

Continuous Multi-objective Zero-touch Network Slicing via Twin Delayed DDPG and OpenAI Gym

Farhad Rezazadeh¹, Hatim Chergui¹, Luis Alonso², and Christos Verikoukis¹

¹ Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain

² Technical University of Catalonia (UPC), Barcelona, Spain

Contact Emails: farhad.rezazadeh, hatim.chergui, cveri@cttc.es, luisg@tsc.upc.edu

Abstract—Artificial intelligence (AI)-driven zero-touch network slicing (NS) is a new paradigm enabling the automation of resource management and orchestration (MANO) in multi-tenant beyond 5G (B5G) networks. In this paper, we tackle the problem of cloud-RAN (C-RAN) joint slice admission control and resource allocation by first formulating it as a Markov decision process (MDP). We then invoke an advanced continuous deep reinforcement learning (DRL) method called twin delayed deep deterministic policy gradient (TD3) to solve it. In this intent, we introduce a multi-objective approach to make the central unit (CU) learn how to re-configure computing resources autonomously while minimizing latency, energy consumption and virtual network function (VNF) instantiation cost for each slice. Moreover, we build a complete 5G C-RAN network slicing environment using OpenAI Gym toolkit where, thanks to its standardized interface, it can be easily tested with different DRL schemes. Finally, we present extensive experimental results to showcase the gain of TD3 as well as the adopted multi-objective strategy in terms of achieved slice admission success rate, latency, energy saving and CPU utilization.

Index Terms—Admission control, B5G, continuous DRL, C-RAN, network slicing, OpenAI Gym, resource allocation, zero-touch

I. INTRODUCTION

NETWORK slicing is a key feature in 5G networks. It enables to run fully or partly isolated logical networks—or tenants—on the same physical network, offering thereby a concrete resource multiplexing gain between slice instances. In this intent, network software-defined networking (SDN) and network functions virtualization (NFV) provide the necessary programmability and flexibility to operate NS by dynamically creating, scaling and terminating chained virtual network functions (VNFs). Multi-access Edge Computing (MEC) [1] is also an important component that—co-located with C-RAN—can bring the high performance computing resources at the edge, paving the way to accommodate low-latency slices. Having said that, zero-touch and fully automated operations and management have become quintessential to harness the potential gain of dynamic resource allocation in SDN/NFV-enabled NS. Besides ETSI’s architecture standardization efforts [2], many algorithms have been presented in the literature to enable the automation of B5G networks as detailed in the sequel.

A. Related Work

In [3], the authors have studied autonomous MANO of VNFs, where the central unit learns to re-configure resources, deploy new VNF instances or offloaded to a central cloud. They have proposed a DRL-based solution dubbed *parameterized action twin* (PAT) Deep Deterministic Policy Gradient (DDPG) which leverages the actor-critic method to learn to provision network resources to the VNFs in an online manner, given the current network state and the requirements of the deployed VNFs. The proposed solution outperforms all benchmark DRL schemes as well as heuristic greedy allocation in a variety of network scenarios. In [4], the authors have developed a data-driven resource scheduling based on DRL for dynamic resource scheduling in networks slicing. They have solved the slicing resource management challenge in an asymmetric information scenario without using the user-related data, due to the model-free and dynamic online learning features. Correspondingly, [5] has proposed a dynamic resource reservation and DRL-based autonomous virtual resource slicing framework for the next generation radio access network. At light load, autonomous radio resource management of the deep Q -network (DQN) algorithm has achieved 100% satisfaction and up to about 80% saturation which is the best compared with other benchmarks. Li *et al.* have studied the application of DRL in some typical resource management scenarios of NS. Their results have shown that compared with the demand prediction-based and some other intuitive solutions, DRL could implicitly incorporate more deep relationship between demand and supply in resource-constrained scenarios [6]. In [7], the authors have proposed vRAIn as a dynamic resource controller for virtualization of RANs (vRAN) based on DRL where vRAN dynamically learns the optimal allocation of computing and radio resources. The proposed solution meets the desired performance targets while minimizing CPU usage and gracefully adapts to shortages of computing resources.

B. Contributions

In this paper, we present the following contributions:

- Given that DDPG algorithm is a limiting case of stochastic policy gradient in actor-critic approaches used for solving continuous tasks, this work adopts and fine-tune an alternative way of updating the actor (policy) in DDPG

algorithm to speed-up convergence and fulfill a stable and robust learning process [8] based on TD3 method [9].

