Deep Reinforcement Learning Based Dynamic Channel Allocation Algorithm in Multibeam Satellite Systems

SHUAIJUN LIU [ID], XIN HU, AND WEIDONG WANG

Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
Information and Electronics Technology Lab, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Xin Hu (huxin2016@bupt.edu.cn)

ABSTRACT Dynamic channel allocation (DCA) is the key technology to efficiently utilize the spectrum resources and decrease the co-channel interference for multibeam satellite systems. Most works allocate the channel on the basis of the beam traffic load or the user terminal distribution of the current moment. These greedy-like algorithms neglect the intrinsic temporal correlation among the sequential channel allocation decisions, resulting in the spectrum resources underutilization. To solve this problem, a novel deep reinforcement learning (DRL)-based DCA (DRL-DCA) algorithm is proposed. Specifically, the DCA optimization problem, which aims at minimizing the service blocking probability, is formulated in the multibeam satellite systems. Due to the temporal correlation property, the DCA optimization problem is modeled as the Markov decision process (MDP) which is the dominant analytical approach in DRL. In modeled MDP, the system state is reformulated into an image-like fashion, and then, convolutional neural network is used to extract useful features. Simulation results show that the DRL-DCA algorithm can decrease the blocking probability and improve the carried traffic and spectrum efficiency compared with other channel allocation algorithms.

INDEX TERMS Dynamic channel allocation (DCA), multibeam satellite systems, Markov decision process (MDP), deep reinforcement learning (DRL), blocking probability.

I. INTRODUCTION

With the increasing demand for high quality and low cost services, multibeam satellite systems have evolved to equip the multibeam transmitters with flexible on-board payloads. High Throughput Satellite (HTS) and High Capacity Satellite (HiCapS) both characterized by a large number of beams and flexible on-board payloads have improved the system performance effectively [1], [2]. The flexibility provided by multibeam satellite systems allows dynamic channel allocations to efficiently exploit the system spectrum resources. As channel reuse technology may bring about severe co-channel interference (CCI) thus degrading the service quality, the appropriate and effective DCA algorithm is needed to further improve the system performance by fully exploiting the potential benefits from multibeam operation [3]–[6].

As an effective technique to improve the spectrum utilization and the service quality, DCA has been studied in many research works for multibeam satellite systems. In [7],

two DCA algorithms are proposed based on pre-defined cost functions. Compared with the fixed channel allocation (FCA) algorithm, the two DCA algorithms of [7] can achieve lower blocking probability at the cost of higher computational overhead. Reference [8] adopts the low-complexity DCA algorithms based on interference measurement (IM-DCA) and user location (L-DCA) to decrease the co-channel interference. Focusing on the service quality of beam edge users, [9] proposes the beam cooperation based DCA algorithm. However, it may result in new challenges on the feed sharing and the combination methods. Recently, [10] improves the IM-DCA with service quality threshold control while [11] improves the L-DCA with channel segregation using priority channels to guarantee the service quality. The proposed DCA algorithms in [10] and [11] tend to allocate the channel with the minimum interference level in a greedy-like manner. However, this greedy-like manner neglects the intrinsic temporal correlation during the DCA sequential decision

making process in multibeam satellite systems, resulting in the spectrum resources underutilization.

In multibeam satellite systems, the DCA algorithms should consider the beam traffic, the UTs distribution and the influence on the latter channel allocation decisions. In addition, the DCA algorithms also subject to the satellite on-board power and co-channel interference (CCI) constraints. The difficulty lies in accurately modeling this temporal correlation characteristic and making better channel allocations decisions, in this way to maximize system performance during a long-term period. Above all, the DCA optimization problem of multibeam satellite systems is a sequential decision making problem embedded in complicated environments. The emerging DRL method shows great prospect in sequential decision making problems [12]–[15] and has been applied in many dynamic resource management problems [16]–[20]. By incorporating both the advantages of the perception of deep learning (DL) and decision making of reinforcement learning (RL), the DRL is able to output control signal directly based on high-dimensional environments. Considering the complexity of DCA optimization problem of multibeam satellite systems, this paper utilizes DRL to deal with the curse of dimensionality problem, which is intractable in traditional RL. Based on the observation that DCA optimization problem is a temporal correlated sequential decision making problem, and the DRL is one of the effective solutions to this problem, the DRL-DCA algorithm is proposed in multibeam satellite systems.

In this paper, the DCA optimization problem, which aims at minimizing the service blocking probability, is formulated in multibeam satellite systems. Then the DCA optimization problem is modeled as the MDP which is a dominant analytical framework for DRL. System state, action and reward are defined in modeled MDP by characterizing the formulated DCA optimization problem. Furthermore, the state reformulation and deep convolutional neural network (CNN) are adopted to extract the useful features. To improve the stability of the DRL-DCA algorithm, the experience replay and target network techniques are adopted. Simulation is conducted to evaluate the performance of the proposed DRL-DCA algorithm. Results demonstrate that the proposed DRL-DCA algorithm can achieve lower blocking probability in various simulation environments. Under the same blocking probability performance, the DRL-DCA algorithm can improve the carried traffic in about $24.4 - 41.7\%$ and 8.5%, and improve the spectrum efficiency in about 21.7% and 5.3%, compared with the traditional FCA and IM-DCA, respectively.

The main contributions of this paper are summarized as follows:

- The DCA optimization problem, which aims at minimizing the service blocking probability, is formulated in multibeam satellite systems.
- The DCA optimization problem is modeled as the MDP, where the system state, action and reward are defined.
- The state reformulation and deep convolutional neural network are used to extract the useful features.

The rest of this paper is organized as follows: Section II describes the multibeam satellite system model and DCA optimization problem formulation. In Section III, the detailed designation and implementation of the proposed DRL-DCA algorithm are addressed. Section IV presents and analyzes the simulation results. Section V concludes the paper.

Notations: Uppercase boldfaces, lowercase boldfaces and normal letters denote the matrices, column vectors and scalars, respectively, such as $\mathbf{H}$, $\mathbf{h}$ and k. Uppercase curlycue represents a set, such as $\mathcal{U}$. $\mathbb{C}^{M \times N}$ and $\mathbb{R}^{M \times N}$ denote the space of complex and real $M \times N$, respectively. $(\cdot)^{-1}$, $(\cdot)^{T}$ and $(\cdot)^{H}$ denote the inversion, the transpose and conjugate transpose operation, respectively.

II. SYSTEM MODEL AND PROBLEM FORMULATION
In this section, we first present the system model on multibeam satellite scenario. Then the DCA optimization problem is formulated.

