IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 29, 2020, accepted March 12, 2020, date of publication March 26, 2020, date of current version April 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983437

A Deep Reinforcement Learning Based Approach for Energy-Efficient Channel Allocation in Satellite Internet of Things

BAOKANG ZHAO[1], (Member, IEEE), JIAHAO LIU[1], ZILING WEI[1], AND ILSUN YOU[2], (Senior Member, IEEE)

[1]College of Computer Science, National University of Defense Technology, Changsha 410073, China
[2]Department of Information Security Engineering, Soonchunhyang University, Asan 31538, South Korea

Corresponding author: Ilsun You (isyou@gmail.com)

This work was supported in part by the National Science Foundation of China under Grant 61972412, and in part by the Soonchunhyang University Fund.

ABSTRACT Recently, Satellite Internet of Things (SIoT), a space network that consists of numerous Low Earth Orbit (LEO) satellites, is regarded as a promising technique since it is the only solution to provide 100% global coverage for the whole earth, without any additional terrestrial infrastructure supports. However, compared with Geostationary Earth Orbit (GEO) satellites, the LEO satellites always move very fast to cover an area within only 5-12 minutes per pass, bringing high dynamics to the network access. Furthermore, to reduce the cost, the power and spectrum channel resources of each LEO satellite are very limited, i.e., less than 10% of GEO. Therefore, to take fully advantage of the limited resource, it is very challenging to have an efficient resource allocation scheme for SIoT. Current resource allocation schemes for satellites are mostly designed for GEO, and these schemes do not consider many LEO specific concerns, including the constrained energy, the mobility characteristic, the dynamics of connections and transmissions etc. Towards this end, we proposed DeepCA, a novel reinforcement learning based approach for energy-efficient channel allocation in SIoT. In DeepCA, we firstly introduce a new sliding block scheme to facilitate the modeling of dynamic feature of the LEO satellite, and formulate the dynamic channel allocation problem in SIoT as a Markov decision process (MDP). We then propose a deep reinforcement learning algorithm for optimal channel allocation. To accelerate the learning process of DeepCA, we utilize the image form to represent the requests of users to reduce the input size, and carefully divide an action into multiple mini-actions to reduce the size of the action set. Extensive simulations show that our proposed DeepCA approach can save at least 67.86% energy consumption compared with traditional algorithms.

INDEX TERMS Energy efficient, channel allocation, artificial intelligence, reinforcement learning, Internet of Things.

I. INTRODUCTION

As one of the most promising technologies, Internet of Things (IoT) has developed a lot in recent years. IoT embeds computational capabilities into each object [1]. It can be applied in many promising and key areas, such as telemedicine, smart cities, and environmental monitoring [2], [3]. To promote the development IoT, numerous technologies and protocols have been designed to be used in IoT. However, most of these technologies only focus on the terrestrial scenario, which are hard to provide valid global connectivity. since at least 70% surface of the earth is unable to be covered by terrestrial networks. Moreover, terrestrial networks are easily destroyed by natural disasters, such as earthquakes and tsunamis. To overcome the above challenges, the LEO satellite network has been introduced to provide the communication capability for the IoT, especially for the Internet of Remote Things (IoRT), including the ocean, desert, etc. Although at present, only two small LEO satellite systems, named Orbcomm, and Argos are already deployed with tens of satellites and few users, many upcoming IoT smallsat constellations with thousands of satellites and billions of users

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Feng.

have been planned and start to launch in recent few years, including SpaceX Starlink, OneWeb, etc. They have shown their ability in extending terrestrial networks to construct global connections for IoT.

There are two kinds of links between satellite and IoT devices on the ground: direct access and indirect access. In direct access mode, ground terminals communicate with satellites directly. For indirect mode, IoT devices communicate with satellites through some relay nodes or sink nodes. In this paper, we consider direct mode, which is used in Direct-to-Satellite IoT (DtS-IoT). Compared with indirect mode, direct connection is a more appealing way in some specific scenarios [4]. For example, in some remote areas with very low node density, it is not worth to set a gateway.

For IoT networks on the ground, some relative studies [5], [6] utilize the cellular networks (e.g., 5G technologies) to build dedicated data gathering networks [7]. However, these approaches are only fit for local networks where network infrastructures are sufficient. When IoT devices are distributed in remote areas, it is uneconomical to deploy base stations in those regions. Thus, to establish a global communication network, more and more researches focus on IoT systems combined with GEO satellites and LEO satellites.

