_{com}$, V_{net} is a set of network nodes that are responsible for the flow access, and V_{com} is a set of compute nodes with computing capabilities. We denote the availability of the compute nodes by a_v, $\forall v \in V$, and assume that the network nodes have availability 1. The capacity of compute nodes is denoted by s_c, $\forall c \in V_{com}$. E is a set of edges, and each edge $e \in E$ has an associated availability a_e, bandwidth capacity b_e, and length d_e.

An SFC request is denoted as $\gamma_i = (s, d, M, b^{sd}, A_{req}, \mathbb{D}^{sd})$, in which s and d represent the source and destination of the SFC respectively. $M = \{m_1, m_2, m_3, .., m_n\}$ is set of VNFs that are required by the SFC, b^{sd} is the bandwidth requirement, A_{req} is the availability requirement, and $\mathbb{D}^{sd}$ is the distance threshold. We assume that the computing requirement of each function m of the SFC over the source-destination pair sd is a function of the bandwidth of the traffic flow, and is given by r_m in units of CPU cycles per unit time. We allow multiple replicas of each VNF to satisfy the availability requirement, and the number of replicas of function m is denoted by q_m. We also assume there can be at most q_{max} replicas of any given function on each path. An individual VNF has an associated availability a_m.

In our problem, we consider that the slice topology can consist of one or more paths selected from a set of candidate paths between the source s and the destination d. Usually, a multi-path slice topology is considered when it is not possible to meet the availability requirements with a single path. Let K denote the number of paths in the slice topology. Each path in the slice topology should include sufficient computing resources to accommodate the M required functions, and their replications (if need be), and the links on each path should include bandwidth flow according to the fraction of flow bandwidth b^{sd} traversing on the path. We define y_k^{sd}, $\{y_k^{sd} : \mathbb{R}^+ | 0 \leq y_k^{sd} \leq 1\}$, as the fraction of requested sd flow bandwidth b^{sd} that will be allocated on path k, and $b_k^{sd} = y_k^{sd} \cdot b^{sd}$ as the total capacity of bandwidth resources allocated on path k. We also accommodate the possibility of over-provisioning of the resources in order to meet the availability requirement. Thus, it is possible that the sum of b_k^{sd} over all K paths may exceed b^{sd}.

Therefore, given the physical infrastructure and the SFC request, we want to determine the slice topology for the SFC, which could possibly have multiple paths between s and d, and then determine the mapping of slice links and slice nodes on the physical infrastructure. Specifically, we need to map the VNFs (and their replications if needed) on the compute nodes, and provide proper dimensioning of bandwidth on the physical links used by the slice path such that the availability requirement is satisfied and cost is minimized.

B. Traffic-Aware Availability Analysis

In this work, we utilize the concept of traffic-weighted availability which gives the long-term fraction of traffic that is supported by the slice, given the bandwidth allocated on each path of the slice, and the physical infrastructure and the deployed VNFs are available [4]. The traffic-weighted availability metric has an advantage over the traditional availability in terms of reduced resource consumption. The traffic-weighted availability of sd traffic is defined as:

$$A_{sd} = \sum_{\mathbf{x}\in\chi} \min\Big(1, \sum_{k=1}^{K} y_k^{sd}\cdot x_k\Big)\pi_{\mathbf{x}}. \tag{1}$$

where, $\mathbf{x} = (x_1, x_2, \dots, x_K)$ is a vector of binary indicators, x_k, where $x_k = 1$ if path k is available, and $x_k = 0$ if path k is unavailable. We denote the set of all possible states as χ, and the fraction of time that the system is in state $\mathbf{x}$ is denoted as $\pi_{\mathbf{x}}$. Note that if the slice topology has more than one path, then the availability of paths may be correlated if they are mapped to the same physical components.

