IEICE TRANS. ??, VOL.Exx–??, NO.xx XXXX 200x

PAPER

Resource-Efficient and Availability-Aware Service Chaining and VNF Placement with VNF Diversity and Redundancy

Takanori HARA[†a)], **Masahiro SASABE**[††b)], ***Members***, **Kento SUGIHARA**[†c)], ***Nonmember***, ***and*** **Shoji KASAHARA**[†d)], ***Fellow***

SUMMARY To establish a network service in network functions virtualization (NFV) networks, the orchestrator addresses the challenge of service chaining and virtual network function placement (SC-VNFP) by mapping virtual network functions (VNFs) and virtual links onto physical nodes and links. Unlike traditional networks, network operators in NFV networks must contend with both hardware and software failures in order to ensure resilient network services, as NFV networks consist of physical nodes and software-based VNFs. To guarantee network service quality in NFV networks, the existing work has proposed an approach for the SC-VNFP problem that considers VNF diversity and redundancy. VNF diversity splits a single VNF into multiple lightweight replica instances that possess the same functionality as the original VNF, which are then executed in a distributed manner. VNF redundancy, on the other hand, deploys backup instances with standby mode on physical nodes to prepare for potential VNF failures. However, the existing approach does not adequately consider the tradeoff between resource efficiency and service availability in the context of VNF diversity and redundancy. In this paper, we formulate the SC-VNFP problem with VNF diversity and redundancy as a two-step integer linear program (ILP) that adjusts the balance between service availability and resource efficiency. Through numerical experiments, we demonstrate the fundamental characteristics of the proposed ILP, including the tradeoff between resource efficiency and service availability.

key words: *Network functions virtualization (NFV), virtual network function (VNF), service chaining, VNF placement, VNF diversity, and VNF redundancy.*

1. Introduction

Thanks to network functions virtualization (NFV), virtual network functions (VNFs) can be deployed with enhanced agility and flexibility, leading to reduced capital expenditure (CAPEX) and operating expenditure (OPEX). This enables the establishment of network services [1]. A certain network service can be considered as a sequence of VNFs. To establish a network service, which involves creating a service chain from a service chain request (SCR) with specific service chain requirements, the VNF placement (VNFP) problem and service chaining (SC) problem needs to be solved. These problems are resource allocation challenges in NFV networks. The VNFP problem aims to map the VNFs required by the SCR onto physical nodes while considering resource constraints [2]. On the other hand, the SC problem aims to establish an appropriate service path from an origin node to a destination node, while executing the VNFs deployed on the physical nodes in the required function order and adhering to resource constraints [3]. More precisely, the SC and VNFP (SC-VNFP) problem involves constructing a service chain forwarding graph, which consists of virtual nodes representing VNFs and virtual links connecting them, and then mapping these virtual nodes and links onto physical nodes and links. It is well-known that efficient resource allocation can be achieved by jointly solving the SC and VNFP problems [3]. Notably, these problems are known to be NP-hard [4].

Unlike traditional networks, network operators in NFV networks face not only hardware failures but also software failures, as NFV networks consist of both physical nodes and software-based VNFs. Consequently, their efforts have increased to address these failures. In this context, it is crucial to establish a resilient network service that can withstand hardware and software failures, ensuring a certain level of network service quality in terms of service availability and fault tolerance. There have been studies on exploring VNF diversity and redundancy to achieve resilient NFV networks [5]–[12]. VNF diversity splits a single VNF into N lightweight replica instances that have the same functionality as the original VNF, and then distributes and executes them [10]. However, this approach incurs load balancer costs and overhead to process the same amount of traffic as the original VNF. On the other hand, VNF redundancy deploys backup instances with standby mode on physical nodes to prepare for VNF failures [8]. While this approach improves service availability, it consumes additional resources for backup instances, which may reduce the resource efficiency of the NFV network. Therefore, there is a tradeoff between resource efficiency and service availability when considering VNF diversity and redundancy.

There have been studies on the SC-VNFP problem [3], [8], [10], [13], [14]. Many of these studies have formulated the SC-VNFP problem as an integer linear program (ILP) with the aim of achieving resource-efficient management of NFV networks, but without taking into consideration service availability in the face of VNF failures [3], [13], [14]. Research has also explored the resilient SC-VNFP problem in terms of VNF diversity and redundancy [8], [10]. Hmaity

Manuscript received January 1, 2015.
Manuscript revised January 1, 2015.

[†]T. Hara, K. Sugihara, and S. Kasahara are with the Division of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan.
[††]M. Sasabe is with the Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, Osaka 569-1095, Japan.
a) E-mail: hara@ieee.org
b) E-mail: m-sasabe@ieee.org
c) E-mail: sugihara.kento.sj6@is.naist.jp
d) E-mail: kasahara@ieee.org
DOI: 10.1587/trans.E0.??.1

et al. proposed three VNF redundancy strategies to mitigate single point failures by preparing primary and backup paths: (1) virtual node protection, (2) virtual link protection, and (3) service chain protection [8]. Alleg et al. integrated VNF diversity and redundancy, and formulated the SC-VNFP problem as a mixed ILP (MILP) by applying both VNF diversity and redundancy to each VNF, ensuring service availability [10]. However, this solution does not allow for adjustment of the diversity level of each VNF with consideration of both resource efficiency and service availability prior to solving the SC-VNFP problem.

In this paper, we present a novel scheme for the SC-VNFP problem that realizes resource efficiency and service availability through VNF diversity and redundancy, formulated as a two-step ILP inspired by [10]. Specifically, the proposed scheme first constructs a service chain forwarding graph from a set of candidate replica instances and virtual links, and then solves the first ILP to select the minimum required replica instances and virtual links from the candidate set for deployment on physical links and nodes. The objective is to satisfy only the availability and assignment constraints. In the next step, the constructed service chain graph is assigned to physical nodes and links, ensuring that the service chain requirements are met by solving the second ILP. Through numerical results, we demonstrate the fundamental characteristics of the proposed ILP, highlighting the tradeoff between resource efficiency and service availability.

The rest of the manuscript is organized as follows. Section 2 gives the related work. Section 3 introduces the system model used in this paper. In Section 4, we formulate the SC-VNFP problem incorporating VNF diversity and redundancy as an ILP. Section 5 shows the fundamental characteristics of the proposed ILP. Finally, Section 6 gives the conclusion.

2. Related Work

There have been prior studies on SC and VNFP problems [1], [2], [14], [15]. Alleg et al. formulated a delay-aware SC-VNFP problem as a mixed integer quadratically constrained program (MIQCP) [15]. Hyodo et al. formulated an SC-VNFP problem as an ILP and devised a heuristic algorithm employing a specialized network model known as the layered network [14]. Bhat et al. developed an algorithm leveraging the similarity between the SC problem and the shortest path tour problem (SPTP) to efficiently find the SPTP for achieving SC [16]. The SPTP is a variant of the shortest path problem that aims to establish the shortest path from an origin node to a destination node while traversing predefined disjoint node subsets $\mathcal{T}_1, \ldots, \mathcal{T}_k$ in the required order [17]. Sasabe and Hara modeled the SC-VNFP problem as the capacitated SPTP (CSPTP) and formulated it as an ILP utilizing an augmented network model, where CSPTP is a generalized version of SPTP that restricts node and link capacities with real values [3]. Furthermore, they proposed the CSPTP-based Lagrangian heuristic algorithm for efficient solving of the SC-VNFP problem [13].

There have been studies on service chain protection to ensure service resilience in terms of VNF redundancy [5], [7], [8] and VNF diversity [9]–[12]. In the context of VNF redundancy, Hmaity et al. proposed a resilient SC-VNFP scheme to address single point failures on physical nodes or links by preparing primary and backup service paths that adhere to service path delay constraints [8]. They categorized VNF redundancy into three types: (1) virtual node protection, (2) virtual link protection, and (3) service chain protection. Virtual node (resp. link) protection ensures that the backup path does not share physical nodes (resp. links) with the primary path. Service chain protection guarantees end-to-end service continuity by preparing the backup path such that it does not share both physical nodes and links, protecting against single point failures on physical nodes or links. Qu et al. formulated a reliability-aware and delay-constrained SC-VNFP problem with VNF redundancy as an MILP [5]. Their algorithm realizes hybrid routing, which combines primary and backup paths, to protect virtual links and service chains while adhering to delay constraints. Yang et al. formulated a delay-sensitive and availability-aware SC-VNFP problem as an integer non-linear program (INLP) and proposed a heuristic algorithm to address the computational complexity [7]. Their formulation takes into consideration virtual link and service chain protection. Carpio and Jukan formulated an availability-aware SC-VNFP problem with combined VNF migration and replication as an ILP [11]. Their solution is based on the virtual node protection and improves service availability by migrating VNFs to alternative physical nodes and replicating VNFs across multiple physical nodes. However, their problem assumes implicit VNF redundancy, where spare (replica) instances are prepared to process traffic in case of a single-point failure on the replica instance within the same group. In this paper, we adopt virtual node protection for VNF redundancy and consider not only VNF redundancy but also VNF diversity.