- We introduce a multi-objective approach in the NS environment to maximize cumulative rewards while minimizing network costs.
- We develop a complete 5G NS environment based on OpenAI Gym to ensure reproducible comparison of DRL algorithms.

II. SYSTEM MODEL

As depicted in Figure 1, we consider a C-RAN architecture according to 3GPP CU-DU functional split. The underlying N single-antenna small-cells ($n = 1, \dots, N$) are connected to a virtual baseband unit (i.e., CUs) pool that runs as a set of VNFs. A total number of J VNFs ($j = 1, \dots, J$) can be deployed on top of the C-RAN datacenter endowed with I active central processing units (CPUs), where each processor i ($i = 1, \dots, I$) has a computing capability of P_i million operations per time slot (MOPTS) [10]. At each time step t , M UEs ($m = 1, \dots, M$) can connect to the N small-cells according to the maximum received power criteria. Each UE m requests a slice and starts its activity, wherein the packet arrival to the CU VNF follows a Poisson distribution with mean rate $\lambda_m^{(t)}$. In this case, let $\Omega = \sum_{m=1}^M \lambda_m^{(t)}$. The mean arrival data rate of all UEs to the CU VNFs is Ω/j , where j is the number of active VNFs.

Computation cost ($K_{Net}^{(t)}$)— The baseband processing procedure at a VNF consists of coding, Fast Fourier Transform (FFT) and modulation. The corresponding computing resources follow that in [11], and is given by:

$$K_{Net}^{(t)} = \sum_{m=1}^M [\theta \log_2(1 + \delta_m)] + MK_0, \quad (1)$$

Where θ is an experimental parameter, δ_m denotes the signal-to-interference-plus-noise ratio (SINR) of UE m and K_0 includes computing resources for FFT function that according to [12] imposes a constant base processing load on the system. Based on the experimental results of [13] we assume that, in each cell n , the computing resource requirements for coding, modulation and FFT are 50%, 10% and 40%, respectively. Moreover, we assume that VNFs have first in first out (FIFO) queues where μ^* is mean service rate for cloud processing and r_m for wireless transmission rate which satisfies according to $r_m = B_m \log(1 + \delta_m)$, where B_m is wireless transmission bandwidth for UE m . In this respect, we further suppose that cloud processing and wireless transmission queues follow an exponential distribution with mean $\frac{1}{\mu^*}$ and $\frac{1}{r_m}$ respectively [14]. Next, we explain each cost functions of this paper.

Latency ($\mathcal{L}_{Net}^{(t)}$)— According to queuing theory [15], the mean processing delay at time step t is $\mathcal{L}_{proc}^{(t)} = \frac{j}{j\mu^* - \Omega}$ and transmission latency in wireless transmission queue is $\mathcal{L}_{trans}^{(t)} = \frac{1}{r_m - \lambda_m^{(t)}}$ so we have:

$$\mathcal{L}_{Net}^{(t)} = j\mathcal{L}_d^{(t)} + \sum_{m=1}^M \left[\frac{j}{j\mu^* - \Omega} + \frac{1}{r_m - \lambda_m^{(t)}} \right] \quad (2)$$

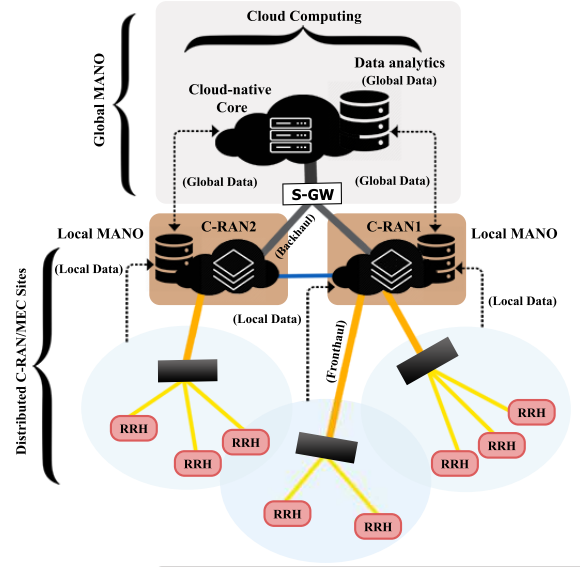


Figure 1: Proposed distributed and hierarchical architecture.

where $\mathcal{L}_d^{(t)}$ denotes latency for creating, booting up and loading new VNFs and j denotes the total number of active VNFs to be deployed. We suppose $\mathcal{L}_{Net}^{(t)} < \eta_m^{(t)}$ where $\eta_m^{(t)}$ is a predefined maximum network delay for UEs which can be viewed as a quality of service (QoS) requirement [14].

