CHALMERS
UNIVERSITY OF TECHNOLOGY

Deep-reinforcement-learning-based RMSCA for space division multiplexing networks with multi-core fibers [Invited Tutorial]

Citation for the original published paper (version of record):
Teng, Y., Natalino Da Silva, C., Li, H. et al (2024). Deep-reinforcement-learning-based RMSCA for space division multiplexing networks with
multi-core fibers [Invited Tutorial]. Journal of Optical Communications and Networking, 16(7): C76-C87. http://dx.doi.org/10.1364/JOCN.518685

Deep-Reinforcement-Learning-based RMSCA for Space Division Multiplexing Networks with Multi-Core Fibers

YIRAN TENG[1], CARLOS NATALINO[2], HAIYUAN LI[1], RUIZHI YANG[1], JASSIM MAJEED[1], SEN SHEN[1], PAOLO MONTI[2], REZA NEJABATI[1], SHUANGYI YAN*[1], AND DIMITRA SIMEONIDOU[1]

[1] *High Performance Networks Group, Smart Internet Lab, University of Bristol, Bristol, United Kingdom.*
[2] *Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden.*
[*] *shuangyi.yan@bristol.ac.uk*

Compiled August 5, 2024

The escalating demands for network capacities catalyze the adoption of space division multiplexing (SDM) technologies. With the continuous advances in multi-core fiber (MCF) fabrication, MCF-based SDM networks are positioned as a viable and promising solution to achieve higher transmission capacities in multi-dimensional optical networks. However, with the extensive network resources offered by MCF-based SDM networks comes the challenge of traditional routing, modulation, spectrum, and core allocation (RMSCA) methods to achieve appropriate performance. This paper proposes an RMSCA approach based on deep reinforcement learning (DRL) for MCF-based elastic optical networks (MCF-EONs). Within the solution, a novel state representation with essential network information and a fragmentation-aware reward function were designed to direct the agent in learning effective RMSCA policies. Additionally, we adopted a proximal policy optimization algorithm featuring an action mask to enhance the sampling efficiency of the DRL agent and speed up the training process. The performance of the proposed algorithm was evaluated with two different network topologies with varying traffic load and fibers with different number of cores. The results confirmed that the proposed algorithm outperforms the heuristics and the state-of-the-art DRL-based RMSCA algorithm in reducing the service blocking probability by around 83% and 51%, respectively. Moreover, the proposed algorithm can be applied to networks with and without core switching capability, and has an inference complexity compatible with real-world deployment requirements.

https://doi.org/10.1364/JOCN.518685

1. INTRODUCTION

Cloud computing, edge computing, and beyond 5G continuously drive network traffic soaring with emerging network applications and technologies, such as industry 4.0, AR/VR and the Metaverse, multi-access edge computing (MEC), and video content distribution network (CDN) [1]. The recent multi-band transmission system that explores beyond C + L band resources in fibers is unable to satisfy the long-term capacity requirements with cumbersome band management and immature/impractical switching and transmission solutions, not to mention the lack of smooth update of the current infrastructure [2]. Therefore, both academia and industry move their focus to MCF-based SDM solutions. The advancements in MCF design and fabrication have made space division multiplexing (SDM) [3] a promising technology in elastic optical networks (EONs) [4] to support an ever-growing network traffic. Demonstrations of high-capacity transmission experiments, networking, and management solutions reassure the availability of essential enabling technologies for MCF-based SDM networks [5–7]. In 2023, the first commercial deployment of MCFs was performed by Google to boost capacities for submarine cables [8]. The potential wide deployment of MCFs raises challenges for the routing, modulation, spectrum, and core allocation (RMSCA) solution in efficiently managing links, fiber cores, and frequency slots to maximize network performance.

Numerous research efforts have been dedicated to developing RMSCA strategies that efficiently manage the spectrum resources to optimize network performance. On the technical front, the RMSCA solutions can be categorized into integer linear programming (ILP)-based [9–12], heuristic-based [13–16], and deep reinforcement learning (DRL)-based [17, 18]. Regarding ILP-based solutions, Yaghubi-Namaad *et al.* formulated the RMSCA problem as ILP in a path-based manner for static traffic scheduling to improve the spectrum utilization [9]. In [10], Zhang *et al.* designed a heterogeneous MCF (HMCF) structure, and ILP models were designed to formulate the process of

virtual optical network embedding over EONs with HMCF to reduce the fragmentation under the crosstalk constraint. While achieving near-optimal results, the computing-intensive and time-consuming property of ILPs in large-scale networks hinders its application in addressing the dynamic RMSCA problems. Therefore, the previous dynamic RMSCA solutions mainly focus on rule-based heuristic algorithms. Zhu *et al.* presented a triangular iterative core allocation strategy to mitigate crosstalk and minimize the overall blocking probability (BP) in MCF-EONs [14]. The authors in [16] proposed a crosstalk-aware and fragmentation-aware score function to evaluate each RMSCA policy, with the approach of always selecting the policy that achieves the highest score. However, rule-based RMSCA heuristics may lead to far-from-optimal solutions in MCF-EONs with high network complexity due to its inability to assess the impact of current decisions on the provisioning results of future connection requests.

Recently, DRL has emerged as a promising solution for complex network optimization problems [19]. Chen *et al.* proposed a DRL-based framework (DeepRMSA) to address the resource management problem in single-core EON. This approach has shown superior performance compared to heuristics in reducing the network blocking probability (BP) [20]. Subsequently, extensive DRL-based algorithms [21–25] have been proposed based on the DeepRMSA framework to further improve the network performance by introducing advanced neural network models [21, 23, 24] or adopting more effective reward functions [22, 25]. Compared to heuristics that rely on fixed, manually designed rules, their DRL-based solutions delve deep to extract essential network information and flexibly adapt their resource management policies to diverse network states. However, research on DRL-based RMSCA in SDM optical networks is still in its early stages. Beghelli *et al.* [17] explored DRL for resource assignment in multiband-EONs (MB-EONs) and MCF-EONs. Still, their solution could not outperform the heuristics due to an unsuitable design of the DRL agent. The DRL-RMSCA algorithm [18] presented by Pinto-Ríos *et al.* extended the DeepRMSA framework [20] to address the RMSCA problem in MCF-EONs with three cores. However, this solution is unsuitable for large-scale MCF-EONs with many cores (e.g., 7 or 12) due to its unawareness of fragmentation. Moreover, their assumption that all fiber cores are aligned to form isolated network planes simplifies the problem and fails to explore the benefit of core switching capabilities of SDM-EONs. As shown in Fig. 1, core switching allows a selected route to choose different fiber cores along the links in the path. Core switching can significantly increase the number of available core paths to be used by flexibly arranging and combining the fiber cores across the links, further enhancing the network performance. Nevertheless, the numerous available core paths introduce massive amounts of information to the DRL agent, significantly increasing the cardinality of both its observation and action spaces.

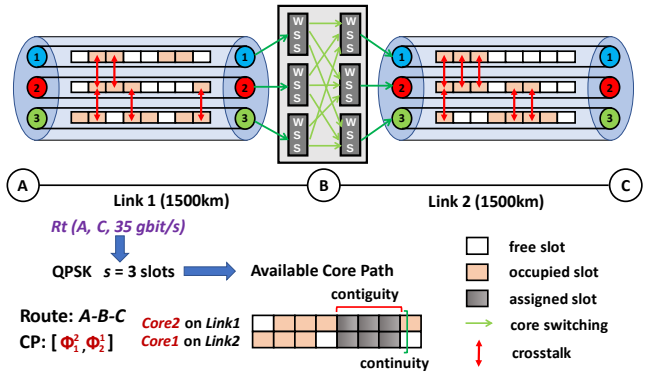


Fig. 1. Routing, modulation, spectrum, and core allocation (RMSCA) in MCF-EONs.

In summary, two critical issues exist regarding DRL-based approaches for large-scale MCF-EONs with core switching capability. Firstly, the DRL agent's scalability and reliability with respect to the number of cores need to be enhanced. For this, both an effective reward function and a detailed state representation are required. Secondly, the significant expansion of the observation and action spaces caused by multiple cores severely impedes the efficient training of the DRL agent.

To address these challenges, this paper extends our research in [26] to develop a DRL-based RMSCA framework that is tailored for MCF-EONs with varying scales, in which a DRL agent jointly solves the routing and core allocation (RCA) problems. To the best of our knowledge, this is the first work in the literature to propose a scalable DRL-based framework for the RCA problem for MCF-EONs with core switching capability. Under the framework, *(i)* the candidate core paths are pre-selected for the DRL agent to restrict its observation space; *(ii)* a detailed state representation that integrates crucial network information and request information is designed to help the DRL agent better perceive the network condition at a given point in time, incorporating a fragmentation-aware reward function to guide the agent in maximizing resource usage, thereby lowering the overall network blocking probability (BP); and *(iii)* an action mask is applied to assist the DRL agent in avoiding selecting invalid actions, improving the training performance, and reducing the training time. Simulation results over the $NSFNET$ and $COST239$ topologies with a different number of cores (i.e., 3, 7, 12) show that the proposed solution reduces BP by up to 83% when compared to RMSCA heuristics and the state-of-the-art DRL-RMSCA algorithm [18]. A sensitivity analysis reveals that the action masking and reward function play an important role in improving the performance of the DRL-based RMSCA solution, as well as broadening the applicability of the solution to scenarios with and without core switching capabilities. Moreover, the inference complexity, measured by the time taken to select an RMSCA decision, is compatible with the requirements of real-world deployments.

