Deep Reinforcement Learning for Provisioning Virtualized Network Function in Inter-Datacenter Elastic Optical Networks

Min Zhu, *Member, IEEE*, Qi Chen, *Graduate Student Member, IEEE*, Jiahua Gu, *Graduate Student Member, IEEE*, and Pingping Gu

***Abstract*—In today's datacenters (DCs), IT resources virtualization is leveraged to realize Network Function Virtualization (NFV) over general-purpose servers. Meanwhile, most of the service providers (SPs) are planning to use Virtual Network Functions (VNFs) to provide agile and flexible network services. In service provisioning, the VNF selection and mapping greatly affect IT resource utilization in DCs and spectrum resource utilization in optical networks. This paper proposes a Deep Reinforce Learning (DRL)-based algorithm for VNF provisioning. By selecting appropriate VNFs for the service requests, the algorithm intelligently guarantees efficient reusing of deployed VNFs while consuming fewer spectrum resources in inter-DC elastic optical networks (EONs). To facilitate the decision-making of the DRL agent, we first decompose the complex VNF-based service chaining (VNF-SC) into several VNF components (VNFCs), which can be solved one-by-one in turn. Then, a feature matrix-based encoding scheme is designed to represent the set of the VNFCs, the available DCs for the VNFCs, and the VNFC being operated, i.e., the input of neural networks. In addition, considering the complexity and difficulty of the VNF-SC provisioning problem, Double Deep Q Network (DDQN) is introduced in the proposed algorithm. Finally, compared with the benchmark heuristics, the extensive simulation results in different network topologies show that the proposed algorithm can reduce the IT and spectrum resource consumption by at least 9.6% and 1.6%, which proves the effectiveness of the proposed DRL-based VNF provisioning algorithm.**

***Index Terms*—NFV, SFC, deep Q-learning, elastic optical networks.**

I. Introduction

With the explosive growth of new networking applications (e.g., cloud/fog computing), network function virtualization (NFV) becomes more and more attractive owing to its agile resource allocation and fast deployment of new services [1]. Decoupling network functions from dedicated hardware devices, Virtual Network Functions (VNFs) could be realized by generic network resources (e.g., bandwidth, CPU cycles, and memory space), which significantly reduces Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) for network operators [2]. Those VNFs will be instantiated and deployed in high performance servers inside DCs to attain the requirements of the network functions. With NFV, each service will be deployed on a set of predefined VNFs in a specific order according to service-level agreement (SLA). Then, to achieve maximum utilization of networks resources, operators need to select appropriate DCs hosting the desired VNFs from the candidate DCs and route data traffic from source to destination through a series of VNFs (i.e., VNF service chaining (VNF-SC) [3]–[7]). For instance, Fig. 1 shows a VNF-SC request, which is abstracted into a directed linear graph. The two ellipses represent the service terminal and the user, respectively. Once the service terminal and user of the SFC request are confirmed, the positions of the source and destination are fixed in the network. The rectangles represent the VNFs such as Dynamic host configuration protocol (DHCP), router, Network Address Translation (NAT) and firewall in the network which are interconnected by directed virtual links $e_1 - e_4$ in a predefined order.

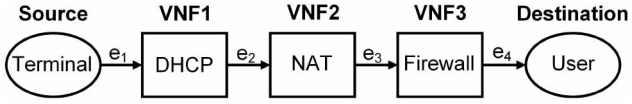

Fig. 1. An example of VNF-SC request [18].

For resource-efficient VNF-SC provisioning, both the IT resource consumption in DC and spectrum resource utilization in fiber links should be optimized. To reduce IT resource consumption in DCs effectively, a VNF reuse mechanism has been proposed to improve IT resource utilization in DCs [16]–[18]. This mechanism allows the VNF instances deployed in the DC to be shared by multiple VNFs of the same type. To meet high transmission capacity and spectral efficiency [8], [9], elastic optical network (EON) has been proposed as one of the most promising networking technologies for the next-generation backbone networks. Compared with wavelength division multiplexing (WDM) technology [10], [11], EON

Manuscript received 12 November 2021; revised 4 March 2022 and 30 April 2022; accepted 30 April 2022. Date of publication 3 May 2022; date of current version 12 October 2022. This work was supported in part by the Key Research and Development Program of Jiangsu Province under Grant BE2020012 and in part by the Transformation Program of Scientific and Technological Achievements of Jiangsu Province under Grant BA2019026. The associate editor coordinating the review of this article and approving it for publication was M. Reisslein. *(Corresponding author: Min Zhu.)*

Min Zhu and Jiahua Gu are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China, and also with the Pervasive Communication Research Center, Purple Mountain Laboratories, Nanjing 211111, China (e-mail: minzhu@seu.edu.cn).

Qi Chen is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Pingping Gu is with the Legal Department, Taicang T&W Electronics Company Ltd., Taicang 215400, China.

Digital Object Identifier 10.1109/TNSM.2022.3172344

TABLE I
ABBREVIATION AND KEY TERMS

Abbreviation	Full Form	Abbreviation	Full Form
BV-OXC	Bandwidth-variable Optical Cross-connect	MFSI	Maximum FS Index
CAPEX	Capital Expenditure	NFV	Network Function Virtualization
DC	Datacenter	NSFNET	National Science Foundation Network
DRL	Deep Reinforcement Learning	OPEX	Operational Expenditure
DL	Deep Learning	O/E/O	Optical/Electronic/Optical
DQN	Deep Q Network	RSA	Routing and Spectrum Assignment
DNN	Deep Neural Network	RMSA	Routing, Modulation, and Spectrum Assignment
DDQN	Double Deep Q Network	SLA	Service-level Agreement
EON	elastic optical network	SC	Service Chaining
FS	Frequency Slot	SFC	Service Function Chain
IT	Information Technology	S-BVTs	Sliceable Bandwidth Variable Transponders
ILP	Integer Linear Programming	SRA	Shortest-path and Random Deployment Algorithm [24]
IoT	Internet of Things	VNF	Virtual Network Functions
LSTM-NN	long/short term memory-based neural network	VNF-SC	VNF-based SC
LCS	Longest Common Subsequence	VNFC	VNF components
LBA	LCS based algorithm [24]	VM	Virtual Machine
LBP	LCS-based Path	WDM	Wavelength Division Multiplexing

can achieve bandwidth-variable super-channels by grooming a series of finer-granularity (e.g., 6.25 GHz) subcarriers and adapting the modulation formats, which achieves agile bandwidth allocation and greatly improves the spectrum efficiency. With the VNF-SC requests provisioned in inter-DC EONs, 'optical SC' has many advantages like high bandwidth capacity and low power consumption, which benefits inter-DC networks exceedingly [12], [13]. However, in addition to a proper selection of DC for VNF deployment, provisioning an optical SC would require a spectrum efficient routing and spectrum assignment (RSA) scheme for connecting source node, intermediate DC nodes and destination node. It should be noted that the optimization of IT resources in DC (through VNF reuse) and spectrum resources in fiber links are interdependent. For example, on the one hand, if we only optimize the IT resources in DC, the spectrum resources in fiber links may not be utilized evenly, which could lead to network congestion and cause some DCs unreachable. On the other, if we only optimize the spectrum resources, some DCs may become overloaded. Therefore, to improve network resource efficiency for optical SC provisioning, joint allocation of IT and spectrum resources should be considered.

Most of the existing works use integer linear programming (ILP) or heuristic algorithms to solve this problem [20]–[21], [24]. Although ILP can be applied to acquire the optimal solution, the running time of ILP would increase exponentially with the scale of the problem [22], [23]. Hence, more time-efficient heuristic algorithms are usually developed to solve the problem in polynomial time. However, the heuristic algorithm always uses the same and fixed set of policies when facing different network scenarios, traffic loads and service requests, which makes the scheduling of network resources inflexible and sometimes inefficient. Hence, a more flexible and efficient algorithm is required.

Recently, DRL methods have been successfully applied to play games [25], realize human-level control [26], and solve multi-resource management problems [27]. Compared with heuristic algorithms that adopt the fixed scheme, DRL agent can take adaptive strategies based on the currently observed network state. In addition, once DRL training is complete, its decisions are made in real time [29]. Therefore, in this paper, we adopt the DRL method for solving the above VNF-SC provisioning issue.