Energy ($\mathcal{E}_{Net}^{(t)}$)— The energy consumption incurred by the VNF instantiation, running processors and wireless transmission power where $\mathcal{E}_v^{(t)} = \psi_j$ refers to energy consumption associated with the deployment of the j^{th} VNF instance where ψ_j is a constant value. The energy consumed by processor i in Watts is $\mathcal{E}_p^{(t)} = \sigma^* P_i^3$ where σ^* is a parameter determined by the processor structure. The wireless transmission power for UE m is given by $\mathcal{E}_w^{(t)} = \frac{1}{\rho} \|W_m\|_2^2$ where W_m is the precoding vector from all cells to UE m , and ρ denotes the efficiency of power amplifier [10] at the cells. Finally we have:

$$\mathcal{E}_{Net}^{(t)} = \sum_{i=1}^I \sigma^* P_i^3 + \sum_{j=1}^J \psi_j + \sum_{m=1}^M \frac{1}{\rho} \|W_m\|_2^2 \quad (3)$$

We define overall network cost as all costs incurred at each time step:

$$\mathcal{N}_T^{(t)} = \frac{\omega_1^* K_{Net}^{(t)} + \omega_2^* \mathcal{L}_{Net}^{(t)} + \omega_3^* \mathcal{E}_{Net}^{(t)}}{M} \quad (4)$$

Where $\omega_1^*, \omega_2^*, \omega_3^* \in \mathbb{R}$ are fixed weights which can be set based on the operator preferences.

III. MDP AND BUILDING NS ENVIRONMENT BY GYM

In this section, we formulate resource allocation optimization problem as Markov Decision Process (MDP). We contemplate the autonomous CU with the goal of improve average return. To this end, we define the observation space and action space that CU can take at each time step.

The MDP for a single agent often is defined by a 5-tuple (S, A, P, γ, R) , consisting of a set of states S (state space), a set of actions A (action space), P denotes the state transition probability for state s and action a . The key term of MDP is decision. In fact, the way that agent make decisions for what actions to do in what states is called a policy which denotes with the symbol π . The notation of return G_t refers to total discounted rewards from time step t and the main goal is to maximize this return, $G_t = \sum_{n=0}^{\infty} (\gamma^n R_{t+n+1})$. Where γ is a real-valued discount factor weighting of future rewards and $\gamma \in [0, 1]$ refers to how much we value rewards right now relative to rewards in the future as short-sighted ($\gamma = 0$) or far-sighted ($\gamma = 1$).

The value function informs the agent how good is at each state or action and how much reward to expect takes a particular $V(s_t, a_t)$. Another value function Q which not only depends on the state s but also action a is action-value function. The policies determine relation between Q and V . The optimal policy in Reinforcement Learning (RL) is the best policy for which there is no greater value function, so for optimal value functions and optimal action-value function we have $\forall s \in S, V_*(s) = \max_{\pi} \{V_{\pi}(s)\}$ and $\forall s \in S, a \in A, Q_*(s, a) = \max_{\pi} \{Q_{\pi}(s, a)\}$ respectively.

A continuous state and action space in OpenAI Gym is defined the action that an agent can take and the input that the agent receives are both continuous values:

- 1) **State Space:** We use Box space as multidimensional continuous spaces with bounds. In telecom environment the state space is the set of possible network configurations. We consider state at time step t consists of :
 - The number of new UEs which connect to network and request services for each slice ($X^{(t)}$).
 - Computing resources allocated to each VNF ($C^{(t)}$)
 - Delay status with respect to latency cost for each slice ($\mathcal{L}^{(t)}$)
 - Energy status with respect to energy cost for each slice ($\mathcal{E}^{(t)}$)
 - Number of users being served in each slice ($m^{(t)}$)
 - Number of VNF instantiations in each slice ($V^{(t)}$)

The network state space or input can be characterized by $S^{(t)} = \{X^{(t)}, C^{(t)}, \mathcal{L}^{(t)}, \mathcal{E}^{(t)}, m^{(t)}, V^{(t)}\}$.