Xie et al. proposed the SC-VNFP problem that considers the heterogeneity between the original VNF and the backup instance, i.e, VNF diversity [9]. In addition to VNF diversity, Alleg et al. formulated an SC-VNFP problem with both VNF diversity and redundancy as an MILP by applying the VNF diversity and redundancy to each VNF in the service chain [10]. Kang et al. formulated a k fault-tolerance SC-VNFP problem with VNF diversity of both primary and backup functions as an MILP [12]. This solution also incorporates VNF diversity into VNF redundancy. However, these studies implicitly pre-select a diversity level for each VNF, making it challenging to appropriately select diversity levels to meet the service chain requirements. In this paper, we determine the minimum VNF diversity level required to meet the service availability requirements, and then achieve a highly-available and resource-efficient SC-VNFP that balances both resource efficiency and service availability.

3. System Model

In this section, we present the system model utilized in this paper, which is inspired by [10]. Table 1 provides the nota-

Table 1: Notations.

Symbol	Description
G_{P}	Physical network $G_{\mathrm{P}} = (\mathcal{V}_{\mathrm{P}}, \mathcal{E}_{\mathrm{P}})$
$\mathcal{V}_{\mathrm{P}}$	Set of physical nodes, $V_{\mathrm{P}} = \lvert\mathcal{V}_{\mathrm{P}}\rvert$
$\mathcal{E}_{\mathrm{P}}$	Set of physical links, $E_{\mathrm{P}} = \lvert\mathcal{E}_{\mathrm{P}}\rvert$
θ_n	Residual capacity of physical node $n \in \mathcal{V}_{\mathrm{P}}$
A_n^{HW}	Availability of physical node $n \in \mathcal{V}_{\mathrm{P}}$
$\delta_{n,m}$	Residual capacity of physical link $(n, m) \in \mathcal{E}_{\mathrm{P}}$
$\mathcal{S}$	Set of SCRs, $S = \lvert\mathcal{S}\rvert$
$\mathcal{F}$	Set of VNFs, $F = \lvert\mathcal{F}\rvert$
A_f^{SW}	Availability of VNF f
$A_{n,f}$	Availability of VNF $f \in \mathcal{F}$ running on the physical node $n \in \mathcal{V}_{\mathrm{P}}$, $A_{n,f} = A_f^{\mathrm{SW}} \cdot A_n^{\mathrm{HW}}$
r_s	Service chain requirements of SCR s
$\mathcal{F}_s$	Set of VNFs required by SCR s
o_s	Origin node of SCR s
d_s	Destination node of SCR s
G_{SC}^s	Service chain forwarding graph, $G_{\mathrm{SC}}^s = (\mathcal{V}_{\mathrm{SC}}^s, \mathcal{E}_{\mathrm{SC}}^s)$
$\mathcal{V}_{\mathrm{SC}}^s$	Set of virtual nodes for SCR s, $\mathcal{V}_{\mathrm{SC}}^s = \{o_s, f_1, \ldots, f_{K_s}, d_s\}$
$\mathcal{E}_{\mathrm{SC}}^s$	Set of virtual links for SCR s, $\mathcal{E}_{\mathrm{SC}}^s = \{(o_s, f_1), (f_1, f_2), \ldots, (f_{K_s-1}, f_{K_s}), (f_{K_s}, d_s)\}$
$A^{\mathrm{th}}(s)$	Availability requirement for SCR s
A_s	Availability of service chain of SCR s
Δ^f	Diversity level of VNF $f \in \mathcal{F}_s$
N	Maximum of diversity level among all VNFs $\forall f \in \mathcal{F}_s$, $N = \max_{f \in \mathcal{F}_s} \Delta^f$
P	Redundancy level
α	Ratio of processing capacity of the backup instance to that of the original VNF
Ψ^f	Processing requirement of VNF $f \in \mathcal{F}_s$
$\Omega^{k,l}$	Bandwidth requirement of virtual link $(k, l) \in \mathcal{E}_{\mathrm{SC}}^s$
G_{DIV}^s	Service chain forwarding graph with VNF diversity, $G_{\mathrm{DIV}}^s = (\mathcal{V}_{\mathrm{DIV}}^s, \mathcal{E}_{\mathrm{DIV}}^s)$
$\mathcal{V}_{\mathrm{DIV}}^s$	Set of virtual nodes in G_{DIV}^s for SCR s, $\mathcal{V}_{\mathrm{DIV}}^s = \{D_i^f\}_{f \in \mathcal{V}_{\mathrm{SC}}^s, i \in \{1,\ldots,\Delta^f\}} \cup \{o_s, d_s\}$
$\mathcal{E}_{\mathrm{DIV}}^s$	Set of virtual links in G_{DIV}^s for SCR s, $\mathcal{E}_{\mathrm{DIV}}^s = \{(D_i^k, D_j^l)\}_{(k,l) \in \mathcal{E}_{\mathrm{SC}}^s, i \in \{1,\ldots,\Delta^k\}, j \in \{1,\ldots,\Delta^l\}}$
D_i^f	The ith replica instance of VNF f
R_i^f	The ith backup instance of VNF f
$\Psi^{D_i^f}$	Processing requirement of ith replica instance of VNF f
$\Psi^{R_i^f}$	Processing requirement of ith backup instance of VNF f
Ψ^{LB}	Processing requirement of the load balancer
$H(f, N)$	Estimated overhead by the VNF diversity

tions employed in this paper. Figure 1 illustrates an overview of the system model.

3.1 Physical Network

A physical network is defined as a directed graph $G_{\mathrm{P}} = (\mathcal{V}_{\mathrm{P}}, \mathcal{E}_{\mathrm{P}})$, where $\mathcal{V}_{\mathrm{P}}$ (resp. $\mathcal{E}_{\mathrm{P}}$) denotes a set of physical nodes (resp. links). Each physical node $n \in \mathcal{V}_{\mathrm{P}}$ has the residual capacity θ_n and the availability $A_n^{\mathrm{HW}} \in [0, 1)$. Each physical link $(n, m) \in \mathcal{E}_{\mathrm{P}}$ has the residual capacity $\delta_{n,m}$. All physical nodes $n \in \mathcal{V}_{\mathrm{P}}$ on the NFV network can support $F = \lvert\mathcal{F}\rvert$ types of VNFs, and each VNF $f \in \mathcal{F}$ has the availability $A_f^{\mathrm{SW}} \in [0, 1)$, where $\mathcal{F}$ is a set of VNFs supported by the NFV network. The bottom layer of Fig. 1 presents an example of a physical network that supports the VNFs f_1, f_2, and f_3.

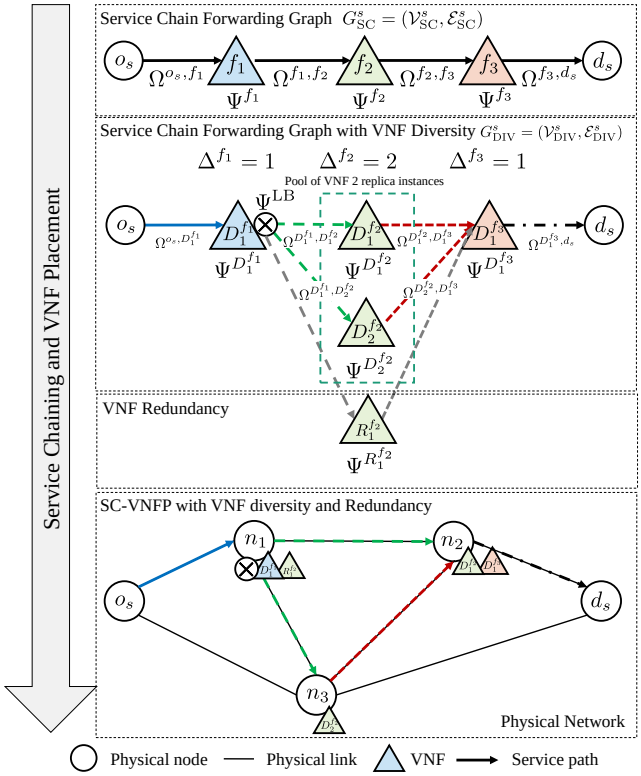


Fig. 1: Overview of the system model.