To analyze the values of $\pi_{\mathbf{x}}$ let us assume $Z = V \cup E$ to be the set of physical resources consisting of nodes and edges, and a_z indicate the availability of $z \in Z$. A mapping for path k is defined by set $U_k \subseteq Z$, which is the set of physical nodes and edges to which path k is mapped. Note that computing nodes along a path are only included if SFC functions are deployed and used by the slice at these nodes. The availability of a single path k is given by:

$$\pi_k = \prod_{u\in U_k} a_u \cdot \prod_{m\in M} (1-(1-a_m)^{q_m}). \tag{2}$$

In this case, we require at least one replica of each VNF to be functional on each slice path at a given time for it to remain functional. If a single path cannot provide the required availability, then the sd flow would need to be directed over two or more paths in the slice topology. To understand how to find the values of $\pi_{\mathbf{x}}$ for multiple paths, we give an example for the case of two paths, p_1 and p_2. For two paths, $\mathbf{x} = (x_1, x_2)$ denotes the availability state of paths p_1 and p_2 ($\mathbf{x} \in \{(0,0),(0,1),(1,0),(1,1)\}$). Let U_k denote the set of network resources used in path p_k. We define $T_{12} = U_1 \cap U_2$ as the set of resources used in both paths, and $T_k = U_k \setminus T_{12}$ as the set of resources only used by path p_k. We assume path p_1 has $q_m^{p_1}$ replications of the function m, and path p_2 has $q_m^{p_2}$ replications of the same function. Now, the fraction of time that both paths are available is given by:

$$\begin{aligned}\pi_{(1,1)} = &\prod_{u\in T_1\cup T_2\cup T_{12}} a_u \cdot \prod_{m\in M_1} 1-(1-a_m)^{q_m^{p1}} \cdot \\ &\prod_{m\in M_2} 1-(1-a_m)^{q_m^{p2}}.\end{aligned} \tag{3}$$

For this, we require all the physical resources used by path p_1 and p_2 to be functional, and at least one replica of each function to be working on each path. Next, we evaluate the cases when just one path out of the two paths is functional, $\pi_{(1,0)}$ and $\pi_{(0,1)}$, given by:

$$\begin{aligned}\pi_{(1,0)} = &\prod_{u\in T_1\cup T_{12}} a_u \cdot \prod_{m\in M_1} 1-(1-a_m)^{q_m^{p1}} \cdot \\ &\Big[\Big(1-\prod_{u\in T_2} a_u\Big) + \Big(1-\prod_{m\in M_2} 1-(1-a_m)^{q_m^{p2}}\Big) \\ &-\Big(1-\prod_{u\in T_2} a_u\Big)\cdot\Big(1-\prod_{m\in M_2} 1-(1-a_m)^{q_m^{p2}}\Big)\Big].\end{aligned} \tag{4}$$

$$\begin{aligned}\pi_{(0,1)} = &\prod_{u\in T_2\cup T_{12}} a_u \cdot \prod_{m\in M_2} 1-(1-a_m)^{q_m^{p2}} \cdot \\ &\Big[\Big(1-\prod_{u\in T_1} a_u\Big) + \Big(1-\prod_{m\in M_1} 1-(1-a_m)^{q_m^{p1}}\Big) \\ &-\Big(1-\prod_{u\in T_1} a_u\Big)\cdot\Big(1-\prod_{m\in M_1} 1-(1-a_m)^{q_m^{p1}}\Big)\Big].\end{aligned} \tag{5}$$

Both paths will be unavailable if either a common physical resource used by both path p_1 and p_2 is not available or when all the common physical resources are available, but at least one physical resource or function used by just p_1 and at least one physical resource or function used by just p_2 is unavailable.

$$\begin{aligned}\pi_{(0,0)} = &\Big(1-\prod_{u\in T_{12}} a_u\Big) + \Big(\prod_{u\in T_{12}} a_u\Big)\Big[\Big[\Big(1-\prod_{u\in T_1} a_u\Big)\cdot \\ &\Big(1-\prod_{u\in T_2} a_u\Big)\Big] + \Big[\Big(\prod_{u\in T_1} a_u\Big)\cdot\Big(\prod_{u\in T_2} a_u\Big)\cdot \\ &\Big(1-\prod_{m\in M_1} 1-(1-a_m)^{q_m^{p1}}\Big)\cdot \\ &\Big(1-\prod_{m\in M_2} 1-(1-a_m)^{q_m^{p2}}\Big)\Big] + \Big[\Big(1-\prod_{u\in T_1} a_u\Big)\cdot \\ &\Big(\prod_{u\in T_2} a_u\Big)\cdot\Big(1-\prod_{m\in M_2} 1-(1-a_m)^{q_m^{p2}}\Big)\Big]+ \\ &\Big[\Big(\prod_{u\in T_1} a_u\Big)\cdot\Big(1-\prod_{u\in T_2} a_u\Big)\cdot \\ &\Big(1-\prod_{m\in M_1} 1-(1-a_m)^{q_m^{p1}}\Big)\Big]\Big].\end{aligned} \tag{6}$$

With similar calculations, $\pi_{\mathbf{x}}$ for a larger number of paths can be determined.