The rest of the paper is organized as follows. Section 2 illustrates the formulation of the RMSCA problem in MCF-EONs. Section 3 introduces the proposed DRL-Based RMSCA framework. The evaluation results of our presented algorithms are presented and analyzed in Section 4. Finally, Section 5 concludes the paper.

2. PROBLEM DESCRIPTION

The objective of an RMSCA algorithm in MCF-EONs is to minimize the total number of blocked service requests. This is done by efficiently managing frequency slots (FSs) while accounting for the crosstalk (XT) constraint. The RMSCA optimization problem addressed in this work is formulated in this section.

A. RMSCA Formulation

The topology of MCF-EONs is modeled as a graph $G(V, E, C)$, where V, E, and C represent the set of nodes, links, and weakly-coupled cores, respectively. F FSs with two states, i.e., free (1)

and occupied (0), are located on each core. Each FS carries a fixed bandwidth. The FSs state of the MCF-EONs is represented by a $E \times C \times F$ matrix denoted as $FSM \in [0,1]$. When a dynamic service request $R_t(s,d,b)$ arrives, an RMSCA policy should be formulated, first determining a route among K pre-computed candidate paths between the source node s and the destination node d. The policy should then assign S FSs on a core path (CP) of the route to establish the lightpath connection for the R_t according to its bit rate requirement b. The CP denotes the cores to be used on each link of the selected route. Each element of CP is presented as Φ_e^c, which refers to the core c on link e. The following expression determines the number of assigned FSs:

$$\mathrm{S} = \frac{b}{B_{\text{slot}} \times m} + 1, \tag{1}$$

where B_{slot} is the bandwidth of each FS (in GHz), and m is the efficiency of the modulation format (in b/Hz/s) as shown in Table 1. Finally, one FS is used as the guard band. The highest-order modulation format that does not exceed the transmission reach [27] for each route is always used to improve the spectral efficiency. When assigning the FSs, both the spectral contiguity and spectral continuity constraints must be met, where the assigned FSs must be contiguous and have the same spectrum across each chosen core.

B. Inter-Core Crosstalk Formulation

In MCF-EONs with weakly-coupled cores, the inter-core XT arises between signals transmitted in an overlapping spectrum segment in neighboring cores (Fig. 1). An elevated XT level results in significant signal distortion, degrading its quality at the receiver. To this end, it is necessary to check the XT level of each lightpath before its establishment. In this paper, the worst-case per-core XT (WCC-XT) estimation is adopted [28], integrating with a widely used analytical model proposed in [29], to evaluate the end-to-end XT of the lightpath lp (denoted as XT_{lp}), as follows:

$$XT_{lp} = \sum_{\Phi_e^c \in CP_{lp}} \frac{n_c - n_c \cdot exp[-(n_c+1) \cdot h \cdot L_e]}{1 + n_c \cdot exp[-(n_c+1) \cdot h \cdot L_e]}, \tag{2}$$

where CP_{lp} denotes the CP for lightpath lp, n_c is the number of neighboring cores of core c, L_e is the length of link e, and h is the increment of XT per unit. The value of h is obtained by:

$$h = \frac{2k^2 r}{\beta w_{tr}}, \tag{3}$$

where k is the fiber coupling coefficient, r is the fiber bending radius, β is the propagation constant and w_{tr} is the core pitch. To guarantee the quality of the signal, only those lightpaths whose XT_{lp} are lower than the XT threshold of their chosen modulation format [30] shown in Table 1 can be established.

C. Core Allocation Schemes

In MCF-EONs, core allocation methods can be categorized into two groups based on whether or not they consider the core continuity constraint. If the core continuity is considered [12, 18], each service request must use the same core across all links in its route. To flexibly utilize the spectrum resources in MCF-EONs, some studies assume core switching at selected nodes [13–15]. As shown in Fig. 1, when core switching is enabled, an incoming wavelength can be switched into any core of the connected links by deploying multiple wavelength selective switches (WSSs) in each optical degree of each node. Consequently, the RMSCA algorithm can select the core to be used at each link in the route. This paper considers the latter (i.e., core switching), which is a more encompassing problem, when developing the proposed DRL-based RMSCA framework.

Table 1. Parameters for Different Modulation Formats [27, 30].

Modulation	m $[b/Hz/s]$	Max Reach $[km]$	XT Threshold $[dB]$
$BPSK$	1	8,000	−14.0
$QPSK$	2	4,000	−18.5
$8QAM$	3	2,000	−21.0
$16QAM$	4	1,000	−25.0
$32QAM$	5	500	−27.0

3. DRL-BASED RMSCA APPROACH FOR MCF-EONS

To minimize the long-term network BP, the RMSCA problem is modeled as a Markov Decision Process (MDP), denoted by the tuple $\langle s_t, a_t, T, r_t, \gamma \rangle$. The state s_t describes the status of the MCF-EON environment, and the action a_t represents an RMSCA decision. The transition distribution $T(s_{t+1}|s_t, a_t)$ defines how the network changes after performing the action a_t. The reward r_t is obtained after applying an RMSCA decision as an incentive for the DRL agent, and the γ is a discounted factor $\in [0,1)$. In resolving the MDP, a DRL-based RMSCA algorithm is employed to find an effective RMSCA policy that maximizes the discounted cumulative reward U_t, as defined by:

$$U_t = \sum_{i=0}^{L-t} \gamma^i r_{t+i} \tag{4}$$

where L is the episode length. The framework of the proposed DRL-based RMSCA approach is shown in Fig. 2. It comprises two principal components: the *RMSCA environment* and the *DRL Agent*. The *DRL Agent* continuously optimizes its RMSCA policy through interaction with the *RMSCA environment*. When lightpath request R_t arrives at timestep t, the Preprocessor traverses the FSM to identify the candidate CPs (step 1). Next, the Feature Extractor generates the state vector s_t according to the candidate CPs (step 2). Then, s_t is fed into the DNNs, which outputs an action a_t (step 3). The RMSCA policy takes a_t and executes it on the MCF-EON (step 4). Subsequently, the Evaluator computes the reward r_t using the feedback of the RMSCA action from the environment (step 5). The tuple (s_t, a_t, r_t, s_{t+1}) is stored in the Experience Buffer as a training sample (step 6). The training is triggered when the Experience Buffer is full. At this step, the Optimizer is used to update the parameters of the DNNs (step 7). Next, we illustrate the specific design of each component within the proposed DRL-RMSCA framework.

A. RMSCA Environment

In our proposed framework, the RMSCA environment includes: *(1)* a simulated MCF-EON with the physical formulation illustrated in Section 2, *(2)* a traffic generator that generates the service request R_t at each timestep, *(3)* a preprocessor to filter the network information, *(4)* a feature extractor to generate the state s_t for the DRL agent, and *(5)* an evaluator to return the reward r_t to the DRL agent as the feedback of the action a_t.