3.2 VNF Diversity and Redundancy

In this section, we introduce the concept of VNF diversity and redundancy. VNF diversity allows for the division of a single VNF into N lightweight replica instances ($\{D_i^f\}_{i \in \{1,\ldots,N\}}$) that possess the same functionality as the original VNF and can be executed in a distributed manner. As a result, these replica instances can concurrently process at least the same amount of traffic as the original VNF. Note that the load balancer is newly required redirect incoming traffic to a pool of N replica instances. The processing requirement $\psi^{\mathrm{diversity}}(f)$ of VNF f employing VNF diversity is defined as follows [10]:

$$\psi^{\mathrm{diversity}}(f) = \Psi^f + \Psi^{\mathrm{LB}} + H(f, N), \tag{1}$$

where Ψ^f denotes the processing requirement originally required by VNF f, and Ψ^{LB} denotes the processing capacity of the load balancer. $H(f, N)$ represents the estimated overhead required to maintain the processing efficiency of N replica instances equivalent to that of the original VNF, and it is defined as follows [10]:

$$H(f, N) = \Psi^f \cdot \frac{N - 1}{2(N + 1)}. \tag{2}$$

In Eq. (2), if VNF diversity is not applied (i.e, $N = 1$), $H(f, N) = 0$, as no overhead of load balancer occurs. When the number N of replica instances increases, we consider

half of the processing requirement required by VNF f as the overhead, $H(f, N)$.

VNF redundancy involves deploying additional instances, known as backup instances, in standby mode on physical nodes to mitigate failures and ensure network service continuity. However, this comes at the cost of increased resource consumption on physical nodes and links. In this paper, we assume that P backup instances of a target VNF f are deployed in the NFV network. When VNF redundancy is applied to VNF f with processing requirement Ψ^f, each of the P backup instances, R_i^f $(1 \le i \le P)$, requires processing requirement $\alpha\Psi^f$, where $\alpha \in [0, 1]$ is the ratio of processing requirement of the backup instance to that of the target VNF. The processing requirement $\psi^{\text{redundancy}}(f, \alpha)$ required by the P backup instances is as follows [10]:

$$\psi^{\text{redundancy}}(f, \alpha) = \alpha\Psi^f \cdot P.$$

In case of traditional redundancy (i.e., $\alpha = 1$), each backup instance requires the same amount of processing requirement as the target VNF.

3.3 Service Chain Request

Each service chain request (SCR) $s \in \mathcal{S}$ is characterized by the following service chain requirements:

$$r_s = (o_s, d_s, \mathcal{F}_s, G^s_{\text{SC}}, \{\Delta^f\}_{f\in\mathcal{V}^s_{\text{SC}}}, P, \alpha, \{\Psi^f\}_{f\in\mathcal{V}^s_{\text{SC}}}, \{\Omega^{k,l}\}_{(k,l)\in\mathcal{E}^s_{\text{SC}}}, \Psi^{\text{LB}}, A^{\text{th}}(s)).$$

o_s and d_s represent the origin and destination nodes, respectively. Let $\mathcal{F}_s \subseteq \mathcal{F}$ denote the sequential set of VNFs required by SCR s (i.e., $\mathcal{F}_s = \{f_{s,1}, \ldots, f_{s,K_s}\}$ and $K_s = |\mathcal{F}_s|$). $G^s_{\text{SC}} = (\mathcal{V}^s_{\text{SC}}, \mathcal{E}^s_{\text{SC}})$ is a directed acyclic graph that represents the service chain forwarding graph for SCR s, where $\mathcal{V}^s_{\text{SC}} = \{o_s, f_{s,1}, \ldots, f_{s,K_s}, d_s\}$ is a sequential set of virtual nodes indicating the K_s VNFs and origin and destination nodes, and $\mathcal{E}^s_{\text{SC}} = \{(f_k, f_{k+1})\}_{k=0,\ldots,K_s}$ is a set of virtual links connecting f_k with f_{k+1}, where $f_0 = o_s$ and $f_{K_s+1} = d_s$. Each virtual link $(k, l) \in \mathcal{E}^s_{\text{SC}}$ is assigned to at least one physical link. Ψ^f is the processing requirement required by VNF $f \in \mathcal{V}^s_{\text{SC}}$ and $\Omega^{k,l}$ is the bandwidth requirement of virtual link $(k, l) \in \mathcal{E}^s_{\text{SC}}$ connecting VNF k and VNF l. The top layer of Fig. 1 illustrates an example of the service chain forwarding graph for SCR s.

For VNF diversity, we calculate the number of replica instances per VNF $f \in \mathcal{F}_s$ as Δ^f, and define N as the upper limit of Δ^f $(1 \le \Delta^f \le N)$. Ψ^{LB} represents the processing requirement for the load balancer. For VNF redundancy, we provision P backup instances for each VNF $f \in \mathcal{F}_s$. $A^{\text{th}}(s)$ denotes the minimum availability requirement for service chain $s \in \mathcal{S}$.

We can construct the service chain forwarding graph $G^s_{\text{DIV}} = (\mathcal{V}^s_{\text{DIV}}, \mathcal{E}^s_{\text{DIV}})$ with VNF diversity, by replacing $\mathcal{V}^s_{\text{SC}}$ and $\mathcal{E}^s_{\text{SC}}$ with $\mathcal{V}^s_{\text{DIV}} = \{D_i^{f_k}\}_{f_k\in\mathcal{V}^s_{\text{SC}}, i\in\{1,\ldots,\Delta^{f_k}\}}$ and $\mathcal{E}^s_{\text{DIV}} = \{(D_i^{f_k}, D_j^{f_{k+1}})\}_{(f_k,f_{k+1})\in\mathcal{E}^s_{\text{SC}}, i\in\{1,\ldots,\Delta^{f_k}\}, j\in\{1,\ldots,\Delta^{f_{k+1}}\}}$, where Δ^f $(f \in \mathcal{V}^s_{\text{SC}})$ denotes the diversity level of VNF f required by SCR r_s. Note that $\Delta^{f_0} = 1$ and $\Delta^{f_{K_s+1}} = 1$ where $f_0 = o_s$ and $f_{K_s+1} = d_s$. $\mathcal{V}^s_{\text{DIV}}$ is a set of replica instances, the origin node, and the destination node. $\mathcal{E}^s_{\text{DIV}}$ is a set of virtual links connecting $D_i^{f_k}$ and $D_j^{f_{k+1}}$.

The middle layer of Fig. 1 illustrates an example of a service chain forwarding graph with VNF diversity and redundancy. In this figure, VNF f_2 is applied with a diversity level of $\Delta^{f_2} = 2$, resulting in 2 replica instances of VNF f_2, each connected to $D_1^{f_1}$ and $D_1^{f_3}$ through virtual links. The load balancer redirects incoming traffic to the replica instances $D_1^{f_2}$ and $D_2^{f_2}$ of VNF f_2, respectively. Additionally, a backup instance of VNF f_2 is prepared with $P = 1$ redundancy in this figure. In case all replica instances of VNF f_2 fail, the backup instance is activated and incoming traffic is redirected to it by the load balancer.

3.4 Availability

The system availability, A, where $A \in [0, 1)$, is defined as the ratio of time during which the system is in an operational state, i.e., $A = \text{MTBF}/(\text{MTBF} + \text{MTTR})$, where MTBF (mean time between failures) and MTTR (mean time to repair) represent the average durations between occurrences of failures and the average time taken to repair the system, respectively.

The service availability can be evaluated by decomposing the service chain into its individual components, including both the physical node and VNF. In the NFV network, the availability $A_{n,f}$ of VNF $f \in \mathcal{F}_s$ running on physical node $n \in \mathcal{V}_{\text{P}}$ can be defined as the product of the availability A_n^{HW} of physical node n and the availability A_f^{SW} of VNF f, assuming software and hardware failures occur independently.

$$A_{n,f} = A_n^{\text{HW}} \cdot A_f^{\text{SW}}, \tag{3}$$

where A_n^{HW} (resp. A_f^{SW}) represents the availability of the hardware (resp. software). Given that the service chain can be interpreted as a sequence of VNFs, the service availability of SCR s can be defined as follows based on Eq. (3):

$$A_s = \prod_{n\in\mathcal{V}_{\text{P}}} \prod_{f\in\mathcal{F}_s} A_{n,f}^{\mathbb{I}(n\in\mathcal{V}_f^{\text{alloc}})}, \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function and $\mathcal{V}_f^{\text{alloc}}$ represents the set of physical nodes on which VNF f is deployed.