IV. Slice Composition Problem with VNF Replication and Bandwidth Allocation

In this section, we explain our approach of slice composition by selecting the number of required paths for an sd pair, finding the number of replications for each VNF, and formulating an optimization problem for bandwidth allocation over the slice paths, while minimizing cost.

A. Required Number of Paths

The number of sd paths required in a slice will depend on the availability of underlying resources in the physical infrastructure and the number of VNF replicas on each path. We can determine if a single path is sufficient by checking if its overall availability considering both the path availability and the availability using the maximum number of replicas of all VNFs meets the availability requirement. However, the maximum-availability path may not necessarily be the minimum-cost path that satisfies the availability requirement.

If a single path is insufficient, then the sd flow would need to be directed over two or more paths in the slice topology. Suppose we consider two paths, p_1 and p_2. In this case, we need to map p_1 and p_2 over the physical infrastructure such that $y_1^{sd}(\pi_{(1,0)} + \pi_{(1,1)}) + y_2^{sd}(\pi_{(0,1)} + \pi_{(1,1)}) \geq A_{req}$, and $\pi_{(1,1)} + y_1^{sd}\pi_{(1,0)} + y_2^{sd}\pi_{(0,1)} \geq A_{req}$. Note that, for any pair of paths p_1 and p_2, if $\pi_{(0,0)} \leq 1 - A_{req}$, then the pair of paths is capable of satisfying the availability requirement.

B. Path Selection

For selecting the routing over for the set of K paths over the physical infrastructure, we utilize the following two algorithms.

- *K shortest paths algorithm*: Calculate K shortest paths (not necessarily disjoint) from s to d using Yen's algorithm [13].
- *K link-disjoint paths algorithm*: Calculate the maximum-availability path first. Edges of this path are then removed from the graph, and the next highest availability path is calculated on the new graph. Continue until K paths are found.

Many real-world applications which are latency sensitive require the distance between the source and destination of the SFC to be less than a given threshold. Therefore, while mapping the paths in the slice topology on the physical infrastructure we only select the paths whose length is less than $\mathbb{D}^{sd}$,

$$\sum_{e \in U_k} d_e \leq \mathbb{D}^{sd}. \tag{7}$$

C. VNF Replication & Bandwidth Allocation

Given the set of K paths, we need to solve for the number of VNF replicas and the bandwidth allocation on each path such that availability requirement is met and the cost is minimized. The overall availability of a path increases if more VNF replicas are placed on the path. In this case, to satisfy the same availability requirement, less bandwidth would need to be allocated on the physical path. Therefore, there is a possible trade-off between meeting the availability requirement with additional bandwidth versus additional VNFs replicas.

Rather than formulating a problem that jointly solves for the number of replicas and the amount of bandwidth, we formulate a linear program (LP) that solves for the optimal amount of bandwidth for a given number of replicas on each path. The idea is to iterate through different values of $q_m^{p_k}$, solving the LP for each case, and then selecting the solution with the minimum cost. For any given set of K paths and a given number of VNF replicas on each path, the LP determines the optimal bandwidth allocation over each path $(y_k^{sd} \cdot b^{sd})$ that results in the lowest cost while meeting the availability requirement. The description of parameters and variable used are shown in Table 1. The objective of the LP is to minimize the cost:

$$\min \mathcal{C}^{sd} = \sum_{k=1}^{K} \Big(\sum_{u \in U_k} y_k^{sd} \cdot b^{sd} \cdot C_u + \sum_{m \in M} q_m^{p_k} \cdot C_m \Big). \tag{8}$$

The problem is subject to the capacity, availability, and bandwidth constraints. We define the capacity constraint for each physical link where the amount of bandwidth allocated on each link should be less than the capacity of the link as,

$$\sum_{k=1}^{K} y_k^{sd} b^{sd} \cdot \beta_k^{\hat{e}} \leq b_{\hat{e}}, \; \forall \hat{e} \in U_k. \tag{9}$$