To establish a global network, the LEO satellite has many advantages over the GEO satellite. For example, compared with the GEO satellite, the LEO satellite provides better signal strength and less propagation delay due to a shorter distance to the IoT devices. Thus, in this paper, we also focus on the LEO satellite IoT. However, compared with GEO Satellites, the LEO satellite has limited on-board resource. For example, due to a small size of the LEO satellite, the energy of the LEO satellite is limited [8]. On the other hand, billions of devices need to be served around the world [9]. Current satellite resource (e.g., energy resource, channel resource, etc.) capability can no longer satisfy such a big number of demands. To address the problem of satellite communication resources shortage, it is important to derive an efficient resource scheduling scheme to take fully use of the limited resources.

By introducing the High-throughput satellite (HTS), a much larger communication capacity is achieved by frequency reuse and spot beam technology. However, different from GEO satellite communication system, LEO satellite moves fast relative to the ground. Beam hopping technology [10], which helps to increase the capacity of broadband satellite communication under the constraint of satellite power resource, cannot be used in such a dynamic environment directly. In this paper, the dynamic feature of the LEO satellite is also well considered and modeled when deriving the channel allocation scheme.

To this end, we first define the system model of LEO satellite communication system and analyze the problem of allocating the limited channel resource to nodes on the ground. Note that the long-term profit of the network is maximized in the formulated problem. Since the formulated problem is too complex to obtain a closed-form result, we convert the formulated problem as an MDP [11]. Then, reinforcement learning (RL) technique is introduced to derive the optimal channel allocation scheme. The RL method can deal with the real-time and dynamic environment. In addition, the satellite-nodes system is complex and difficult to accurately model with pure mathematical expressions. Meanwhile, the RL method does not need to obtain the accurate and prior information of the environment. Therefore, we introduce the RL method in this work. Inspired by model-free RL approaches which can achieve a good performance without knowing the dynamics of network environment, we select a model-free method for this sequential decision-making problem.

The main contributions of this paper are described as follows:

1. Compared with the existing channel allocation work, the dynamic feature of the LEO satellite is well modeled by introducing a sliding block scheme in our work. Thus, our proposed channel allocation scheme is suitable for the practical scenario.
2. Considering the conflict between the limited resource of the LEO satellite and massive data transmission demands, the channel allocation problem with green communication for the Direct-to-Satellite IoT is formulated. In addition, the long-term profit of the LEO satellite IoT system is maximized in the formulated problem. To the best of our knowledge, this paper is the first to study the energy-efficient communication problem for the Direct-to-Satellite IoT.
3. Considering the complicated environment of the LEO satellite IioT system, the formulated problem is much complex. In addition, the prior information of the system is unknown to the LEO satellite. Thus, the formulated problem is hard to solve. To address this challenge, we formulate this problem to an MDP, and then, a deep RL algorithm is proposed to derive the optimal channel allocation scheme.
4. In the proposed algorithm, we propose two tricks to speed the learning process. The first one utilizes the image form to represent the requests of users, and thus, the size of the algorithm input is decreased. To reduce the large size of the action set, another trick is proposed by deriving each action of the MDP to multiple mini-actions.
5. The proposed channel allocation scheme is evaluated by extensive simulation. It shows that our proposed scheme can largely improve the performance compared with the classic schemes.

The remainder part of this paper is organized as follows. Part II briefly reviews the related work. In Part III, the model of LEO satellite communication system is presented. Part IV describes our proposed DeepCA scheme. Part V presents experiment results and analyses. Part VI concludes our work.

II. RELATED WORK

Several studies have focused on energy-saving resource allocation methods in satellite-based IoT. Related work can be

broadly categorized into three parts: physical layer, link layer, and network layer.

A. PHYSICAL LAYER

Studies on physical layer mainly try to improve the spectrum efficiency [12], [13]. Huang *et al.* [12] presented a real-time algorithm for energy-efficient data uploading in a terrestrial distributed sensor networks, which aims to improve the network throughput and also save the energy consumption for gateways. However, it only focuses on uplinks with data gathering gateways, and does not support energy saving in the satellite. Fu et.al discussed the problem of optimal power allocation and admission control for satellite networks using a dynamic programming approach [13]. However, the energy consumption of each user is produced with a prior known probability distribution, while in reality, it is a stochastic event and usually unknown in advance. Thus, the proposed approach is not feasible in the above-mentioned scenarios.