As in [10], we regard the availability of VNF f with VNF diversity as the probability of having at least one replica instance in an operational state among Δ^f replica instances.

$$A_f^{\text{parallel}} = 1 - \prod_{k=1}^{\Delta^f} \prod_{n\in\mathcal{V}_{\text{P}}} (1 - A_n^{\text{HW}} \cdot A_{D_k^f}^{\text{SW}})^{\mathbb{I}(n\in\mathcal{V}_f^{\text{alloc}})}. \tag{5}$$

The service availability A_s^{parallel} of SCR s with VNF diversity can be obtained from Eqs. (4) and (5) as follows:

$$A_s^{\text{parallel}} = \prod_{f \in \mathcal{F}_s} A_f^{\text{parallel}}. \quad (6)$$

Please note that this approach aims to sustain the network service even when up to $N-1$ failures of replica instances per VNF occur, albeit at the expense of performance degradation that scales with the number of replica instance failures.

3.5 Existing Strategies for Deploying Replica Instances

In [10], the authors proposed three strategies to adjust Δ^f of each VNF $f \in \mathcal{F}_s$. All diversity strategy (ALLDIV) sets Δ^f of all VNFs $f \in \mathcal{F}_s$ to N. Random diversity strategy (RANDDIV) randomly selects Δ^f of each VNF $f \in \mathcal{F}_s$ from the range of $[1, N]$. Selective diversity strategy (SELEDIV) applies VNF diversity only to VNF f' with the lowest availability A'_f among all VNFs $f \in \mathcal{F}_s$, i.e., $f' = \arg\min_{f \in \mathcal{F}_s} A_f$, and sets $\Delta^{f'}$ to N. In [10], they construct G^s_{SC} before formulating the ILP according to one of the three strategies. The bottom layer of Fig. 1 depicts an example of the assignment result for SC-VNFP given the service chain forwarding graph in the middle layer of Fig. 1.

3.6 Existing Strategies for Deploying Backup Instances

The strategies for VNF redundancy can be categorized into three types: (1) virtual node protection, (2) virtual link protection, and (3) service chain protection, which ensure the availability of the entire service path, including both primary and backup paths [8]. Virtual node protection deploys additional VNFs on physical nodes that are not included in the primary path. Virtual link protection maps virtual links onto one or more physical links that are not included in the primary path. Service chain protection constructs a backup path that does not share any physical nodes or links with the primary path. In this paper, we adopt virtual node protection as VNF redundancy strategy. The middle layer of Fig. 1 illustrates an example of a backup instance $R_1^{f_2}$, which is deployed on physical node n_1 at a different location from the replica nodes $D_i^{f_2}$ $(i = 1, 2)$ in the primary path.

4. Proposed Scheme

4.1 Overview

The service chain forwarding graph G^s_{DIV} of SCR s can be constructed by pre-determining the values of Δ^f and $\Psi^{D_i^f}$ for each VNF $f \in \mathcal{F}_s$. However, to achieve an SC-VNFP solution that balances resource efficiency and service availability, the appropriate number and processing capacity of replica instances must be determined based on availability and processing requirements. For example, deploying more than the minimum required number of replica instances on the physical network can result in degraded resource efficiency due to the overhead from VNF diversity. On the other hand, deploying a small number of replica instances may not meet the availability requirement. It is a challenging problem to determine the appropriate number and processing capacity of replica instances prior to mapping the service chain forwarding graph onto physical nodes and links.

Assuming that the processing capacity of VNF f is evenly distributed among the Δ^f replica instances, we can derive the appropriate processing capacity $\Psi^{D_i^f}$ of replica instance D_i^f from Eq. (1):

$$\Psi^{D_i^f} = \frac{\psi^{\text{diversity}} - \Psi^{\text{LB}}}{\Delta^f}.$$

From this equation, we can see that the value of $\Psi^{D_i^f}$ depends on the number Δ^f of replica instances that are deployed on the physical nodes. Therefore, we can confirm that it is not possible to simultaneously determine both the appropriate number and processing capacity of replica instances, as well as conducting their mapping, in the context of linear programming.

To address this issue, we propose a two-step ILP for the SC-VNFP scheme that takes into account VNF diversity and redundancy to achieve a balance between resource efficiency and service availability. Figure 2 illustrates an overview of the proposed ILP. The first ILP, which will be described in Section 4.3, generates a service chain forwarding graph by selecting an appropriate number Δ^f of replica instances from a candidate set of N replica instances for each VNF f to satisfy the availability and assignment requirements. Subsequently, it maps the service chain forwarding graph onto the physical nodes and links. We initially develop the service chain forwarding graph $G^s_{\text{DIV}} = (\mathcal{V}^s_{\text{DIV}}, \mathcal{E}^s_{\text{DIV}})$ to represent the candidate set of N replica instances by replacing $\mathcal{V}^s_{\text{DIV}}$ with $\{D_i^{f_k}\}_{f_k \in \mathcal{V}^s_{\text{SC}}, i \in \{1,\ldots,N\}}$ and $\mathcal{E}^s_{\text{DIV}}$ with $\{(D_i^{f_k}, D_j^{f_{k+1}})\}_{(f_k, f_{k+1}) \in \mathcal{E}^s_{\text{SC}}, i \in \{1,\ldots,N\}, j \in \{1,\ldots,N\}}$ (the upper left side of Fig. 2). Let N denote the maximum number of replica instances among all VNF $f \in \mathcal{V}^s_{\text{SC}}$ (i.e., $1 \leq \Delta^{f*} \leq N$). After solving the first ILP, we obtain the service chain forwarding graph (the upper center of Fig. 2) and the temporal assignment result (the lower center of Fig. 2) that maps the minimum required number Δ^{f*} of replica instances, chosen from the candidate set, to physical nodes and the virtual links connecting them to physical links, which only hold the availability and assignment requirements.

Using the constructed service chain forwarding graph and temporal assignment result generated in the first ILP as input, the second ILP, which will be described in Section 4.4, reassigns virtual nodes and links to the physical nodes and links, ensuring that the service chain requirements r_s i.e., availability, assignment, processing, and capacitated constraints are satisfied, as shown in the right of Fig. 2. As an example, in Fig. 2, after solving the second ILP, $D_1^{f_3}$ will be reassigned to n_3 from n_2 if n_2 cannot meet the residual capacity constraint for $D_1^{f_3}$.

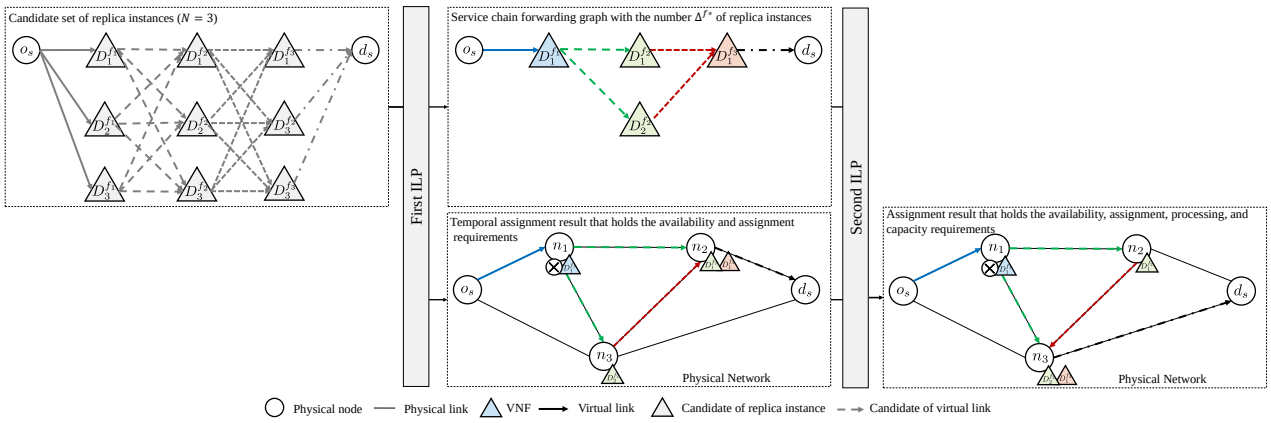


Fig. 2: Overview of the proposed two-step ILP.