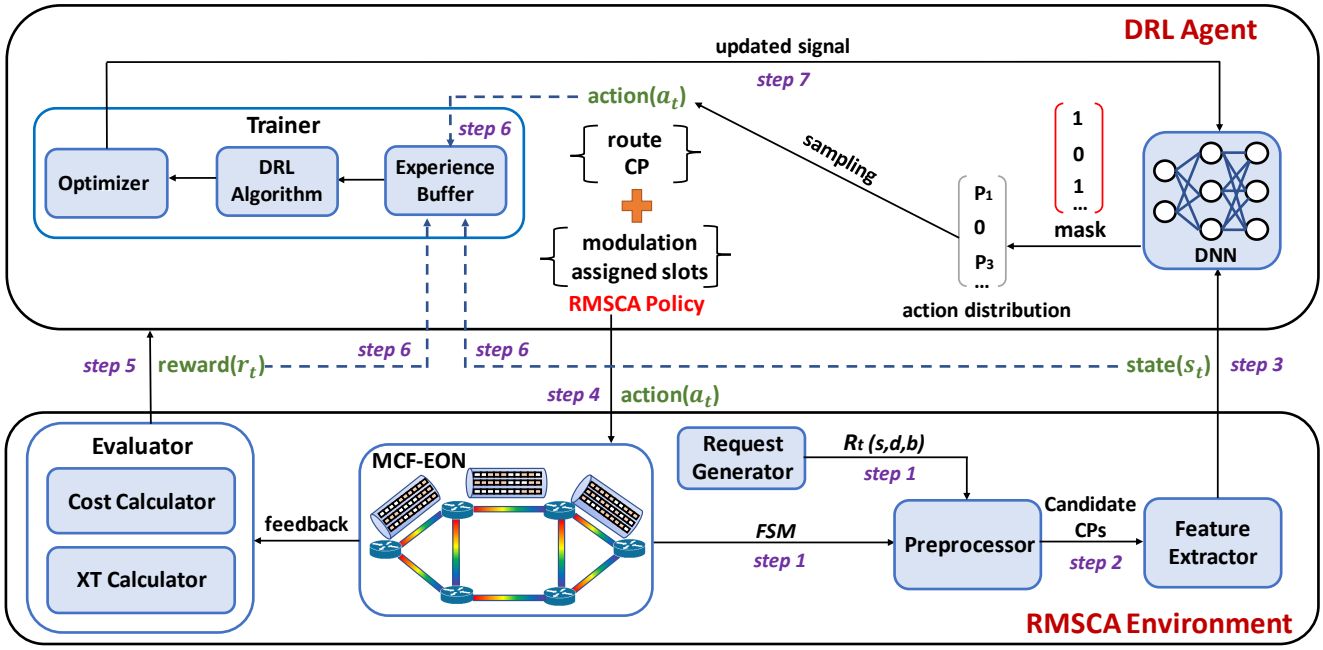


Fig. 2. Proposed DRL-based RMSCA framework.

A.1. Preprocessor

In MCF-EONs with core switching, the total number of CPs varies with the number of hops, and it is extremely large when the number of cores C is high ($|C|^j$ for a route with j hops). Exposing all the CPs to the DRL agent may lead to excessive training time due to the highly complex observation and action spaces. Moreover, it becomes challenging for the DRL agent to extract adequate information from massive features within the observation. To restrict the size of the observation/action space, we employ a preprocessor to pre-select M candidate CPs that possess the lowest starting frequency in their first available FS-block for each of the K shortest routes. The preprocessing procedure traverses the CP options and selects the M options with the suitable FS-block with the lowest frequency available.

B. DRL Agent

In the DRL-based RMSCA framework, the DRL agent is responsible for provisioning the request and learning an effective RMSCA policy by training. The main elements, i.e., state (s_t), action (a_t), and reward (r_t), are discussed next.

B.1. State Representation

To help the DRL agent efficiently perceive the environment, a clear state representation containing important information about the environment is needed. In this model, the state s_t is designed based on the request information and the spectrum state of the candidate CPs. Specifically, s_t is a $(2 \times V + (7 + |C_{cat}|) \times M \times K)$ vector, where $2 \times V$ elements in one-hot format represent the request source and destination. For each of the M candidate CPs on each of the K candidate routes, three parameters are considered: ***(i)*** number of assigned FSs when using this CP, ***(ii)*** number of hops in the route, and ***(iii)*** total number of adjacent links for all links in the route. Then, the **AND** operation is performed on the FSs with the same frequency across the cores of the CP to generate an aggregated FSs vector. In this way, the aggregated FS vector represents the available FSs for a CP (i.e., respecting the spectrum continuity constraint). Four essential features are extracted from the aggregated FSs: ***(i)*** the number of free FSs, ***(ii)*** the start index of the first available FS-block that can accommodate the request, ***(iii)*** the length of the first available FS-block, and ***(iv)*** the average number of free FSs with same spectrum as the candidate assigned FSs in first available FS-block on the adjacent links of each core. Additionally, all fiber cores are classified into $|C_{cat}|$ categories based on their number of neighboring cores. For each core category, one element is used to represent the proportion of the physical distance transmitted through cores of that category in the current CP to the total length of the CP. Figure. 3a shows the resource usage at a specific timestep in a 3-core MCF-EON, and Fig. 3b shows all available CPs that have adequate contiguous and continuous FSs to accommodate the service request R_t depicted in Fig. 3a. For the vertical core layout used in Fig. 3a, the *core* 1 and *core* 3 with one neighboring core and *core* 2 with two neighboring cores are classified into different core categories so the $|C_{cat}|$ is 2. For the CP_1 $[\Phi_1^1, \Phi_3^2]$ of the route *A-B-C* shown in Fig. 3b, the signal will be transmitted 1200 km via *core* 1 on *link* 1, and 1800 km via *core* 2 on *link* 3. Hence, the values of the corresponding elements for category 1 with one neighboring core and category 2 with two neighboring cores are 0.4 and 0.6, respectively.

Table 2 presents the features extracted from each available CP depicted in Fig. 3b. Note that if a candidate CP cannot provision the R_t, all its corresponding features are set to -1, enabling the DRL agent to distinguish it from the available CPs.

Table 2. Features of candidate core-paths (CPs).

route	CP	aggregated FSs	features
A-C	$[\Phi_2^1]$	[0 0 0 0 0 1 0 1 0]	[-1 -1 -1 -1 -1 -1 -1 -1 -1]
A-C	$[\Phi_2^3]$	[0 1 1 0 0 1 0 0 1]	[2 1 2 4 2 2 2.5 1 0]
A-B-C	$[\Phi_1^1, \Phi_3^2]$	[0 0 0 0 1 1 1 1 1]	[3 2 4 5 5 5 1.75 0.4 0.6]
A-B-C	$[\Phi_1^2, \Phi_3^3]$	[1 1 1 0 0 0 0 0 0]	[3 2 4 3 1 3 1.5 0.6 0.4]

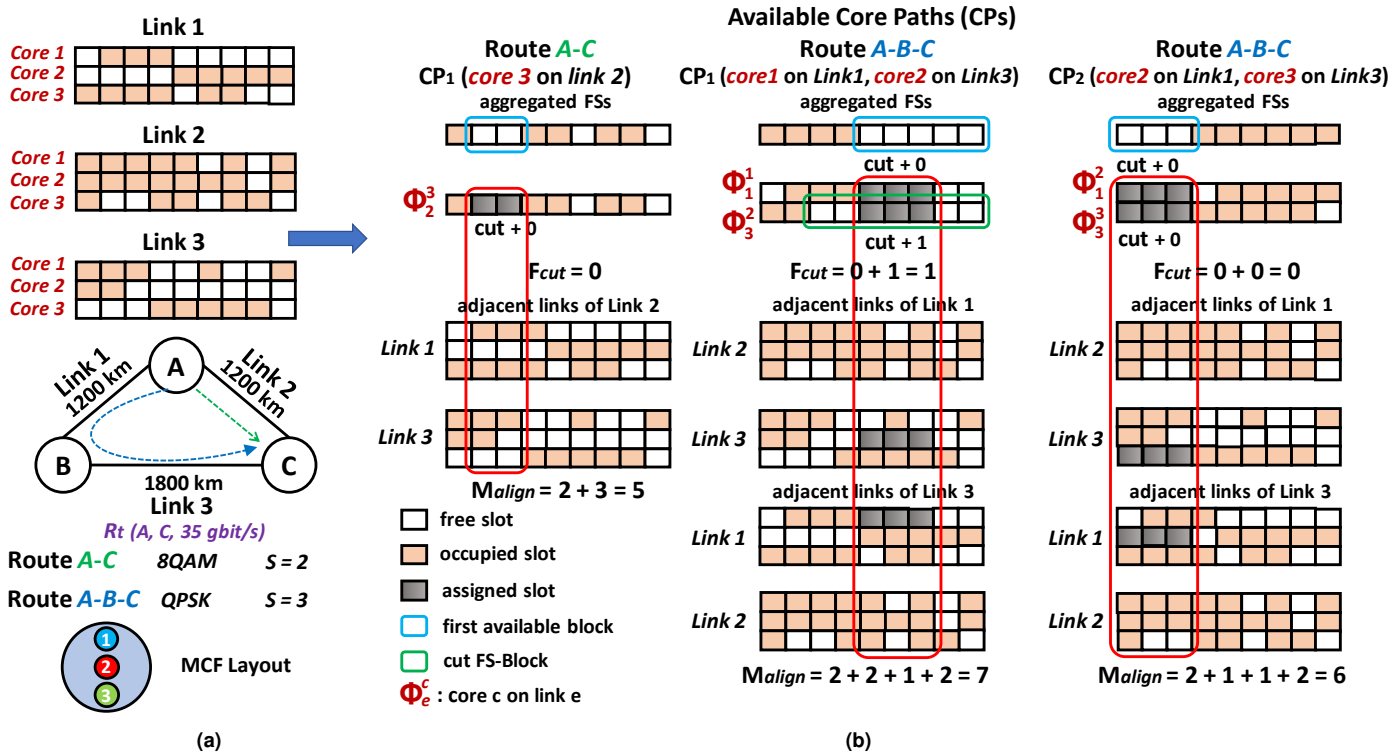


Fig. 3. (a) The resource usage in a 3-core MCF-EON, (b) available core paths (CPs) and the calculation of F_{cut} and M_{align} related to fragmentation and misalignment of spectrum slots

B.2. Action Space

The action space A is a $M \times K + 1$ vector. Each action represents either selecting one of the M CPs on one of the K shortest routes or rejecting the request. The modulation scheme is determined by the transmission reach of the route (Table 1). The spectrum is always allocated on the first available FS-block using the first-fit method.

