

# DRL-based Slice Placement Under Non-Stationary Conditions

Jose Jurandir Alves Esteves<sup>\*†</sup>, Amina Boubendir<sup>\*</sup>, Fabrice Guillemin<sup>\*</sup> and Pierre Sens<sup>†</sup>

<sup>\*</sup>Orange Labs, France

<sup>†</sup>Sorbonne Université / CNRS / Inria, LIP6, France

{josejurandir.alvesesteves, amina.boubendir, fabrice.guillemin}@orange.com, pierre.sens@lip6.fr

**Abstract**—We consider online learning for optimal network slice placement under the assumption that slice requests arrive according to a non-stationary Poisson process. We propose a framework based on Deep Reinforcement Learning (DRL) combined with a heuristic to design algorithms. We specifically design two pure-DRL algorithms and two families of hybrid DRL-heuristic algorithms. To validate their performance, we perform extensive simulations in the context of a large-scale operator infrastructure. The evaluation results show that the proposed hybrid DRL-heuristic algorithms require three orders of magnitude of learning episodes less than pure-DRL to achieve convergence. This result indicates that the proposed hybrid DRL-heuristic approach is more reliable than pure-DRL in a real non-stationary network scenario.

**Index Terms**—Network Slicing, Deep Reinforcement Learning, Placement, Optimization.

## I. INTRODUCTION

The promise of network Slicing is to enable a high level of customization of network services in future networks (5G and beyond) leveraged by virtualization and software defined networking techniques. These key enablers transform telecommunications networks into programmable platforms capable of offering virtual networks enriched by Virtual Network Functions (VNFs) and IT resources tailored to the specific needs of certain customers (e.g., companies) or vertical markets (automotive, e-health, etc.) [1], [2].

From an optimization theory perspective, the Network Slice Placement problem can be viewed as a specific case of Virtual Network Embedding (VNE) or VNF Forwarding Graph Embedding (VNF-FGE) problems [3]. It is then generally possible to formulate Integer Linear Programming (ILP) problems [4], which, however, turn out to be  $\mathcal{NP}$ -hard [5] with very long convergence time.

With regard to network management, there are specific characteristics related to network slicing: slices are expected to share resources and coexist in a large and distributed infrastructure. Moreover, slices have a wide range of requirements in terms of resources, quality objectives and lifetime. In practice, these characteristics bring additional complexity as the placement algorithms need to be highly scalable with low response time even under varying network conditions.

As an alternative to optimization techniques and the development of heuristic methods, Deep Reinforcement Learning (DRL) has recently been used in the context of VNE and Network Slice Placement [6]–[13].

DRL techniques are seen as very promising since they allow, at least theoretically, to learn optimal decision policies only based on experience [14]. However, from a practical point of view, especially in the context of non-stationary environments, ensuring that a DRL agent converges to an optimal policy is still a challenge.

As a matter of fact, when the environment is continually changing the rules, the algorithm has trouble in using the acquired knowledge to find optimal solutions. The usage of the DRL algorithm in an online fashion can then become impractical. In fact, most of the existing works applying DRL to solve the Network Slice Placement or VNE problem assume a stationary environment, i.e., with static network load. However, traffic conditions in networks are basically non-stationary with daily and weekly variations and subject to drastic changes (e.g., traffic storm due to an unpredictable event).

To cope with traffic changes, we propose in the present paper to extend a hybrid DRL-heuristic algorithm we have recently introduced in [15] (namely Heuristically Assisted DRL, HA-DRL) to evaluate the performance under non-stationary network loads. We apply this strategy to a fully online learning scenario with time-varying network loads to show how this strategy can be used to accelerate and stabilize the convergence of DRL techniques when applied to the Network Slice Placement problem.

The contributions of the present paper are threefold:

- 1) We propose a network load model to network slice infrastructure conditions with time-varying network loads;
- 2) We propose a framework combining Advantage Actor Critic and a Graph Convolutional Network (GCN) for conceiving DRL-based algorithms adapted to the non-stationary case;
- 3) We show how the DRL learning process can be accelerated by using the proposed HA-DRL technique to control the algorithms convergence.

The organization of this paper is as follows: In Section II, we review related work. In Section III, we describe the Network Slice Placement problem modeling. The learning framework for slice placement optimization is described in Section IV. The adaptation of the pure-DRL approaches and its control by using heuristic is introduced in Section IV-B.

The experiments and evaluation results are presented in Section V, while conclusions and perspectives are presented in Section VI.

## II. RELATED WORK ANALYSIS

We review, in this section, recent studies on DRL-based approaches for network slice placement. The reader interested in a more detailed and comprehensive discussion of those works cited in the present paper may refer to [15].

### A. On Pure-DRL approaches for slice placement

There are only a few recent works on DRL for network slice placement and VNE related problems in the literature and the majority of them are pure-DRL approaches [6]–[13]. In those works, only the knowledge acquired by the learning agent via training is used as a basis for taking placement decisions. The drawback of this approach is that the learning agent needs extensive exploration of the state and action spaces to learn an appropriate policy for decision making; such a process takes a lot of time during which the agent takes bad placement decisions; this leads to rejection of slices and bad performance.

Furthermore, existing works consider static or fixed average network load regimes in which network slices arrive and exit the system always at fixed rates. In reality, however, network load conditions vary over time as the demand for network services depend on many factors (unpredictable events, day-night variations, etc.). In this kind of non-stationary environment, the learning agent might have trouble in using its knowledge acquired via training as the rules of the environment are constantly changing.

### B. On Hybrid DRL-heuristic approaches for slice placement

Two recent works on network slice placement and VNE have been proposed and combine DRL with heuristic methods to speed up the convergence and to increase the reliability of DRL algorithms [16], [17]. However, these hybrid DRL-heuristic approaches have some drawbacks explained in more details in [15]. In particular, the approach proposed in [16] adopts an infinite action space formulation that adds some overhead reducing the applicability of the algorithm, and the approach proposed in [17] strongly depends on the quality of the initial solution provided by a heuristic. Moreover, these two works do not consider the non-stationarity assumption discussed in the present paper.

In [15], we proposed to adapt heuristically accelerated reinforcement learning to the network slice placement problem. The approach is inspired by [18] and addresses the shortcomings of [16] and [17]. In this paper, we extend this approach by including the non-stationarity assumption.

### C. On AI/ML approaches considering dynamic network load

A recent body of research considers dynamic network load scenarios when applying AI/ML-based approaches to support optimization of network slice life cycle management. For instance, the authors of [19] propose a deep learning-based data analytics tool to predict dynamic network slice traffic demands to help avoid SLA violations and network over-provisioning. The paper [20] also adopts neural networks to predict network slice traffic demands but, in this case, to perform proactive resource provisioning and congestion control.

To the best of our knowledge, the only paper considering placement optimization using DRL in dynamic network load scenarios is [21].