Deep reinforcement learning (DRL) has emerged as a promising technique to deal with optimization problems. Accordingly, resource allocation in satellite systems by DRL receive more and more attention. In [14], a novel DRL-based Dynamic Channel Allocation (DRL-DCA) algorithm, which focuses on decreasing the service blocking probability, was proposed for GEO satellite communication. However, it only considers a GEO satellite, which remains fixed in the same position from the perspective of the earth. Different from the GEO satellite, the LEO satellite moves fast, resulting in a fast variation of channel condition.

B. LINK LAYER

To improve the energy efficiency, TDMA is a widely adopted multiple channel access scheme. However, in satellite IoT which deploys a huge number of devices, only a limited number of devices have data transmission demands at any moment. Thus, a fixed-allocation multiple access scheme (e.g., TDMA) can no longer satisfy transmission demands. Therefore, some novel protocols suitable for satellite IoT are proposed recently, including contention-free direct access protocols [15] and contended direct access protocols [16], [17]. Although these algorithms can achieve a higher using efficiency of bandwidth compared with TDMA-based schemes or slotted ALOHA schemes, the energy constraint of the satellite is seldom considered in these algorithms.

C. NETWORK LAYER

Since the infrastructure of LEO constellation is moving with respect to nodes on the ground, the context of communications on the network layer is changing all the time. Consequently, to minimize energy consumption in the complex environment of satellite networks on the IP level is challenging. Authors in [18] dealt with the routing problem in satellite networks. A green satellite routing scheme to save power and extend the lifetime of satellite was proposed. By switching proper nodes into sleeping mode and distribute the network traffic properly, the lifespan of battery cells can be extended. On the other hand, there are also studies on applying information-centric networking (ICN) [19] to integrated satellite networks. [20] shows that combining ICN paradigm with geostationary satellite networks can enhance the utilization efficiency of the bandwidth, whose effect is more significant than those achieved by HTTP ways.

Compared with existing studies, we particularly focus on the energy-efficient resource allocation problem for the globally distributed IoT networks in LEO satellite system.

III. DESIGN OF DEEP REINFORCEMENT LEARNING BASED CHANNEL ALLOCATION METHOD

In this section, we will firstly build the system model on LEO satellite system and describe the channel allocation problem. Thereafter, we formulate the problem as an MDP.

A. SYSTEM MODEL

Considering a modern LEO satellite network such as SpaceX Starlink, the altitude of each satellite is generally around 1000 km, and one LEO satellite can cover thousands of kilometers on the ground. The coverage area of a satellite, which is also called footprint, is a circular area located directly below it. To realize frequency reuse and avoid interference among different transmissions, the footprint is covered by small cells.

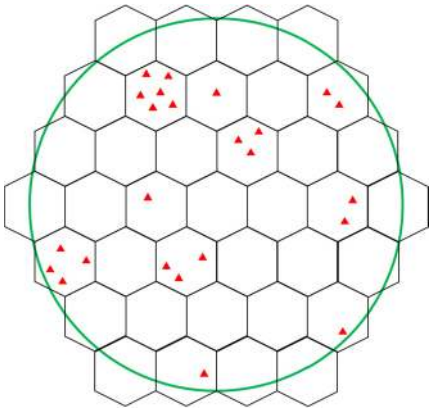


FIGURE 1. Spot beams in a footprint.

Figure 1 shows a geographical footprint of a satellite, in which the green circle is the footprint. The same frequencies can be 'reused' in multiple spot beams if the distance between the spot beams is far enough [21]. Each red point refers to one user node on the ground. The geographical distribution of users is unknown for the satellite. Some parts of the coverage area are crowded with users while some other spot beams contain few nodes.

LEO satellites move at very high speeds with respect to the communication nodes on the earth. Due to this dynamic characteristic, the footprint of a LEO satellite changes continuously. In the geo-distributed network system, all channel resources can be utilized by each user. In general, the number

of requests is larger than that of the channel resources. In addition, the energy of the LEO satellite is also limited. Thus, to improve the spectrum efficiency, the satellite needs to allocate its resource to each beam according to the number of requests in the beam.

We consider the total energy supply as the critical resource for the satellite that limits the power allocation scheme. The main constraint of the throughput of satellite is Co-channel Interference (CCI). When the adjacent beams use the same frequency, these beams will interfere with each other, consequently decreasing the transmitting rate. Thus, we use beam hopping, which illuminates cells with a small number of active beams.

As shown in Figure 2, since the LEO satellite coverage area is moving, it is divided into rectangle blocks in a row [8], in which every block is a region. For the tractability of modelling, we assume that the coverage area is rectangle. Each region i is an area of nodes where the satellite covers for a duration of T_r at each time when the satellite passes overhead. T_r denotes the time when a communication node on the earth stays in the coverage area of a satellite, which means that the terminal can communicate with the satellite directly. The requests in each region is randomly distributed.