- 2) **Action Space:** We consider vertical scaling action space. The vertical scaling can be classified into scale up and scale down that is related to increasing or decreasing capacity respectively. The CU select continuous value action with respect to traffic fluctuation and learn to decide to increase/decrease computing resources allocated to each VNF. In OpenAI Gym, It takes an action as input and provides observation, reward, done and an optional info object as output at each step. Let consider vertical scaling action for CPU resources as $\zeta_{CPU}^{(t)}$. Therefore, due to change the allocation resources according to time slot, we have:

$$\zeta_{CPU}^{(t)} \in \{z | z \in \mathbb{R}, -K_{Net}^{(t)} \leq z \leq K_I^{(t)} - K_{Net}^{(t)}\} \quad (5)$$

One may note that vertical scaling is limited by the amount of free computing resources available on the physical server hosting the virtual machine [16].

- 3) **Reward:** The main objective of this work is minimize the total network cost where the agent learns to increase the expected return. To this end we define the return as follows:

$$R^{(t)} = \frac{1}{\mathcal{N}_T^{(t)}} \quad (6)$$

We pursue an experimental approach because maybe the total network cost ($\mathcal{N}_T^{(t)}$) is a general and imprecise metric to guide the agent for learning and leading to good results. Consequently, tuning the hyperparameters, Deep Neural Networks (DNNs) architecture and designing training steps are very tricky.

IV. TWIN DELAYED DDPG

The basic idea behind policy-based algorithms is to adjust the parameters ϕ of the policy in the direction of the performance gradient $\nabla_{\phi} J(\pi_{\phi})$. The fundamental result underlying these algorithms is the policy gradient theorem [17]: $\nabla_{\phi} J(\pi_{\phi}) = \int_S p_{\pi}(s) \int_A \nabla_{\phi} \pi_{\phi}(a|s) Q^{\pi}(s, a) da ds$, which equals to $\mathbb{E}_{s \sim p_{\pi}} [\nabla_{\phi} \pi_{\phi}(s) \nabla_a Q_{\pi}(s, a)|_{a=\pi_{\phi}(s)}]$ in deterministic policy gradient theorem..

We can parameterize policy like value function and the goal is to find the optimal policy π_{ϕ} where ϕ includes updating the weight of the policy. The expected return can be approximated in many ways. We calculate the gradient of expected return according to parameters of ϕ as $\nabla_{\phi} J(\phi)$. We use gradient ascent as opposed of gradient descent for updating the parameters, $\phi_{t+1} = \phi_t + \alpha \nabla_{\phi} J(\pi_{\phi})|_{\phi_t}$.

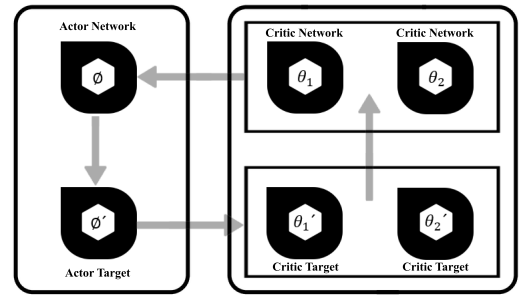


Figure 2: Training flow process between different DNNs.

In actor-critic method, we have two models that work concurrently where the actor is a policy taking state as input and delivering actions as output, while the critic takes states and actions concatenated together and return the Q-value and a policy that can be updated through the deterministic policy gradient[9], $\nabla_{\phi} J(\phi) = \mathbb{E}_{s \sim p_{\pi}} [\nabla_a Q^{\pi}(s, a)|_{a=\pi(s)} \nabla_{\phi} \pi_{\phi}(s)]$, where $Q^{\pi}(s, a) = \mathbb{E}_{s_f \sim p_{\pi} \sim a_f \sim \pi} [R_t | s, a]$ is known as value function or critic.