In this paper, our main contributions can be summarized as follows. 1) This paper proposes a DRL-based algorithm to solve the VNF-SC provisioning issue considering the VNF instance use in inter-DC EONs. 2) A VNF-SC decomposition scheme is proposed to facilitate the decision-making of the DRL agent. In this scheme, each VNF-SC is divided into several VNF components (VNFCs), which are provisioned in turn. 3) A feature matrix-based encoding scheme is introduced to solve the problem that the length of neural network input constantly changes with the variable VNF-SC length and the stochastic source and destination nodes of VNF-SCs. It produces a fixed-length input to facilitate the DRL training. 4) Extensive simulations are conducted in various network topologies to prove the effectiveness of our proposed DRL-based VNF-SC provisioning scheme. Compared with the benchmarks, better IT and spectrum resource utilization and lower average cost are achieved.

The rest of this paper is organized as follows. Section II briefly overviews the related researches. We describe the system model in Section III. Section IV introduces the proposed DRL-based algorithm. The simulation results in two network topologies and related discussions are presented in Section V. Finally, Section VI summarizes the paper. The major abbreviations are provided in Table I.

II. RELATED WORKS

The aforementioned optical SC provisioning issue has attracted massive research in recent years. The authors in [14] studied cooperative RSA for multi-domain service provisioning in software-defined network and experimentally demonstrated the proposed system, which achieves low blocking probability and provision latency. In [15], authors designed a new Deep Learning (DL) model based on the long/short term memory-based neural network (LSTM-NN) for realizing more

accurate VNF-SC pre-deployment in inter-DC EONs. A better tradeoff between resource utilization and blocking probability is achieved. The above works mainly addressed the VNF-SC provisioning issue in inter-DC EONs, but how to effectively improve the IT resource utilization in DC was not considered.

To further improve resource utilization in DCs, many works have adopted the VNF reuse mechanism. In [16], the authors studied the relation between the link and virtual machine (VM) usage, and proposed a VNF-SC deployment algorithm that better utilizes the limited resources to serve a larger size of demands. Through a two-stage latency-aware VNF deployment scheme, the authors in [17] jointly optimized the resource utilization of both edge servers and physical links under the latency limitations. In [18], an energy-aware routing and adaptive delayed shutdown algorithm for dynamic service function chain (SFC) deployment was proposed to decrease the number of open servers and reduce the idle/switching energy consumption of these servers. The above works have considered the VNF instance reuse but only addressed the VNF-SC provisioning problem in a network scenario where the bandwidth resources of fiber links are viewed as a resource pool. However, when provisioning VNF-SCs in an inter-DC EON, the spectrum contiguity, continuity and non-overlapping constraints need to be satisfied [19].

Some works have addressed the VNF-SC provisioning in inter-DC EONs considering the VNF reuse. By setting a resource threshold for each DC and fiber link, [20] investigated how to realize joint placement of SCs and RSA efficiently in EON. The literature [21] considered load balancing of IT and spectrum resources in inter-DC EONs and proposed the concept of joint balancing factors to quantify the impact of different VNF selections on network load. In [24], to deploy the VNF-SCs in inter-DC EONs, an ILP model was first formulated to get the optimal solution, and then a longest common subsequence (LCS) based algorithm (LBA) was proposed to jointly optimize spectrum and IT resources. In LBA, LCS-based path (LBP) is selected from several shortest paths between source and destination nodes, and VNF-SC could be provisioned in the LBP. Although the heuristic method is efficient, in practical cases, we often find that the LCS path is not unique and even in the same LCS path, provisioning schemes show the variety. Therefore, how to choose the best LCS to achieve optimum allocation in both spectrum and IT resources becomes a quite difficult task for the heuristic algorithm. This again demonstrates the need to design a more efficient algorithm for solving the above VNF-SC provisioning issue.

Recently, network resource orchestration via DRL has attracted a lot of attention from academia and industry. The literature [28] proposed a DRL-based scheme to manage heterogeneous VNF nodes and IoT network devices, but it does not involve the optical layer. A DRL-based framework is proposed for routing, modulation, and spectrum assignment (RMSA) in EONs, but it only considered spectrum resources for optical path provisioning [29]. To the best of our acknowledge, this is the first paper that adopts the DRL method to solve the VNF-SC provisioning issue in inter-DC EONs while considering the VNF reuse.

III. System Model

A. Network Scenario

1) Substrate EON: We consider the inter-DC EON as a directed linear graph G (V, E), where V represents the set of nodes and E denotes the set of fiber links, respectively. We assume that each node $v \in V$ includes a bandwidth-variable optical cross-connect (BV-OXC) for setting up inter-DC communications, some of which are locally attached to a DC and any type of VNF could be deployed in the DC. The set of DCs would be denoted as N. Specially, we assume that, on each of its fiber ports, a BV-OXC is equipped with several sliceable bandwidth variable transponders (S-BVTs) [9] that can cover the full spectra on a fiber. Therefore, if we intend to set up a lightpath, spectrum contiguous and continuous constraints [30] ought to be considered as well as the spectrum capacity constraint. For simplicity, the one-domain inter-DC EON is considered in this paper, and the cross-domain issues will be taken into account in our future works.

2) VNF-SC request: We denote a VNF-SC request as a directed linear graph R (s, d, T, B), in which s and d are the source and destination nodes, respectively. $T = (t_1, t_2, \ldots, t_J)$ represents the requested VNF sequence, where t_j is the type of the j-th VNF and J means the length of this service chain. Similarly, $B = (b_0, b_1, \ldots, b_J)$ indicates bandwidth requirements in terms of frequency slot (FS), where b_0 is the initial bandwidth requirement and b_j is the bandwidth requirement after steering through VNF t_j. Here, we address the practical scenario in which the bandwidth requirement of a VNF-SC can change after steering through a VNF. For a VNF-SC request, source and destination nodes namely s and d could be any node of a network but VNFs can only be deployed in DCs. It is allowed that multiple adjacent VNFs may be placed in the same DC or in different DCs. In fact, the above definition of VNF-SC request (i.e., a directed linear graph) can be viewed as a special case of the typical VNF forwarding graph [31].

Note that, we only address the static network planning problem, hence all the VNF-SC requests are assumed to be known in advance. The static network requests, once provisioned successfully, does not expire in the network. Hence, for static network planning problem, the requests only need to be provisioned once and after provision, the network state will remain unchanged.

B. Optimization Objective

We set boolean variable $z_{e,f}$ and $h_{n,t}$, in which $z_{e,f} = 1$ if FS $f \in \mathrm{F}_e$ is occupied and $h_{n,t} = 1$ if DC $n \in N$ has deployed VNF t, where F_e is the set of FSs in the fiber $e \in E$. Note that, because we only study scheduling tasks in a static network, all demands are known in advance. As for optimization objectives, there are two factors we need to consider. On the one hand, since VNF instances are a kind of reusable resource [24], we could attempt to reuse them as much as possible rather than instantiating a new one. On the other hand, if we reuse VNFs excessively, VNF-SCs may have to be routed on a longer lightpath to a remote DC deployed with the needed VNF types, leading to plenty of wastage of spectrum resources. Generally