To prevent the agent from choosing unavailable CPs, we adopt an action mask [31] (Fig. 2). The mask g_t is a $M \times K + 1$ vector $\in \{0,1\}$ corresponding to the availability of each action. After masking, the original logit z_i of each invalid action is replaced by a very small negative number (i.e., -1×10^{-8}), making their output probability p_i after the *softmax* activation function in Eq. (5) near 0, as explained next:

$$\mathrm{p_i} = \mathrm{Softmax}(\hat{z}_i) = \frac{e^{\hat{z}_i}}{\sum_{j=1}^{M\times K+1} e^{\hat{z}_j}}, \tag{5}$$

where $\hat{z}_i$ is the masked logit of the i-th action and p_i represents the probability of choosing action a_i.

B.3. Reward Function

To reduce the long-term BP of the MCF-EONs, the reward r_t should be associated with whether R_t is accepted or not, and the impact of the RMSCA decision on the MCF-EONs. An appropriate reward r_t can steer the agent towards efficient exploration of the environment, enabling it to learn an effective RMSCA strategy. However, most of the previous DRL-based RMSA/RMSCA models have not delved deeply into the design of the reward function, often relying on a basic binary reward system: +1 if a service is accepted and -1 otherwise. This reward function cannot accurately evaluate the impact of action a_t on the MCF-EON environment, as all actions that successfully establish a lightpath connection receive the same reward. When addressing the complex RMSCA problem, the performance of the DRL agent with this reward function is further degraded because the significantly expanded state vector increases the difficulty for the value DNNs to establish a clear relationship between the observation and the U_t. Therefore, evaluating the specific impact of each action a_t on the network is necessary. To this end, we developed a fragmentation-aware and load-balance-aware reward function to assess the impact of action a_t comprehensively.

Specifically, after the RMSCA action is deployed in the MCF-EON, a factor Q that represents the network cost caused by taking action a_t is calculated as follows:

$$\mathrm{Q} = \frac{S_o + \lambda \times F_{cut} + \frac{M_{align}}{|C|}}{S_a}, \tag{6}$$

where $|C|$ is the number of cores, S_o represents the total number of assigned FSs for the request R_t, and S_a is the number of available aggregated FSs on the selected CP. F_{cut} and M_{align} are related to slot fragmentation and misalignment in the network [32], respectively.

As shown in Fig. 3b, F_{cut} denotes the total number of FS-blocks on the CP that are cut into sub-blocks by the assigned FSs, and M_{align} indicates the total number of free FSs on the adjacent links of each link across the selected route that have overlapping spectrum with the assigned FSs. For example, if the CP_1 $[\Phi_1^1, \Phi_3^2]$ of the route A-B-C is selected, the FS_5, FS_6, and FS_7 will be assigned. After the RMSCA process, the FS-block FS_{5-9} on Φ_1^1 is cut to FS_{8-9} so the number of FS-blocks is not increased. However, the FS_{3-9} on Φ_3^2 is cut into two blocks: FS_{3-4} and FS_{8-9}, thus the increased cut is 1. The increase in F_{cut} may potentially lead to the rise of unassignable small-size

FS-blocks, resulting in the waste of the spectrum resources. The M_{align} represents the cumulative number of free FSs from FS_5, FS_6, and FS_7 located in the adjacent links of *link* 1 (i.e., *link* 2 and *link* 3) and *link* 3 (i.e., *link* 1 and *link* 2). The misalignment of the FSs on two adjacent links reduces commonly available spectra, which is detrimental to the provision of future requests. Additionally, the cost Q becomes higher if large number of FSs (S_o) are assigned, or if a CP with less available FSs (S_a) is used. Given the value of F_{cut} is relatively small (the cut on each link $\in$ {0,1}) compared to other parameters, we amplify its impact by multiplying it with a specific factor λ to ensure its magnitude is comparable to those of the other parameters. The Q value can comprehensively evaluate the potential impact of the current action on the network. A smaller Q value indicates efficient utilization of the spectrum resources, which is beneficial for future service provisioning.

The reward r_t is designed based on the cost Q as outlined in the following:

$$r_t = \begin{cases} -1, & \text{if } R_t \text{ is rejected} \\ 0.67 + \dfrac{N}{M \times K} \times 0.33, & \text{if } R_t \text{ is accepted, } Q = Q_{min} \\ \max(-0.1 \times Q + 0.6,\ 0), & \text{if } R_t \text{ is accepted, } Q \neq Q_{min} \end{cases} \tag{7}$$

where N is the number of available CPs among $M \times K$ candidate CPs. If there are insufficient FSs to accommodate R_t or if the XT of the candidate lightpath is over the threshold, R_t is rejected, and -1 will be given as a penalty. If R_t is provisioned, a positive reward ranging from 0 to 1 is assigned. Specifically, a large reward ranging from 0.67 to 1 is given to the agent if the selected CP leads to the lowest cost Q_{min} among all candidates, encouraging the agent to learn an efficient policy in the short term. The specific reward depends on the ratio of number of available CPs (N) to the total number of candidate CPs ($M \times K$). If the agent can select the CP with the smallest Q from a larger number of available CPs, it will receive a higher reward. When the selected CP is not the lowest one among all available CPs, the value of r_t ranges from 0 to 0.6 and is negatively correlated to the Q value.

B.4. Training

A trainer is set for the DRL agent to optimize its RMSCA policy to maximize the discounted cumulative reward U_t in Eq. (4). We use the proximal policy optimization (PPO) algorithm [33] to train the DRL agent. PPO has high stability and reliability with simple implementation. PPO adopts the actor-critic framework including a policy DNN $\pi(a_t|s_t,\theta)$ and a value DNN $V(s_t,w)$. θ and w are the sets of the parameters of the policy DNN and the value DNN, respectively. In our training algorithm, the policy DNN is represented as $\pi(a_t|s_t,g_t,\theta)$ due to implementing the action mask g_t. The $\pi(a_t|s_t,g_t,\theta)$ takes s_t as input and outputs a probability distribution over all actions after masking. The $V(s_t,w)$ estimates the expectation of U_t at state s_t. In order to efficiently utilize the training samples, a policy $\pi(a_t|s_t,g_t,\theta_k)$ is employed to interact with the environment and collect the training samples. These samples are utilized to iteratively update the $\pi(a_t|s_t,g_t,\theta)$ through importance sampling. The policy gradient $d\theta$ and the value gradient dw are computed by Eq. (8) and Eq. (9), respectively.

Algorithm 1. Training of the DRL agent

initialize experience buffer $\mathcal{D} = \emptyset$ with size Z, set epochs T
initialize θ for policy DNN and w for value DNN, $\theta_k \leftarrow \theta$
$t \leftarrow 0$, $t_0 \leftarrow 0$
for $R_t(s,d,b)$ **do**
  get state s_t and mask g_t based on candidate CPs and R_t
  obtain $\pi(a_t|s_t,g_t,\theta_k)$
  obtain action a_t by sampling $\pi(a_t|s_t,g_t,\theta_k)$
  RMSCA for MCF-EONs based on a_t
  receive reward r_t from the RMSCA environment
  store X_t $(s_t,a_t,r_t,g_t,s_{t+1})$ into $\mathcal{D}$
  if $|\mathcal{D}| == Z$ **then**
    for $t \in \{t_0, t_0+1, \ldots, t_0+Z-1\}$ **do**
      calculate U_t and A_t by Eq. (4) and Eq. (10)
      add U_t and A_t into X_t
    for $epoch = 1$ **to** T **do**
      for each mini-batch $\mathcal{B}$ in $\mathcal{D}$ **do**
        compute $d\theta$ and dw by Eq. (8) and Eq. (9)
        $\theta \leftarrow \theta + d\theta$, $w \leftarrow w + dw$
    $\theta_k \leftarrow \theta$, $t_0 \leftarrow t+1$, empty $\mathcal{D}$
  $t \leftarrow t+1$

$$d\theta = \alpha \nabla_\theta \frac{1}{|\mathcal{B}|} \sum_{X_t \in \mathcal{B}} \min(r_t(\theta)A_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right) A_t) \tag{8}$$

$$dw = \beta \nabla_w \frac{1}{|\mathcal{B}|} \sum_{X_t \in \mathcal{B}} (U_t - V(s_t,w))^2. \tag{9}$$

Here, α and β are the learning rate of the policy DNN and the value DNN, respectively. $r_t(\theta)$ is the ratio of the probability density of a_t under the updated policy $\pi(a_t|s_t,g_t,\theta)$ to that under the sampling policy $\pi(a_t|s_t,g_t,\theta_k)$. The clip factor ϵ ranges from 0 to 1 to avoid large policy updates. $|\mathcal{B}|$ represents the number of training samples X_t in the mini-batch $\mathcal{B}$. A_t denotes the advantage of taking action a_t, defined as follows:

$$A_t = U_t - V(s_t,w). \tag{10}$$

Alg. 1 outlines the entire training procedure.