The availability constraint in which the traffic-weighted availability of the slice should be greater than equal to the availability requirement of the SFC is given by,

$$\sum_{\mathbf{x} \in \chi} \min \Big(1, \sum_{k=1}^{K} y_k^{sd} \cdot x_k \Big) \pi_{\mathbf{x}} \geq A_{req}. \tag{10}$$

Furthermore, the fraction of bandwidth on each path should be between 0 and 1. The case in which the fraction of bandwidth is 1 on each path is equal to the dedicated protection approach. This constraint is given by,

$$0 \leq y_k^{sd} \leq 1, \; \forall k \in \{1.....K\}. \tag{11}$$

Iterating through all possible combinations for the number of VNF replicas on K paths would require running the LP $q_{max}^{K \cdot M}$ times. In this paper, we make the simplifying assumption that $q_m^{p_k} = q_n^{p_k} \; \forall \; m, \, n \in M$ and $\forall \; k \in K$, reducing the number of combinations to q_{max}^{K}.

D. Mapping of VNFs on the Compute Nodes

As a part of the overall mapping procedure, we need to map the VNFs on the slice nodes, and then map the slice nodes on the compute nodes. We assume that each VNF instance is mapped to a different slice node, and that different slice nodes may be mapped to the same compute node. Note that if VNF replications are needed, they could be mapped to the same slice node as the original VNF.

For a slice path to be available, we need all physical links and compute nodes on the path to be available, and at least one replica of each function should be available. In this case, similar to [14], the placement of VNFs along the path does not affect the availability of the SFC. Thus we select compute nodes randomly on the path to map the required slice nodes, following the capacity constraint of the physical nodes,

$$\sum_{m \in M} \sum_{k=1}^{K} r_m \cdot q_m^{p_k} \cdot \alpha_k^{mi} \leq s_i, \; \forall c_i \in U_k, \tag{12}$$

where α_k^{mi} is 1 if function m (or slice node with function m) is mapped to compute node i on path k, 0 otherwise.

TABLE I
DESCRIPTION OF VARIABLES

Parameters	Description
C^{sd}	cost of establishment for all paths between s and d.
b^{sd}	bandwidth requirement between s and d.
C_u	cost of physical component u.
M	set of VNFs in the SFC.
$q_m^{p_k}$	number of replicas of function m on path k.
C_m	cost of VNF m.
K	total number of paths in the slice.
$\beta_k^{\hat{e}}$	1 if physical link $\hat{e}$ is on path k, 0 otherwise.
$b_{\hat{e}}$	bandwidth capacity of physical link $\hat{e}$.
U_k	physical components used in path k.
x_k	state, 1 if path k is available, 0 otherwise.
χ	set of all possible states.
$\pi_{\mathbf{x}}$	the fraction of time that the system is in state $\mathbf{x}$.
A_{req}	availability requirement of the slice.
Variable	**Description**
y_k^{sd}	the amount of bandwidth on path k between s and d.

V. Numerical Evaluation

A. Network Setting

We evaluate the performance of our approach over a 24-node USNET network topology. In each experiment, the availability of all physical edges in the topology is set to one of these two values: [0.99999, 0.999999]. The distance of the edges is in the range 350-1200 km, which is fixed for the network topology. We set the cost of allocating 1 Gb/s to each physical resource as 400 cost units, and we assume that the capacity of a physical link is 100 Gb/s. To host VNFs, 20 nodes are arbitrarily selected as compute nodes. The capacity of all compute nodes is kept constant at 100 units. The arrival times of the SFC requests follow a Poisson process with 2 requests/hour and an SFC holding time follows an exponential distribution with an average holding time of 1/2 hours. For each data point in our experiments, we generate a set of 10000 trials consisting of individual SFC requests between randomly selected source-destination pairs in the topology. For each SFC request, we randomly select its bandwidth requirement from the range 25-60 Gb/s, and randomly select the number of VNFs in the SFC request to be between 2 and 6. We consider that the VNFs require a number of computing resource units ranging from 1 to 3. We assume that the availability of each VNF is 0.99999. Our assumption that every VNF has the same availability holds even when the number of replications is increased on the slice paths. We run our experiments for six different levels of SFC availability requirement, ranging from 0.9 to 0.999999.