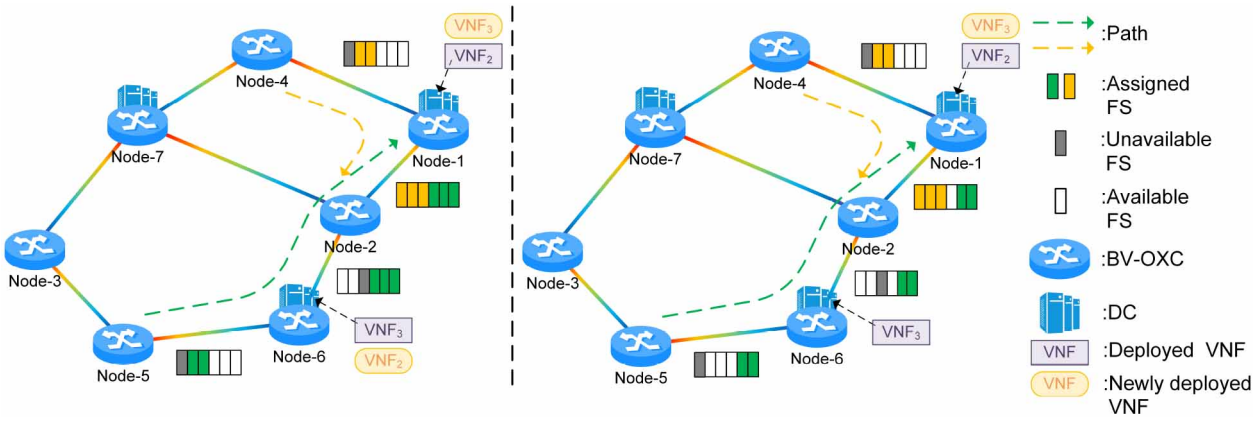

Fig. 2. Example of two provisioning schemes for VNF-SCs in inter-DC EONs.

speaking, reuse means consuming more spectrum resources, hence it is a trade-off between spectrum and IT resources, obviously. Therefore, our objective is to minimize the spectrum utilization ratio and the number of deployed VNFs in the network jointly, as

$$\text{Minimize}\left(\alpha\cdot\frac{\sum_e\sum_f z_{e,f}}{|E|\cdot|F|}+\beta\cdot\frac{\sum_n\sum_t h_{n,t}}{|N|\cdot|T|}\right) \quad (1)$$

where $|E|$, $|F|$, $|N|$, $|T|$ represent the number of fibers in the network, number of FSs in one fiber, number of DCs, and number of all possible VNF types, respectively. The α and β are factors introduced to adjust the importance of these two terms. The first term in Eq. (1) reflects the spectrum utilization ratio, and the second term denotes the normalized number of deployed VNFs.

C. Problem Description

Fig. 2 shows an intuitive example of provisioning VNF-SCs in a 7-node EON, where *Node* 1, 6, 7 are locally attached to DCs and each fiber contains 6 FSs. In the network, there are three types of VNFs (i.e., VNF1, VNF2, and VNF3) and every VNF-SC could demand 2 different types of VNFs and 3 FSs at most. VNF2 and VNF3 have already been deployed on *Node* 1 and *Node* 6, respectively. There are two requests, $R_1\{5,1,(\text{VNF2},\text{VNF3}),(2,2,3)\}$ and $R_2\{4,2,(\text{VNF3}),(2,3)\}$. This example presents two different schemes for VNF selection and RSA. The left scheme of Fig. 2 shows that R_1 chooses the path $5{\rightarrow}6{\rightarrow}2{\rightarrow}1$, deploys VNF2 and reuses existing VNF3 on *Node* 6, while R_2 goes through $4s{\rightarrow}1{\rightarrow}2$ and deploys VNF3 on *Node* 1. For this scheme, we deploy 2 VNFs and consume 13 FSs in total. However, if we take an optimized scheme, things are different. In the right scheme of Fig. 2, R_1 still takes the path $5{\rightarrow}6{\rightarrow}2{\rightarrow}1$ but instantiates VNF3 and reuses existed VNF2 on *Node* 1, correspondingly, and R_2 goes through $4{\rightarrow}1{\rightarrow}2$ and reuses the VNF3 on *Node* 1 that has been deployed by R_1. After provisioning, we utilize 1 VNFs and 11 FSs totally.

In fact, path $5{\rightarrow}6{\rightarrow}2{\rightarrow}1$ is the LCS-based path (LBP) for R_1 in this network. Two available DCs exist in the LBP, i.e., *Node* 6 and *Node* 1. R_1 could either choose to instantiate VNF2 and reuse VNF3 on *Node* 6 (i.e., left scheme in Fig. 2) or instantiate VNF3 and reuse VNF2 on *Node* 1 (i.e., right scheme in Fig. 2). Compared with the left scheme, the right scheme effectively reduces spectrum resource consumption. IT resource is also saved when deploying R_2 using the right scheme, which was not considered in [24] and [30]. It illustrates that, even if in the same LBP, the selection of VNF greatly affects the load status of networks. Hence, we can figure that to efficiently provision VNF-SCs in an inter-DC EON, it is meaningful to allocate spectrum and IT resources jointly to optimize VNF selections further.

It is notable that in the two schemes above, the spectrum allocation from source to destination will not have to meet the continuity constraint. With O/E/O conversion equipped in DC, an end-to-end path can be divided into several lightpaths, where spectrum allocation can be performed independently. For instance, as shown in the left scheme of Fig. 2, for the FSs assigned to R_1 colored with green, link 6-2 and link 2-1 satisfy the continuity constraint to form a lightpath but link 5-6 serves as another lightpath with one FS assigned, owing to the *Node* 6 is locally attached to a DC, in which the VNF2 and VNF3 are deployed.

The notations of all variables and parameters are listed in Table II.

IV. DRL-Based Algorithm

Deep Q-learning or deep Q network (DQN) can effectively solve complex decision-making problems [29], which can be used for the VNF-SC provisioning in inter-DC EONs. In DQN, a deep neural network (DNN) is used to approximate Q function that is used to evaluate how good an action is for each state. Experience replay is introduced to overcome the problems of correlation and non-stationary distribution of the empirical data [32]. However, DQN will overestimate Q-values. To overcome this problem, the double DQN (DDQN) is introduced, which includes two neural networks named online neural network and target neural network, respectively.

In this section, we first illustrate how to find the LBP for a VNF-SC and maximum FS index (MFSI) based spectrum resources provisioning scheme. Then, the modeling of the

TABLE II
NOTATIONS FOR THE PARAMETERS AND VARIABLES

Parameters	Description
Network related	
G	Substrate network graph
V	Set of nodes
E	Set of fiber links
F_e	Set of FSs in the fiber $e \in E$
N	Set of DCs
$\|E\|$	Number of fibers in the network
$\|F\|$	Number of FSs in a fiber link
$\|N\|$	Number of DCs
Request related	
R	Set of VNF-SC request
s and d	Source and destination node
T=(t_1,t_2,…,t_J)	The requested VNF sequence, where t_j is the j-th requested VNF
$\|T\|$	Number of all possible VNF types
J	Length of requested VNF sequence
J_{max}	Maximum length of VNF-SC
B=(b_0,b_1,…,b_J)	Bandwidth requirements (BR), where b_0 is the initial BR and b_j is the BR after steering through VNF t_j.
b_{max}	Maximum bandwidth requirement
Others	
α and β	Weighting factors
$P_{s,d}\{p_1,…,p_K\}$	the K shortest paths between s and d
φ_k	Set of deployed VNFs on p_k
p^{LBP}	The LCS path between φ_k and T
I_e^{max}	Index of the last used FS on link e
I_p^{max}	Index of the last used FS on path p
D=$\{n_1,n_2,…,n_M\}$	available DCs on p^{LBP}, where n_m means the m-th DC in p^{LBP}, $m \in [1, M]$
M	Number of available DCs on p^{LBP}
M_{max}	Number of maximum available DCs in D
C	Set of VNFCs
CI	Index of the VNFC $CI \in [1, J_i]$
DRL Modeling and Training related	
$\mathbf{SC}_1$	State matrix 1 with size of (M_{max}+2, $\|V\|$+$\|T\|$)
$\mathbf{SC}_2$	State matrix 2 with size of (J_{max}+1, b_{max}+$\|T\|$)
$\mathbf{SC}_3$	State matrix 3 with size of (1, J_{max})
$\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_t^-$	Parameters of online network and target network at time t
$\boldsymbol{S}_t$	State at time t
$Q(S_t,a_t;\theta_t)$	Q value of action a_t under state $\boldsymbol{S}_t$ using DQN with parameter $\boldsymbol{\theta}_t$
ε	Probability to select a random action
Λ	Replay buffer
α	Learning rate
γ	Discount rate

Variables	Description
$z_{e,f}$	Binary, 1 if FS $f \in F_e$ is occupied
$h_{n,t}$	Binary, 1 if DC $n \in N$ has deployed VNF t
a_t	Action at time t
$\boldsymbol{\theta}_t^{'}$	Updated online network parameters

DRL-based algorithm for VNF-SC provisioning, including the encoding scheme, action space, and reward are introduced. Finally, we present the training mechanism.