A. SYSTEM MODEL
This paper considers a scenario where multibeam satellite systems provide direct-to-user services for ground user terminals (UTs). The multibeam satellite generates a geographical footprint subdivided into N beams, represented as $\mathcal{B} = \{n \,|\, n = 1, 2, \dots, N\}$. Figure.1 illustrates a geographical footprint of $N = 37$ multibeams. System available bandwidth B_{tot} is divided in M channels providing frequency granularity of $B_c = B_{tot}/M$, constituting the available channel set $\mathcal{C} = \{m \,|\, m = 1, 2, \dots, M\}$. Number of K UTs, denoted as $\mathcal{U} = \{k \,|\, k = 1, 2, \dots, K\}$ are served by the multibeam satellite system. Let the $\mathbf{w}_k = [w_{k,1}, w_{k,2}, ..., w_{k,M}]^T$ be the channel allocation (CA) vector for UT k representing the transmitting power on each allocated channel. Then, all UTs' CA vectors form the satellite system CA matrix per-UT basis,
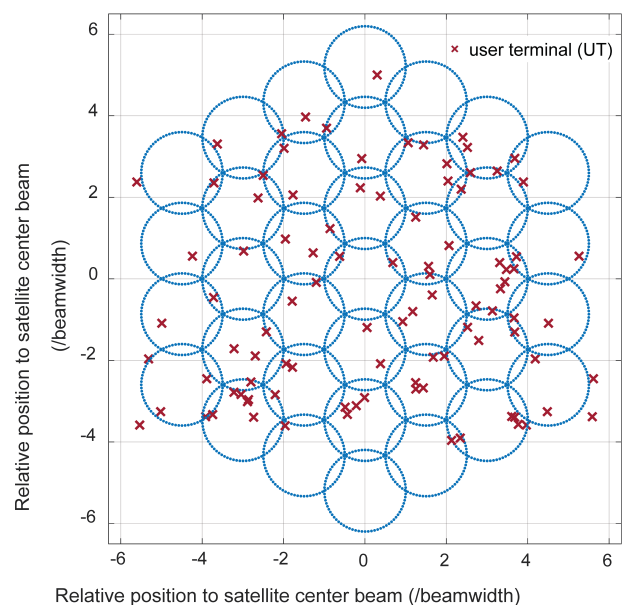

FIGURE 1. Multibeam satellite scenario with 37 beams.

denoted as $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ... \mathbf{w}_K]$, $\mathbf{W} \in \mathbb{R}^{M \times K}$. Satellite on-board maximum power is denoted as P_{tot}, and the maximum power for each beam is denoted as P_b.

Let $\mathbf{s}_k = [s_{k,1}, s_{k,2}, \dots, s_{k,M}]^T$ be symbols transmitted to UT k, assuming normalized amplitude $|s_{k,m}|^2 = 1, \forall k, m$ without loss of generality. Signal attenuation from satellite multibeam transmitter to UT receiver is represented as $\mathbf{H} = \{h_{i,j} | 1 \le i, j \le K\} \in \mathbb{R}^{K \times K}$ taking into consideration of the channel gain, transmitting and receiving antenna gain, etc. Specifically, let $\mathbf{A} = diag\{\alpha_1, \alpha_2, \cdots, \alpha_K\}$ denote the channel gain matrix, $\mathbf{G}_B = \{g_{k,n}^b | 1 \le k \le K, 1 \le n \le N\}$ denote the N multibeam transmitting antenna gain towards K UTs, and $\mathbf{G}_U = diag\{g_1^u, g_2^u, \cdots, g_K^u\}$ denote K UTs receiving antenna gain. UT k is served by the beam which can provide maximum receiving signal level. Thus, the UT-beam association matrix $\mathbf{X}$ is expressed as (1):

$$\mathbf{X} = \{x_{k,n} | x_{k,n} \in \{0, 1\}, \quad \sum_{n=1}^{N} x_{k,n} = 1, \forall k, n\} \quad (1)$$

In eq.(1), $x_{k,n} = 1$ means the UT k is associated with beam n, and $x_{k,n} = 0$ otherwise. Then the total signal attenuation $\mathbf{H}$ from multibeam transmitters to UTs receivers is expressed as eq.(2):

$$\mathbf{H} = \mathbf{A} \cdot \mathbf{G}_U \cdot \mathbf{G}_B \cdot \mathbf{X}^T \quad (2)$$

Through the eq.(1) and eq.(2), the received signal can be expressed as the desired signal and the CCI signal plus noise as eq.(3):

$$\mathbf{y}_k = h_{k,k} \cdot \mathbf{w}_k \odot \mathbf{s}_k + \sum_{i=1, i \ne k}^{K} h_{k,i} \cdot \mathbf{w}_i \odot \mathbf{s}_i + \sigma_k \quad (3)$$

where the operator $\odot$ means the Hadamard product. The $h_{i,j}$ represents the channel gain from transmitting signal j to receiving signal i, $h_{i,j} \in \mathbf{H}$, and σ_k is the noise mainly due to the UT receiving antenna thermal noise.

To calculate the CCI, we define the $\mathbf{V}$ as the CA matrix per-channel basis, where $\mathbf{V} = \mathbf{W}^T$, $\mathbf{V} \in \mathbb{R}^{K \times M}$. Each vector $\mathbf{v}_m = [v_{m,1}, v_{m,2}, \cdots, v_{m,K}]^T$ represents the transmitting power for each UT on the channel m. From the eq.(3), the desired signal power $\mathbf{D}_k$ and CCI power $\mathbf{I}_k$ for UT k on each channel can be calculated as eq.(4) and eq.(5), respectively.

$$\mathbf{D}_k = |h_{k,k}|^2 \cdot diag\{\mathbf{w}_k\} \cdot [diag\{\mathbf{w}_k\}]^H \quad (4)$$

$$\mathbf{I}_k = diag\{[\mathbf{g}_k \cdot \mathbf{v}_m \cdot \mathbf{v}_m^H \cdot \mathbf{g}_k^H]_{m=1,2,\cdots,M}\} \quad (5)$$

where $\mathbf{g}_k = [h_{k,1}, h_{k,2}, \cdots, h_{k,K}]|_{(h_{k,k}=0)}$.

Through the eq.(3) and eq.(5), the interference plus the noise $\mathbf{U}_k$ for UT k is expressed as (6):

$$\mathbf{U}_k = \mathbf{I}_k + |\sigma_k|^2 \cdot \mathbf{E}_M \quad (6)$$

where $\mathbf{E}_M$ is the M-order identity matrix.

Based on the desired signal $\mathbf{D}_k$ calculated through eq.(4) and the interference plus noise $\mathbf{U}_k$ calculated through eq.(6), the service quality for UT k in terms of Shannon capacity can

be given as eq.(7).

$$C_k = B_c \cdot det[log_2(\mathbf{E}_M + \Gamma_k)] \quad (7)$$

where $\Gamma_k = \mathbf{D}_k \cdot \mathbf{U}_k^{-1}$ is the signal and interference plus noise ratio (SINR) of UT k receiving signal.

To guarantee the service quality of UT k, a minimum capacity C_{Th} is required, namely $C_k \ge C_{Th}$. Otherwise, the service is dropped or blocked.