The authors propose a Double Deep Q Network (DDQN) algorithm for re-optimizing an initial VNF placement. They consider that the network load changes periodically with a time cycle  $T$ . They then separate time cycle  $T$  in time intervals  $\Delta t$ , and train a DDQN model to specifically take charge of VNF placement re-optimization in each time interval  $\Delta t$ . Despite its originality, the approach proposed in [21] presents two drawbacks: 1) it depends on offline learning, which is not applicable to online optimization scenarios; 2) it does not use DRL to optimize placement directly as the DRL algorithm selects the region of the network to re-optimize and delegates the optimization to threshold policy procedure. The heuristic calculation of placement decisions can lead to sub-optimal solutions. Contrary to [21], the present contribution is applicable to offline and online learning and directly learns a placement optimization policy.

## III. NETWORK SLICE PLACEMENT OPTIMIZATION PROBLEM

This section presents the various elements composing the model for slice placement. Slices are placed on a substrate network, referred to as Physical Network Substrate (PSN) and described in Section III-A. Slices give rise to as Network Slice Placement Requests (Section III-B), generating a network load defined in Section III-C. The optimization problem is formulated in Section III-D.

### A. Physical Substrate Network Modeling

The Physical Substrate Network (PSN) is composed of the infrastructure resources, namely IT resources (CPU, RAM, disk, etc.) needed for supporting the Virtual Network Functions (VNFs) of network slices together with the transport network, in particular Virtual Links (VLs) for interconnecting the VNFs of slices. The PSN is divided into three components: the Virtualized Infrastructure (VI) corresponding to IT resources, the Access Network (AN), and the Transport Network (TN).

The Virtual Infrastructure (VI) hosting IT resources is the set of Data Centers (DCs) interconnected by network elements (switches and routers). We assume that data centers are distributed in Points of Presence (PoP) or centralized (e.g., in a big cloud platform). As in [22], we define three types of DCs with different capacities: Edge Data Centers (EDCs) close to end users but with small resources capacities, Core Data Centers (CDCs) as regional DCs with medium resource capacities, and Central Cloud Platforms (CCPs) as national DCs with big resource capacities.

We consider that slices are rooted so as to take into account the location of those users of a slice. We thus introduce an Access Network (AN) representing User Access Points (UAPs) such as Wi-Fi APs, antennas of cellular networks, etc. and Access Links. Users access slices via one UAP, which may change during the life time of a communication by a user (e.g., because of mobility).

The Transport Network (TN) is the set of routers and transmission links needed to interconnect the different DCs and the UAPs. The complete PSN is modeled as a weighted undirected graph  $G_s = (N, L)$  with parameters described in Table I, where  $N$  is the set of physical nodes in the PSN, and  $L \subset \{(a, b) \in N \times N \wedge a \neq b\}$  refers to a set of substrate links. Each node has a type in the set {UAP, router, switch, server}. The available CPU and RAM capacities on each node are defined as  $cap_n^{cpu} \in \mathbb{R}$ ,  $cap_n^{ram} \in \mathbb{R}$  for all  $n \in N$ , respectively. The available bandwidth on the links are defined as  $cap_{(a,b)}^{bw} \in \mathbb{R}, \forall (a, b) \in L$ .

TABLE I  
PSN PARAMETERS

Parameter	Description
$G_s = (N, L)$	PSN graph
$N$	Network nodes
$S \subset N$	Set of servers
$DC$	Set of data centers
$S_{dc} \subset S, \forall dc \in DC$	Set of servers in data center $dc$
$SW_{dc}, \forall dc \in DC$	Switch of of data center $dc$
$L = \{(a, b) \in N \times N \wedge a \neq b\}$	Set of physical links
$cap_{(a,b)}^{bw} \in \mathbb{R}, \forall (a, b) \in L$	Bandwidth capacity of link $(a, b)$
$cap_s^{cpu} \in \mathbb{R}, \forall s \in S$	available CPU capacity on server $s$
$M_s^{cpu} \in \mathbb{R}, \forall s \in S$	maximum CPU capacity of server $s$
$cap_s^{ram} \in \mathbb{R}, \forall s \in S$	available RAM capacity on server $s$
$M_s^{ram} \in \mathbb{R}, \forall s \in S$	maximum RAM capacity of server $s$
$M_s^{bw} \in \mathbb{R}, \forall s \in S$	maximum outgoing bandwidth from $s$

### B. Network Slice Placement Requests Modeling

We consider that a slice is a chain of VNFs to be placed and connected over the PSN. VNFs of a slice are grouped into a request, namely a Network Slice Placement Request (NSPRs), which has to be placed on the PSN. A NSPR is represented as a weighted undirected graph  $G_v = (V, E)$ , with parameters described in Table II, where  $V$  is the set of VNFs in the NSPR, and  $E \subset \{(\bar{a}, \bar{b}) \in V \times V \wedge \bar{a} \neq \bar{b}\}$  is a set of VLs to interconnect the VNFs of the slice. The CPU and RAM requirements of each VNF of a NSPR are defined as  $req_v^{cpu} \in \mathbb{R}$  and  $req_v^{ram} \in \mathbb{R}$  for all  $v \in V$ , respectively. The bandwidth required by each VL in a NSPR is given by  $req_{(\bar{a}, \bar{b})}^{bw} \in \mathbb{R}$  for all  $(\bar{a}, \bar{b}) \in E$ .

TABLE II  
NSPR PARAMETERS

Parameter	Description
$G_v = (V, E)$	NSPR graph
$V$	Set of VNFs of the NSPR
$E = \{(\bar{a}, \bar{b}) \in N \times N \wedge \bar{a} \neq \bar{b}\}$	Set of VLs of the NSPR
$req_v^{cpu} \in \mathbb{R}$	CPU requirement of VNF $v$
$req_v^{ram} \in \mathbb{R}$	RAM requirement of VNF $v$
$req_{(\bar{a}, \bar{b})}^{bw} \in \mathbb{R}$	Bandwidth requirement of VL $(\bar{a}, \bar{b})$

### C. Network Load Modeling

We consider the case when the load offered by NSPRs is time varying. We specifically assume that there are several classes

of NSPRs and two sets of classes. A first set is composed of NSPR classes (referred to as static) with constant arrival rates, creating a background traffic. A second set is composed of NSPR classes, with time-varying arrival rates so as to reflect some volatility in the NSPRs. Those NSPRs are said dynamic.

1) *Network Load for Static NSPR Classes:* Let  $J$  be the set of resources in the network (i.e., CPU, RAM, bandwidth). Let  $\mathcal{K}_c \subset \mathbb{N}$  be the set of static NSPR classes. We compute the load generated by arrivals of NSPRs of class  $k$  in  $\mathcal{K}_c$  for resource  $j$  in  $J$  as in [23]:

$$\rho_j^k = \frac{1}{C_j} \frac{\lambda^k}{\mu^k} A_j^k, \quad (1)$$

where  $C_j$  is the total capacity of resource  $j$ ,  $A_j^k$  is the number of resource units requested by an NSPR of class  $k$ ,  $\lambda^k$  is the average arrival rate for an NSPR of class  $k$  and  $1/\mu^k$  is the average lifetime of an NSPR of class  $k$ .