B. Experiments and Discussions

In the experiment shown in Fig. 2, we compare the cost of establishing the SFC requests for the two different path selection algorithms for increasing values of A_{req}. We construct a slice topology using K = 2, set q_{max} = 5, and map the SFC requests to a pair of paths using the path-selection algorithms mentioned in Section IV.B. We use a modified version of the graph where link distances are replaced by $-log(a_e)$, where a_e is the availability of edge e. For the construction of slice topology using the dedicated 1+1 approach, we use the K link-disjoint path selection algorithm to calculate two disjoint paths between s and d. In dedicated 1+1 protection, the full capacity of a single path is used without a second path as long as the SFC requirements can be satisfied by one path. Bandwidth is only allocated on the second path, utilizing the

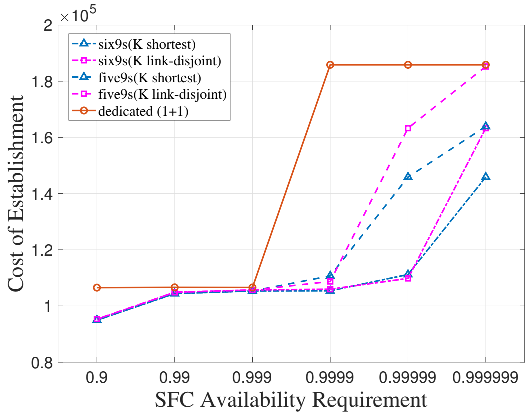

Fig. 2. Cost of establishment versus availability requirement.

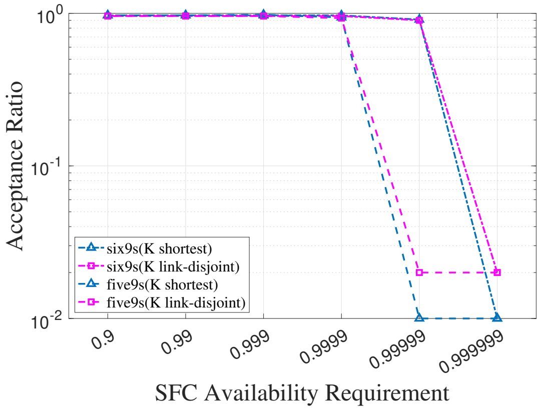

Fig. 3. Acceptance ratio versus availability requirement.

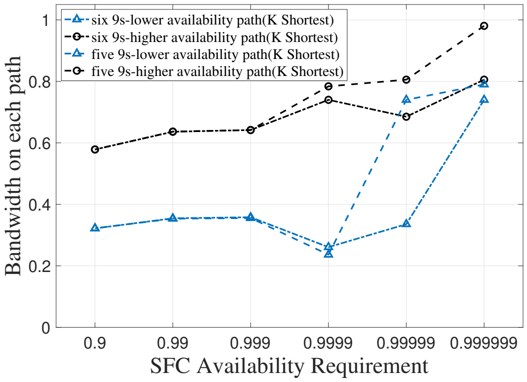

Fig. 4. Bandwidth ratio on both paths versus availability requirement for K-shortest path algorithm.

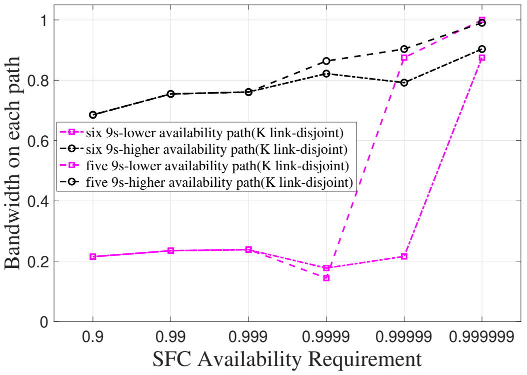

Fig. 5. Bandwidth ratio on both paths versus availability requirement for K link-disjoint path algorithm.

full capacity of the second path, if the availability requirement cannot be satisfied with just one path.

For both path selection algorithms, we observe a gradual increase in the cost with an increase in the value of A_{req}. We also observe that the cost of establishment is higher when the availability of the physical resources is lower since more VNF replicas and bandwidth is required to meet the same availability requirement. For dedicated 1+1 protection, the cost is constant throughout because the amount of bandwidth utilized by these paths remains constant. We observe a sudden increase in cost between A_{req} 0.999 to 0.9999, which is due to the increase in the number of paths from one to two in that interval.