B. PROBLEM FORMULATION

The channel allocation problem in multibeam satellite systems can be viewed as the sequential decision making problems. From this aspect, the multibeam satellite system is modeled as a discrete-time event system, driven by the new service arrival events. At each time step t, denote the u^t as the event UT and the b^t as the event beam where the new service arrives. From the eq.(1), we know the $b^t = \arg\max_n x_{u^t, n}$. Once new arrival service event occurs, the satellite system checks whether if there exist available channels for the new arrival u^t to provide the required service quality satisfying the satellite on-board power constraints meanwhile not degrading other existing services quality. If there exist available channels, the system makes decisions on the CA vector $\mathbf{w}_{u^t}$ for the new arrival UT u^t. Otherwise, the service is blocked. An performance indicator Φ^t representing whether the new service is blocked or not is defined as eq.(8):

$$\Phi^t = \begin{cases} 1, & \text{new arrival UT } u^t \text{ is blocked} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Considering the CA decisions should satisfy the satellite on-board power constraints, we define the CA matrix per-beam basis $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N] \in \mathbb{R}^{M \times N}$, where the vector $\mathbf{f}_n$ represents the transmitting power on each channel for beam n. Given the UT-beam association matrix $\mathbf{X}$ through eq.(1) and the CA matrix per-UT basis $\mathbf{W}$, the CA matrix per-beam basis $\mathbf{F}$ can be calculated as eq.(9).

$$\mathbf{F} = \mathbf{W} \cdot \mathbf{X} \quad (9)$$

Focusing on minimizing the service blocking probability, this paper aims at finding the optimal DCA algorithm so that the number of blocked service is minimized during a long-term period T. To further decrease the optimization problem complexity, we assume the equal transmitting power on each channel and the maximum number of one channel is allowed for each UT. Then, the DCA optimization problem can be modeled as eq.(10) subject to eq.(11)-(14).

$$opt. \mathcal{P}(\mathcal{U}^t, \mathbf{W}^t, u^t) = \min_{\mathbf{w}_{u^t}} \sum_{t=1}^{T} \Phi^t \quad (10)$$

$$s.t. \sum_{n=1}^{N} \mathbf{f}_n^t \cdot (\mathbf{f}_n^t)^H \le P_{tot}, \quad \forall t \quad (11)$$

$$\mathbf{f}_n^t \cdot (\mathbf{f}_n^t)^H \le P_b, \quad \forall n, t \quad (12)$$

$$C_k \ge C_{Th}, \quad \forall k \in \mathcal{U}^t \quad (13)$$

$$\sum_{m=1}^{M} |w_{u^t, m}|^2 \le P_c, \quad |w_{u^t, m}|^2 \in \{0, P_c\}, \forall m \quad (14)$$

where $\mathbf{W}^t$, $\mathbf{f}_n^t$ and $\mathcal{U}^t$ represent the CA matrix per-UT basis, the CA vector per-beam basis for beam n and the serving UTs set at the time step t, respectively.

For the optimization problem $\mathcal{P}$ in eq.(10), constraint in eq.(11) means the total transmitting power should not exceed total satellite on-board power, and constraint in eq.(12) means that any beam power should not exceed the beam power limitation. Constraint in eq.(13) means current CA decision should not degrade the existing service quality. Constraint in eq.(14) means the equal power on each channel and the maximum number of one channel is allowed for each UT. From the aforementioned description, we can know that the formulated problem $\mathcal{P}$ can be viewed as a temporal correlated sequential decision making optimization problem with complex constraints. While the DRL method is one of the effective ways to solve this problem, the DRL-DCA algorithm is proposed which will be further described in detail in the next section.

Notes: The notations and corresponding descriptions in this section are summarized in Table 1.

TABLE 1. Notations and corresponding descriptions.

Notation	Description
$\mathbf{A}$	channel gain, $\mathbf{A} \in \mathbb{C}^{K \times K}$
$\mathcal{B}$	The set of beams, $\mathcal{B} = \{n \,|\, n = 1, 2, ..., N\}$
$\mathcal{C}$	The set of channels, $\mathcal{C} = \{m \,|\, m = 1, 2, ..., M\}$
C_k	Shannon capacity of UT k, $C_k \in \mathbb{R}$
C_{Th}	minimum traffic capacity threshold, $C_{Th} \in \mathbb{R}$
$\mathbf{D}_k$	desired signal power of UT k on each channel, $\mathbf{D}_k \in \mathbb{R}^{M \times M}$
$\mathbf{F}$	CA matrix per-beam basis, $\mathbf{F} \in \mathbb{R}^{M \times N}$
$\mathbf{f}_n$	CA vector of beam n, $\mathbf{f}_n \in \mathbb{R}^{M \times 1}$
$\mathbf{G}_B$	transmitting multibeam antenna gain, $\mathbf{G}_B \in \mathbb{R}^{K \times N}$
$\mathbf{G}_U$	receiving UT antenna gain, $\mathbf{G}_U \in \mathbb{R}^{K \times K}$
$\mathbf{H}$	total gain between transmitter and receiver, $\mathbf{H} \in \mathbb{R}^{K \times K}$
$\mathbf{I}_k$	CCI power of UT k on each channel, $\mathbf{I}_k \in \mathbb{R}^{M \times M}$
P_b	satellite beam power limitation, $P_b \in \mathbb{R}$
P_c	satellite saturation power per channel, $P_c \in \mathbb{R}$
P_{tot}	satellite total power limitation, $P_{tot} \in \mathbb{R}$
$\mathbf{s}_k$	transmitting signal for UT k, $\mathbf{s}_k \in \mathbb{C}^{M \times 1}$
$\mathcal{U}$	The set of UTs, $\mathcal{U} = \{k \,|\, k = 1, 2, ..., K\}$
$\mathbf{U}_k$	CCI plus noise power of UT k on each channel, $\mathbf{U}_k \in \mathbb{R}^{M \times M}$
$\mathbf{V}$	CA matrix per-channel basis, $\mathbf{V} \in \mathbb{R}^{K \times M}$
$\mathbf{v}_m$	CA vector of channel m, $\mathbf{v}_m \in \mathbb{R}^{K \times 1}$
$\mathbf{W}$	CA matrix per-UT basis, $\mathbf{W} \in \mathbb{R}^{M \times K}$
$\mathbf{w}_k$	CA vector of UT k, $\mathbf{w}_k \in \mathbb{R}^{M \times 1}$
$\mathbf{X}$	UT-beam association matrix, $\mathbf{X} \in \{0, 1\}^{K \times N}$
$\mathbf{y}_k$	receiving signal of UT k, $\mathbf{y}_k \in \mathbb{C}^{M \times 1}$
$\mathbf{\Gamma}_k$	SINR of UT k on each channel, $\mathbf{\Gamma}_k \in \mathbb{R}^{M \times M}$
Φ	service blocked indicator, $\Phi \in \{0, 1\}$

III. PROPOSED DRL-DCA ALGORITHM

In this section, we describe the proposed DRL-DCA algorithm in multibeam satellite systems. We first illustrate the DRL-DCA architecture. Then, the optimization problem $\mathcal{P}$ is modeled as the MDP, where the state, action and reward are defined. Then, the state is reformulated into an image-like fashion. At last, the detailed implementation of the DRL-DCA algorithm is illustrated.