2) *Network Load for Dynamic NSPR Classes:* Let  $\mathcal{K}_\Delta \subset \mathbb{N}$  be the set of dynamic NSPR classes. We consider a periodic average arrival rate  $\lambda^k(t)$  for class  $k$  in  $\mathcal{K}_\Delta$  given by

$$\lambda^k(t) = f^k \sin^2 \left( \frac{\pi t}{\tau^k} \right) \quad (2)$$

where  $\tau^k$  is the period of  $\lambda^k(t)$  in time units and  $f^k$  is a parameter used to control the amplitude of  $\lambda^k(t)$ . We then adapt Eq. (1) to compute the network load for dynamic NSPR class  $k$  and resource  $j$  as  $\rho_j^k(t) = \frac{1}{C_j} \frac{\lambda^k(t)}{\mu^k} A_j^k$ . It is worth noting that to preserve  $\rho_j^k(t)$  in the  $[0, 1]$  interval,  $f^k$  must be between 0 and  $\frac{C_j \mu^k}{A_j^k}$ .

3) *Global Network Load:* Finally, we define the global network load  $\rho_j(t)$  for each resource  $j$  in  $J$  in the simulated time instant  $t$  as the sum of the network loads generated by static NSPR classes ( $\mathcal{K}_c$ ) and dynamic NSPR classes ( $\mathcal{K}_\Delta$ ), that is,

$$\rho_j(t) = \sum_{k \in \mathcal{K}_\Delta} \rho_j^k(t) + \sum_{k \in \mathcal{K}_c} \rho_j^k \quad (3)$$

It is worth noting  $0 \leq \rho_j(t) \leq 1$ .

### D. Network Slice Placement Optimization Problem Statement

- *Given:* a NSPR graph  $G_v = (V, E)$  and a PSN graph  $G_s = (N, L)$ ,
- *Find:* a mapping  $G_v \rightarrow \bar{G}_s = (\bar{N}, \bar{L})$ ,  $\bar{N} \subset N$ ,  $\bar{L} \subset L$ ,
- *Subject to:* the VNF CPU requirements  $req_v^{cpu}, \forall v \in V$ , the VNF RAM requirements  $req_v^{ram}, \forall v \in V$ , the VLs bandwidth requirements  $req_{(\bar{a}, \bar{b})}^{bw}, \forall (\bar{a}, \bar{b}) \in E$ , the server CPU available capacity  $cap_s^{cpu}, \forall s \in S$ , the server RAM available capacity  $cap_s^{ram}, \forall s \in S$ , the physical link bandwidth available capacity  $cap_{(a,b)}^{bw}, \forall (a, b) \in L$ .
- *Objective:* maximize the network slice placement request acceptance ratio, minimize the total resource consumption and maximize load balancing.

#### IV. LEARNING FRAMEWORK FOR NETWORK SLICE PLACEMENT OPTIMIZATION

We describe in this section the machine learning framework used to solve the optimization formulated in Section III. We adopt the same approach as in [15] but to cope with the non stationary behavior of NSPR arrivals, we introduce an additional set of states to describe network load.

Other methods could be considered to deal with the cyclic nature of NSPR arrivals considered in this paper (e.g., LSTM techniques able to infer the periodic characteristics of the NSPRs arrival process). However, our goal in this paper is to set up a method of dealing with non stationary NSPR arrivals, independently of underlying periodic structures. We test the proposed approach for periodic NSPR arrivals to validate the approach, keeping in mind that the method has to be robust against non stationary traffic variations. This point will be addressed in further studies.

##### A. Learning framework

1) *Policy*: We reuse the framework introduced in [15]. We denote by  $\mathcal{A}$  the set of possible actions (namely placing VNFs on nodes) and by  $\mathcal{S}$  the set of all states. We adopt a sequential placement strategy so that we choose a node  $n \in N$  where to place a specific VNF  $v \in \{1, \dots, |V|\}$ . The VNFs are sequentially placed so that placement starts with the VNF  $v = 1$  and ends for the VNF  $v = |V|$ .

At each time step  $t$ , given a state  $\sigma_t$ , the learning agent selects an action  $a$  with probability given by the Softmax distribution given by

$$\pi_\theta(a_t = a | \sigma_t) = \frac{e^{Z_\theta(\sigma_t, a)}}{\sum_{b \in N} e^{Z_\theta(\sigma_t, b)}}, \quad (4)$$

where the function  $Z_\theta : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  yields a real value for each state and action calculated by a Deep Neural Network (DNN) as detailed in Section IV-B1. The notation  $\pi_\theta$  is used to indicate that policy depends on  $Z_\theta$ . The control parameter  $\theta$  represents the weights in the DNN.

2) *State representation*: As in [15], the **PSN state** is characterized by the occupancy of servers:  $cap^{cpu} = \{cap_n^{cpu} : n \in N\}$ ,  $cap^{ram} = \{cap_n^{ram} : n \in N\}$  and  $cap^{bw} = \{cap_n^{bw} = \sum_{(n,b) \in L} cap_{(n,b)}^{bw} : n \in N\}$ . In addition, we keep track of the placement of the outstanding NSPR (under placement) via the vector  $\chi = \{\chi_n \in \{0, \dots, |V|\} : n \in N\}$ , where  $\chi_n$  is the number of VNFs of the current NSPR placed on node  $n$ .

Accordingly, the **NSPR state** is a view of the current placement and is composed of four characteristics, three related to resource requirements (see Table II for the notation) of the current VNF  $v$  to be placed:  $req_v^{cpu}$ ,  $req_v^{ram}$  and  $req_v^{bw} = \sum_{(v,\bar{b}) \in E} req_{(v,\bar{b})}^{bw}$ . In addition, let  $m_v = |V| - v + 1$  be the number of VNFs of the outstanding NSPR still to be placed.

To deal with the non stationary environment, we introduce the **Load state** that represents the network load forecast used to learn the network load variations. It is defined by a set  $\rho_j$  of 100 features for each resource  $j$  calculated using the network

load formula given by Eq. (3) as follows :  $\rho_j = \{\rho_j(t) : t_a \leq t \leq t_a + 100, t \in \mathbb{N}\}$ , where  $t_a$  is the simulated time instant in which the current NSPR arrives.