Algorithm 1 LBP Policy

Input: $R(s, d, T, B)$;
Output: p^{LBP};

```
Calculate P_{s,d}{p_1, p_2, ..., p_K} according to s and d;
for k = 1 to K do
    Get VNFs already deployed on p_k as sequence φ_k;
    Calculate LCS degree between φ_k and T;
end
Select p_k with maximum LCS degree;
p^LBP = p_k;
```

A. LBP Policy and MFSI Based Spectrum Resources Provisioning Scheme

By definition, sequence $Y = (y_1, y_2, \ldots, y_k)$ is a subsequence of sequence $X = (x_1, x_2, \ldots, x_m)$ if there is a strictly increasing index sequence $(i_1, i_2, \ldots, i_k)$ such that for $j \in \{1, 2, \ldots, k\}$, we have $X_{i_j} = Y_j$. Sequence Z is the longest common subsequence (LCS) of the two sequences X and Y if Z is the longest sequence that is the subsequence of both X and Y [24]. Specially, the length of Z is called LCS degree. Receiving a VNF-SC request $R(s, d, T, B)$, we calculate the K shortest paths between s and d as $P_{s,d}\{p_1, p_2, \ldots, p_K\}$ and get the VNFs already deployed on p_k as sequence φ_k. Then the path p^{LBP} that has maximum LCS degree between φ_k and T is selected as LBP. The process is shown below.

In fact, the optimization of IT resources usually causes traffic congestion in some DCs and increases the probability of request blockage. Therefore, a maximum FS index (MFSI) based spectrum resources provisioning scheme is applied. For each link $e \in E$, define I_e^{max} as the index of the last used FS on the link. Similarly, define I_p^{max} as the index of the last used FS on the path p, which means $I_p^{\max} = \max_{e \in p} I_e^{\max}$. In the scheme, for two DCs n_i and $n_j \in N$, we first calculate the K shortest paths between them, denoted as $P_{n_i,n_j}\{p_1, p_2, \ldots, p_K\}$. Then the path which has minimum I_p^{max} is selected and b required FSs are assigned on fibers in the path. The process is shown below.

B. Modeling

1) Encoding scheme: Before provisioning a VNF-SC request $R(s, d, T, B)$, we firstly find the path p^{LBP} according to Algorithm 1. Then, available DCs on p^{LBP} are picked out as $D = \{n_1, n_2, \ldots, n_M\}$, where n_m means the m-th DC in p^{LBP}, $m \in [1, M]$. Generally, one action is taken at each step, which means that multiple steps are required to provision a VNF-SC request. Thus, to facilitate the decision making of the DRL-agent, a VNF-SC is decomposed into J VNFCs as $C = \{\{b_0, t_1\}, \{b_1, t_2\}, \ldots, \{b_{J-1}, t_J, b_J\}\}$, which will be solved in J steps. The j-th IT resource requirement and the j-th spectrum resource requirement in R are combined as the j-th ($j \neq J$) element in C. Specially, the J-th element in C (i.e., the last one) is constituted with the J-th IT resource requirement, the J-th and the $(J + 1)$-th spectrum resource requirements in R. If variable CI is set as the index of the VNFC that is being operated, $CI \in [1, J_i]$, we were able to obtain the array

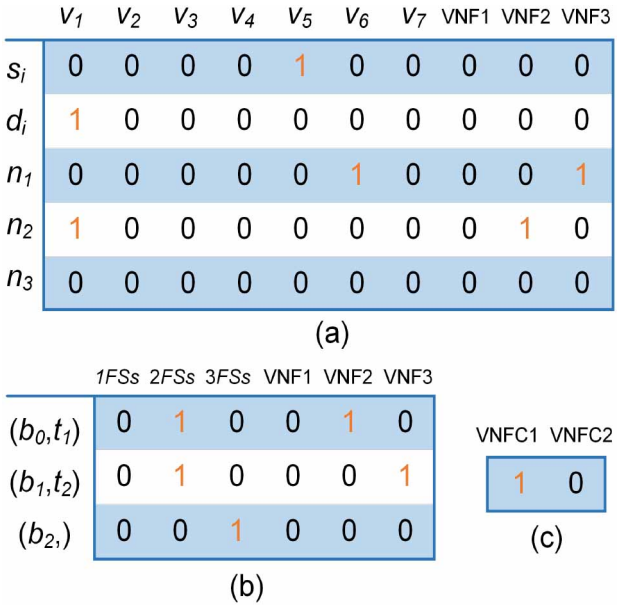

Fig. 3. An example of feature matrix based encoding scheme (a) $\mathbf{SC}_1$ (b) $\mathbf{SC}_2$ (c) $\mathbf{SC}_3$.

$\{s, d, D, C, CI\}$ as the state representation of the DRL environment. However, in practice, the number of available DCs and the number of elements in C always change with different LBPs and VNF-SCs. Hence, a feature matrix-based encoding scheme is designed to solve such a problem.

We use three binary matrixes $\mathbf{SC}_1$, $\mathbf{SC}_2$ and $\mathbf{SC}_3$ to represent the state array $\{s, d, D, C, CI\}$. As shown in Fig. 3(a), $\mathbf{SC}_1$ of size $(M_{max}+2)\times(|V|+|T|)$ is used to represent $\{s, d, D\}$. Take the VNF-SC provisioning in Fig. 2 as an example, there are 7 nodes in the network (i.e., $|V|=7$) and the requested VNF-SC is $R_i\{5, 1, (\text{VNF2}, \text{VNF3}), (2, 2, 3)\}$. Meanwhile, we assume that $M_{max}=3$, $|T|=3$, and the LBP of R_i (i.e., *pLBP i*) is $5\rightarrow 6\rightarrow 2\rightarrow 1$. Then, the size of example matrix $\mathbf{SC}_1$ in Fig. 3(a) becomes 5×10. (s_i, V_5) and (d_i, V_1) are set to 1 because the source node and destination node of R_i are *Node* 5 and *Node* 1, respectively. Also, (n_1, V_6) and (n_1, VNF3) are set to 1 because the first DC node on *pLBP i* is *Node* 6 and the already deployed VNF instance (or instances) is VNF3. Similarly, (n_2, V_1) and (n_2, VNF2) are set to 1. The second matrix $\mathbf{SC}_2$ of size $(J_{max}+1)\times(b_{max}+|T|)$ is used to represent $\{C\}$. Here, we assume that every VNF-SC could request 2 different types of VNFs and 3 FSs at most (i.e., $J_{max}=2$ and $b_{max}=3$), then the size of example $\mathbf{SC}_2$ in Fig. 3(b) becomes 3×6. Because the first VNF in VNF-SC is VNF2, (b_0, t_1, VNF2) is set to 1. Meanwhile, since the bandwidth requirement between source node and VNF2 node is 2 FSs, (b_0, t_1, 2FSs) is set to 1. Similarly, (b_1, t_2, VNF3), (b_1, t_2, 2FSs) and (b_2, 3FSs) are set to 1. The third matrix $\mathbf{SC}_3$ of size $1\times J_{max}$ is used to represent $\{CI\}$. The j-th element in $\mathbf{SC}_3$ is set to 1 if the j-th VNFC is being provisioned. The three matrixes, $\mathbf{SC}_1$, $\mathbf{SC}_2$ and $\mathbf{SC}_3$, together represent the state of our proposed DRL model, which is used as the input of the DNNs. With the fixed tubular form, we can ensure the number of neurons on the input layer of DNNs maintain constant no matter how VNF-SC and LBP vary.

Algorithm 2 MFSI Based Spectrum Resource Provisioning Scheme

Input: DCs n_i and n_j, required FS number b;

1: Calculate $P_{n_i,n_j}\{p_1, p_2, \ldots, p_K\}$ according to n_i and n_j;
2: **for** $k=1$ to K **do**
3: Calculate $I_{p_k}^{\max}$;
4: **end**
5: Select p_k with minimum $I_{p_k}^{\max}$;
6: Assign FSs indexed from $I_{p_k}^{\max}+1$ to $I_{p_k}^{\max}+b$ on all fibers in p_k;

2) Action & Reward: As mentioned above, each VNFC contains a VNF, which needs to be assigned a DC for it. In the DRL-based algorithm, the DRL agent will decide which DC in the $D=\{n_1, n_2, \ldots, n_M\}$ should be assigned to the VNFC. We assume that there are M available DCs, and the DRL agent executes one action at each step.