4. PERFORMANCE ASSESSMENT

A. Setup and Configuration of the Simulation Environment

The MCF-EON environment was simulated by extending the basic MCF environment available in the Optical RL-Gym [34]. The evaluations were conducted over the *NSFNET* topology (Fig. 4a) with 14 nodes and 22 links, and the *COST*239 topology (Fig. 4b) with 11 nodes and 26 links. The MCF-EONs featuring 3,7 and 12 cores with the fiber layout depicted in Fig. 5 were used for simulation. We utilized five parallel environments running on separate central processing units (CPUs) for sampling to accelerate training. We employed the state-of-the-art MaskablePPO algorithm provided by *Stable Baselines* [35] for training. The parameters adopted for the RMSCA environments and the DRL agent are shown in Table 3 and Table 4, respectively. The coefficients related to the Q and the reward function were set as follows. The slot fragmentation factor λ in Eq. (6) was set to 10 to amplify the original F_{cut} value, which has a range of $(0, H_{max})$, ensuring its magnitude is comparable with other terms in the

Table 3. RMSCA Environment Settings.

Parameter	env1	env2	env3
number of cores per link	3	7	12
number of FSs per core	100	320	320
FS capacity (GHz)	12.5	12.5	12.5
required capacity (Gbps)	25-100	25-100	25-100
traffic load (Erlang)	425	4000	7500
K	5	5	5
M	1	2	2
fiber coupling coefficient	$6{\cdot}10^{-4}$	$1{\cdot}10^{-3}$	$1{\cdot}10^{-3}$
fiber bending radius (mm)	50	55	65
core pitch (μm)	45	40	35

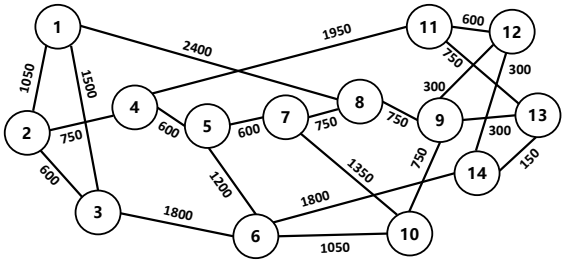
Table 4. DRL Agent Settings.

Hyperparameter	Value
learning rate for policy DNN (α)	$1e^{-4}$
learning rate for value DNN (β)	$1e^{-4}$
discounted factor (γ)	0.95
clip factor (ϵ)	0.2
epoch (T)	10
episode length (L)	1,000
buffer length (Z)	1,000
mini-batch size ($\lvert\mathcal{B}\rvert$)	500
DNN architecture	5x128 & 8x256

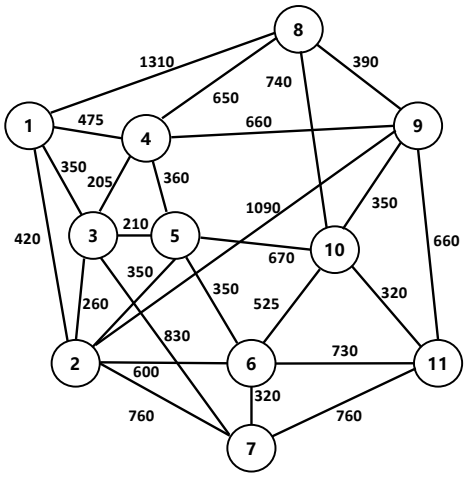
equation. H_{max} has value equal to the maximum number of hops among all candidate routes. The reward function coefficients in Eq. (7) were set based on the desired range of reward, i.e., $(0.67, 1)$ when the cost is minimum, and $(0, 0.6)$ when the cost is not minimum. These values can build a reasonable connection between Q and r_t, and keep the r_t value within an appropriate range to enhance the stability of the training.

As for the MCF-EONs under the core continuity constraint, M was set as the number of cores to ensure that all possible core allocation methods are considered. For MCF-EONs that allow core switching, the selection of the M value affects the performance of the DRL agent. Therefore, we evaluated the performance of the DRL agents across a range of M values to identify the optimal parameter for simulation. Specifically, M was set to 1, 2, 3, 4, 5, 7, 10, and 20 for the different DRL agents. We assessed the BP from deploying these DRL agents for RMSCA in *NSFNET* with 3, 7, and 12 cores. Figure. 6 demonstrates the boxplot of BP achieved by each DRL agent upon training converges at different values of M using two classical DNN architectures of sizes 5×128 (5 layers with 128 neurons per layer) and 8×256. In the boxplots, the central line marks the median, the box edges show the 25th and 75th percentiles, and the whiskers extend up to 1.5 times the interquartile range.

Figure. 6a shows the performance of DRL agents executing RMSCA in a 3-core MCF-EON, which deteriorates as M increases. When the value of M is relatively low (up to 4), the DRL agent with a 5×128 DNN demonstrates slightly better performance. Notably, the best performance is achieved when



(a) 14-node NSFNET



(b) 11-node COST239

Fig. 4. MCF-EONs topologies with link distance [km].

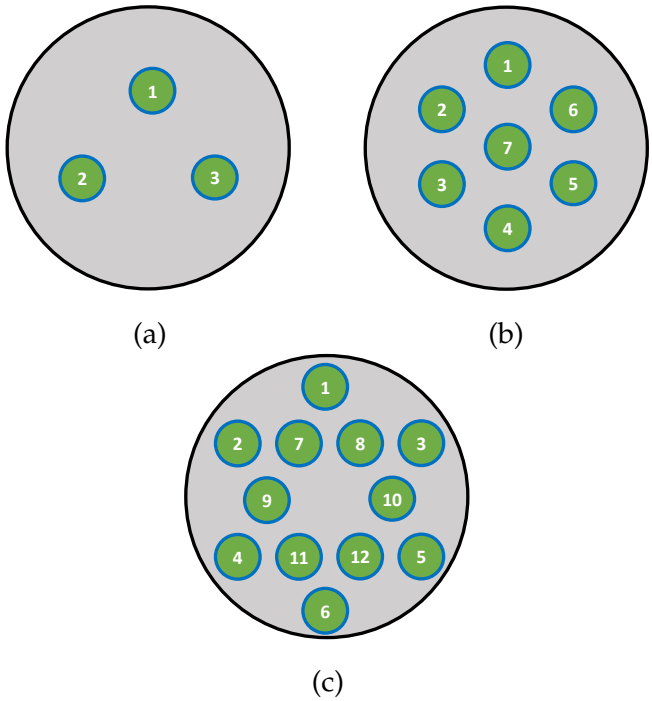
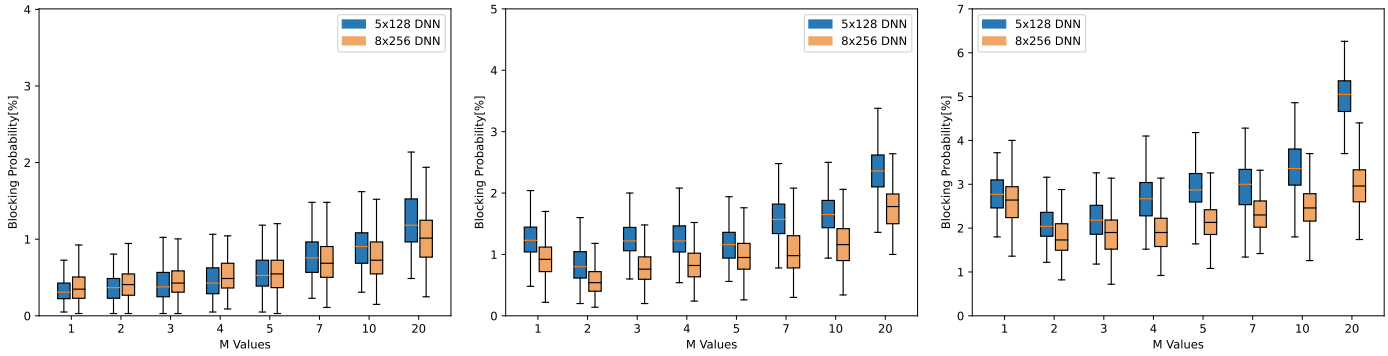
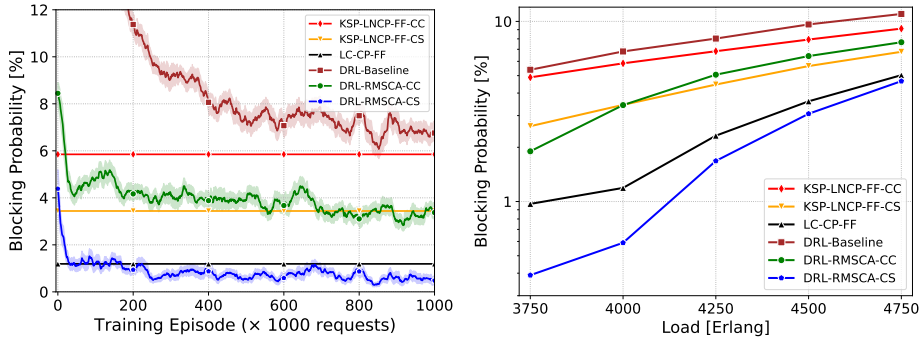


Fig. 5. The layout of MCF with (a) 3, (b) 7, and (c) 12 cores.