4.2 Two-Step ILP Formulation

In this section, we propose a two-step ILP for the SC-VNFP problem that incorporates VNF diversity and redundancy. Let $[M_n^{D_i^f}]$ $(\forall n \in \mathcal{V}_\mathrm{P}, \forall f \in \mathcal{V}_\mathrm{SC}^s, \forall i \in \{1,\ldots,N\})$, $[M_n^{R_i^f}]$ $(\forall n \in \mathcal{V}_\mathrm{P}, \forall f \in \mathcal{V}_\mathrm{SC}^s, \forall i \in \{1,\ldots,P\})$, $[M_n^{\mathrm{LB}(f)}]$ $(\forall n \in \mathcal{V}_\mathrm{P}, \forall f \in \mathcal{V}_\mathrm{SC}^s)$, $[M_{n,m}^{D_i^k,D_j^l}]$ $(\forall (n,m) \in \mathcal{E}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}_\mathrm{DIV}^s)$, and $[a^{D_i^k,D_j^l}]$ $(\forall (D_i^k, D_j^l) \in \mathcal{E}_\mathrm{DIV}^s)$ denote the binary decision variables as

$$M_n^{D_i^f} = \begin{cases} 1 & \text{if the } i\text{th replica instance } D_i^f \text{ is deployed} \\ & \text{on the physical node } n, \\ 0 & \text{otherwise,} \end{cases}$$

$$M_n^{R_i^f} = \begin{cases} 1 & \text{if the } i\text{th backup instance } R_i^f \text{ is deployed} \\ & \text{on the physical node } n, \\ 0 & \text{otherwise,} \end{cases}$$

$$M_n^{\mathrm{LB}(f)} = \begin{cases} 1 & \text{if the load balancer for VNF } f \text{ is} \\ & \text{deployed on the physical node } n, \\ 0 & \text{otherwise,} \end{cases}$$

$$M_{n,m}^{D_i^k,D_j^l} = \begin{cases} 1 & \text{if the virtual link } (D_i^k, D_j^l) \text{ is (partly)} \\ & \text{established by the physical link } (n,m), \\ 0 & \text{otherwise,} \end{cases}$$

$$a^{D_i^k,D_j^l} = \begin{cases} 1 & \text{if the virtual link } (D_i^k, D_j^l) \text{ is used,} \\ 0 & \text{otherwise,} \end{cases}$$

respectively.

4.3 First Step: Construction and Assignment of Service Chain Forwarding Graph to Meet Service Availability Requirement

Given the candidate set of the service chain forwarding graph, we formulate the ILP to construct the service chain forwarding graph G_DIV^s and map it onto the physical network such that the availability and assignment requirements hold as follows:

$$\min \quad \sum_{n\in\mathcal{V}_\mathrm{P}} \sum_{f\in\mathcal{V}_\mathrm{SC}^s} \sum_{i=1}^{N} M_n^{D_i^f}, \tag{7}$$

$$\text{s.t.} \quad M_n^{D_i^f} \in \{0,1\}, \quad \forall n \in \mathcal{V}_\mathrm{P}, \forall f \in \mathcal{V}_\mathrm{SC}^s, \forall i \in \{1,\ldots,N\}, \tag{8}$$

$$M_n^{R_i^f} \in \{0,1\}, \quad \forall n \in \mathcal{V}_\mathrm{P}, \forall f \in \mathcal{V}_\mathrm{SC}^s, \forall i \in \{1,\ldots,P\}, \tag{9}$$

$$M_n^{LB(f)} \in \{0,1\}, \quad \forall n \in \mathcal{V}_\mathrm{P}, \forall f \in \mathcal{V}_\mathrm{SC}^s, \tag{10}$$

$$M_{(n,m)}^{D_i^k,D_j^l} \in \{0,1\}, \quad \forall (n,m) \in \mathcal{E}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}_\mathrm{SC}^s, \tag{11}$$

$$a^{D_i^k,D_j^l} \in \{0,1\}, \quad \forall (D_i^k, D_j^l) \in \mathcal{E}_\mathrm{SC}^s, \tag{12}$$

$$1 \le \sum_{i=1}^{N} \sum_{n\in\mathcal{V}_\mathrm{P}} M_n^{D_i^f} \le N, \quad \forall f \in \mathcal{F}_s, \tag{13}$$

$$\sum_{m\in\mathcal{V}_\mathrm{P}} M_{n,m}^{D_i^k,D_j^l} - \sum_{m\in\mathcal{V}_\mathrm{P}} M_{m,n}^{D_i^k,D_j^l} = a^{D_i^k,D_j^l}(M_n^{D_i^k} - M_n^{D_j^l}), \quad \forall (n,m) \in \mathcal{E}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}_\mathrm{DIV}^s, \tag{14}$$

$$\sum_{i=1}^{N} M_n^{D_i^f} + \sum_{i=1}^{P} M_n^{R_i^f} \le 1, \quad \forall f \in \mathcal{F}_s, \forall n \in \mathcal{V}_\mathrm{P}, \tag{15}$$

$$\sum_{n\in\mathcal{V}_\mathrm{P}} M_n^{D_i^f} \le 1, \quad \forall f \in \mathcal{F}_s, \forall i \in \{1,\ldots,N\}, \tag{16}$$

$$\sum_{i=1}^{N} M_n^{D_i^k}(1-\varepsilon_k)\varepsilon_l = M_n^{LB(l)}, \quad \forall n \in \mathcal{V}_\mathrm{P}, \forall (k,l) \in \mathcal{E}_\mathrm{SC}^s, \tag{17}$$

$$\prod_{f \in \mathcal{V}_{\mathrm{SC}}^s} \left(1 - \prod_{i=1}^{N} \prod_{n \in \mathcal{V}_{\mathrm{P}}} \left(1 - A_f^{\mathrm{SW}} A_n^{\mathrm{HW}}\right)^{M_n^{D_i^f}}\right) \geq A^{\mathrm{Th}}(s), \tag{18}$$

$$M_{n,m}^{D_i^k, D_j^l} \leq a^{D_i^k, D_j^l}, \quad \forall (n,m) \in \mathcal{E}_{\mathrm{P}}, \forall (D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^s, \tag{19}$$

$$\sum_{j=1}^{N} a^{D_i^k, D_j^l} \leq \Delta^l, \quad \forall (k,l) \in \mathcal{E}_{\mathrm{SC}}^s, j \in \{1, \ldots, N\}, \tag{20}$$

$$a^{D_i^k, D_j^l} \leq \sum_{n \in \mathcal{V}_{\mathrm{P}}} M_n^{D_j^l}, \forall (D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^s, \tag{21}$$

$$a^{D_i^k, D_j^l} \leq \sum_{n \in \mathcal{V}_{\mathrm{P}}} M_n^{D_i^k}, \forall (D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^s, \tag{22}$$

$$\sum_{n \in \mathcal{V}_{\mathrm{P}}} M_n^{D_i^k} + \sum_{n \in \mathcal{V}_{\mathrm{P}}} M_n^{D_j^l} - 1 \leq a^{D_i^k, D_j^l}, \quad \forall (D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^s, \tag{23}$$

$$M_{o_s}^{D_i^{f_0}} = \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{otherwise,} \end{cases} \quad \forall i \in \{1, \ldots, N\}, \tag{24}$$

$$M_{d_s}^{D_i^{f_{K_s+1}}} = \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{otherwise,} \end{cases} \quad \forall i \in \{1, \ldots, N\}. \tag{25}$$

The objective function (7) minimizes the number of replica instances deployed on physical nodes. Constraints (8)–(12) define the domain of binary decision variables. Constraint (13) is associated with the maximum number of replica instances for each VNF f. Constraint (14) enforces the standard flow conservation rules. In case of $a^{D_i^k, D_j^l} = 1$, virtual link (D_i^k, D_j^l) is mapped to one or more physical links. As constraint (14) is nonlinear, we use its linear equivalent, which will be discussed in Appendix A. Constraint (15) indicates that each physical node n can support either a replica or backup instance of VNF f. Constraint (16) limits the number of the ith replica instances of VNF f deployed on the physical nodes to at most one. Constraint (17) represents the assignment strategy of the load balancer. If VNF diversity is applied to the VNF l, $\varepsilon_l = 1$, otherwise $\varepsilon_l = 0$.