3) *Reward function*: We reuse the reward function introduced in [15]. We precisely consider

$$r_{t+1} = \begin{cases} 0, & \text{if } t < T \text{ and } a_t \text{ is successful} \\ \sum_{i=0}^T \delta_{i+1}^a \delta_{i+1}^b \delta_{i+1}^c, & \text{if } t = T \text{ and } a_t \text{ is successful} \\ \delta_{t+1}^c, & \text{otherwise} \end{cases} \quad (5)$$

where  $T$  is the number of iterations of a training episode and where the rewards  $\delta_{i+1}^a$ ,  $\delta_{i+1}^b$ , and  $\delta_{i+1}^c$  are defined as follows:

- An Action  $a_t$  may lead to a successful or unsuccessful placement. We then define the Acceptance Reward value due to action  $a_t$  as

$$\delta_{t+1}^a = \begin{cases} 100, & \text{if } a_t \text{ is successful,} \\ -100, & \text{otherwise.} \end{cases} \quad (6)$$

- The Resource Consumption Reward value for the placement of VNF  $v$  via action  $a_t$  is defined by

$$\delta_{t+1}^c = \begin{cases} \frac{req_{(v-1,v)}^{bw}}{req_{(v-1,v)}^{bw}|P|} = \frac{1}{|P|}, & \text{if } |P| > 0, \\ 1, & \text{otherwise.} \end{cases} \quad (7)$$

where  $P$  is the path used to place VL  $(v-1, v)$ . Note that a maximum  $\delta_{t+1}^c = 1$  is given when  $|P| = 0$ , that is, when VNFs  $v-1$  and  $v$  are placed on the same server.

- The Load Balancing Reward value for the placement of VNF  $v$  via  $a_t$

$$\delta_{t+1}^b = \frac{cap_{a_t}^{cpu}}{M_{a_t}^{cpu}} + \frac{cap_{a_t}^{ram}}{M_{a_t}^{ram}}. \quad (8)$$

##### B. Adaptation of DRL and Introduction of a Heuristic Function

1) *Proposed Deep Reinforcement Learning Algorithm* : As in [15], we use a single thread version of the A3C Algorithm introduced in [24]. This algorithm relies on two DNNs that are trained in parallel: i) the Actor Network with the parameter  $\theta$ , which is used to generate the policy  $\pi_\theta$  at each time step, and ii) the Critic Network with the parameter  $\theta_v$  which generates an estimate  $\nu_{\theta_v}^{\pi_\theta}(\sigma_t)$  for the State-value function defined by

$$\nu_\pi(t|\sigma) = \mathbb{E}_\pi \left[ \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1} | \sigma_t = \sigma \right],$$

for some discount parameter  $\gamma$ .

As depicted in Fig. 1 both Actor and Critic Networks have almost identical structure. As in [6], we use the GCN formulation proposed by [25] to automatically extract advanced characteristics of the PSN. The characteristics produced by the GCN represent semantics of the PSN topology by encoding and accumulating characteristics of neighbour nodes in the PSN graph. The size of the neighbourhood is defined by the order-index parameter  $K$ . As in [6], we consider in the following  $K = 3$  and perform automatic extraction of 60 characteristics per PSN node. Both the NSPR state and Network Load characteristics are separately transmitted to fully connected layers with 4 and 100 units, respectively.

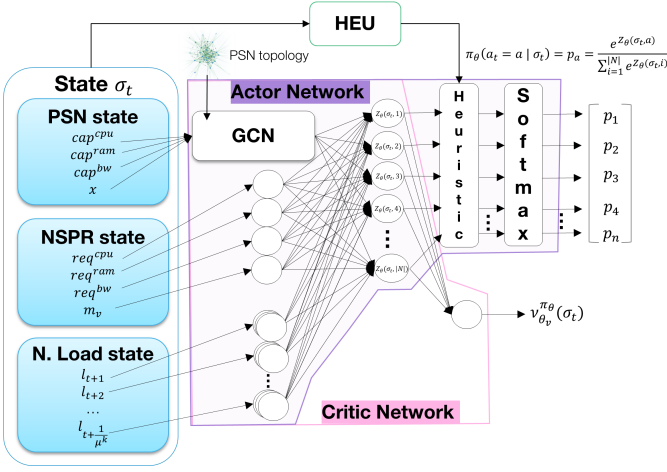


Fig. 1. Reference framework for the proposed learning algorithms.

The characteristics extracted by both layers and the GCN layer are combined into a single column vector of size  $60|N| + 104$  and passed through a full connection layer with  $|N|$  units.

In the Critic Network, the outputs are forwarded to a single neuron, which is used to calculate the state-value function estimation  $v_{\theta_v}^{\pi_{\theta}}(\sigma_t)$ . In the Actor Network, the outputs represent the values of the function  $Z_{\theta}$  introduced in Section IV-A. These values are injected into a Softmax layer that transforms them into a Softmax distribution that corresponds to the policy  $\pi_{\theta}$ .

During the training phase, at each time step  $t$ , the A3C algorithm uses the Actor Network to calculate the policy  $\pi_{\theta}(\cdot|\sigma_t)$ . An action  $a_t$  is sampled using the policy and performed on the environment. The Critic Network is used to calculate the state-value function approximation  $v_{\theta_v}^{\pi_{\theta}}(\sigma_t)$ . The learning agent receives then the reward  $r_{t+1}$  and next state  $\sigma_{t+1}$  from the environment and the placement process continues until a terminal state is reached, that is, until the Actor Network returns an unsuccessful action or until the current NSPR is completely placed. At the end of the training episode, the A3C algorithm updates parameters  $\theta$  and  $\theta_v$  by using the same rules as in [15].

2) *Introduction of a Heuristic Function:* To guide the learning process, we use as in [15] the placement heuristic introduced in [26]. This yields the HA-DRL algorithm. More precisely, from the reference framework shown in Fig. 1, we proposed to include in the Actor Network the Heuristic layer that calculates a Heuristic Function  $H : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  based on external information provided by the heuristic method, referred to as HEU.

Let  $Z_{\theta}$  be the function computed by the fully connected layer of the Actor Network that maps each state and action to a real value which is after converted by the Softmax layer into the selection probability of the respective action (see Section IV-A). Let  $\bar{a}_t = \operatorname{argmax}_{a \in \mathcal{A}} Z_{\theta}(\sigma_t, a)$  be the action with the highest  $Z_{\theta}$  value for state  $\sigma_t$ . Let  $a_t^* = \text{HEU}(\sigma_t)$  be the action derived by the HEU method at time step  $t$  and the preferred action to be chosen.  $H(\sigma_t, a_t^*)$  is shaped to allow the value of  $Z_{\theta}(\sigma_t, a_t^*)$  to become closer to the value of  $Z_{\theta}(\sigma_t, \bar{a}_t)$ .

The aim is to turn  $a_t^*$  into one of the likeliest actions to be chosen by the policy.

The Heuristic Function is then formulated as

$$H(\sigma_t, a_t) = \begin{cases} Z_{\theta}(\sigma_t, \bar{a}_t) - Z_{\theta}(\sigma_t, a_t) + \eta, & \text{if } a_t = a_t^* \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where  $\eta$  parameter is a small real number. During the training process the Heuristic layer calculates  $H(\sigma_t, \cdot)$  and updates the  $Z_{\theta}(\sigma_t, \cdot)$  values by using the following equation:

$$Z_{\theta}(\sigma_t, \cdot) = Z_{\theta}(\sigma_t, \cdot) + \xi H(\sigma_t, \cdot)^{\beta} \quad (10)$$

The Softmax layer then computes the policy using the modified  $Z_{\theta}$ . Note the action returned by  $a_t^*$  will have a higher probability to be chosen. The  $\xi$  and  $\beta$  are parameters used to control how much HEU influence the policy.