(a) 3 cores **(b)** 7 cores **(c)** 12 cores

Fig. 6. Blocking probability of DRL agents under different M values in NSFNET topology with 3, 7, and 12 cores.



(a) **(b)**

Fig. 7. Blocking probability of different algorithms in 7-core NSFNET topology under (a) 4000 Erlangs, (b) different traffic loads.

the DNN size is set to 5×128 and M is 1. In the MCF-EON with 7 cores, the 5×128 DNN is insufficient for the DRL agents. The performance is worse than the 8×256 DNN (Fig. 6b). The DRL agent achieves the best performance with M=2 and an 8×256 DNN architecture. For the DRL agents operating in the 12-core MCF-EON, M=2 continues to enable the DRL agent to achieve the lowest BP (Fig. 6c). The results indicate that further increasing the value of M does not necessarily enhance the performance of the DRL agent. We speculate that this is because an increase of M introduces more information to the DNNs, thereby increasing the complexity of the policy that needs to be learned by the DRL agent.

We developed two RMSCA algorithms based on the proposed framework for different core allocation rules: *DRL-RMSCA with core-continuity (DRL-RMSCA-CC)* and *DRL-RMSCA with core-switching (DRL-RMSCA-CS)*. The effectiveness of the proposed algorithms is evaluated based on the comparison with the state-of-the-art DRL-based RMSCA algorithm proposed in [18], referred to as DRL-Baseline. The DRL-Baseline adopts a straightforward binary reward function (+1 if service is accepted, −1 otherwise) and does not consider action masking. In addition, we adopt three rule-based heuristics, including *(1)* the K-shortest-path least-neighbors-core-path first-fit with core continuity (KSP-LNCP-FF-CC), *(2)* the KSP-LNCP-FF allowing core switching (KSP-LNCP-FF-CS) and *(3)* the least-cost core-path first-fit (LC-CP-FF). For the heuristics, the first two always assign FSs to the first available spectrum of the CP with the least number of neighboring cores along the shortest route to prevent the XT of the lightpath from exceeding the threshold, and the last one always allocates FSs on the available CP that results in the minimal cost Q, as calculated by Eq. (6).

In this study, we assume that the service request R_t will be blocked under two conditions: *(1)* there are no sufficient contiguous and continuous FSs to accommodate the R_t, and *(2)* the XT_{lp} of the selected lightpath lp for provisioning R_t exceeds the XT threshold.

B. Performance Evaluation

First, we assessed the performance of the proposed algorithms on the *NSFNET* topology. Figure 7a illustrates the training results of the DRL agent in an MCF-EON with 7 cores and 320 FSs (i.e., *env2* in Table 3). The x-axis represents the number of training episodes, with each episode encompassing the arrival of 1,000 requests. The y-axis indicates the average BP value for each episode. The value of the confidence interval with a 95% confidence level is presented as a shaded region around the training curve. During the initial phase of the training, the DRL agent is inefficient due to its random parameter initialization, which results in a random policy. However, the DRL agent incrementally refines its RMSCA policy through interactions with the environment, manifesting as a significant decline in BP. After processing approximately 30,000 requests, the DRL-RMSCA-CS reaches a local optimum, equivalent to the performance achieved by LC-CP-FF, followed by a plateau in performance. As for the DRL-RMSCA-CC, its performance reaches a local minimum after processing 37,000 requests, followed by a temporary degradation in performance. After 180 and 155 training episodes, respectively, the DRL agents of the

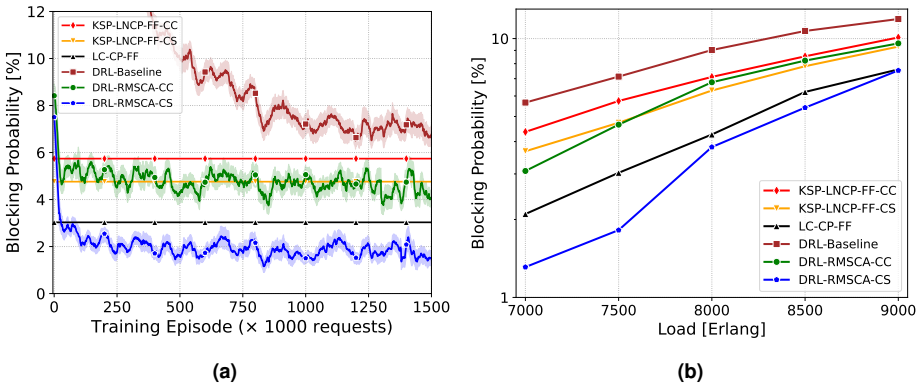


Fig. 8. Blocking probability of different algorithms in 12-core NSFNET topology under (a) 7500 Erlangs, (b) different traffic loads.

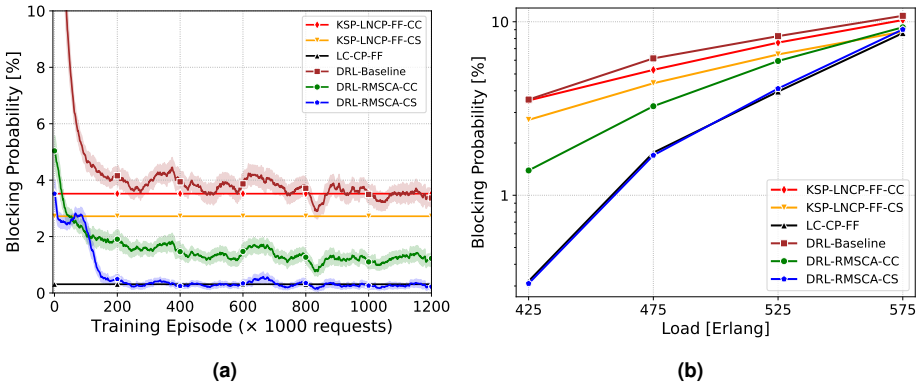


Fig. 9. Blocking probability of different algorithms in 3-core NSFNET topology under (a) 425 Erlangs, (b) different traffic loads.

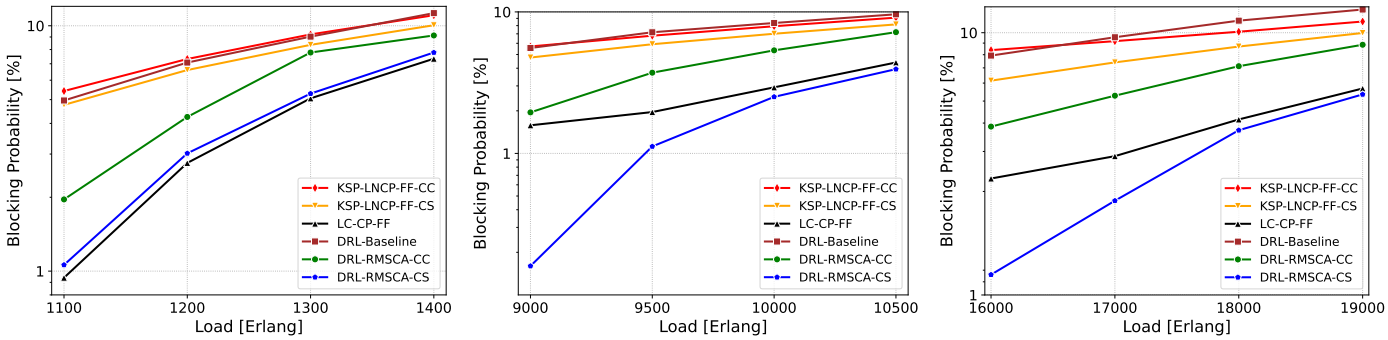


Fig. 10. Blocking probability in COST239 topology with 3, 7, and 12 cores under different traffic loads.