Constraint (18) specifies the minimum availability requirement imposed by the SCR s. Due to its nonlinear nature, we linearize constraint (18) following the approach in [10]. As the service chain forwarding graph of SCR s comprises K_s VNFs, the service availability becomes $\prod_{f \in \mathcal{F}_s} A_f$ from Eq. (6). The availability requirement $A^{\mathrm{th}}(s)$ should meet $\prod_{f \in \mathcal{F}_s} A_f \geq A^{\mathrm{th}}(s)$ under the assumption of independent failures of each VNF. Assuming that the availability requirement A_f for each VNF f in the service chain is homogeneous, we can confirm that $A_f \geq A^{\mathrm{th}}(s)^{1/K_s}$ should be established for each VNF f. Therefore, constraint (18) can be transformed into the following constraint:

$$1 - \prod_{i=1}^{\Delta^f} \prod_{n \in \mathcal{V}_{\mathrm{P}}} \left(1 - A_n^{\mathrm{HW}} \cdot A_{D_i^f}^{\mathrm{SW}}\right)^{M_n^{D_i^f}} \geq A^{\mathrm{th}}(s)^{1/K_s}, \quad \forall f \in \mathcal{F}_s. \tag{26}$$

Since constraint (26) is nonlinear, we linearize it by taking the logarithm of both sides.

$$\sum_{i=1}^{\Delta^f} \sum_{n \in \mathcal{V}_{\mathrm{P}}} M_n^{D_i^f} \log\left(1 - A_n^{\mathrm{HW}} \cdot A_{D_i^f}^{\mathrm{SW}}\right) \leq \log\left(1 - A^{\mathrm{th}}(s)^{1/K_s}\right), \quad \forall f \in \mathcal{F}_s. \tag{27}$$

Constraint (27) ensures that the availability of VNF f is greater than or equal to $A^{\mathrm{th}}(s)^{1/K_s}$ when its replica instances D_i^f ($i \in \{1, \ldots, N\}$) are deployed on physical node $n \in \mathcal{V}_{\mathrm{P}}$. In this paper, we adopt constraint (27) instead of constraint (18).

Constraint (19) prohibits the assignment of virtual link (D_i^k, D_j^l) with no need for mapping to physical link (n, m). Constraint (20) limits the total number of virtual links between replica instances of VNF k and VNF l to Δ^l. Constraints (21)–(23) ensure that virtual link (D_i^k, D_j^l) is assigned to the physical link if and only if replica instances D_i^k and D_j^l are deployed on the physical nodes. Constraints (24) and (25) specify the origin and destination nodes.

4.4 Second Step: Reassignment of Service Chain Forwarding Graph to Meet Service Chain Requirements

The formulation for the second ILP is based on [10]. However, we have detected errors in the formulation presented in [10] and have rectified them. Using the service chain forwarding graph, constructed in Section 4.3, as input, the following ILP reassigns virtual nodes and links in the service chain forwarding graph to physical nodes and links, in order to ensure all the service chain requirements, i.e., availability, assignment, processing, and capacity requirements.

$$\begin{aligned} \min \quad & \sum_{n \in \mathcal{V}_{\mathrm{P}}} \sum_{f \in \mathcal{V}_{\mathrm{SC}}^s} \sum_{i=1}^{N} \Psi^{D_i^f} M_n^{D_i^f} \\ & + \sum_{n \in \mathcal{V}_{\mathrm{P}}} \sum_{f \in \mathcal{V}_{\mathrm{SC}}^s} \sum_{i=1}^{P} \Psi^{R_i^f} M_n^{R_i^f} \\ & + \sum_{n \in \mathcal{V}_{\mathrm{P}}} \sum_{f \in \mathcal{V}_{\mathrm{SC}}^s} \Psi^{LB} M_n^{LB(f)}, \end{aligned} \tag{28}$$

$$\begin{aligned} \text{s.t.} \quad & (8)\text{–}(13), (19)\text{–}(25), (27), \\ & \sum_{m \in \mathcal{V}_{\mathrm{P}}} M_{n,m}^{D_i^k, D_j^l} - \sum_{m \in \mathcal{V}_{\mathrm{P}}} M_{m,n}^{D_i^k, D_j^l} \\ & = M_n^{D_i^k} - M_n^{D_j^l}, \\ & \forall (n,m) \in \mathcal{E}_{\mathrm{P}}, \forall (D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^s, \end{aligned} \tag{29}$$

$$\sum_{f\in\mathcal{V}_{\mathrm{SC}}^{s}}\sum_{i=1}^{N}\Psi^{D_i^f}M_n^{D_i^f}+\sum_{f\in\mathcal{V}_{\mathrm{SC}}^{s}}\sum_{i=1}^{P}\Psi^{R_i^f}M_n^{R_i^f}+\sum_{f\in\mathcal{V}_{\mathrm{SC}}^{s}}\Psi^{LB}M_n^{LB(f)}\leq\theta_n,\quad\forall n\in\mathcal{V}_{\mathrm{P}},\tag{30}$$

$$\sum_{(D_i^k,D_j^l)\in\mathcal{E}_{\mathrm{SC}}^{s}}\Omega^{D_i^k,D_j^l}M_{n,m}^{D_i^k,D_j^l}\leq\delta_{n,m},\quad\forall(n,m)\in\mathcal{E}_{\mathrm{P}},\tag{31}$$

$$\sum_{i=1}^{N}\sum_{n\in\mathcal{V}_{\mathrm{P}}}\Psi^{D_i^f}M_n^{D_i^f}\geq\psi^{\mathrm{diversity}}(f)-\Psi^{\mathrm{LB}},\quad\forall f\in\mathcal{F}_s,\tag{32}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\Omega^{D_i^k,D_j^l}\cdot a^{D_i^k,D_j^l}\geq\Omega^{k,l},\quad\forall(k,l)\in\mathcal{E}_{\mathrm{SC}}^{s},\tag{33}$$

$$\sum_{n\in\mathcal{V}_{\mathrm{P}}}\Psi^{R_i^k}M_n^{R_i^k}\geq\frac{\psi^{\mathrm{redundancy}}(k,\alpha)}{P},\quad\forall k\in\mathcal{F}_s,\forall i\in\{1,\ldots,P\}.\tag{34}$$

The objective function (28) aims to minimize resource utilization of physical nodes and links. It should be noted that the service chain forwarding graph G_{DIV}^{s}, used in the constraints (8)–(13), (19)–(25), and (27), is derived by solving the first ILP presented in Section 4.3. Constraint (29) reflects the standard flow conservation rules, ensuring that each virtual link in the service chain forwarding graph, denoted by $\forall(D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^{s}$, is assigned to one or more physical links. Constraint (30) (resp. (31)) represents the capacity constraint of physical node $n \in \mathcal{V}_{\mathrm{P}}$ (resp. physical link $(n,m) \in \mathcal{E}_{\mathrm{P}}$). Constraint (32) (resp. (33)) ensures that the processing (resp. bandwidth) requirements of replica instances of VNF f (resp. virtual links $(D_i^k, D_j^l) \in \mathcal{E}_{\mathrm{DIV}}^{s}$) meet the minimum required standards. Constraint (34) ensures that the backup instance has the higher processing requirement than the SCR s.

The proposed ILP assumes the SC-VNFP problem in an online processing manner, where the SCR arriving at the NFV network is immediately served one by one. We can extend the online processing to the offline one by simultaneously serving all SCRs $\mathcal{S}$, adding $\forall s \in \mathcal{S}$ to each constraint of the ILP.

5. Numerical Results

5.1 Evaluation Scenario

We use an NSFNET topology to present the physical network, comprising of 14 physical nodes and 21 physical links. Each physical node $n \in \mathcal{V}_{\mathrm{P}}$ is equipped with processing resources of an 8-core CPU (i.e., $\theta_n = 8$) and has availability A_n^{HW} in the range of $[0.99, 0.999]$. Each physical link $(n,m) \in \mathcal{E}_{\mathrm{P}}$ possesses a bandwidth capacity of $\delta_{n,m} = 10$ Gbps. For our service chain demand and requirements, we refer to Table 2, where each service chain request $s \in \mathcal{S}$ serves 500 aggregated users. We assume the existence of six VNF types $F = 6$ and four service types, each of which includes five functions (i.e., $K_s = 5$). Each VNF f has processing requirements as described in Table 3, and availability A_f^{SW} in the range of $[0.9, 0.99]$.

Table 2: Service chain demand and requirements (NAT: Network Address Translator, FW: Firewall, TM: Traffic Monitor, WOC: WAN Optimization Controller, IDPS: Intrusion Detection Prevention System, and VOC: Video Optimization Controller).

Service	Sequence of functions	Demand	Bitrate
Web service	NAT-FW-TM-WOC-IDPS	18.2%	1 Mbps
VoIP	NAT-FW-TM-FW-NAT	11.8%	0.64 Mbps
Video streaming	NAT-FW-TM-VOC-IDPS	69.9%	40 Mbps
Online gaming	NAT-FW-VOC-WOC-IDPS	0.1%	40 Mbps

Table 3: Relationship between function type and processing requirements per SCR [18].