A. DRL-DCA ARCHITECTURE

The main idea of the DRL-DCA algorithm is to model the multibeam satellite system as the agent and the service event as the environment. The architecture of the DRL-DCA algorithm is shown in Figure.2.
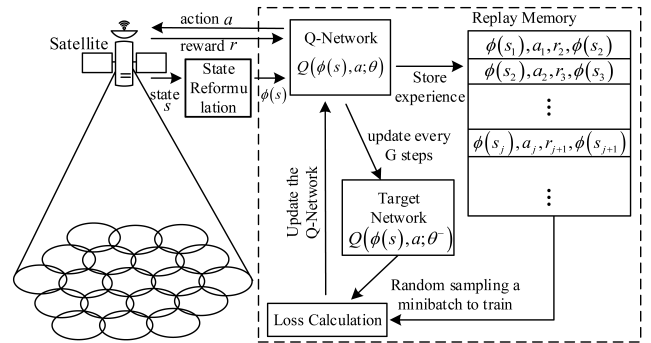

FIGURE 2. DRL-DCA Architecture.

For the DRL-DCA architecture, the state s, action a and reward r are defined in the modeled MDP. Then the state s is reformulated into an image tensor $\phi(s)$ to take full advantage of the CNN. The Q-Network $Q(\phi(s), a; \theta)$ with parameters θ is the action-value function in charge of mapping the input environments to output CA decisions. During each mapping, Q-Network generates an history result consisting of current state $\phi(s_j)$, current action a_j, instant reward r_{j+1}, the next state $\phi(s_{j+1})$ and stores them into the replay memory $\mathcal{D}$. The target network $\hat{Q}$ with parameter θ^- is copied from the Q-Network every G steps. At each time step, a minibatch randomly sampled from the replay memory $\mathcal{D}$ together with the target network $\hat{Q}$ is used to calculate the loss and train the Q-Network. A detailed description on the MDP model, the state reformulation and the implementation of the proposed DRL-DCA algorithm will be given in the following subsections.

B. MDP MODEL

The MDP is set of sequential decision making process with Markov property. The MDP contains a set of states $s \in \mathcal{S}$, a set of actions $a \in \mathcal{A}$, a reward function $r \in \mathbb{R}$, and a series of transition probabilities $p(s_{t+1}|s_t, a_t)$ of moving from the current state s_t to the next state s_{t+1} given an action a_t. The goal of an MDP is to find a policy that maximizes the expected accumulated rewards $R_t = \sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i}$, where r_{t+i} is the immediate reward at the $t+i$ time stamp and $\gamma \in [0, 1]$ is the discount factor.

We adopt the model-free method meaning that there is no knowledge of the transition probabilities $p(s_{t+1}|s_t, a_t)$. Based on the formulated DCA optimization problem $\mathcal{P}$ as eq.(10), we define the state, action and reward in the modeled MDP as following. An illustration of the modeled MDP is showed in Figure.3.

1) STATE

The state is an abstraction of the environment based on which the agent makes the action decisions. From the optimization
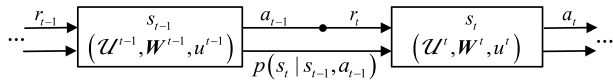
FIGURE 3. MDP Model.

problem $\mathcal{P}$, CA vector $\mathbf{w}_{u^t}$ decision depends on the current UTs set $\mathcal{U}^t$, the CA matrix per-UT basis $\mathbf{W}^t$ and the new arrival UT u^t. Based on this observation, we define the system state as eq.(15).

$$s_t = (\mathcal{U}^t, \mathbf{W}^t, u^t) \qquad (15)$$

2) ACTION
In modeled MDP, the agent should make decisions to take actions based on the system state s_t defined in eq.(15). For DCA optimization problem $\mathcal{P}$, the decision is to determine the CA vector $\mathbf{w}_{u^t}$ for the new arrival UT u^t while satisfying the constraints (11)-(14). From the eq.(14), we can know that at most one element in $\mathbf{w}_{u^t}$ is nonzero. In the case that there is available channels which can provide required service for u^t while satisfying (11)-(14), we denote such channels set as $\mathcal{A}(s_t)$, where $\mathcal{A}(s_t) \subset \mathcal{C}$. Otherwise, the new arrival service is blocked, $\mathcal{A}(s_t) = \emptyset$. Based on this observation, we define the action a_t as an index indicating which channel is allocated at this time step t as eq.(16).

$$a_t = \{m \mid m \in \mathcal{A}(s_t)\} \qquad (16)$$

3) REWARD
In modeled MDP, the agent tries to maximize the accumulated rewards. In the optimization problem $\mathcal{P}$, the goal is to minimize the number of the blocked services. Thus, we can define the reward in the principle that a positive reward r_{SF} is set when the new service is satisfied while a negative reward r_{BL} is set when the service is blocked. Based on this observation, the reward is defined as eq.(17).

$$r_t = \begin{cases} r_{BL}, & \Phi^t = 1 \\ r_{SF}, & \text{otherwise} \end{cases} \qquad (17)$$

C. STATE REFORMULATION
In this section, we propose a state reformulation allowing the system state s_t to be represented in an image-like fashion, here we call it image tensor, $\phi(s_t)$. The reason for reformulating the state into an image-like fashion lies in the following three aspects.

- Constructing the input data into a certain structured format. As the deep neural network (DNN) has many layers and each layer has the specific input and output format, the input data for DNN should be regular-structured.
- Extracting the spatial features of the DCA optimization problem. In multibeam satellite systems, the defined system state s_t has the spatial correlation features mainly due to the co-channel interference (CCI) mechanism. This is because the CCI depends on the geographical relative location of UTs which occupied the same channels.

- Pre-processing the state to decrease the computational complexity. In multibeam satellite systems, the number of beams can be hundreds to thousands. In fact, the useful features of state s_t for decision making mainly depend on the new arrival u^t's surrounding beams.

Based on the aforementioned factors, we reformulate the state into an image-like fashion through two steps. The first step can be done by extracting the partial useful beams $\mathcal{B}_\phi$ information from all beams $\mathcal{B}$ as eq.(18).

$$\mathcal{B}_\phi = \{n | \theta_{n,b^t} \le \theta_{Th}, n \in \mathcal{B}\} \qquad (18)$$

where b^t is the event beam, $\theta_{i,j}$ means the angle isolation between beam i and beam j. Here, we set the $\theta_{Th} = \sqrt{3}\theta_{BW}$ meaning the neighboring two layers of beams is remained.