DRL-RMSCA-CS and DRL-RMSCA-CC begin to improve their RMSCA policies further. After processing around 280,000 and 700,000 requests, respectively, the DRL-RMSCA-CS and DRL-RMSCA-CC training curves converge, suggesting that the performance of their DRL agents has stabilized. After convergence, DRL-RMSCA-CS demonstrates superior performance among the core-switching-based algorithms. Specifically, it achieves a 83% and 50% reduction in BP compared with KSP-LNCP-FF-CS and LC-CP-FF, respectively. For the algorithms under core continuity constraint, DRL-RMSCA-CC decreases BP by 41% and 51% compared with KSP-LNCP-FF-CC and DRL-Baseline, respectively. Meanwhile, the results show that enabling core switching provides additional flexibility, leading to improved utilization of network resources. Specifically, the DRL-RMSCA-CS can achieve 83% lower BP than DRL-RMSCA-CC. However, although undergoing appropriate training, the performance of DRL-Baseline cannot surpass the KSP-LNCP-FF-CC and shows significant disadvantages compared to DRL-RMSCA-CC. Additionally, DRL-Baseline requires 200 more training episodes to reach convergence than DRL-RMSCA-CC. We believe there are two primary reasons for these discrepancies. Firstly, the DRL-Baseline requires a large number of samples to learn how to avoid selecting unavailable CPs, which increases the complexity and difficulty of the training. This is not the case in our proposed approaches due to the action masking. Secondly, when the observation space and the action space are large, the binary reward

function used by DRL-Baseline leads to a long exploration by the DRL agent, making it struggle to learn behaviors beneficial for reducing BP (e.g., mitigate fragmentation and minimize the use of bottleneck links). As shown in Fig. 7b, DRL-RMSCA-CS performs well under various traffic loads. The advantages are particularly significant under low traffic loads, where core choice plays a more substantial role in blocking requests.

Next, the evaluations were conducted in a large-scale 12-core MCF-EON with high network complexity (i.e., *env3* in Table 3). As shown in Fig. 8a, DRL-RMSCA-CC and DRL-RMSCA-CS can effectively train their DRL agent. Specifically, the reductions in the BP achieved by DRL-RMSCA-CS over KSP-LNCP-FF-CS and LC-CP-FF are 63% and 40%, respectively. The DRL-RMSCA-CC realizes 35% and 19% lower BP than DRL-Baseline and KSP-LNCP-FF-CC, respectively. Meanwhile, DRL-RMSCA-CS maintains superior performance across various traffic loads (Fig. 8b).

In a small-scale MCF-EON with 3 cores (i.e., *env1* in Table 3) the results shown in Fig. 9 indicate that DRL-RMSCA-CS still achieves a significant advantage compared to KSP-LNCP-FF-CS. However, its performance is very close to that of LC-CP-FF. This is because when XT is negligible, selecting actions with lower costs allows the agent to obtain a higher discounted reward. As a result, the DRL agent eventually learns an RMSCA policy similar to LC-CP-FF.

To verify the generality of the proposed algorithms, the evaluations were also conducted in the *COST*239 topology (Fig. 4b). The appropriate values for M and the DNN architectures for MCF-EONs with 3, 7 and 12 cores were determined to be (M=2, 5×128 DNNs), (M=3, 8×256 DNNs), and (M=3, 8×256 DNNs), respectively, following an analysis similar to the one for *NSFNET* shown in Fig. 6. As shown in Fig. 10, the DRL-RMSCA-CS still demonstrates effective performance as it can obtain the lowest BP in MCF-EONs with 7 cores and 12 cores across different traffic loads compared to other algorithms. With 3 cores, where the XT has small impact, DRL-RMSCA-CS follows closely the performance of the LC-CP-FF heuristic. Meanwhile, the DRL-RMSCA-CC outperforms the other RMSCA algorithms under the core continuity constraint.

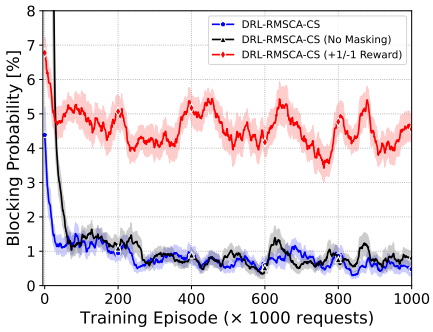
C. Sensitivity Analysis on the DRL Design

This work proposes the use of two novel components to the DRL framework: *(i)* action masking and *(ii)* fragmentation-aware reward function. These two components are crucial in augmenting the performance of our proposed DRL-based RMSCA algorithm. In this section, we performed a sensitivity analysis to validate the effectiveness of these components. This involved a comparison between the proposed DRL-based RMSCA algorithm that adopts both the two components with solutions without either one of them. In the case of action masking, we trained an agent without this feature. In the case of the fragmentation-aware reward function, we trained an agent using a simple +1/-1 reward function. All other components of the DRL agent are kept unchanged. The simulations were conducted in the NSFNET topology with 7 cores (i.e., *env2* in Table 3). The performance of the three DRL-based RMSCA algorithms was evaluated in both MCF-EONs with core switching capability (Fig. 11a) and under core continuity constraint (Fig. 11b).

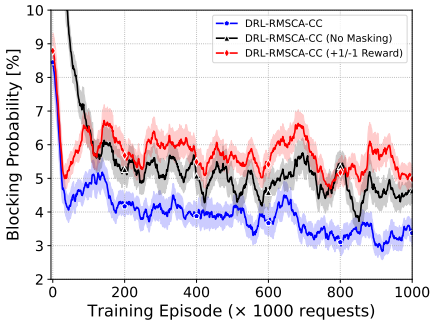
In the case of action masking, the gains vary depending on the ability of performing core switching. When core switching is available (Fig. 11a), the gains of applying action masking are moderate but noteworthy. Firstly, we can see that the agent with action masking is able to reduce BP with less training episodes than the one without it. Moreover, the average BP after convergence is reduced by 8%, from 0.65% to 0.6%. When core switching is not available (Fig. 11b) the gains obtained by action masking are more substantial. Specifically, The average BP after convergence is reduced by 28% from 4.7% to 3.4 %. This is partially due to the large difference in the number of actions available for each scenario, i.e., 11 actions for the core-switching based and 36 actions for the core-continuity based. Again, the agent with action masking is able to more quickly learn how to reduce BP as it does not expend the training samples for learning to avoid the selection of unavailable CPs.

When it comes to the fragmentation-aware reward function, in MCF-EONs with core switching capability (Fig. 11a), its adoption yields an approximate 86% BP reduction over the traditional $+1/-1$ reward. However, in the scenario under core continuity constraint (Fig. 11b), the gains are not as substantial, but still show up to 38 % lower BP after convergence.

These results validate that the proposed action masking and fragmentation-aware reward function are crucial in enhancing the performance of our DRL-based RMSCA solution. Owing partially to these two key components, our solution is versatile enough to be applicable in MCF-EONs with and without core switching capability.



(a) Core-switching based scenario



(b) Core-continuity based scenario

Fig. 11. DRL design sensitivity analysis in the 7-core NSFNET with 4,000 Erlangs.

D. Complexity Analysis

The scalability and the feasibility of our algorithm were also evaluated based on the training time and the RMSCA decision time of the DRL agent. We chose to train the DNNs on the CPU, as the current DNN architectures that are not very deep (5×128 and 8×256), and the training speed on the CPU is slightly faster than that observed on the Graphics Processing Unit (GPU). Specifically, the simulations were conducted on a 12th-generation Intel

i7 CPU operating at 2.2GHz. In the MCF-EONs with 3, 7, and 12 cores (described in Table 3), the DRL agents of DRL-RMSCA-CC converge after approximately 1.1, 2.7, and 3.1 hours of training, respectively, while for DRL-RMSCA-CS, the training times are 1.2, 2.6, and 6.0 hours, respectively. After the training is completed, we assessed the average RMSCA decision time taken by the DRL agent in provisioning each request (Fig. 12), which includes the total time spent on pre-selecting candidate CPs, generating the observation, and performing forward propagation through the DNN to determine the RMSCA action. The results indicate that as the number of cores increases, the DRL agent's training time and request response time remain within a reasonable range.

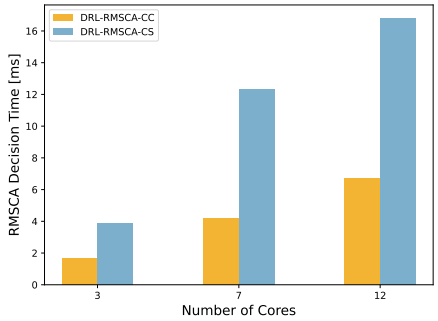


Fig. 12. RMSCA time of the DRL agent in NSFNET topology with 3, 7, and 12 cores.

5. CONCLUSIONS

In this paper, we proposed a DRL-based solution for the dynamic RMSCA problem in SDM-EONs with MCFs. A DRL agent was developed and trained based on the proposed DRL-based framework to handle requests while minimizing the long-term service blocking probability. In the proposed DRL-based framework, a comprehensive path-level state representation that contains both the network spectrum availability and the core allocation information was developed to assist the DRL agent in capturing the essential network information. Additionally, a fragmentation-aware reward function capable of precisely evaluating the impact of the RMSCA decision on the MCF-EONs was designed to help the DRL effectively explore the environment. These elements were integrated with the state-of-the-art Maskable PPO algorithm, guiding the DRL agent towards an efficient RMSCA strategy. The scenarios where core allocation with and without core continuity constraint were considered, and corresponding algorithms were developed based on the proposed framework. The performance assessments under various scales of MCF-EONs and traffic loads demonstrated the advantage of our proposed solution in reducing the BP compared with both the heuristics and the other DRL-based algorithms from the literature. The sensitivity analysis highlighted the contribution of the proposed innovative components in improving the performance of the DRL-based RMSCA solution. The complexity analysis validated the feasibility of our proposed approach for use in a real-world operational environment due to its suitable decision time.