Nevertheless, M would change with different LBP and VNF-SC requests, where the maximum possible M is denoted as M_{max}. Therefore, in this work, to ensure that DRL-agent can select all possible existing DCs in LBP, the action space includes M_{max} actions, i.e., $action\ \epsilon\ \{n_1, n_2, \ldots, n_{Mmax}\}$. However, this design might cause $actions\epsilon\{n_{M+1}, n_{M+2}, \ldots, n_{Mmax}\}$ to be selected, where the DCs denoted by these *actions* are nonexistent. On the other hand, VNF-SC generally needs to place VNFs sequentially in a path, but DRL-agent would not naturally select DCs in sequence. Both these two conditions would make the VNFC being handled rejected, which leads to the blockage of the whole VNF-SC, and for this, we will give DRL agent a large penalty. Inversely, if the DRL agent selects a feasible DC, the required VNF will be deployed or reused, and Algorithm 2 will be applied for the following spectrum resource allocation. For this, we will set the negative value of occupied resources by the VNFC, which could be derived based on Eq. (1), as the reward of the step.

C. Training Mechanism

Fig. 4 shows the training process of the DDQN mechanism. DDQN includes two neural networks. The neural network with the latest parameters that takes charge of the action selection is called the online network. The neural network with parameters from a period of time ago that conducts Q-value evaluation is called the target network. Decoupling action selection and Q-value evaluation, DDQN can efficiently enhance stability and mitigate overestimation during model training.

For the training process of a DDQN model, the loss function, which is used to indicate the estimation performance, is defined in Eq. (2) as follows:

$$loss = E\Big[\Big(r_t+\gamma Q\Big(\mathbf{S}_{t+1}, \arg\max_a Q(\mathbf{S}_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t^-\Big) - Q(\mathbf{S}_t, a_t; \boldsymbol{\theta}_t)\Big)^2\Big] \quad (2)$$

we use $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t^-$ to represent the parameters of online network and target network at time t, respectively. Specially, the first part $r_t+\gamma Q(\mathbf{S}_{t+1}, \arg\max_a Q(\boldsymbol{S}_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t^-)$

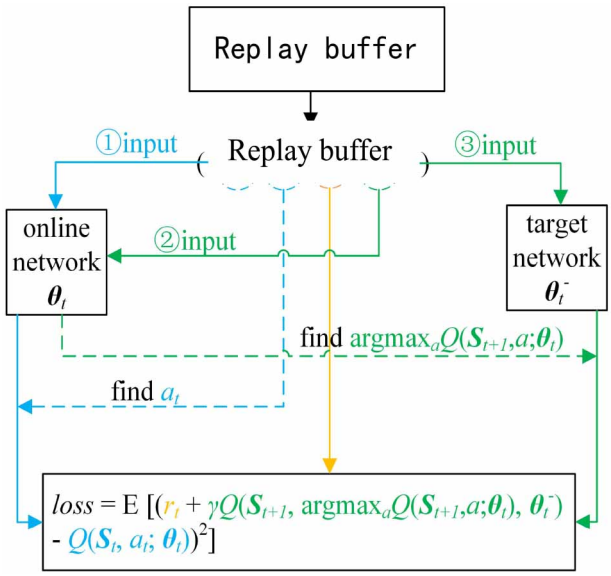

Fig. 4. The training process of DDQN mechanism.

denotes a target that the Q-value needs to move, where γ is the discount rate, and the second part $Q(\boldsymbol{S}_t, a_t; \boldsymbol{\theta}_t)$ denotes the estimation of Q-value. Note that, in Eq. (2), the first part needs to be calculated in two steps. The first step is to calculate $\arg\max_a Q(\boldsymbol{S}_{t+1}, a; \boldsymbol{\theta}_t)$, which means that the action a with the highest Q-value according to online network under state $\boldsymbol{S}_{t+1}$. The second part is to evaluate the Q-value of action a under state $\boldsymbol{S}_{t+1}$ using the target network. Therefore, the loss function reflects the estimation error of a DDQN model, and the smaller the loss function is, the better estimation performance a DDQN model will be. To train the DDQN model, a mini-batch of sequences $(s_t,\ a_t,\ r_t,\ s_{t+1})$ is firstly selected from replay buffer, then gradient descent algorithm is utilized to update the weights of online neural network, which is also to optimize the loss function in Eq. (2), as follows:

$$\boldsymbol{\theta}_t' = \boldsymbol{\theta}_t + \alpha\Big(r_t + \gamma Q\Big(\boldsymbol{S}_{t+1}, \arg\max_a Q(\boldsymbol{S}_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t^-\Big) - Q(\mathbf{S}_t, a_t; \boldsymbol{\theta}_t)\Big)\nabla Q(\mathbf{S}_t, a_t; \boldsymbol{\theta}_t) \tag{3}$$

where $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t'$ indicate online network before and after update and α is the step-size parameter.

The DRL-based VNF-SC provisioning algorithm is summarized in Algorithm 3. In the beginning, the agent initiates an empty experience buffer Λ with capacity A and sets exploration rate ε and discount rate γ. In lines 2, we initialize DNNs online network and target network with parameters $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t^-$, respectively, and set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^-$. The DNNs are trained iteratively. In each training iteration, a state $\boldsymbol{s}_t$ is initialized (lines 3-4). In lines 5-6, we find the LBP for each VNF-SC. To facilitate the decisions of DRL-agent, line 7 decomposes a VNF-SC request into J VNFC(s). After that, we generate $\boldsymbol{S}_t$ with the proposed encoding scheme in line 8. In lines 9-18, for each VNFC, we first select a_t with ε-greedy strategy, and if a_t is executable, we will deploy or reuse VNFs, then allocate FSs requested by the VNFC with Algorithm 2, otherwise, we return a large penalty for this action. Lines 19-21 show the training phase when mini-batch of sequences $(\boldsymbol{S}_j, a_j, r_j, \boldsymbol{S}_{j+1})$ is selected randomly and used to update $\boldsymbol{\theta}_t$ with Eq. (3). Specially, we will set target network parameters $\boldsymbol{\theta}_t^-$ as online network parameters $\boldsymbol{\theta}_t$ periodically.

Algorithm 3 Deep Q-Learning Based VNF-SC Provision Algorithm

Require: discount rate γ, exploration rate ε, memory capacity A for experience replay;

1: Initialize replay buffer Λ with capacity A;
2: Initialize online and target networks with random weights $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^-$
3: **for** each iteration **do**
4: Initialize state $\boldsymbol{S}_t$ with the proposed encoding scheme;
5: **for** each VNF-SC request R **do**
6: Find p^{LCS} with **Algorithm 1**;
7: Decompose R into J VNFC(s);
8: Generate $\boldsymbol{S}_t$ with proposed encoding scheme;
9: **for** each VNFC **do**
10: With probability ε select a random action a_t, otherwise select $a_t = \arg\max_a Q(\mathbf{S}_t, a; \boldsymbol{\theta}_t)$
11: **if** a_t is executable **then**
12: Deploy or reuse VNFs and allocate FSs required by the VNFC with **Algorithm 2**;
13: Get reward r_t and new state $\boldsymbol{S}_{t+1}$;
14: **else**
15: Get penalty r_t and new state $\boldsymbol{S}_{t+1}$;
16: **end if**
17: Record sequence $(\boldsymbol{S}_t, a_t, r_t, \boldsymbol{S}_{t+1})$ in Λ;
18: **end for**
19: Randomly select mini-batch of sequences $(\boldsymbol{S}_j, a_j, r_j, \boldsymbol{S}_{j+1})$ from Λ;
20: Update $\boldsymbol{\theta}_t$ with Eq. (3);
21: Periodically reset $\boldsymbol{\theta}_t^- = \boldsymbol{\theta}_t$
22: **end for**
23: **end for**

V. Performance Evaluation

A. Simulation Setup

Considering the applicability and generality of our proposed VNF-SC provisioning algorithm, we evaluate the performance with the 28-node US Backbone topology and the 14-node NSFNET topology as shown in Fig. 5. In the 28-node US Backbone topology, the number of nodes that have local DCs is set to be 15 and these DC nodes are selected randomly from the topology. We assume that there are $|T| = 8$ types of VNFs and 250 FSs in each link. The number of VNF requests is randomly distributed between 1 and 3, the initial bandwidth requests are 6, 7, or 8 FSs and when data go through different kinds of VNFs, the bandwidth request will increase or decrease according to its type. In the 14-node NSFNET topology, the number of nodes that have local DCs is set to 8. Besides, we set $|T| = 6$ and there are 200 FSs in each fiber link due to the smaller network size.