Initially we should store random experience in the buffer β . In the other words, we store (s_t, a_t, r_t, s_{t+1}) to train Deep Q-Network. We take a random batch B and for all transitions

Algorithm 1: TD3-based NS with OpenAI Gym

```

Initialize actor network  $\phi$  and critic networks  $\theta_1, \theta_2$ 
Initialize (copy parameters) target networks  $\phi', \theta'_1, \theta'_2$ 
Initialize replay buffer  $\beta$ 
Import network slicing environment ('smartechn-v0')
while  $t < \text{max\_timesteps}$  do
  if  $t < \text{start\_timesteps}$  then
     $a = \text{env.action\_space.sample}()$ 
  else
     $a \leftarrow \pi_\phi(s) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma)$ 
  end
  next_state, reward, done, _ = env.step(a)
  store the new transition  $(s_t, a_t, r_t, s_{t+1})$  into  $\beta$ 
  if  $t \geq \text{start\_timesteps}$  then
    sample batch of transitions  $(s_{t_B}, a_{t_B}, r_{t_B}, s_{t_B+1})$ 
     $\tilde{a} \leftarrow \pi_{\phi'}(s') + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \tilde{\sigma}), -c, c)$ 
     $Q_t = r + \gamma * \min(Q_{t1}, Q_{t2})$ 
     $L = l_{MSE}(Q_1, Q_t) + l_{MSE}(Q_2, Q_t)$ 
     $\theta_f \leftarrow \text{argmin}_{\theta_f} N^{-1} \sum (L - Q_{\theta_f(s,a)})^2$ 
    if  $t \% \text{policy\_freq} == 0$  then
       $\nabla_\phi J(\phi) = N^{-1} \sum [\nabla_a Q_{\theta_1}(s, a)|_{a=\pi(\phi)} \nabla_\phi \pi_\phi(s)]$ 
       $\theta'_f \leftarrow \tau \theta_f + (1 - \tau) \theta'_f$ 
       $\phi' \leftarrow \tau \phi + (1 - \tau) \phi'$ 
    end
  end
end
if done then
  obs, done = env.reset(), False
end
t=t+1
end

```

$(s_{t_B}, a_{t_B}, r_{t_B}, s_{t_B+1})$ of β , the predictions are $Q(s_{t_B}, a_{t_B})$ and the targets consider as optimal immediate return that are exactly first part of temporal difference learning (TD) error as $R(s_{t_B}, a_{t_B}) + \gamma \max_a (Q(s_{t_B+1}, a))$, and over the whole batch B , we calculate the loss between predictions and the targets in the batch B . Another target network is used instead of using Q-network to calculate the target to fulfill more stability for learning algorithm. As shown in figure 2, the TD3 is based on the actor-critic model that it leverages three tricks to improve algorithm:

1) **Clipped double Q-learning with pair of critic networks:**

We use two DNNs as two actor networks and denote them by ϕ as actor network and ϕ' as actor target. In addition, we create two pair of critic networks and denote them by θ_1, θ_2 for parameterization of value network and θ'_1, θ'_2 as critic targets. Indeed, two learnings happen simultaneously, namely, Q-learning and Policy learning, and they address approximation error, reduce the bias, and find the highest Q-value. This was inspired by the technique seen in [18] as Double-Q Learning. For each element and transition of batch, the actor target plays a' based on s' while we add Gaussian noise to this a' . The critic targets takes the couple (s', a') and return two Q-values Q'_{t1} and Q'_{t2} as output. Then, the $(\min Q'_{t1}, Q'_{t2})$ is considered as an approximated value for critic networks. In [19] has proposed using the target network as one of the value estimates. Given that we calculate the final target of the two value networks, we have:

$$Q_t = r + \gamma * \min(Q'_{t1}, Q'_{t2}) \quad (7)$$

then the two critic networks return two Q-values as $Q_1(s, a)$ and $Q_2(s, a)$. Next, we calculate the loss based on two critic networks and with Mean Squared Error (MSE). To

minimize the loss over iterations via back-propagation technique, we use an efficient optimizer called Adaptive Moment Estimation (Adam) [20] in our code:

$$L = l_{MSE}(Q_1, Q_t) + l_{MSE}(Q_2, Q_t) \quad (8a)$$

$$\nabla_\phi J(\phi) = N^{-1} \sum [\nabla_a Q_{\theta_1}(s, a)|_{a=\pi(\phi)} \nabla_\phi \pi_\phi(s)] \quad (8b)$$

In the next step, we explain how we update the target networks.

- 2) **Delayed policy updates and target networks:** The main idea is to update the policy network less frequently than the value network since we need to estimate the value with lower variance[21]. The update rule is given by Polyak Averaging, so we update parameters by:

$$\theta'_f \leftarrow \tau \theta_f + (1 - \tau) \theta'_f \quad (9a)$$

$$\phi' \leftarrow \tau \phi + (1 - \tau) \phi' \quad (9b)$$

where $\tau \leq 1$ is an hyperparameter to tune the speed of updating.