Through the extracted beam information in eq.(18), the corresponding UTs set in the $\mathcal{B}_\phi$ can be denoted as eq.(19).

$$\mathcal{U}_\phi = \{k | x_{k,n} = 1, k \in \mathcal{U}, n \in \mathcal{B}_\phi\} \qquad (19)$$

Second step is to reformulate the extracted information as the image tensor $\phi(s_t)$ so that we can take advantage of CNN to deal with the reformulated image. The image tensor is of size $L_w \times L_h \times (M+1)$, where the $M+1$ represents the number of images while the L_w and L_h are the width and height of each image. The useful information on beams and UTs are broken down into $M + 1$ images, where each image of $m \in [1, M]$ represents the CA matrix consisting of the beams $\mathcal{B}_\phi$ and serving UTs $\mathcal{U}_\phi$ for the channel m, while the image $M+1$ represents the arrival UT. The pixel value of each image is in range of $\{0, 1\}$, where value 1 means there is an existing(/new arrival) UT for an image $m \in \{1, \cdots, M\}(/m = M + 1)$. Then the reformulated image tensor $\phi(s_t)$ can be represented as $\phi(s_t) \in \{0, 1\}^{L_w \times L_h \times (M+1)}$.

The state reformulation process is illustrated in Figure.4, where the marker "∘" represents the new arrival UT while the marker "×" represents the existing UTs. For existing UTs, different colors represent the different allocated channels. For example, the red color means the channel $m = 1$ and the green color means the channel $m = M$. In the DRL-DCA algorithm, we set the $L_w = L_h = 10$.
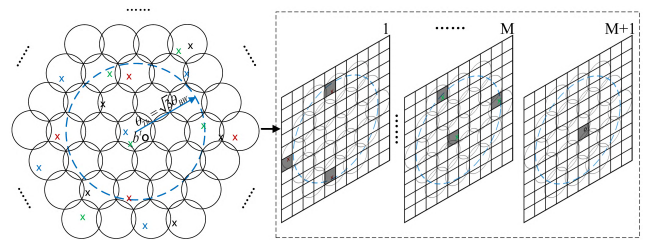

FIGURE 4. Illustration of state reformulation process.

D. DRL-DCA IMPLEMENTATION
The DRL-DCA algorithm implementation mainly refers the Deep Q-Network (DQN) algorithm proposed in [13].

Three main aspects about the Q-Network architecture, the Q-Network update and the action selection policy are addressed in the following.

1) Q-NETWORK ARCHITECTURE

The Q-Network acts as the decision making functionality mapping the input $\phi(s_t)$ to the output action value, $Q : \phi(s_t) \rightarrow Q(\phi(s_t), a_t; \theta)$, which represents the expected accumulated rewards for taking action a_t under the situation $\phi(s_t)$.

In the DRL-DCA algorithm, the Q-Network adopts the CNN as the non-linear function approximator, which consists of two convolutional (Conv) layers and two fully-connected (FC) layers. The first convolutional layer, Conv1, consists of 16 kernel each size is 5×5, with 'sigmoid' activation function. The second convolutional layer, Conv2, consists of 32 kernel each size is 3×3, with a same nonlinear activation function. The first fully-connected layer, FC1, reshapes the Conv2 output and takes as the input. The second layer is backward connected to M possible actions. The Q-Network architecture is illustrated in Table.2.

TABLE 2. Architecture of the Q-Network.

layer	Input	Kernel	Activation	Output
Conv1	$10 \times 10 \times (M+1)$	$5 \times 5, 16$	Sigmoid	$6 \times 6 \times 16$
Conv2	$6 \times 6 \times 16$	$3 \times 3, 32$	Sigmoid	$4 \times 4 \times 32$
FC1	512	NA	Sigmoid	512
FC2	512	NA	Sigmoid	M

2) Q-NETWORK UPDATE

Traditional RL is known to be unstable or even to diverge when a nonlinear function approximator such as a neural network is used to represent the action-value function [21]. This instability has several causes: the correlations present in the sequence of observations, the fact that small updates to action-value functions may significantly change the policy and therefore change the data distribution, and the correlations between the action-values and the target values. To solve this problem, we adopt the experience replay and target network to improve the Q-Network stability.

In the DRL-DCA algorithm, the replay memory $\mathcal{D}$ with capacity N_{ep} is emptied in the initialization stage. Then during the training and operating process, new generated experience tuple $(\phi(s_t), a_t, r_{t+1}, \phi(s_{t+1}))$ is stacked into the $\mathcal{D}$. Once the size of stored experiences achieves the number N_{st}, the DRL-DCA algorithm starts training the Q-Network. During the training, a minibatch data with size of N_{mb} is randomly sampled from the $\mathcal{D}$. For each experience tuple of the N_{mb} sampled minibatch, the target network $\hat{Q}(\phi(s), a; \theta^-)$ is used to calculate the loss as eq.(20).

$$L(\theta) = E[(y_j - Q(\phi(s), a; \theta_j))^2] \quad (20)$$

where y_j is the target value and calculated as eq.(21).

$$y_j = \begin{cases} r_{j+1}, & \text{if } \mathcal{A}(s_{j+1}) = \emptyset \\ r_{j+1} + \gamma \max_{a \in \mathcal{A}(s_{j+1})} \hat{Q}(\phi(s_{j+1}), a; \theta^-), & \text{else} \end{cases} \quad (21)$$

Through the calculated loss $L(\theta)$, the stochastic gradient descent (SGD) method is adopted to train the Q-Network [23]. During the training process, batch normalization (BN) technology is adopted to accelerate the training by reducing internal covariate shift effect [24].

3) ACTION SELECTION

On the action selection strategy, the $\epsilon - greedy$ policy is adopted to balance the exploration and exploitation, i.e. to balance the reward maximization based on the knowledge already known with trying new actions to obtain knowledge unknown. In the DRL-DCA algorithm implementation, we linearly decrease the exploration rate ϵ from initial value ϵ_i to final value ϵ_f during the training.

The DRL-DCA algorithm process is illustrated in Table.3.

TABLE 3. The DRL-DCA Algorithm.