Function type	NAT	FW	TM	IDPS	VOC	WOC
Ψ^f	0.0092	0.009	0.133	0.107	0.054	0.054

For the comparison purpose, we adopt the three schemes, i.e., ALLDIV, RANDDIV, and SELEDIV, each of which was described in Section 3.5. In terms of VNF diversity, we initially set the maximum diversity level N in the range of $[1, 5]$, and then adjust the diversity level of each VNF based on the diversity strategy, i.e., ALLDIV, RANDDIV, and SELEDIV. Each of these schemes is formulated as an ILP in the same manner presented in Section 4.4.

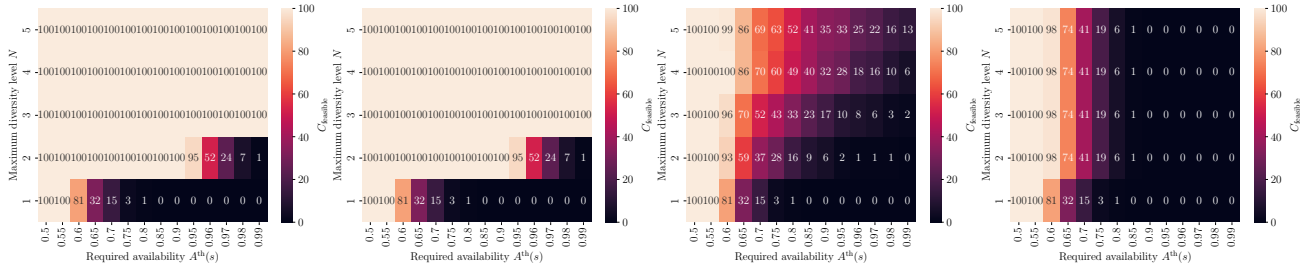
The processing (resp. bandwidth) requirement of each replica instance (resp. virtual link) is set as $\Psi^{D_i^f} = (\Psi^f + H(f, \Delta^f))/\Delta^f$ (resp. $\Omega^{D_i^k, D_j^l} = \Omega^{k,l}/\Delta^k\Delta^l$). On the other hand, with regards to VNF redundancy, we prepare a backup instance ($P = 1$) of each VNF $f \in \mathcal{F}_s$ by applying traditional redundancy ($\alpha = 1$) to each of them.

As for the evaluation metrics, we define the mean service availability as $A_{\mathrm{mean}} = |\mathcal{S}|^{-1}\sum_{s\in\mathcal{S}} A_s$. Furthermore, it is desirable to establish service paths with a lower level of resource utilization in terms of resource efficiency. Since it is challenging to calculate the resource efficiency of an NFV network directly, we use the objective value, i.e., Eq. (28). We define the number C_{feasible} of trials in which an optimal solution is found. In the following, we present the average value of each evaluation metric from 100 independent simulation runs in which an optimal solution was found.

5.2 Fundamental Characteristics

5.2.1 Impact of Diversity Level and Required Availability

In this section, we examine the effects of diversity level and required availability on solution optimality in terms of C_{feasible} when $S = 1$, as shown in Fig. 3. Specifically, for the



(a) Proposed ILP. (b) ALLDIV. (c) RANDDIV. (d) SELEDIV.

Fig. 3: Number C_{feasible} of trials in which an optimal solution is found.

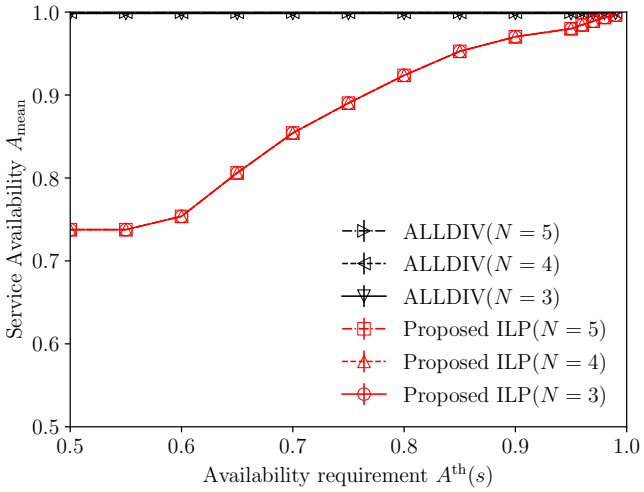


Fig. 4: Relationship between the availability requirement $A^{\mathrm{th}}(s)$ and service availability A_{mean}.

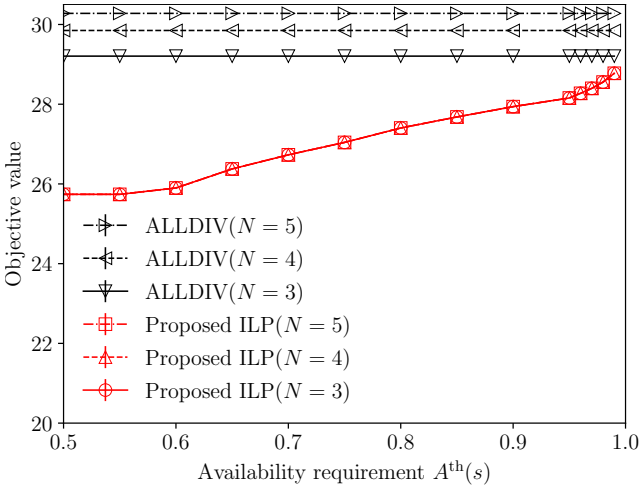


Fig. 5: Relationship between the availability requirement $A^{\mathrm{th}}(s)$ and objective value.

case of no VNF diversity ($N = 1$), we observe that C_{feasible} decreases as $A^{\mathrm{th}}(s)$ increases, while it increases with the maximum diversity level N due to the ability of VNF diversity to improve service availability and avoid availability constraint violations. Regarding the differences among the schemes, we confirm that SELEDIV fails to improve C_{feasible} by increasing N when $A^{\mathrm{th}}(s) \geq 0.9$. In addition, the improvement ratio of C_{feasible} by VNF diversity degrades as $A^{\mathrm{th}}(s)$ increases. Notably, the proposed ILP demonstrates competitive C_{feasible} performance compared to ALLDIV. Specifically, both the proposed ILP and ALLDIV provide optimal solutions for all trials when $N = 3$, regardless of $A^{\mathrm{th}}(s)$.

5.2.2 Tradeoff between Resource Efficiency and Service Availability

In this section, we present a comparison of the tradeoff between resource efficiency and service availability under the condition $S = 1$. Specifically, we examine the proposed ILP and ALLDIV, because the remaining schemes cannot achieve the optimal solution under high availability requirement, as shown in Section 5.2.1. Our results show that ALLDIV consistently achieves the service availability A_{mean} and objective value, regardless of the availability requirement $A^{\mathrm{th}}(s)$. On the other hand, the proposed ILP increases A_{mean} with $A^{\mathrm{th}}(s)$, but exhibits a smaller objective value compared to ALLDIV. This is due to the overhead incurred by applying VNF diversity to satisfy the availability requirement.

Furthermore, we investigate the impact of maximum diversity level N on the performance. Our findings show that the objective value of ALLDIV increases with N, since ALLDIV applies VNF diversity to all VNFs. In contrast, the objective value of proposed ILP does not depend on N, as it deploys the minimum required number of replica instances to satisfy the availability requirement on the physical nodes. As a result, the proposed ILP yields a smaller A_{mean} but satisfies the condition $A_{\mathrm{mean}} \geq A^{\mathrm{th}}(s)$.

6. Conclusion

In this paper, we have formulated a two-step integer linear program (ILP) for the service chaining and virtual network function placement (SC-VNFP) problem, incorporating VNF diversity and redundancy. The proposed ILP adjusts resource efficiency and service availability based on availability constraints, as demonstrated through numerical results. We have analyzed the tradeoff between resource efficiency and service availability by comparing the proposed and existing ILPs under different diversity levels and required availability, using the number of trials that achieve optimal

solutions as the metric. Furthermore, we have examined the fundamental characteristics of the proposed ILP, including the service availability and resource usage. In future work, we plan to develop a heuristic algorithm to efficiently address the computational complexity of SC-VNFP problem with VNF diversity and redundancy.

Acknowledgments

This work was supported in part by the JSPS KAKENHI (B) under Grant 22H03586 and the JSPS Grant-in-Aid for Young Scientists under Grant 23K16869, Japan.