- 3) **Target policy smoothing and noise regularisation:**

When updating the critic, a learning target using a deterministic policy is highly susceptible to inaccuracies induced by function approximation error, increasing the variance of the target. This induced variance can be reduced through regularization [9] to be sure for the exploration of all possible continuous parameters. We add Gaussian noise to the next action a' to prevent two large actions played and disturb the state of the environment:

$$\tilde{a} \leftarrow \pi_{\phi'}(s') + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \tilde{\sigma}), -c, c) \quad (10)$$

where the noise ϵ is sampled from a Gaussian distribution with zero and certain standard deviation and clipped in a certain range of value between $-c$ and c to encourage exploration. Due to avoid the error of using the impossible value of actions, we clip the added noise to the range of possible actions ($\text{min_action}, \text{max_action}$). The TD3-based NS method is summarized in Algorithm 1.

V. NUMERICAL RESULTS

To evaluate our method described in section IV, we generate six deep neural networks which work together based on actor-critic model. The implementation is written in PyTorch, fol-

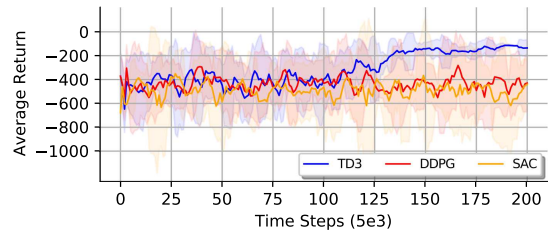


Figure 3: Learning curves for the gym NS environment.

lowed by the experimental parameters used in the simulations.