Algorithm 1: The DRL-DCA Implementation

1. **Initialization**
2.   Initialize scenario parameters $\mathbf{W} = \emptyset, \mathcal{U} = \emptyset$
3.   Initialize the statistic parameters $N_{block} = 0, N_{arrival} = 0$
4.   Initialize replay memory $\mathcal{D}$ to capacity N_{ep}
5.   Initialize action-value function Q with random weights θ
6.   Initialize target action-value function $\hat{Q}$ with weights $\theta^- = \theta$
7. **Training and Operating**
8.   **For** event_time $t = 1, T$ **do**
9.   Update the parameter $N_{arrival} = N_{arrival} + 1$
10.   Observe the new arrival UT u^t and form the state s_t as eq.(15)
11.   Calculate the available action set $\mathcal{A}(s_t)$
12.   **If** $\mathcal{A}(s_t)$ is $\emptyset$ **do**
13.   Update the parameter $N_{block} = N_{block} + 1$
14.   Set the immediate reward $r_t = r_{BL}$
15.   **Else do**
16.   Set the immediate reward $r_t = r_{SF}$
17.   Reformulate the state s_t to $\phi(s_t)$
18.   Store the experience tuples $(\phi(s_{t-1}), a_{t-1}, r_t, \phi(s_t))$ in $\mathcal{D}$
19.   With probability ϵ select random action $a_t \in \mathcal{A}(s_t)$
20.   Otherwise select $a_t = \arg\max_{a \in \mathcal{A}(s_t)} Q(\phi(s_t), a; \theta)$
21.   Execute the CA decision action a_t
22.   Update the parameters $\mathbf{W}$ and $\mathcal{U}$
23.   Sample a minibatch of $(\phi(s_j), a_j, r_{j+1}, \phi(s_{j+1}))$ from $\mathcal{D}$
24.   Calculate the target value y_j through eq.(21)
25.   Calculate the loss $L(\theta)$ as eq.(20) and train the $Q(\phi(s), a; \theta)$
26.   **If** $mod(t, G) = 0$
27.   Update the Target Network $\hat{Q}$ with weights $\theta^- = \theta$
28.   **End If**
29.   **End If**
30.   **End For**

IV. SIMULATION RESULTS AND ANALYSIS

In this section, we use computer simulations to show the performance of the proposed DRL-DCA algorithm in multibeam satellite systems. We first present the simulation parameters. Then, the performance under different traffic distribution and system bandwidth are simulated and analyzed. At last, the convergence performance of the proposed DRL-DCA algorithm is illustrated.

A. SIMULATION PARAMETERS

For simulation parameters, we consider the typical L-band with downlink frequency of 1542 MHz multibeam satellite

system as present simulation scenario. Simulation parameters mainly refer to the GEO-Mobile Radio (GMR) Interface Specifications [22] which is adopted in the Thuraya systems. Channel gain mainly considers the free space propagation loss, and multibeam transmitting antenna gain depends on the antenna radiation pattern with beamwidth $\theta_{BW} = 1°$ and maximum gain 41.6 dBi. The G/T value of UTs receiving antenna is $G/T = -22$ dB/K. Service arrival is supposed to obey poisson distribution with arrival rate λ, and the service duration obeys an exponential distribution with mean duration μ. Service quality with required minimum capacity $C_{Th} = 500$ kbps is supposed. Simulation is conducted based on Matlab 2017 platform and a MATLAB toolbox named DeepLearnToolbox [23] is selected to implement the CNN.

The simulation parameters are summarized in Table 4.

TABLE 4. Simulation Parameters.

Parameters	Values
Scenario parameters	
Frequency Band f_c	1542 MHz
System bandwidth B_{tot}	5 MHz
Number of beams N	37 (illustrated as Figure.1)
Number of channels M	16
Channel bandwidth B_c	312.5 kHz
Multibeam antenna beamwidth	$\theta_{BW} = 1°$
Maximum antenna gain	41.6 dBi
Total satellite power P_{tot}	23 dBW
Maximum beam power P_b	20 dBW
Maximum channel power P_c	0.35 dBW
Propagation loss	187.3 dB
UT receiving G/T	-22 dB/K
Required service capacity C_{Th}	500 kbps
traffic arrival model	Poisson arrival with λ
traffic duraion model	Exponential with μ
Algorithm parameters	
Replay memory capacity N_{ep}	20000
Replay start size N_{st}	5000
Minibatch size N_{mb}	16
Target network update frequency G	100
Activation function	sigmoid
reward of blocked r_{BL}	0.3
reward of satisfied r_{SF}	0.05
Discount factor γ	0.99
Initial exploration rate ϵ_i	1.0
Final exploration rate ϵ_f	0.01

We compared the proposed DRL-DCA algorithm with the following two algorithms:

- *FCA*: fixed channel allocation algorithm where a set of channels is permanently allocated to each beam with the frequency reuse factor $FR = 4$.
- *IM-DCA*: interference measurement based dynamic channel allocation algorithm with threshold proposed in [10], where the channel with minimum CCI is allocated.

The blocking probability, carried traffic and spectrum efficiency are used as the performance metrics. In this paper, these three metrics are defined as follows:

- *Blocking probability*: defined as the ratio of blocked service numbers and the arrived service numbers, $p_{bl} = N_{block}/N_{arrival}$.

- *Carried traffic*: defined as the maximum traffic (in terms of traffic arrival rate λ) that the system can carry subject to a given blocking probability $p_{bl} = 0.10$.
- *Spectrum efficiency*: defined as a variable inversely proportional to the required bandwidth (in terms of the number of channels M) so that the system can achieve the carried traffic subject to a given blocking probability $p_{bl} = 0.10$.

B. PERFORMANCE UNDER DIFFERENT TRAFFIC DISTRIBUTION

1) UNIFORM TRAFFIC DISTRIBUTION

In this case, the traffic of each beam follows the uniform distribution with traffic duration $\mu = 3$ with unit minutes. We evaluate the performance under different beam traffic in terms of poisson arrival rate λ with unit times per hour. Simulation result is illustrated in Figure.5.
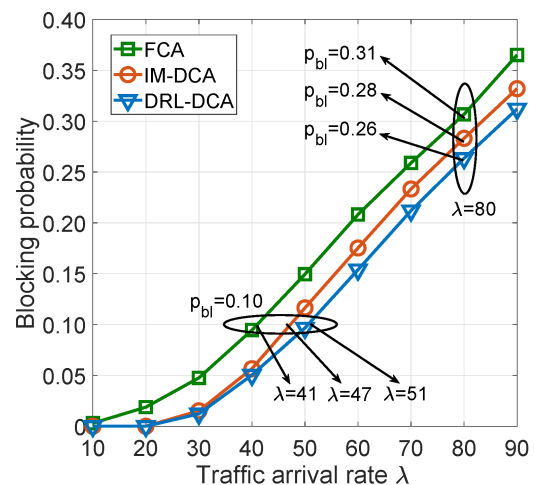

FIGURE 5. Blocking probability versus beam traffic arrival rate λ under uniform traffic distribution.

It can be seen from Figure.5 that the blocking probability of the three algorithms increases as the traffic arrival rate increases. The proposed DRL-DCA algorithm achieves lower blocking probability compared with the FCA and IM-DCA. For example, under the scenario of traffic arrival rate $\lambda = 80$, the blocking probability of the FCA, IM-DCA and DRL-DCA algorithm is $p_{bl} = 0.31$, $p_{bl} = 0.28$ and $p_{bl} = 0.26$, respectively. For the interesting range of blocking probability $p_{bl} = 0.10$, the FCA, IM-DCA and DRL-DCA algorithm can carry the traffic with arrival rate $\lambda = 41$, $\lambda = 47$, $\lambda = 51$, respectively. In other words, the DRL-DCA algorithm can improve the carried traffic in about 24.4% and 8.5%, compared to the FCA and IM-DCA, respectively.