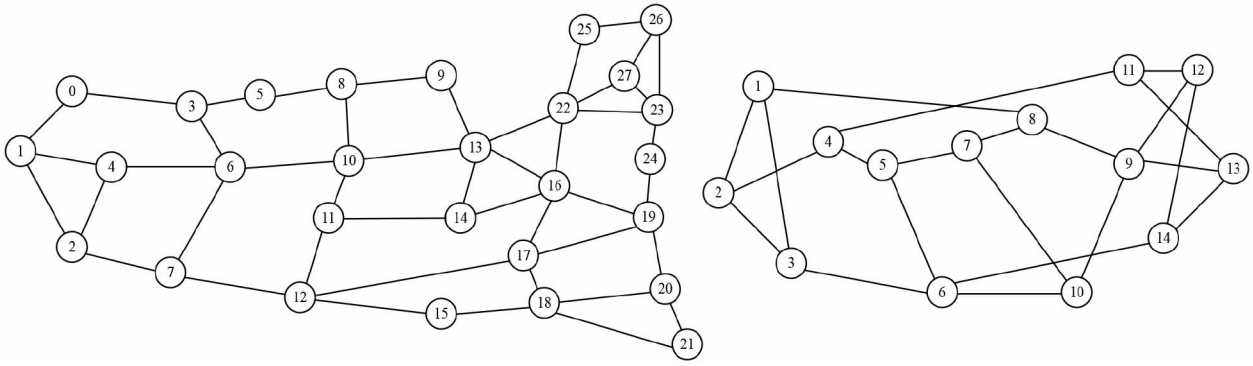

Fig. 5. 28-node US Backbone topology and 14-node NSFNET topology.

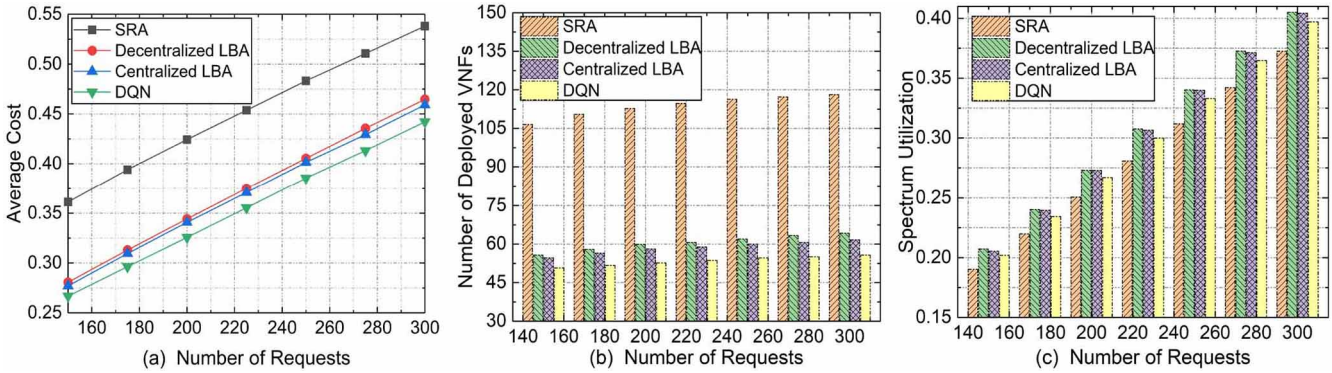

Fig. 6. (a) Average Cost, (b) Number of deployed VNFs and (c) Spectrum utilization in the US Backbone.

Eq. (1) generally uses $\alpha \leq \beta$ because the spectrum resources are more difficult to be optimized than IT resources. It is because that the spectrum resource allocation is not only affected by the DRL's decision, and more importantly, also affected by the scheduling of Algorithm 2. Meanwhile, the IT resource allocation is only determined by the DRL's decision, and hence the feedback of IT resource allocation would be more direct. In addition, the spectrum resource allocation is also affected by many other factors, such as the network topology and the already allocated spectrum resources, which makes the effective spectrum resource allocation more difficult. The proposed DRL-based algorithm is compared against some benchmark algorithms. The first one is the shortest-path and random VNF deployment algorithm (SRA), which selects the shortest path and maps VNFs randomly [33]. The second and the third are decentralized and centralized LBA [24], which choose the most decentralized and centralized LCSs to reuse VNFs as many as possible, respectively. Specifically, centralization means more VNFs are deployed in the same DCs.

B. Results and Discussions

Fig. 6 (a) shows the results of the average cost (AC) calculated by Eq. (1) in the US backbone. As shown, AC increases linearly with the number of requests because more IT and spectrum resources are required for provisioning more requests (as indicated by Fig. 6(b) and Fig. 6(c)). Furthermore, from Fig. 6(a), it can be observed that the proposed DQN algorithm achieves the lowest AC compared with the benchmark heuristics, which proves the effectiveness of the proposed DQN algorithm. To reveal the fundamental reasons, we compare the results of IT resource consumption and spectrum resource consumption in Fig. 6(b) and Fig. 6(c), respectively. In Fig. 6(b), with the increasing number of requests, all four algorithms consume more IT resources, where IT resources consumption is the lowest with the proposed DQN. SRA consumes the most IT resources, because it lacks the ability to proactively reuse the already deployed VNFs. Simulation demonstrated that the reuse of the existing VNFs can reduce the required number of VNFs significantly, and the performance of the decentralized/centralized LBA approximate to that of the proposed DQN. Nonetheless, with DRL training that aims at minimizing the AC, DQN deploys 13.2% and 9.6% less VNFs than decentralized/centralized LBA on average. Fig. 6 (c) shows the utilization of spectrum resources in the US backbone. Thanks to the DRL training, the proposed DQN algorithm still outperforms the decentralized/centralized LBA by 2.3% and 2.0% on average. Note that SRA outperforms other algorithms in terms of spectrum resource consumption because it always chooses the shortest path. However, the overall AC of SRA remains the worst because of too much wastage of IT resources. It is because that DQN achieves a better tradeoff between VNF reuse and spectrum utilization by the aid of the effective DRL training.

To further validate the effectiveness of our proposed DQN algorithm, the simulation results under the NSFNET is shown in Fig. 7 in terms of AC (Fig. 7(a)), IT resource consumption

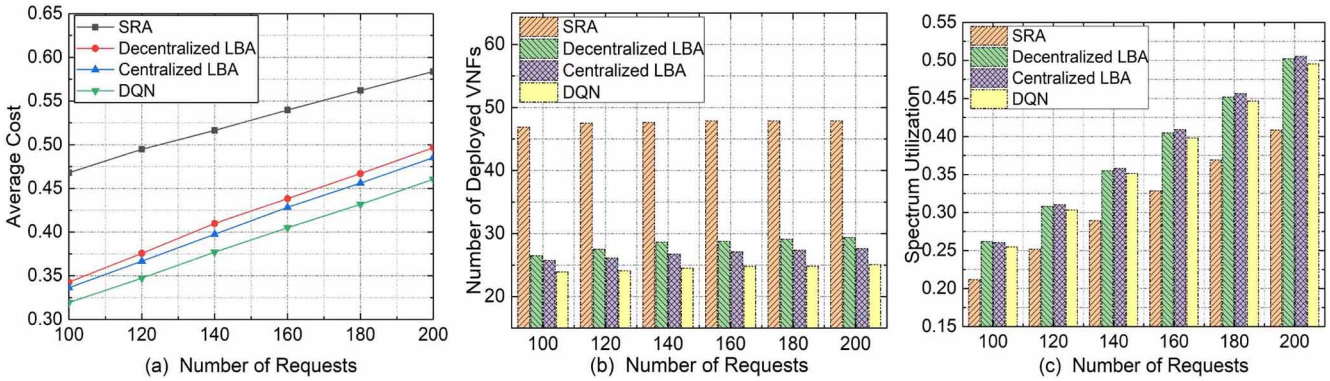

Fig. 7. (a) Average Cost, (b) Number of deployed VNFs and (c) Spectrum utilization in the NSFNET.