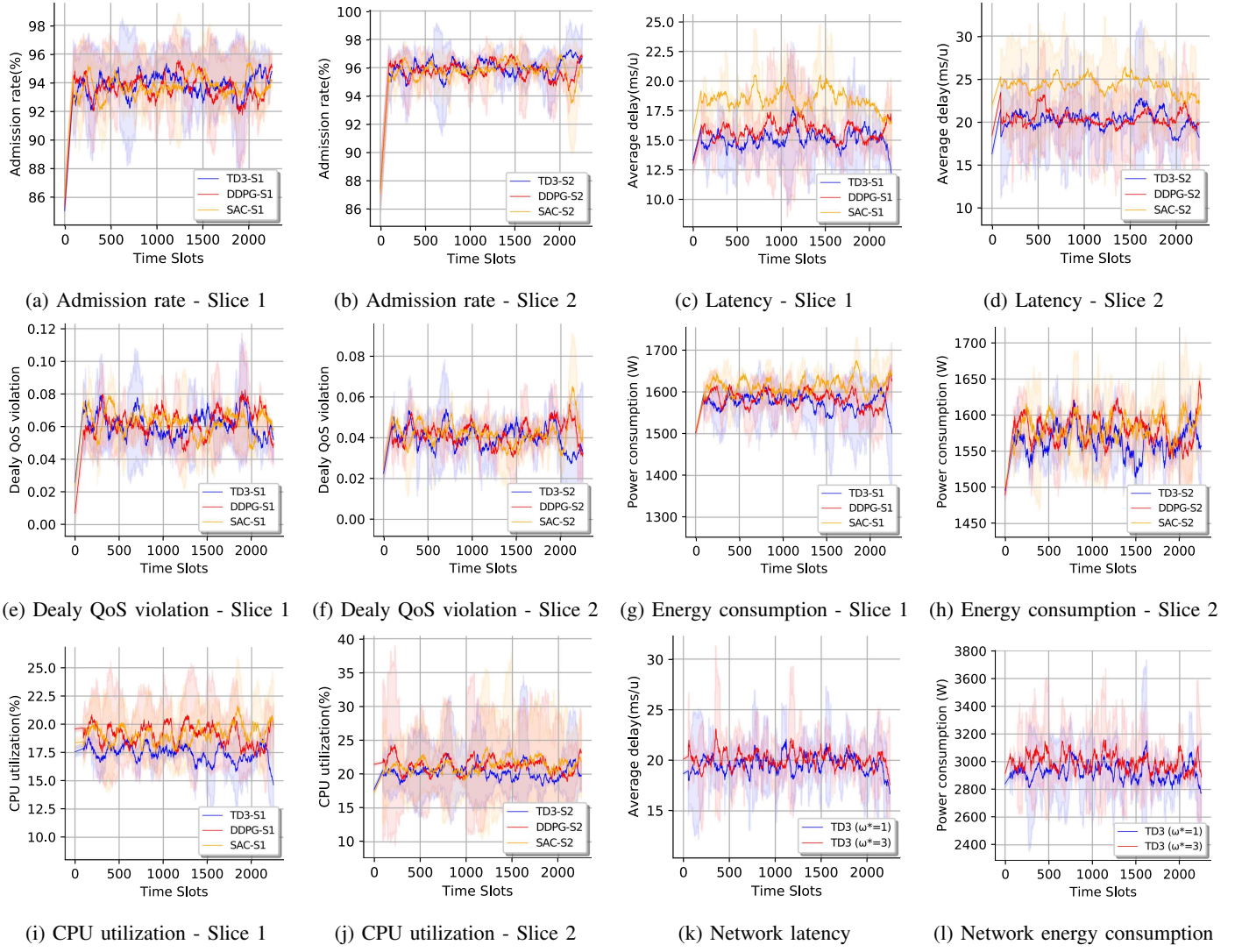


Figure 4: Network performance and costs comparison between TD3 and other DRL benchmarks. The curves are smoothed for visual clarity with respect to confidence bands and standard deviation. [$\beta_{\text{Initialization}} = 20000$, $B_{\text{size}} = 128$, $\text{Policy_freq} = 2$, $\text{Policy_noise} = 0.2$, $\text{Noise_clip} = 0.5$, $\text{Exploration_noise} = 0.1$, $\tau = 0.005$, $\text{Optimizer} : \text{Adam}$ and SGD , $\text{Activation_function} : \text{ReLU}$ and tanh , $\text{Discount_factor} = 0.99$, $N = 10$, $\text{QoS_S1 requirement } \eta_m^{(t)} = 20(\text{ms})$, $\text{QoS_S2 requirement } \eta_m^{(t)} = 40(\text{ms})$, $\sigma^* = 10^{-26}$, $P = 10^9$]

In fact, the values of parameters depend highly on capability, scenario and technology used. We measure the performance on our customized NS environment, interfaced through OpenAI Gym to fulfill reproducible comparison. In this environment, the mobile network operator (MNO) collects the free and unused resources from the tenants and when slices need more resources can receive new resources. It is done either periodically to avoid over-heading or based on requests of tenants. We consider a two-tenants scenario, i.e., two slices with different QoS requirements in terms of latency and CPU constraints. For each time step, users packets arrive into the network and the algorithm computes the computing requirements to allocate to the relevant VNF. We compare the performance of TD3 method against a fine-tuned version of the DDPG presented in [21] and [9] as well as Soft Actor-Critic (SAC) [22] to keep all

algorithms consistent. Table I presents the number of DNNs for each method. As shown in Figure 3, the learning curve of TD3 outperforms all other algorithms in final performance with respect to our big and complex state space. However, the problem formulation is general, we use constraints as penalty in implementation to leads agent to the good results and this is the reason of negative values in the learning curves.

Table I: Number of networks for algorithms.

	SAC	DDPG	TD3
Policy	1	1	1
Value	2	1	2
Target Policy	1	1	1
Target Value	2	1	2

In Figure 4, we present the results of comparison with respect to network performance and cost functions in (II) for each slice

under similar traffic patterns and also network measures based on different weights that are shown in Figures 4-(k) and 4-(l). **Admission rate:** Figures 4-(a) and 4-(b) show that the re-tuned TD3 algorithm outperforms the other approaches with respect to resource availability and constraints. The algorithm learns according to iterations or interact with the environment with different configurations of network so this is the reason of high fluctuation. The main goal for the algorithm is to find the best policy during the time. The similarity between results of slice 1 and slice 2 is because the MNO is trading-off between slices while there is a high resource availability. This approach enables slices to request more resources and results in provisioning services to more users and increasing admission rate.

Latency: As shown in figure 4-(c) and 4-(d), our solution leads to less average delay per user compared to DDPG and SAC.

Dealy QoS violation: The comparison between Delay QoS metric of TD3 and other schemes are presented in figure 4-(e) and 4-(f).

Energy consumption: Figures 4-(g) and 4-(h) show that the performance of our scheme and other methods where agent learns to satisfy another objective and minimize power consumption by decreasing VNFs instantiation and tuning wireless transmission power.

CPU utilization: As depicted in figures 4-(i) and 4-(j), the TD3 deployment leads to more efficient usage of CPU compared to other methods.