### C. Implementation Remarks

1) *Algorithms considered:* We consider four learning algorithms based on the reference framework presented in Fig. 1.

- **DRL:** It is the pure-DRL algorithm we initially proposed in [15]. The state representation does not include the network load state and the Actor Network does not contain the Heuristic layer.
- **eDRL:** This algorithm is an enhanced version of DRL in which the state representation includes the network load state and the Actor Network does not contain the Heuristic layer;
- **HA-DRL:** This algorithm embeds the heuristic but the state representation does not include the network load state while the Actor Network contain the Heuristic layer;
- **HA-eDRL:** The state representation includes the network load state and the Actor Network contains the Heuristic layer.

2) *Implementation details:* All resource-related characteristics are normalized to be in  $[0, 1]$ . This is done by dividing  $cap^j$  and  $req^j$ ,  $j \in \{\text{cpu}, \text{ram}, \text{bw}\}$ , by  $\max_{n \in N} M_n^j$ . With regard to the DNNs, we have implemented the Actor and Critic as two independent Neural Networks. Each neuron has a bias assigned. We have used the hyperbolic tangent (tanh) activation for non-output layers of the Actor Network and Rectified Linear Unit (ReLU) activation for all layers of the Critic Network. We have normalized positive global rewards to be in  $[0, 10]$ . During the training phase, we have considered the policy as a Categorical distribution and used it to sample the actions randomly.

## V. IMPLEMENTATION AND EVALUATION RESULTS

In this section, we present the implementation and experiments we conducted to evaluate the proposed algorithms.

### A. Implementation Details & Simulator Settings

1) *Experimental setting:* We developed a simulator in Python containing: i) the elements of the Network Slice Placement Optimization problem (i.e., PSN and NSPR); ii) the DRL, eDRL, HA-DRL and HA-eDRL algorithms. We used the PyTorch framework to implement the DNNs. Experiments were run in a 2x6 cores @2.95Ghz 96GB machine.

TABLE III  
NETWORK LOAD CALCULATION FOR BOTH NSPR CLASSES

NSPR Classes ( $k$ )	# of VNFs per NSPR ( $ V^k $ )	CPU requested per VNF ( $req^{cpu,k}$ )	NSPR lifetime ( $\frac{1}{\mu^k}$ )	NSPR arrival rate ( $\lambda^k(t)$ )	Total CPU capacity ( $C^{cpu}$ )	Network Load ( $\rho^k(t)$ )
Volatile	5	25	20	$1.5 \sin^2(\frac{\pi t}{96})$	6300	$1.5 \sin^2(\frac{\pi t}{96}) \frac{2500}{6300}$
Long term	10	25	500	0.02	6300	$0.02 \frac{125000}{6300}$

2) *Physical Substrate Network Settings*: We consider a PSN that could reflect the infrastructure of an operator as discussed in [23]. In this network, three types of DCs are introduced as in Section III. Each CDC is connected to three EDCs which are 100 km apart. CDCs are interconnected and connected to one CCP that is 300 km away. We consider 15 EDCs each one with 4 servers, 5 CDCs each with 10 servers and 1 CCP with 16 servers. The CPU and RAM capacities of each server are 50 and 300 units, respectively. A bandwidth capacity of 100 Gbps is given to intra-data center links inside CDCs and CCP—10Gbps being the case for intra-data center links inside EDCs. Transport links connected to EDCs have 10Gbps of bandwidth capacity. Transport links between CDCs have 100Gbps of bandwidth capacity as well for the ones between CDCs and the CCP.

3) *Network Slice Placement Requests Settings*: We consider NSPRs to have the Enhanced Mobile Broadband (eMBB) setting described in [26]. Each NSPR is composed of 5 VNFs. Each VNF requires 25 units of CPU and 150 units of RAM. Each VL requires 2 Gbps of bandwidth.

## B. Algorithms & Experimental Setup

1) *Training Process & Hyper-parameters*: We consider a training process with maximum duration of 55 hours for the considered algorithms. We perform seven independent runs of each algorithm to assess their average and maximal performance in terms of metrics introduced below (see Section V-C). After performing Hyper-parameter search, we set the learning rates for the Actor and Critic networks of DRL and HA-DRL algorithms to  $\alpha = 5 \times 10^{-5}$  and  $\alpha' = 1.25 \times 10^{-3}$ , respectively. For eDRL and HA-eDRL, we set the learning rates for the Actor and Critic networks to  $\alpha = 5.7 \times 10^{-5}$  and  $\alpha' = 1.4 \times 10^{-3}$ , respectively. We implement four versions of HA-DRL and HA-eDRL agents, each with a different value for the  $\beta$  parameter of the heuristic function formulation (see Section IV-B2). We set in addition the parameters  $\xi = 1$  and  $\eta = 0$ .

2) *Network load calculation*: We implement the network load model introduced in Section V-B2 considering two NSPR classes: i) a volatile class (referred to as Volatile), with a dynamic arrival rate and ii) a static class (referred to as Long term), with a static arrival rate. Table III presents the network load models proposed for both classes. We consider that each simulation time unit corresponds to 15 minutes in reality. We set the period for the network load function to 96 simulation time units, i.e., one day. The network global network load  $\rho(t)$  varies then between 0.3 and 1.0.

## C. Evaluation Metrics

To characterize the performance of the placement algorithms, we consider two performance metrics:

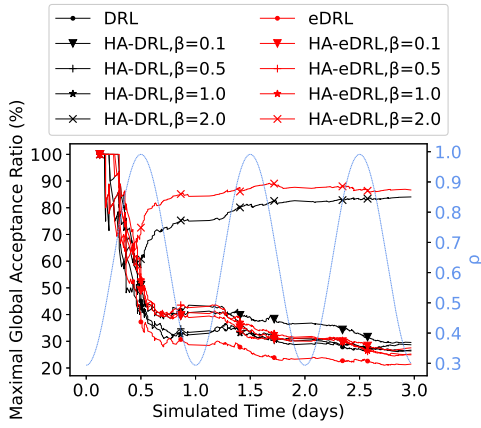
- 1) **Global Acceptance Ratio (GAR)**: The Acceptance Ratio of the different tested algorithms during training computed after each arrival as follows:  $\frac{\#\text{accepted NSPRs}}{\#\text{arrived NSPRs}}$ . This metric is used to evaluate the accumulated acceptance ratio of the different algorithms as the learning process progresses.
- 2) **Acceptance Ratio per training phase (TAR)**: The Acceptance Ratio obtained in each training phase, i.e., each part of the training process, corresponding to  $10^4$  NSPR arrivals or  $10^4$  episodes. It is calculated as follows:  $\frac{\#\text{accepted NSPRs}}{10^4}$ . This metric allows us to better observe the evolution of algorithm performance over time since it measures algorithm performance in independent parts (phases) of the training process without accumulating the performance of previous training phases.