For the next experiment, as shown in Fig. 3, we compare the acceptance ratio for both path selection algorithms. We define the acceptance ratio as the ratio between the number of SFC requests that were accepted in the network and the total number of SFC requests. We observe that with the increase in A_{req}, the acceptance ratio for both the path selection algorithms decreases. This is due to the fact that the requests cannot be accepted if the A_{req} is not satisfied even if we allocate maximum available bandwidth on those paths.

For the next set of experiments, as shown in Fig. 4 and Fig. 5, we compare the fraction of bandwidth that is allocated on both of the paths. For both path selection algorithms, we observe an increase in the bandwidth allocation on the paths as the availability requirement increases. For lower availability requirements, the algorithm will allot maximum bandwidth on the higher availability path. However, as A_{req} increases, the fraction of bandwidth on both paths for both path selection algorithms approaches 1. Also, for lower availability requirements (A_{req} = 0.9 and 0.999) we do not have to provision the total required bandwidth of the SFC as long as the availability requirement could be satisfied; thus, $\sum_{k=1}^{2} y_k^{sd} \leq 1$ for 0.9 and 0.99. This helps in saving some resources and only allocating enough bandwidth on both paths.

VI. Conclusion

We studied the problem of network slice composition for SFCs where the objective is to minimize the total cost of establishment for a network operator while satisfying the availability constraint of the SFC. For designing a slice topology, we consider that a slice could have multiple paths (which could be non-disjoint) in case the availability value is below the requirement. For each path in the slice topology, we calculate the proper dimensioning of resources on the computing nodes where multiple replicas of VNFs can be deployed and on physical links where bandwidth resources are allocated. The effectiveness of our proposed approach is demonstrated in terms of the cost of establishment from the perspective of a network operator.

Acknowledgment

This work was supported in part by the National Science Foundation under Grant No. CNS-2008856.

References

[1] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, “Network function virtualization: Challenges and opportunities for innovations,” *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, Feb 2015.

[2] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, “Network slicing to enable scalability and flexibility in 5G mobile networks,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[3] J. Fan, C. Guan, Y. Zhao, and C. Qiao, “Availability-aware mapping of service function chains,” in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.

[4] R. Gour, G. Ishigaki, J. Kong, and J. P. Jue, “Availability-guaranteed slice composition for service function chains in 5G transport networks,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 13, no. 3, pp. 14–24, 2021.

[5] M. Jalalitabar, Y. Wang, and X. Cao, “Branching-aware service function placement and routing in network function virtualization,” in *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2019, pp. 1–6.

[6] J. Li, W. Shi, P. Yang, and X. Shen, “On dynamic mapping and scheduling of service function chains in SDN/NFV-enabled networks,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[7] G. Sun, G. Zhu, D. Liao, H. Yu, X. Du, and M. Guizani, “Cost-efficient service function chain orchestration for low-latency applications in NFV networks,” *IEEE Systems Journal*, vol. 13, no. 4, pp. 3877–3888, 2019.

[8] A. Laghrissi, T. Taleb, M. Bagaa, and H. Flinck, “Towards edge slicing: VNF placement algorithms for a dynamic realistic edge cloud environment,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[9] J. Fan, M. Jiang, and C. Qiao, “Carrier-grade availability-aware mapping of service function chains with on-site backups,” in *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*, 2017, pp. 1–10.

[10] M. Wang, B. Cheng, and J. Chen, “Joint availability guarantee and resource optimization of virtual network function placement in data center networks,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 821–834, 2020.

[11] P. M. Mohan and M. Gurusamy, “Resilient VNF placement for service chain embedding in diversified 5G network slices,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[12] M. Wang, B. Cheng, S. Wang, and J. Chen, “Availability- and traffic-aware placement of parallelized SFC in data center networks,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 182–194, 2021.

[13] J. Y. Yen, “Finding the K shortest loopless paths in a network,” 2007.

[14] J. Kong, I. Kim, X. Wang, Q. Zhang, H. C. Cankaya, W. Xie, T. Ikeuchi, and J. P. Jue, “Guaranteed-availability network function virtualization with network protection and VNF replication,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.