From Figure.5, we can see when the traffic load is light to moderate, $\lambda \in [10, 50]$, the performance of DCA (the IM-DCA and DRL-DCA) algorithms shows great advantages over FCA algorithm. That is mainly because the DCA can achieve more efficient utilization by dynamic scheduling the available channels. While when the traffic is very heavy,

$\lambda \geq 80$, the performance advantage of DCA algorithms over FCA becomes less obvious. This can be explained by the fact that relatively small number of channels are available under the heavy traffic load and thus optimal channel allocation decision becomes meaningless. We can also observe that the DRL-DCA algorithm performs more prominently than IM-DCA, it can be explained that the DRL-DCA algorithm focuses on the performance optimization during a long-term period while the IM-DCA makes the channel allocation with only current situation considered.

2) NON-UNIFORM TRAFFIC DISTRIBUTION

In this case, the traffic of each beam follows non-uniform distribution with traffic duration $\mu = 3$ with unit minutes. We evaluate the performance under different beam traffic in terms of average arrival rate $\overline{\lambda}$ with unit times per hour. Figure.6 illustrates the blocking probability performance.
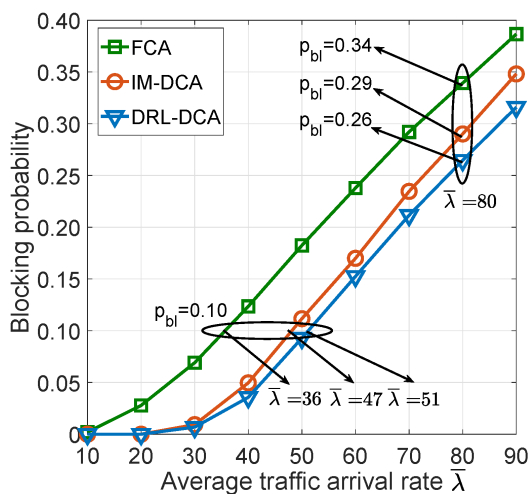

FIGURE 6. Blocking probability versus average beam traffic arrival rate $\overline{\lambda}$ under non-uniform traffic distribution.

Figure.6 clearly shows that the blocking probability of the three algorithms increases as the traffic arrival rate increases. The proposed DRL-DCA algorithm achieves lower blocking probability compared with the FCA and IM-DCA. From Figure.5 and Figure.6, we observe that the blocking probability of FCA becomes larger in the non-uniform distribution ($p_{bl} = 0.34$) compared with that in uniform distribution case ($p_{bl} = 0.31$) under the same beam traffic ($\lambda/\overline{\lambda} = 80$). While the DCA algorithms show almost the same performance in the uniform and non-uniform distribution, and the DRL-DCA algorithm performs better than IM-DCA. For the interesting range of blocking probability $p_{bl} = 0.10$, the FCA, IM-DCA and DRL-DCA algorithm can carry the traffic with arrival rate $\overline{\lambda} = 36$, $\overline{\lambda} = 47$, $\overline{\lambda} = 51$, respectively. In other words, the DRL-DCA algorithm can achieve the performance improvement in carried traffic of 41.7% and 8.5%, compared to the FCA and IM-DCA, respectively.

3) DIFFERENT TRAFFIC DURATION

In this case, the traffic of each beam follows non-uniform distribution with average arrival rate $\overline{\lambda} = 40$. We evaluate the performance under different traffic duration in terms of exponential duration parameter μ with unit minutes. Figure.7 illustrates the blocking probability performance.
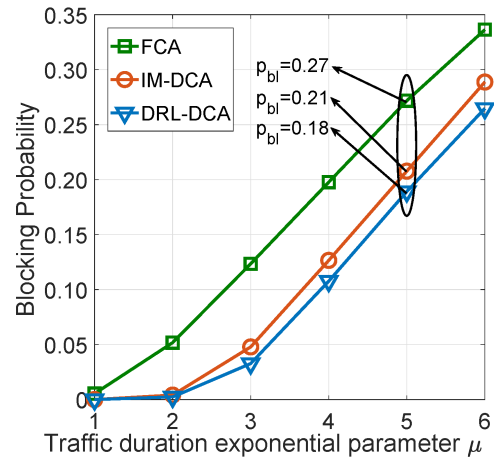

FIGURE 7. Blocking probability versus beam traffic duration μ under non-uniform traffic distribution.

It can be seen from Figure.7 that the blocking probability of the three algorithms increases as the traffic duration increases. The proposed DRL-DCA algorithm achieves lower blocking probability compared with the FCA and IM-DCA algorithms. For example, under the scenario of traffic duration $\mu = 5$, the blocking probability of the FCA, IM-DCA and DRL-DCA algorithm is $p_{bl} = 0.27$, $p_{bl} = 0.21$ and $p_{bl} = 0.18$, respectively. From Figure.7, the proposed DRL-DCA algorithm is effective under different traffic duration scenarios.

C. PERFORMANCE UNDER DIFFERENT SYSTEM BANDWIDTH

In this case, the system bandwidth, $B_{tot} = B_c \times M$, is represented by the number of channels M with fixed channel bandwidth $B_c = 312.5$ kHz. Figure.8 shows the blocking probability with different number of channels M and average arrival rate $\overline{\lambda}$ under non-uniform traffic distribution scenario.

It can be seen from Figure.8 that the blocking probability decreases with the increase of channel numbers M and the decrease of traffic average arrival rate $\overline{\lambda}$. Under the same beam traffic and blocking probability performance, the DRL-DCA algorithm needs the smaller bandwidth, i.e. it can achieve a higher spectrum efficiency. From Figure.8, under the scenario of traffic arrival rate $\overline{\lambda} = 60$, the needed number of channels is $M = 23$, $M = 19$ and $M = 18$ for the FCA, IM-DCA and DRL-DCA algorithm, respectively. That is to say, the DRL-DCA algorithm can achieve the spectrum efficiency improvement of about 21.7% and 5.3%, compared to the FCA and IM-DCA.
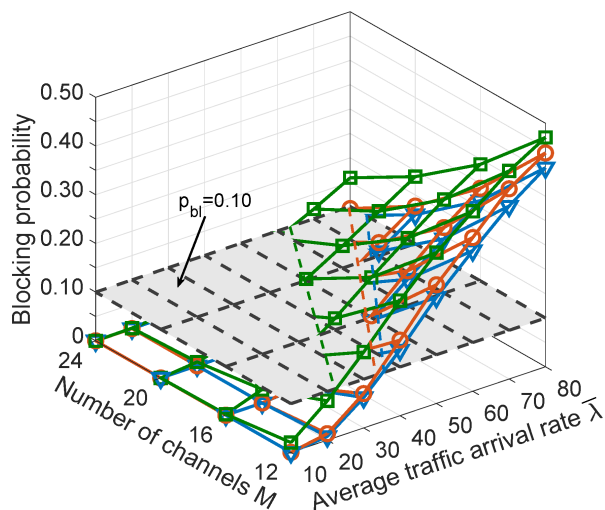
FIGURE 8. Blocking probability versus number of channels *M* and average beam traffic arrival rate $\bar{\lambda}$ under non-uniform traffic distribution.
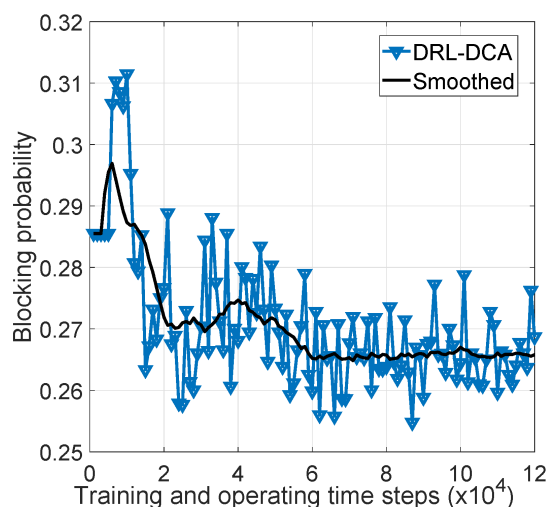