References

[1] B. Yi, X. Wang, K. Li, S. k. Das, and M. Huang, "A Comprehensive Survey of Network Function Virtualization," Computer Networks, vol.133, pp.212–262, March 2018.

[2] S. Demirci and S. Sagiroglu, "Optimal Placement of Virtual Network Functions in Software Defined Networks: A Survey," Journal of Network and Computer Applications, vol.147, p.102424, Dec. 2019.

[3] M. Sasabe and T. Hara, "Capacitated Shortest Path Tour Problem-Based Integer Linear Programming for Service Chaining and Function Placement in NFV Networks," IEEE Transactions on Network and Service Management, vol.18, no.1, pp.104–117, March 2021.

[4] S. Yang, F. Li, S. Trajanovski, R. Yahyapour, and X. Fu, "Recent Advances of Resource Allocation in Network Function Virtualization," IEEE Transactions on Parallel and Distributed Systems, vol.32, no.2, pp.295–314, Feb. 2021.

[5] L. Qu, C. Assi, K. Shaban, and M.J. Khabbaz, "A Reliability-Aware Network Service Chain Provisioning With Delay Guarantees in NFV-Enabled Enterprise Datacenter Networks," IEEE Transactions on Network and Service Management, vol.14, no.3, pp.554–568, Sept. 2017.

[6] D. Li, P. Hong, K. Xue, and J. Pei, "Availability Aware VNF Deployment in Datacenter Through Shared Redundancy and Multi-Tenancy," IEEE Transactions on Network and Service Management, vol.16, no.4, pp.1651–1664, Feb. 2019. Conference Name: IEEE Transactions on Network and Service Management.

[7] S. Yang, F. Li, R. Yahyapour, and X. Fu, "Delay-Sensitive and Availability-Aware Virtual Network Function Scheduling for NFV," IEEE Transactions on Services Computing, vol.15, no.1, pp.188–201, Jan. 2022.

[8] A. Hmaity, M. Savi, F. Musumeci, M. Tornatore, and A. Pattavina, "Protection Strategies for Virtual Network Functions Placement and Service Chains Provisioning," Networks, vol.70, no.4, pp.373–387, Dec. 2017.

[9] J. Xie, P. Yi, Z. Zhang, C. Zhang, and Y. Gu, "A Service Function Chain Deployment Scheme Based on Heterogeneous Backup," Proc. of IEEE 18th International Conference on Communication Technology (ICCT), pp.1096–1103, Oct. 2018.

[10] A. Alleg, T. Ahmed, M. Mosbah, and R. Boutaba, "Joint Diversity and Redundancy for Resilient Service Chain Provisioning," IEEE Journal on Selected Areas in Communications, vol.38, no.7, pp.1490–1504, July 2020.

[11] F. Carpio and A. Jukan, "Improving Reliability of Service Function Chains with Combined VNF Migrations and Replications," Nov. 2017.

[12] R. Kang, F. He, and E. Oki, "Fault-Tolerant Resource Allocation Model for Service Function Chains with Joint Diversity and Redundancy," Computer Networks, vol.217, p.109287, Nov. 2022.

[13] T. Hara and M. Sasabe, "Speedy and Efficient Service Chaining and Function Placement Based on Lagrangian Heuristics for Capacitated Shortest Path Tour Problem," Journal of Network and Systems Management, vol.31, no.1, pp.1–34, Dec. 2022.

[14] N. Hyodo, T. Sato, R. Shinkuma, and E. Oki, "Virtual Network Function Placement for Service Chaining by Relaxing Visit Order and Non-Loop Constraints," IEEE Access, vol.7, pp.165399–165410, 2019.

[15] A. Alleg, T. Ahmed, M. Mosbah, R. Riggio, and R. Boutaba, "Delay-Aware VNF Placement and Chaining Based on a Flexible Resource Allocation Approach," Proc. of 13th International Conference on Network and Service Management (CNSM), Tokyo, pp.1–7, IEEE, Nov. 2017.

[16] S. Bhat and G.N. Rouskas, "Service-Concatenation Routing with Applications to Network Functions Virtualization," Proc. of 26th International Conference on Computer Communication and Networks (ICCCN), Vancouver, BC, Canada, pp.1–9, IEEE, July 2017.

[17] P. Festa, "The Shortest Path Tour Problem: Problem Definition, Modeling, and Optimization," Proc. of INOC, pp.1–7, 2009.

[18] M. Savi, M. Tornatore, and G. Verticale, "Impact of Processing-Resource Sharing on the Placement of Chained Virtual Network Functions," IEEE Transactions on Cloud Computing, vol.9, no.4, pp.1479–1492, 2019. arXiv:1710.08262 [cs].

[19] N. Shahriar, R. Ahmed, S.R. Chowdhury, M.M.A. Khan, R. Boutaba, J. Mitra, and F. Zeng, "Virtual Network Embedding With Guaranteed Connectivity Under Multiple Substrate Link Failures," IEEE Transactions on Communications, vol.68, no.2, pp.1025–1043, Feb. 2020.

Appendix A: Linearization of Standard Flow Constraint

Since constraint (14) is nonlinear, according to the approach used in [19], we replace it with the following linear constraints.

$$w_n^{D_i^k,D_j^l} \in \{0,1\}, \forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 1)$$

$$z_n^{D_i^k,D_j^l} \in \{0,1\}, \forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 2)$$

$$\sum_{m\in\mathcal{V}_\mathrm{P}} M_{n,m}^{D_i^k,D_j^l} - \sum_{m\in\mathcal{V}_\mathrm{P}} M_{m,n}^{D_i^k,D_j^l} = w_n^{D_i^k,D_j^l} - z_n^{D_i^k,D_j^l},$$
$$\forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 3)$$

$$w_n^{D_i^k,D_j^l} \le a^{D_i^k,D_j^l},$$
$$\forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 4)$$

$$w_n^{D_i^k,D_j^l} \le M_n^{D_i^k},$$
$$\forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 5)$$

$$w_n^{D_i^k,D_j^l} \le a^{D_i^k,D_j^l} + M_n^{D_i^k} - 1,$$
$$\forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 6)$$

$$z_n^{D_i^k,D_j^l} \le a^{D_i^k,D_j^l}, \forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 7)$$

$$z_n^{D_i^k,D_j^l} \le M_n^{D_j^l}, \quad \forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}, \quad (\mathrm{A}\cdot 8)$$

$$z_n^{D_i^k,D_j^l} \le a^{D_i^k,D_j^l} + M_n^{D_j^l} - 1,$$
$$\forall n \in \mathcal{V}_\mathrm{P}, \forall (D_i^k, D_j^l) \in \mathcal{E}^s_\mathrm{DIV}. \quad (\mathrm{A}\cdot 9)$$

Constraints (A· 1) and (A· 2) represent binariy variables. Constraint (A· 3) enforces the linearized flow con-

servation rule. In case of $a^{D_i^k,D_j^l} = 0$, $w_n^{D_i^k,D_j^l}$ and $z_n^{D_i^k,D_j^l}$ are also set to 0 according to constraints (A· 4) and (A· 7), respectively. On the other hand, in case of $a^{D_i^k,D_j^l} = 1$, $M_n^{D_i^k}$ (resp. $M_n^{D_j^l}$) is equivalent to $w_n^{D_i^k,D_j^l}$ (resp. $z_n^{D_i^k,D_j^l}$) from the constraints (A· 5) and (A· 6) (resp. (A· 8) and (A· 9)). Thus, these constraints are equivalent to constraint (14).

Takanori HARA received the M.Eng. and Ph.D. degrees from Nara Institute of Science and Technology, Japan, in 2018 and 2021. He is currently an Assistant Professor with the Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan. His research interests include AI/ML empowered networking, network virtualization, eBPF/XDP, pedestrian navigation, and game-theoretic approaches.

Masahiro SASABE received the B.S., M.E., and Ph.D. degrees from Osaka University, Japan, in 2001, 2003, and 2006, respectively. He is currently a Professor of Faculty of Informatics, Kansai University, Japan. His research interests include P2P/NFV networking, game-theoretic approaches, human-harmonized network systems, and network optimization.

Kento Sugihara received the M. Eng. degree from Nara Institute of Science and Technology, Japan, in 2023. He is currently working at Recruit Co., Ltd.

Shoji KASAHARA received the B. Eng., M. Eng., and Dr. Eng. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 1996, respectively. Currently, he is a Professor of Division of Information Science, Nara Institute of Science and Technology, Nara, Japan. His research interests include stochastic modeling and analytics of large-scale complex systems based on computer/communication networks.