## D. Global Acceptance Ratio Evaluation

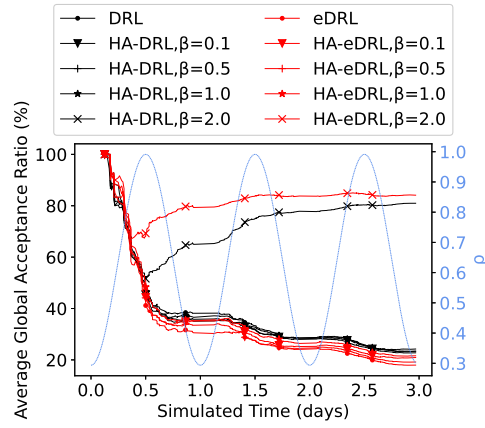
Fig. 2 and 3 show the progression of the Global Acceptance Ratio (GAR) over time for the considered algorithms. Fig. 2a and Fig. 3a show that after 3 simulated days, the average and maximal GARs of HA-DRL and HA-eDRL with  $\beta = 2.0$  are convergent and higher than 80% while for the other algorithms, the GARs remain between 20% and 40% and are not stable. Fig. 2b and Fig 3b exhibit the progression of the GARs over a longer simulated time horizon of 4500 simulated days.

We observe that while GARs of HA-DRL and HA-eDRL with  $\beta = 2.0$  remain stable after the convergence reached in the first 3 simulated days, for the other algorithms, these performance metrics start to stabilize only after 2000 simulated days. Table IV shows that both the maximal and average final GARs, i.e., maximal and average GARs achieved at the end of training, are close to 82% for both HA-DRL and HA-eDRL with  $\beta = 2.0$ . For the other algorithms, maximal final GARs are between 61% and 71% and average final GARs are between 52% and 65%, except for HA-DRL and HA-eDRL with  $\beta = 1.0$ . Table IV also shows that maximal and average final GARs of HA-DRL algorithms are generally higher than the equivalent HA-eDRL versions, the gap being never higher than 6%.

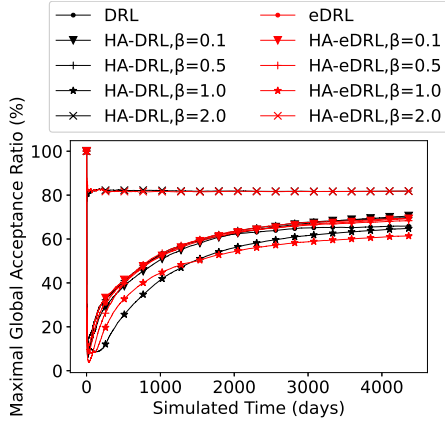
The above results show that HA-DRL and HA-eDRL with  $\beta = 2.0$  exhibit more robust performance with faster convergence and handle better network load variations than the other evaluated algorithms, notably when compared with the classical DRL approach.



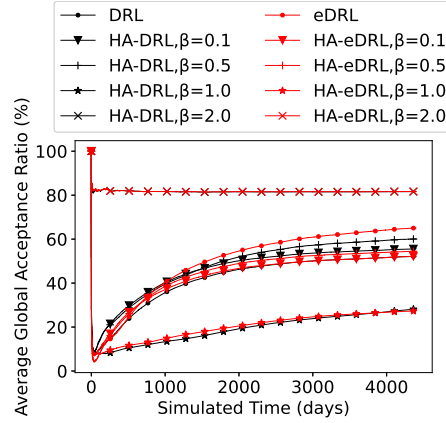
(a) Maximal performance in 3 simulated days.



(a) Average performance in 3 simulated days.



(b) Maximal performance in 4500 simulated days.



(b) Average performance in 4500 simulated days.

Fig. 2. Maximal Global Acceptance ratio results.

Fig. 3. Average Global Acceptance ratio results.

 TABLE IV  
 SUMMARY OF EVALUATION RESULTS

Algorithm	Maximal Final GAR (%)	Average Final GAR (%)	Maximal Final TAR (%)	Average Final TAR (%)
DRL	65.90	52.15	77.69	55.82
HA-DRL, $\beta = 0.1$	70.56	55.57	81.45	59.83
HA-DRL, $\beta = 0.5$	69.80	60.12	73.56	65.33
HA-DRL, $\beta = 1.0$	64.89	28.23	72.14	41.69
HA-DRL, $\beta = 2.0$	81.84	81.62	83.91	82.21
eDRL	69.36	65.10	77.39	58.61
HA-eDRL, $\beta = 0.1$	69.25	52.22	76.34	48.45
HA-eDRL, $\beta = 0.5$	68.40	54.49	71.91	57.57
HA-eDRL, $\beta = 1.0$	61.46	27.33	67.05	33.10
HA-eDRL, $\beta = 2.0$	81.82	81.68	81.37	81.95

Indeed, when setting  $\beta = 2.0$ , the Heuristic Function computed on the basis of the HEU algorithm has strong influence on the actions chosen by the agent. Since the HEU algorithm often indicates a good action, this strong influence of the heuristic function helps the algorithms to become stable more quickly.

The addition of network load related features to the state observed by the eDRL algorithm helps improve the maximal and average final GAR when compared with DRL. However, this improvement is less significant than the improvement

brought by the Heuristic Function acceleration as the eDRL algorithm does not achieve the same fast convergence and robustness than HA-DRL and HA-eDRL with  $\beta = 2.0$ .

### E. Acceptance Ratio per Training Phase Evaluation

Fig. 4 and Fig. 5 show that all algorithms need at least 16 training phases to reach a convergent maximal and average Acceptance Ratio per Training Phase (TAR), except for HA-DRL and HA-eDRL with  $\beta = 2.0$  which need only one training phase (see Fig 4a and Fig 5a). The fast convergence of HA-DRL and HA-eDRL with  $\beta = 2.0$  is due to the strong influence of the Heuristic Function in the choice of actions as explained in Section V-D. Fig. 4a and Fig. 4b illustrate that algorithms HA-DRL and HA-eDRL with  $\beta = 0.5$ , HA-DRL with  $\beta = 0.1$  and DRL need around 16 phases to reach a convergent maximal TAR while Fig. 4b and Fig. 4c demonstrate that eDRL, HA-eDRL with  $\beta \in \{0.1, 1.0\}$  and HA-DRL with  $\beta = 1.0$  need around 34 phases. Finally, Fig. 5a and Fig. 5b show that HA-DRL and HA-eDRL with  $\beta \in \{0.1, 0.5\}$ , eDRL, and DRL need around 16 phases to obtain a convergent average TAR while Fig. 5b and Fig. 5c show that HA-DRL and HA-eDRL with  $\beta = 1.0$  need around 34 phases.