FIGURE 2. Satellite coverage area division [8].

As shown in Figure 3, we divide each footprint into a number of rectangular blocks. The blocks of each footprint form a matrix with I rows and J columns. Each block illuminated by a spot beam is denoted as blk_{ij}. Similarly, to simplify the model, we assume that the shape of each beam is rectangle, and the size of each beam is that of one block. Although some small overlaps exist between beams, this kind of approximation reduces the complexity of the model greatly, which does not affect the effectiveness of our method in general. Our future work might investigate how to modify the model for the scenario where the coverage area is circle or oval.

FIGURE 3. A footprint divided into blocks [8].

B. PROBLEM DEFINITION

We describe the channel allocation scenario as follows: an LEO satellite creates N_b beams which are denoted as $B = \{n | n = 1, 2, \ldots, N_b\}$. Available channels are denoted as $C = \{c | c = 1, 2, \ldots, N_c\}$. Resource allocation state of beam n is described as a vector $\overrightarrow{v_n} = [v_{n,1}, v_{n,2}, \ldots, v_{n,M}]^T$, in which $v_{n.M} \in \{0, 1\}$ indicates whether beam n is using channel M. $v_{n.M} = 1$ means channel M is under occupancy in beam n. Otherwise, channel M is not utilized in beam n. Channel allocation vectors of all beams constitute the channel allocation matrix of the satellite, which is denoted as $V = [\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_N}]$.

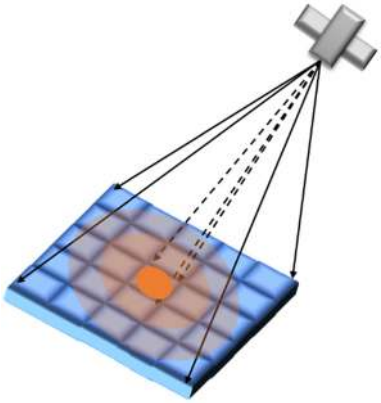


FIGURE 4. Levels of transmission power.

Since the channel state information (CSI) relates to the transmission distance, the transmission power for the nodes in different beams should be different under a constant transmission rate. For example, in Figure 4, the power consumption of the data transmission between the satellite and the nodes that locates in the centre of the footprint is assumed as the lowest. Since the nodes at the edge of the footprint are much far away from the satellite, intuitively, the satellite needs to cost a large amount of energy for the transmission to these nodes.

Important notations are explained in TABLE 1.

TABLE 1. Notations and variables.

Symbol	Quantity
N_c	The number of channels
N_b	The number of beams
T	The number of time slots
R_i	The power that satellite allocate to beam i
P_{tot}	Maximum transmission power of the satellite
P_b	Maximum transmission power of each beam
P_{ch}	Transmission power of one channel
A	Set of actions
dur_k	Duration of service for node k
$Tsat_k$	Time remained for node k in the satellite's coverage
$\mathrm{br}_{\max}$	Maximum blocking rate of service

Our target is to serve as many as possible users under the energy constraint. Thus, the constraints limited by power supply of satellite and CCI are described as follows:

1) Total transmitting power of the satellite is limited to P_{tot}.
2) Power allocated to beam b is limited to P_b.

3) Adjacent beams are not allowed to utilize the same channels.

To take fully advantage of the satellite energy and serve as many as possible ground users, we define a utility function that relates to the transmission power. On the other hand, we also consider the basic quality of service (QoS) for all users by guaranteeing the block rate. Blocking rate is defined as the ratio of requests that are rejected by the satellite due to limited resource and the total number of requests on the ground. Thus, the utility function is defined as follows:

$$maximize \sum_{t=1}^{T} r_t \tag{1}$$

$$s.t. \sum_{b=1}^{N} P_b < P_{tot}, \tag{2}$$

$$0 \leq P_{bi} \leq P_b, i = 1, \ldots, N_b, \tag{3}$$

$$0 \leq P_{cj} \leq P_{ch}, j = 1, \ldots, N_c, \tag{4}$$

$$r_{blo} < br_{max} \tag{5}$$

For modelling tractability, we assume that the duration of satellite scanning the diameter of one beam is a time step. This is also practical in real LEO satellite networks. In the formulas above, r_t is an immediate reward at time step t, which is defined by prior regulation. The value of r_t depends on the power consuming and the number of node services rejected by the satellite. P_{bi} is the allocated power to beam i, which should not exceed P_b. For each channel, the sum of the power allocated to all beams in a certain channel j (i.e. P_{cj}) is assumed to be smaller than a certain amount P_{ch}, which is the constraint (4). To guarantee the QoS of users, the blocking rate, denoted as r_{blo}, is set as smaller than a threshold br_{max}. The value of br_{max} is determined by the QoS requirement of users. The total bandwidth of the network is set as B_{tot}. Therefore, the bandwidth of each channel is $B_f = B_{tot}/N_c$.