VI. CONCLUSION

In this paper, we have presented a new continuous multi-objective zero-touch NS solution. In this intent, we have developed an OpenAI Gym NS environment and used advanced DRL-based algorithm for resource allocation problem to enable CU to learn how to re-configure computing resources autonomously, aiming at minimizing the latency, energy consumption and VNF instantiation for each slice. This method leverages 3 techniques to fulfill more stability of learning algorithm. In this respect, we have compared the network performance and costs between TD3 and other DRL benchmarks. We have shown that the proposed solution outperforms other DRL methods.

ACKNOWLEDGEMENT

This work has been supported in part by the research projects 5G STEP FWD (722429), MonB5G (871780), 5G-SOLUTIONS (856691), AGAUR(2017-SGR-891) and SPOT5G (TEC2017-87456-P).

REFERENCES

- [1] Y.C. Hu, M. Patel *et al.*, "Mobile Edge Computing: A Key Technology Towards 5G," in *ETSI White Paper*, vol. 11, no. 11, pp. 1-6, 2015.
- [2] ETSI, "Zero-touch network and Service Management (ZSM); Reference Architecture", *White Paper*, Aug. 2019.
- [3] J.S. Pujol Roig *et al.*, "Management and Orchestration of Virtual Network Functions via Deep Reinforcement Learning," in *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 2, pp. 304-317, 2020.
- [4] H. Wang *et al.*, "Data-driven dynamic resource scheduling for network slicing: A Deep reinforcement learning approach," in *Inf. Sci.*, vol. 498, pp. 106-116, 2019.
- [5] G. Sun *et al.*, "Dynamic Reservation and Deep Reinforcement Learning Based Autonomous Resource Slicing for Virtualized Radio Access Networks," in *IEEE Access*, vol. 7, pp. 45758-45772, Apr. 2019.
- [6] R. Li *et al.*, "Deep reinforcement learning for resource management in network slicing," in *IEEE Access*, vol. 6, pp. 74429-74441, 2018.
- [7] Jose A. Ayala-Romero *et al.*, "vrAIn: A Deep Learning Approach Tailoring Computing and Radio Resources in Virtualized RANs," in *ACM Mobicom*, 2019.
- [8] H.P. Singh *et al.*, "Hybrid Policy Gradient for Deep Reinforcement Learning," in *The 32nd Annual Conference of the Japanese Society for Artificial Intelligence*, 2018.
- [9] S. Fujimoto *et al.*, "Addressing Function Approximation Error in Actor-Critic Methods," in *CoRR*, 2018, Available: <http://arxiv.org/abs/1802.09477>.
- [10] Y. Sun *et al.*, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," in *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1960-1971, Apr. 2019.
- [11] Y. Liao *et al.*, "How much computing capability is enough to run a cloud radio access network?" in *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 104-107, Jan. 2017.
- [12] S. Bhaumik *et al.*, "CloudIQ: A framework for processing base stations in a data center," in *Proc. ACM MobiCom*, pp. 125-136, Aug. 2012.
- [13] E. Aqeeli *et al.*, "Power-aware optimized RRH to BBU allocation in C-RAN," in *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1311-1322, Feb. 2018.
- [14] J. Tang *et al.*, "System cost minimization in cloud RAN with limited fronthaul capacity," in *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371-3384, May 2017.
- [15] N. Chee-Hock and S. Boon-Hee, "Queuing Modelling Fundamentals with Applications in Communication Networks," in *Wiley*, Second Edition, 2008.
- [16] S. Dutta *et al.*, "Smartscale: Automatic application scaling in enterprise clouds," in *IEEE 5th International Conference on Cloud Computing*, pp. 221-228, 2012.
- [17] D. Silver *et al.*, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning (ICML)*, 2014.
- [18] H. van Hasselt, "Double Q-learning," in *Advances in Neural Information Processing Systems*, vol. 23, pp. 2613-2621, 2010.
- [19] H. Van Hasselt *et al.*, "Deep reinforcement learning with double Q-learning," in *AAAI*, pp. 2094-2100, 2016.
- [20] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in *ICRL*, 2015, Available: <https://arxiv.org/abs/1412.6980>.
- [21] T. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *ICRL*, 2016, Available: <https://arxiv.org/abs/1509.02971>.
- [22] T. Haarnoja *et al.*, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," Available: <https://arxiv.org/abs/1801.01290>, 2018.