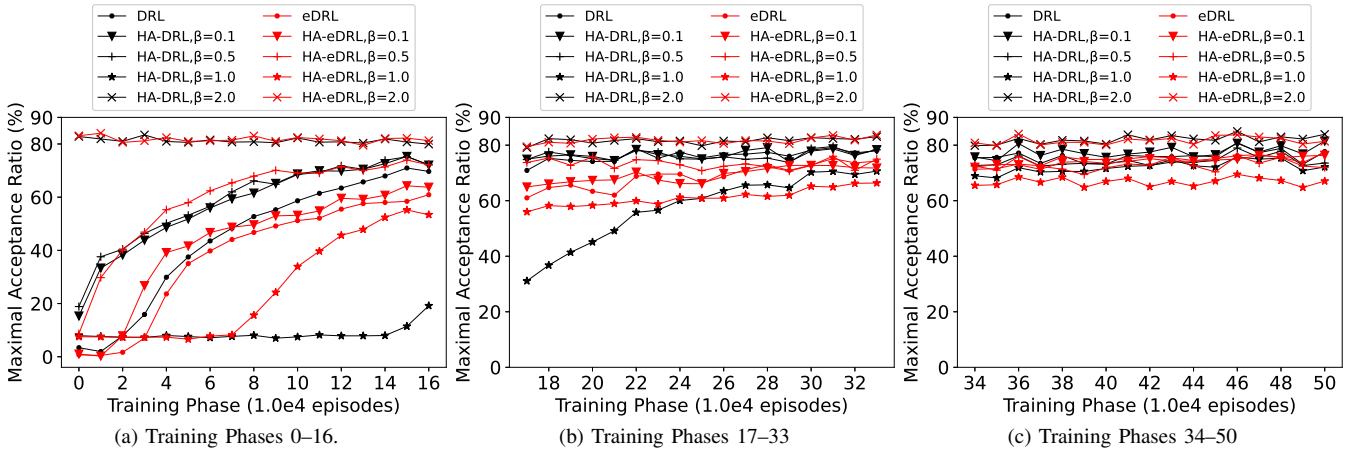


Fig. 4. Maximal Acceptance Ratio per training phase results.

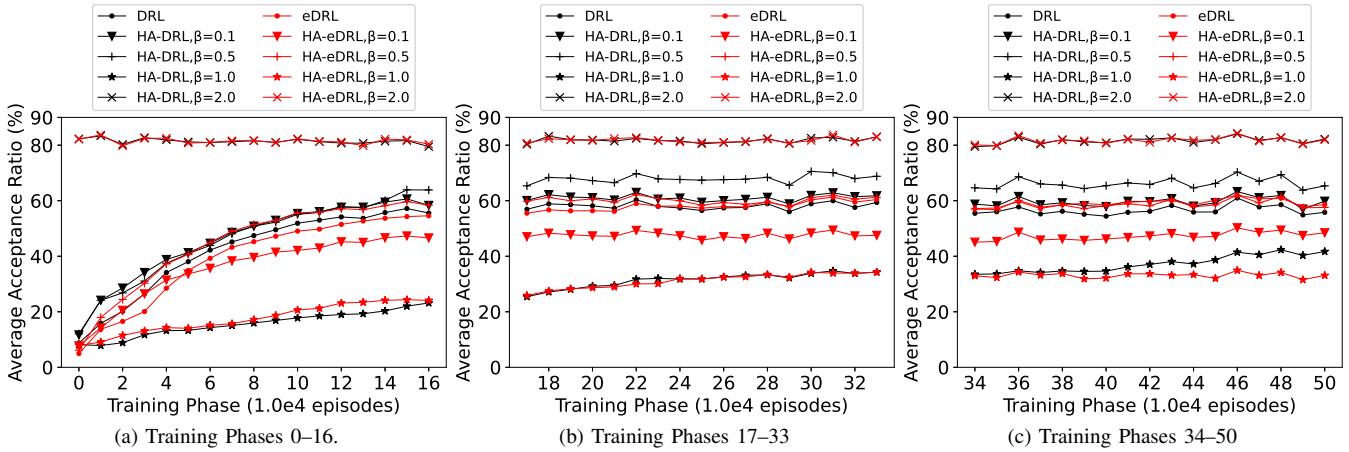


Fig. 5. Average Acceptance Ratio per training phase results.

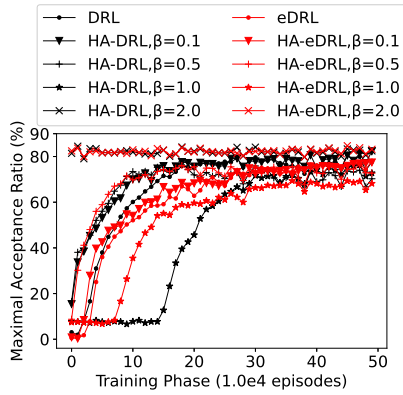
Table IV shows that the only algorithms that reach a maximal and average final TAR (i.e., maximal and average TAR achieved at the end of training) higher than 80% are HA-DRL and HA-eDRL with  $\beta = 2.0$ . HA-DRL with  $\beta = 0.1$  also attains a maximal final TAR higher than 80% but its average final TAR is less than 60%. HA-DRL with  $\beta = 0.5$  has average final TAR higher than 65%, but its maximal final TAR is around 74%. In addition, Table IV shows that: HA-DRL algorithms have maximal final TAR generally higher than the equivalent HA-eDRL versions, the gap being never higher than 6%; eDRL and DRL have equivalent maximal final TAR but eDRL has average final TAR 3% higher. Fig. 6b and Fig. 6a show that all algorithms have better maximal and average TAR on volatile NSPRs than on long-term NSPRs. This difference is arguably related to the number of arrivals of each requests and, therefore, the amount of training performed on each class of requests, since many more volatile requests arrive to be placed during the simulated period than long-term requests. These results reinforce the conclusion introduced in Section V-D. The maximal TAR progression results (i.e., Fig. 4) allow us to observe that in the best case, if a large number of training phases is given, all the algorithms attain a good performance.

However, the average TAR progression results (see Fig. 5) show that only HA-DRL and HA-eDRL with  $\beta = 2.0$  have robust performance. These algorithms are also the only one to have quick convergence being, among those evaluated, the most adapted to be used in practice.

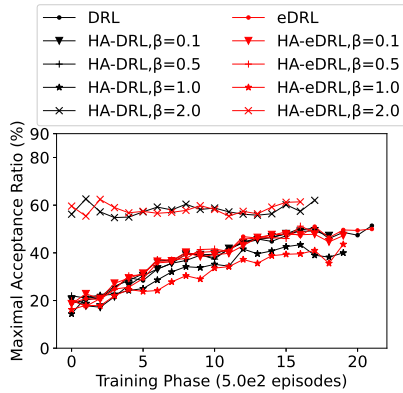
## VI. CONCLUSION

We have introduced a DRL-heuristic algorithm that supports the variations of network load in the placement of network slice requests as a follow-up of our work in [15]. We have considered four families of algorithms, pure-DRL and eDRL techniques and their variants combined with a heuristic whose efficiency was investigated in a previous work [26]. The influence of the heuristic on the DRL algorithm can be tuned by means of a parameter (namely, parameter  $\beta$  (introduced in Section IV-B2)). We assume network load has periodic fluctuations. We have shown how introducing network load states into the DRL algorithm together with a combination with the heuristic function yields very good results in terms of GAR and TAR which are stable in time. This study aims at proving that coupling DRL and heuristic functions yields good and stable results even under non stationary conditions.