ACKNOWLEDGEMENT

This work was partially supported by the CSA Catapult - UoB collaboration project, and by Sweden's innovation agency VINNOVA, within the framework of the EUREKA cluster CELTIC-NEXT project AI-NET-PROTECT (2020-03506). The authors also acknowledge the support of the China Scholarship Council (CSC)/University of Bristol joint-funded scholarships program.

REFERENCES

1. K. Samdanis and T. Taleb, "The road beyond 5G: A vision and insight of the key technologies," IEEE Netw. **34**, 135–141 (2020).
2. A. Napoli, N. Costa, J. K. Fischer, J. ao Pedro, S. Abrate, N. Calabretta, W. Forysiak, E. Pincemin, J. P.-P. Gimenez, C. Matrakidis, G. Roelkens, and V. Curri, "Towards multiband optical systems," in *Advanced Photonics (BGPP, IPR, NP, NOMA, Sensors, Networks, SPPCom, SOF)*, (2018), p. NeTu3E.1.
3. D. J. Richardson, J. M. Fini, and L. E. Nelson, "Space-division multiplexing in optical fibres," Nat. photonics **7**, 354–362 (2013).
4. O. Gerstel, M. Jinno, A. Lord, and S. B. Yoo, "Elastic optical networking: A new dawn for the optical layer?" IEEE Commun. Mag. **50**, s12–s20 (2012).
5. B. Puttnam, R. Luis, W. Klaus, J. Sakaguchi, J.-M. Delgado Mendinueta, Y. Awaji, N. Wada, Y. Tamura, T. Hayashi, M. Hirano, and J. Marciante, "2.15 Pb/s transmission using a 22 core homogeneous single-mode multi-core fiber and wideband optical comb," in *European Conference on Optical Communication (ECOC)*, (2015), pp. 1–3.
6. R. Yang, L. Liu, S. Yan, and D. Simeonidou, "A programmable ROADM system for SDM/WDM networks," Appl. Sci. **11**, 4195 (2021).
7. N. Amaya, S. Yan, M. Channegowda, B. R. Rofoee, Y. Shu, M. Rashidi, Y. Ou, E. Hugues-Salas, G. Zervas, R. Nejabati, D. Simeonidou, B. Puttnam, W. Klaus, J. Sakaguchi, T. Miyazawa, Y. Awaji, H. Harai, and N. Wada, "Software defined networking (SDN) over space division multiplexing (SDM) optical networks: Features, benefits and experimental demonstration," Opt. Express **22**, 3638–3647 (2024).
8. B. Quigley and M. Cantono, "Delivering multi-core fiber technology in subsea cables," .
9. M. Yaghubi-Namaad, A. G. Rahbar, and B. Alizadeh, "Adaptive modulation and flexible resource allocation in space-division-multiplexed elastic optical networks," J. Opt. Commun. Netw. **10**, 240–251 (2018).
10. Q. Zhang, X. Zhang, X. Gong, and L. Guo, "Crosstalk-avoid virtual optical network embedding over elastic optical networks with heterogeneous multi-core fibers," J. Light. Technol. **40**, 7687–7700 (2022).
11. K. Takeda, T. Sato, B. C. Chatterjee, and E. Oki, "Joint inter-core crosstalk- and intra-core impairment-aware lightpath provisioning model in space-division multiplexing elastic optical networks," IEEE Transactions on Netw. Serv. Manag. **19**, 4323–4337 (2022).
12. S. Zhang and K. L. Yeung, "Efficient embedding of service function chains in space-division multiplexing elastic optical networks," Comput. Networks **233**, 109869 (2023).
13. H. Tode and Y. Hirota, "Routing, spectrum and core assignment for space division multiplexing elastic optical networks," in *International Telecommunications Network Strategy and Planning Symposium (Networks)*, (2014), pp. 1–7.
14. R. Zhu, A. Samuel, P. Wang, S. Li, B. K. Oun, L. Li, P. Lv, M. Xu, and S. Yu, "Protected resource allocation in space division multiplexing-elastic optical networks with fluctuating traffic," J. Netw. Comput. Appl. **174**, 102887 (2021).
15. A. Mahmoudi, A. Ghaffarpour Rahbar, and M. Jafari-Beyrami, "QoS-aware routing, space, and spectrum assignment in space division multiplexing networks," Comput. Networks **208**, 108920 (2022).
16. J. L. Ravipudi and M. Brandt-Pearce, "Impairment- and fragmentation-aware, energy-efficient dynamic RMSCA for SDM-EONs," J. Opt. Commun. Netw. **15**, D10–D22 (2023).
17. A. Beghelli, P. Morales, E. Viera, N. Jara, D. Bórquez-Paredes, A. Leiva, and G. Saavedra, "Approaches to dynamic provisioning in multiband elastic optical networks," in *2023 International Conference on Optical Network Design and Modeling (ONDM)*, (IEEE, 2023), pp. 1–6.
18. J. Pinto-Ríos, F. Calderón, A. Leiva, G. Hermosilla, A. Beghelli, D. Bórquez-Paredes, A. Lozada, N. Jara, R. Olivares, and G. Saavedra, "Resource allocation in multicore elastic optical networks: A deep reinforcement learning approach," Complexity. (2023).

19. N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," IEEE Commun. Surv. & Tutorials **21**, 3133–3174 (2019).
20. X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. B. Yoo, "DeepRMSA: A deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks," J. Light. Technol. **37**, 4155–4163 (2019).
21. M. Shimoda and T. Tanaka, "Mask RSA: End-to-end reinforcement learning-based routing and spectrum assignment in elastic optical networks," in *European Conference on Optical Communication (ECOC),* (IEEE, 2021), p. Th1E.4.
22. B. Tang, Y.-C. Huang, Y. Xue, and W. Zhou, "Heuristic reward design for deep reinforcement learning-based routing, modulation and spectrum assignment of elastic optical networks," IEEE Commun. Lett. **26**, 2675–2679 (2022).
23. L. Xu, Y.-C. Huang, Y. Xue, and X. Hu, "Deep reinforcement learning-based routing and spectrum assignment of EONs by exploiting GCN and RNN for feature extraction," J. Light. Technol. **40**, 4945–4955 (2022).
24. J. Momo Ziazet and B. Jaumard, "Deep reinforcement learning for network provisioning in elastic optical networks," in *ICC 2022 - IEEE International Conference on Communications,* (2022), pp. 4450–4455.
25. E. Etezadi, C. Natalino, R. Diaz, A. Lindgren, S. Melin, L. Wosinska, P. Monti, and M. Furdek, "Deep reinforcement learning for proactive spectrum defragmentation in elastic optical networks," J. Opt. Commun. Netw. **15**, E86–E96 (2023).
26. Y. Teng, R. Yang, C. Natalino, S. Shen, P. Monti, R. Nejabati, S. Yan, and D. Simeonidou, "DRL-based RMSCA for sdm networks with core switching in multi-core fibres," in *2023 International Conference on Photonics in Switching and Computing (PSC),* (2023), pp. 1–3.
27. Ítalo Brasileiro, L. Costa, and A. Drummond, "A survey on challenges of spatial division multiplexing enabled elastic optical networks," Opt. Switch. Netw. **38**, 100584 (2020).
28. M. Klinkowski and G. Zalewski, "Dynamic crosstalk-aware lightpath provisioning in spectrally-spatially flexible optical networks," J. Opt. Commun. Netw. **11**, 213–225 (2019).
29. T. Hayashi, T. Taru, O. Shimakawa, T. Sasaki, and E. Sasaoka, "Design and fabrication of ultra-low crosstalk and low-loss multi-core fiber," Opt. express **19**, 16576–16592 (2011).
30. A. Muhammad, G. Zervas, and R. Forchheimer, "Resource allocation for space-division multiplexing: Optical white box versus optical black box networking," J. Light. Technol. **33**, 4928–4941 (2015).
31. S. Huang and S. Ontañón, "A closer look at invalid action masking in policy gradient algorithms," The Int. FLAIRS Conf. Proc. (2022).
32. Y. Yin, M. Zhang, Z. Zhu, and S. J. B. Yoo, "Fragmentation-aware routing, modulation and spectrum assignment algorithms in elastic optical networks," in *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC),* (2013), pp. 1–3.
33. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347 (2017).
34. C. Natalino and P. Monti, "The Optical RL-Gym: An open-source toolkit for applying reinforcement learning in optical networks," in *International Conference on Transparent Optical Networks (ICTON),* (2020).
35. A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," The J. Mach. Learn. Res. **22**, 12348–12355 (2021).