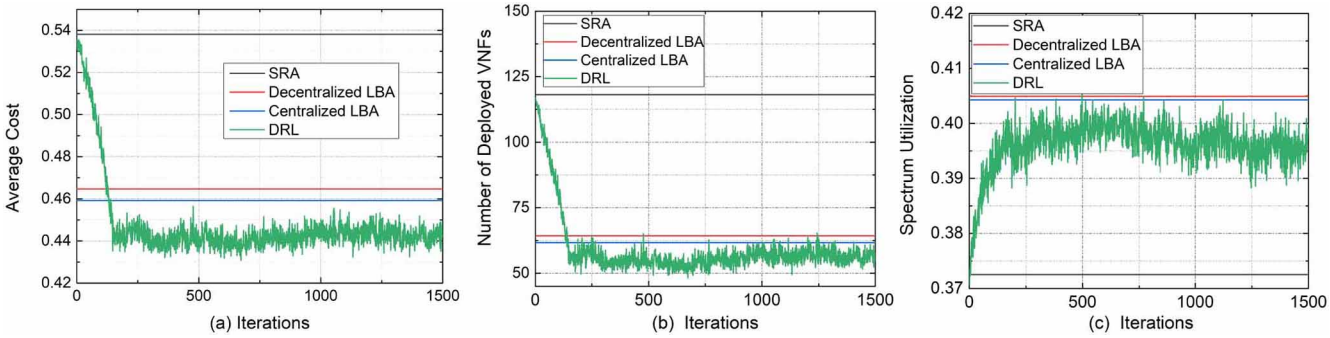

Fig. 8. Training results of RL vs. benchmark heuristics in the US Backbone: (a) Average Cost, (b) Number of deployed VNFs and (c) Spectrum utilization.

(Fig. 7(b)) and spectrum utilization (Fig. 7(c)). In Fig. 7(a), the proposed DQN achieves the lowest AC. In Fig. 7(b), DQN consumes 16.3% and 9.5% less VNFs than decentralized LBA and centralized LBA. In Fig. 7(c), besides SRA, the proposed DQN consumes 1.6% and 2.2% less spectrum resource than decentralized LBA and centralized LBA. These results are similar to those in Fig. 6, which suggest that the proposed DQN algorithm can well adapt to different network scenarios.

Fig. 8 shows the training process of DQN in US backbone with iteration-by-iteration AC (Fig. 8(a)), IT resource consumption (Fig. 8(b)) and spectrum utilization (Fig. 8(c)). It demonstrates how the DQN is able to adapt during the DRL training phase. At the beginning, the DRL agent has no prior knowledge about the system. It randomly selects DCs for VNF deployment, and hence the results of DQN in Figs. 8(a), (b) and (c) are similar to SRA. With the DRL training, the agent quickly learns that it can significantly reduce AC by reusing existing VNF instances in DC. By iteration 150, the number of deployed VNFs obtained by DQN is lower than other benchmark heuristics. Meanwhile, during the same training period (from iteration 1 to 150), the proposed DQN algorithm consumes more spectrum resources as shown in Fig. 8(b). It is because to gain higher AC, the DRL agent decide to trade some of the spectrum resource for saving more IT resources (i.e., the trade-off problem described in Section III). After iteration 150, the training curves of DRL in Figs. 8(a), (b) and (c) become flatten and converged. Note that after convergence, the DRL still consumes less spectrum resources than centralized LBA and decentralized LBA. It demonstrates that the proposed DRL is able to reduce the consumption of IT and spectrum resources, simultaneously, which is why the proposed DRL can achieve the lowest AC. Although LBAs also reuse VNFs proactively, it cannot gain better results because it only considers the DC that hosts the most or the least VNF instances in LBP, while the DRL-based algorithm can choose from among all the DCs in LBP flexibly to make a better provisioning decision.

Fig. 9 shows the training process of DQN in NSFNET with iteration-by-iteration AC (Fig. 9(a)), IT resource consumption (Fig. 9(b)) and spectrum utilization (Fig. 9(c)). It is observed that by iteration 150, the number of deployed VNFs obtained by DQN is lower than the other benchmark heuristics and the training curves in Figs. 9(a)–(c) gradually converges after iteration 150. Compared with Fig. 8, we notice that the convergence speed in Fig. 9 is basically unaffected by the size of the network scale. It could be a result of the proposed VNF-SC decomposition scheme and feature matrix-based encoding scheme, which reduce the size of the state space and action space, and thereby accelerates the DRL training process.

VI. Conclusion

This paper presents a DRL-based VNF-SC provisioning algorithm in inter-DC EON. We elaborate that the joint allocation of IT and spectrum resources is essential for the VNF-SC provisioning, and based on it, the optimization objective that considers both the deployed VNFs and the spectrum utilization is formulated to evaluate the performance of the proposed algorithm. To illustrate the necessity of optimizing

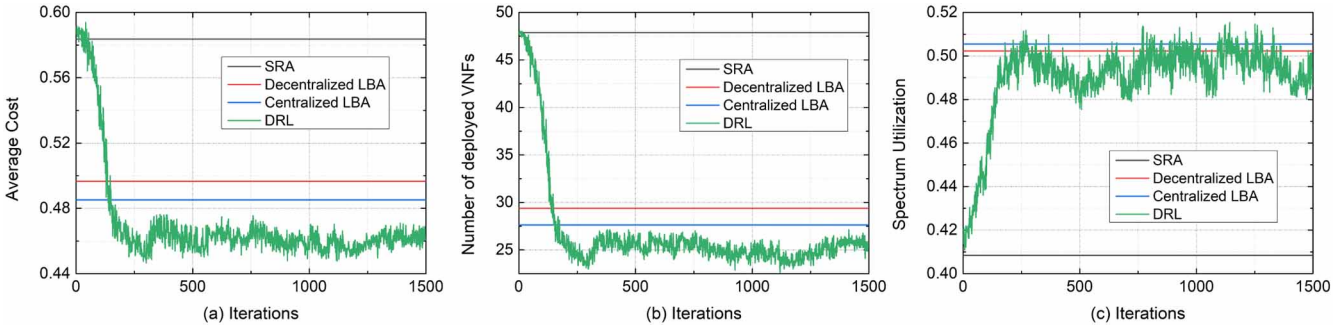

Fig. 9. Training results of RL vs. benchmark heuristics in the NSFNET: (a) Average Cost, (b) Number of deployed VNFs and (c) Spectrum utilization.

VNF selection, an intuitive example that provisions VNF-SCs in a 7-node network is introduced. For our proposed DRL-based algorithm, in order to solve the problem that state representation is not in a fixed form due to the variable length of VNF-SCs and the constantly changing number of available DCs, a feature matrix-based encoding scheme is introduced. Then, we propose the decomposition of VNF-SC into several VNFCs. With it, any VNF-SC could be provisioned within finite steps. Finally, to improve the accuracy of the proposed algorithm, DDQN that decouples action selection and Q-value evaluation is applied.

Both large network topology (US Backbone) and small network topology (NFSNET) of physical networks are considered in the simulations. The simulation results show that the proposed algorithm can achieve better performance than the benchmark heuristics and have great adaptability for different network topologies.

REFERENCES

[1] "Network Functions Virtualization (NFV)." Jan. 2012. [Online]. Available: https://portal.etsi.org/portal/server.pt/community/NFV/367

[2] M. S. Yoon and A. E. Kamal, "NFV resource allocation using mixed queuing network model," in *Proc. GLOBECOM*, 2016, pp. 1–6.