(a) Maximal performance for volatile NSPRs.



(b) Maximal performance for longterm NSPRs.

Fig. 6. Acceptance Ratio per training phase results (per NSPR class).

The next step is to study the behavior of the proposed algorithm in case of an unpredictable network load disruption.

#### ACKNOWLEDGMENT

This work has been performed in the framework of 5GPPP MON-B5G project ([www.monb5g.eu](http://www.monb5g.eu)). The experiments were conducted using Grid'5000, a large scale testbed by Inria and Sorbonne University ([www.grid5000.fr](http://www.grid5000.fr)).

#### REFERENCES

- [1] 3GPP, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3 (Release 17)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 28.541, Dec. 2020, version 17.1.0.
- [2] ETSI NFV ISG, "Network Functions Virtualisation (NFV); Evolution and Ecosystem; Report on Network Slicing Support, ETSI Standard GR NFV-EVE 012 V3.1.1," ETSI, Tech. Rep., 2017. [Online]. Available: <https://www.etsi.org/technologies-clusters/technologies/nfv>
- [3] J. Gil Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 518–532, Sep. 2016.
- [4] J. J. A. Esteves, A. Boubendir, F. Guillemin, and P. Sens, "Location-based data model for optimized network slice placement," in *Proc. 2020 6th IEEE Conf. Netw. Softwarization (NetSoft)*, 2020, pp. 404–412.
- [5] E. Amaldi, S. Coniglio, A. M. Koster, and M. Tieves, "On the computational complexity of the virtual network embedding problem," *Electron. Notes Discrete Math.*, vol. 52, pp. 213–220, Jun. 2016.
- [6] Z. Yan, J. Ge, Y. Wu, L. Li, and T. Li, "Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 6, pp. 1040–1057, Jun. 2020.
- [7] M. Dolati, S. B. Hassanpour, M. Ghaderi, and A. Khonsari, "DeepViNE: Virtual network embedding with deep reinforcement learning," in *Proc. IEEE INFOCOM 2019 - IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2019, pp. 879–885.
- [8] H. Yao, X. Chen, M. Li, P. Zhang, and L. Wang, "A novel reinforcement learning algorithm for virtual network embedding," *Neurocomputing*, vol. 284, pp. 1–9, Apr. 2018.
- [9] H. Wang, Y. Wu, G. Min, J. Xu, and P. Tang, "Data-driven dynamic resource scheduling for network slicing: A deep reinforcement learning approach," *Inf. Sci.*, vol. 498, pp. 106–116, Sep. 2019.
- [10] Y. Xiao, Q. Zhang, F. Liu, J. Wang, M. Zhao, Z. Zhang, and J. Zhang, "NFVdeep: Adaptive online service function chain deployment with deep reinforcement learning," in *Proc. 2019 IEEE/ACM 27th Int. Symp. Qual. Service (IWQoS)*, 2019, pp. 1–10.
- [11] P. T. A. Quang, A. Bradai, K. D. Singh, and Y. Hadjadj-Aoul, "Multi-domain non-cooperative VNF-FG embedding: A deep reinforcement learning approach," in *Proc. IEEE INFOCOM 2019 - IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2019, pp. 886–891.
- [12] S. Schneider, A. Manzoor, H. Qarawlus, R. Schellenberg, H. Karl, R. Khalili, and A. Hecker, "Self-driving network and service coordination using deep reinforcement learning," in *Proc. 2020 IEEE 16th Int. Conf. Netw. Service Manag. (CNSM)*, 2020, pp. 1–9.
- [13] F. Habibi, M. Dolati, A. Khonsari, and M. Ghaderi, "Accelerating virtual network embedding with graph neural networks," in *Proc. 2020 IEEE 16th Int. Conf. Netw. Service Manag. (CNSM)*, 2020, pp. 1–9.
- [14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press, 2015.
- [15] J. J. Alves Esteves, A. Boubendir, F. Guillemin, and P. Sens, "A heuristically assisted deep reinforcement learning approach for network slice placement," *arXiv preprint arXiv:2105.06741*, 2021.
- [16] P. T. A. Quang, Y. Hadjadj-Aoul, and A. Outtagarts, "A deep reinforcement learning approach for vnf forwarding graph embedding," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 4, pp. 1318–1331, Dec. 2019.
- [17] A. Rkhami, Y. Hadjadj-Aoul, and A. Outtagarts, "Learn to improve: A novel deep reinforcement learning approach for beyond 5G network slicing," in *Proc. 2021 IEEE 18th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2021, pp. 1–6.
- [18] R. A. Bianchi, C. H. Ribeiro, and A. H. Costa, "Accelerating autonomous learning by using heuristic selection of actions," *J. Heuristics*, vol. 14, no. 2, pp. 135–168, 2008.
- [19] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "Deepcog: Cognitive network management in sliced 5g networks with deep learning," in *Proc. IEEE INFOCOM 2019 - IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2019, pp. 280–288.
- [20] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, C. Viho, and D. Darche, "An efficient and lightweight load forecasting for proactive scaling in 5g mobile networks," in *Proc. 2018 IEEE Conf. Standards Commun. Netw. (CSCN)*, 2018, pp. 1–6.
- [21] J. Pei, P. Hong, M. Pan, J. Liu, and J. Zhou, "Optimal vnf placement via deep reinforcement learning in sdn/nfv-enabled networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 263–278, Feb. 2020.
- [22] F. Slim, F. Guillemin, and Y. Hadjadj-Aoul, "CLOSE: A costless service offloading strategy for distributed edge cloud," in *Proc. 2018 15th IEEE Annu. Cons. Commun. Netw. Conf. (CCNC)*, 2018, pp. 1–6.
- [23] F. Slim, F. Guillemin, A. Gravey, and Y. Hadjadj-Aoul, "Towards a dynamic adaptive placement of virtual network functions under ONAP," in *Proc. 2017 IEEE Conf. on Netw. Function Virtualization Softw. Defined Netw. (NFV-SDN)*, 2017, pp. 210–215.
- [24] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Int. Conf. Mach. Learn.* PMLR, 2016, pp. 1928–1937.
- [25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations (ICLR)*, 2017, pp. 1–14.
- [26] J. J. Alves Esteves, A. Boubendir, F. Guillemin, and P. Sens, "Heuristic for edge-enabled network slicing optimization using the 'power of two choices'," in *Proc. 2020 IEEE 16th Int. Conf. Netw. Service Manag. (CNSM)*, 2020, pp. 1–9.