FIGURE 9. Blocking probability performance during the training and operating time steps.

D. PERFORMANCE OF CONVERGENCE

To show the convergence of the DRL-DCA algorithm, we take a deeper look into one specific data-point of Figure.5 under $\lambda = 80$. The blocking probability performance is illustrated as Figure.9. From Figure.9, the performance of the DRL-DCA algorithm keeps constant during the first 0.5×10^4 time steps. This is mainly because the Q-Network parameters θ starts updating only when the replay memory achieves the N_{st} experience tuples. Furthermore, we can see the DRL-DCA algorithm converged after about 6×10^4 time steps. From a practical point of view, CA decisions made by multibeam satellite systems are often highly repetitive, thus generating an abundance of training data available for implementing the DRL-DCA algorithm.

V. CONCLUSION

In this paper, a novel deep reinforcement learning based dynamic channel allocation (DRL-DCA) algorithm is proposed in multibeam satellite systems. Simulation result demonstrates that the proposed DRL-DCA algorithm can achieve lower blocking probability, compared with the traditional FCA and IM-DCA algorithms. The proposed DRL-DCA algorithm can also improve the carried traffic and the spectrum efficiency. Joint channel and power allocation algorithm based on DRL method remains to be further studied.

REFERENCES

[1] C. Balty, J. Gayrard, and P. Agnieray, "Communication satellites to enter a new age of flexibility," *Acta Astronautica*, vol. 65, nos. 1–2, pp. 75–81, 2009.
[2] Y. Vasavada, R. Gopal, C. Ravishankar, G. Zakaria, and N. BenAmmar, "Architectures for next generation high throughput satellite systems," *Int. J. Satellite Commun. Netw.*, vol. 34, no. 4, pp. 523–546, 2016.
[3] K. Kaneko, H. Nishiyama, N. Kato, A. Miura, and M. Toyoshima, "An evaluation of flexible frequency utilization in high throughput satellite communication systems with digital channelizer," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.
[4] L. D. C. H. R. Gaytan, Z. Pan, J. Liu, and S. Shimamoto, "Dynamic scheduling for high throughput satellites employing priority code scheme," *IEEE Access*, vol. 3, pp. 2044–2054, 2015.
[5] A. I. Aravanis, B. S. M. R., P. D. Arapoglou, G. Danoy, P. G. Cottis, and B. Ottersten, "Power allocation in multibeam satellite systems: A two-stage multi-objective optimization," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3171–3182, Jun. 2015.
[6] F. Li, K.-Y. Lam, X. Liu, J. Wang, K. Zhao, and L. Wang, "Joint pricing and power allocation for multibeam satellite systems with dynamic game model," *IEEE Trans. Veh. Technol.*, to be published.
[7] E. D. Re, R. Fantacci, and G. Giambene, "Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 2, pp. 397–405, Feb. 1995.
[8] M. Umehira and F. Naito, "Centralized dynamic channel assignment schemes for multi-beam mobile satellite communications systems," in *Proc. AIAA ICSSC*, 2012, p. 15123.
[9] H. Li, Z. Zhu, Z. Gao, J. Wang, and M. Umehira, "Dynamic channel assignment scheme with cooperative beam forming for multi-beam mobile satellite networks," in *Proc. 6th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Hefei, China, Oct. 2014, pp. 1–5.
[10] M. Umehira, S. Fujita, Z. Gao, and J. Wang, "Dynamic channel assignment based on interference measurement with threshold for multi-beam mobile satellite networks," in *Proc. 19th Asia–Pacific Conf. Commun. (APCC)*, Denpasar, Indonesia, Aug. 2013, pp. 688–692.
[11] M. Umehira, S. Fujita, Z. Gou, and J. Wang, "Location-based centralized dynamic channel assignment with channel segregation for multi-beam mobile satellite networks," in *Proc. 6th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Hefei, China, Oct. 2014, pp. 1–5.
[12] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," *Comput. Sci.*, 2013.
[13] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.
[14] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.
[15] L. Li, Y. Lv, and F.-Y. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 3, pp. 247–254, Apr. 2016.
[16] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. ACM Workshop Hot Topics Netw.*, 2016, pp. 50–56.
[17] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, 2017, pp. 1–6.
[18] N. Liu *et al.*, "A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Atlanta, GA, USA, Jun. 2017, pp. 372–382.
[19] J. Zhu, Y. Song, D. Jiang, and H. Song, "A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of Things," *IEEE Internet Things J.*, to be published.

[20] T. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, vol. 1. Cambridge, MA, USA: MIT Press, 1988.

[22] *GEO-Mobile Radio Interface Specifications (Release 1)*, document ETSI TS 101 376-5-5, V1.3.1, 2005.

[23] R. Palm. (2014). *DeepLearnToolbox, a MATLAB Toolbox for Deep Learning*. [Online]. Available: https://github.com/rasmusbergpalm/DeepLearnToolbox

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

XIN HU was born in 1985. He received the Ph.D. degree from the Institute of Electrics, Chinese Academy of Sciences, in 2012. He is currently an Associate Professor with the Information and Electronics Technology Lab, Beijing University of Posts and Telecommunications. His research interests include smart signal processing, space and ground information integration, and aerospace electronic information synthesis.

SHUAIJUN LIU was born in 1988. He received the B.E. degree in electronic and information engineering from Chongqing University, China, in 2012. He is currently pursuing the M.E. and Ph.D. degrees in electronics science and technology from the Beijing University of Posts and Telecommunications, China. His research interests include mobile satellite communications, dynamic resource management, and machine learning.

WEIDONG WANG was born in 1967. He received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2002. He is currently a Professor and the Vice President of the School of Electronic Engineering, Beijing University of Posts and Telecommunications. His research interests include satellite communication, radio resource management, Internet of Things, and signal processing. He is an Expert of the National Natural Science Foundation and a member of the China Association of Communication.

• • •