[3] M. Xia, M. Shirazipour, Y. Zhang, H. Green, and A. Takacs, "Network function placement for NFV chaining in packet/optical datacenters," *J. Lightw. Technol.*, vol. 33, no. 8, pp. 1565–1570, Apr. 15, 2015.

[4] M. Sasabe and T. Hara, "Shortest path tour problem based integer linear programming for service chaining in NFV networks," in *Proc. 6th IEEE Conf. Netw. Softwarization (NetSoft)*, 2020, pp. 114–121.

[5] D. Zheng, H. Gu, W. Wei, C. Peng, and X. Cao, "Network service chaining and embedding with provable bounds," *IEEE Internet Things J.*, vol. 8, no. 9, pp. 7140–7151, May 2021.

[6] D. Li, P. Hong, K. Xue, and J. Pei, "Availability aware VNF deployment in datacenter through shared redundancy and multi-tenancy," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 4, pp. 1651–1664, Dec. 2019.

[7] M. M. Tajiki, S. Salsano, L. Chiaraviglio, M. Shojafar, and B. Akbari, "Joint energy efficient and QoS-aware path allocation and VNF placement for service function chaining," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 1, pp. 374–388, Mar. 2019.

[8] H. N. Tan, T. Inoue, K. Tanizawa, T. Kurosu, and S. Namiki, "All-optical Nyquist filtering for elastic OTDM signals and their spectral defragmentation for inter-datacenter networks," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2014, pp. 1–3.

[9] Y. Ou, A. Hammad, S. Peng, R. Nejabati, and D. Simeonidou, "Online and offline virtualization of optical transceiver," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 7, no. 8, pp. 748–760, Aug. 2015.

[10] A. Muhammad, N. Skorin-Kapov, and L. Wosinska, "Manycast and anycast routing for replica placement in datacenter networks," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2015, pp. 1–3.

[11] Y. Ou *et al.*, "Demonstration of virtualizeable and software-definedoptical transceiver," *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1916–1924, Apr. 15, 2016.

[12] S. Clayman, E. Maini, A. Galis, A. Manzalini, and N. Mazzocca, "The dynamic placement of virtual network functions," in *Proc. IEEE Netw. Oper. Manage. Symp. (NOMS)*, 2014, pp. 1–9.

[13] W. Lu, L. Liang, and Z. Zhu, "On vNF-SC deployment and task scheduling for bulk-data transfers in inter-DC EONs," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2017, pp. 1–4.

[14] Z. Zhu *et al.*, "Demonstration of cooperative resource allocation in an OpenFlow-controlled multidomain and multinational SD-EON testbed," *J. Lightw. Technol.*, vol. 33, no. 8, pp. 1508–1514, Apr. 15, 2015.

[15] B. Li, W. Lu, S. Liu, and Z. Zhu, "Designing deep learning model for accurate vNF service chain pre-deployment in inter-DC EONs," in *Proc. Asia Commun. Photon. Conf. (ACP)*, Hangzhou, China, 2018, pp. 1–3.

[16] T.-W. Kuo, B.-H. Liou, K. C.-J. Lin, and M.-J. Tsai, "Deploying chains of virtual network functions: On the relation between link and server usage," *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1562–1576, Aug. 2018.

[17] P. Jin, X. Fei, Q. Zhang, F. Liu, and B. Li, "Latency-aware VNF chain deployment with efficient resource reuse at network edge," in *Proc. INFOCOM Conf. Comput. Commun.*, 2020, pp. 267–276.

[18] G. Sun, R. Zhou, J. Sun, H. Yu, and A. V. Vasilakos, "Energy-efficient provisioning for service function chains to support delay-sensitive applications in network function virtualization," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6116–6131, Jul. 2020.

[19] T. Gao, X. Li, W. Zou, and S. Huang, "Survivable VNF placement and scheduling with multipath protection in elastic optical datacenter networks," in *Proc. Opt. Fiber Commun. Conf. Exhibition (OFC)*, 2019, Art. no. Th3J.2.

[20] A. Khatiri and G. Mirjalily, "Resource balanced service chaining in NFV-enabled inter-datacenter elastic optical networks," in *Proc. 12th Int. Conf. Knowl. Smart Technol. (KST)*, 2020, pp. 168–171.

[21] Y. Li *et al.*, "Joint balancing of IT and spectrum resources for selecting virtualized network function in inter-datacenter elastic optical networks," *Opt. Exp.*, vol. 27, no. 11, pp. 15116–15128, 2019.

[22] M. Zeng, W. Fang, J. J. P. C. Rodrigues, and Z. Zhu, "Orchestrating multicast-oriented NFV trees in inter-DC elastic optical networks," in *Proc. ICC*, 2016, pp. 1–6.

[23] S. Zhao and Z. Zhu, "On virtual network reconfiguration in hybrid optical/electrical datacenter networks," *J. Lightw. Technol.*, vol. 38, no. 23, pp. 6424–6436, Dec. 2020.

[24] W. Fang, M. Zeng, X. Liu, W. Lu, and Z. Zhu, "Joint spectrum and IT resource allocation for efficient VNF service chaining in inter-datacenter elastic optical networks," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1539–1542, Aug. 2016.

[25] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," in *Proc. NIPS Deep Learn. Workshop*, 2013, pp. 1–9.

[26] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[27] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. HotNets*, New York, NY, USA, Nov. 2016, pp. 50–56.

[28] X. Fu, F. R. Yu, J. Wang, Q. Qi, and J. Liao, "Dynamic service function chain embedding for NFV-enabled IoT: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 507–519, Jan. 2020.
[29] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. Yoo, "DeepRMSA: A deep reinforcement learning framework for routing modulation and spectrum assignment in elastic optical networks," *J. Lightw. Technol.*, vol. 37, pp. 4155–4163, Aug. 15, 2019.
[30] K. Christodoulopoulos, I. Tomkos, and E. A. Varvarigos, "Elastic bandwidth allocation in flexible OFDM-based optical networks," *J. Lightw. Technol.*, vol. 29, no. 9, pp. 1354–1366, May 1, 2011.
[31] O. Houidi, O. Soualah, W. Louati, and D. Zeghlache, "Dynamic VNF forwarding graph extension algorithms," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 3, pp. 1389–1402, Sep. 2020.
[32] D. Zhao, Q. Zhang, D. Wang, and Y. Zhu, "Experience replay for optimal control of nonzero-sum game systems with unknown dynamics," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 854–865, Mar. 2016.
[33] P. Lu, L. Zhang, X. Liu, J. Yao, and Z. Zhu, "Highly efficient data migration and backup for big data applications in elastic optical inter-data-center networks," *IEEE Netw.*, vol. 29, no. 5, pp. 36–42, Sep./Oct. 2015.

Min Zhu (Member, IEEE) is currently an Associate Professor and a Doctoral Supervisor with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China, and the Principal Investigator of Purple Mountain Laboratories, Nanjing, China. He has published more than 100 articles in refereed journals and conferences of IEEE and OSA. His research interests include design, optimization and performance analysis of 5G RAN transport networks, network function virtualization, software-defined radio, and optical access networks, with emphasis on applications of deep reinforcement learning to networking.

Qi Chen (Graduate Student Member, IEEE) received the B.Sc. degree from the School of Information Science and Engineering, Southeast University, Nanjing, China, in June 2021. He is currently pursuing the M.S. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include network function virtualization and machine learning.

Jiahua Gu (Graduate Student Member, IEEE) received the B.Sc. degree from the School of Electrical Engineering and Automation, Nanjing Normal University, Nanjing, China, in June 2016. He is currently pursuing the Ph.D. degree with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, and the Pervasive Communication Research Center, Purple Mountain Laboratories, Nanjing. His research interests include C-RAN and machine learning.

Pingping Gu is from Taicang T&W Electronics Company Ltd., Taicang City, which is established on April 1, 2008. Its business scope includes R&D, production, processing and sales of broadband communication equipment, wireless communication equipment, network equipment, set-top box, computer board, and adapter.