In this scenario, the energy efficient problem is defined as finding the optimal channel allocation strategy, which allocates the limited number of channels to nodes on the ground with the goal of saving transmitting power in the long term. We consider that a demand resource is properly allocated if there is enough available capacity for it in the beam where the demand is located, i.e., there is a free channel for the request in that beam. The channel demand does not expire until the end of the current time step. This implies a challenging task for the DRL agent because it has to not only identify critical resource (channels) but also deal with the uncertainty in the generation of future channel demands.

C. MDP FORMULATION

Conventional channel allocation approaches usually utilize a prior knowledge to make a decision and do not perform well in a complex system. In contrast, RL solves these problems very well. Unlike existing solutions that allocate channels in a fixed way or a heuristic way, we design a learning-based approach. Specifically, we investigate reinforcement learning technique, which is one branch of machine learning (ML) that focuses on decision making. RL studies how to teach an agent to find an optimal behavior for a specific target in a complex

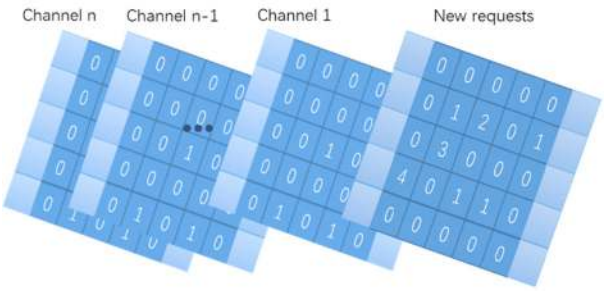


FIGURE 5. State representation.

environment [22]. An agent observes previous state-action pairs and rewards, and then takes an action that is thought to be the best. Recently, RL has shown its ability in many environments with the development of deep neural networks (DNN). Inspired by these results, we take RL technique to solve the above formulated energy-saving channel allocation problem.

MDP provides a mathematical framework to model serial decision-making problems [23]. It is used to describe an environment in RL. In MDP, an agent selects the best action based on current state. The history of states that agent has experienced does not affect the decision to be made. Mathematically, a decision process has the Markov property when it follows the formula:

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, \ldots, S_t] \tag{6}$$

where S_{t+1} denotes the state at time step $t+1$, $P[S_{t+1}|S_t]$ is the transition probability from state S_t to S_{t+1} when given the information S_t. The formula shows that the previous state S_t contains all necessary information from history $(S_1, \ldots, S_t)$ for next decision. The current state contains three fundamental elements in general: a set of state S, action A, and reward R. Specifically, at time step t, the agent observes state s_t, and takes an action a_t. Then, the agent can receive reward r_t and the state transits to s_{t+1} at the next time step (i.e., step $t+1$)[24]. The target of training the agent is to maximize the expected cumulative discounted reward $E[\sum_{t=0}^{\infty} \gamma^t r_t]$ by taking an appropriate action for each step, where $\gamma \in (0, 1]$ is the discounting factor [22], r_t is the immediate reward at time step t.

The main idea of our solution is to translate the model into a Markov Decision Process. Since the channel allocation action at each time step depends on the current state of channel resource instead of the history of states, it can be formalized as an MDP. In our approach, an LEO satellite is an agent, and the target of the agent is to maximize the utility function Eq. (1).

Now we define the state space, action space, and reward function to formulate the problem.

State space: In our setting model, we divide the footprint of beams into a $X * Y$ grid area. The value of each grid represents the number of task requests at each time step. Since we only consider whether there is a task request in

the area, the state space simply includes information about user tasks, such as the size and location of tasks. Moreover, some space relevance between nodes on the ground may do effect on decision. Thus, to take fully consideration of the whole information of a state, we take a creative way to deal with the state input, which is stated as follows. The state input is taken as the form of matrices, which is similar to the image. As is shown in Figure 5, $n+1$ images make up a state representation, where n refers to the number of channels. Each image is a matrix of size $X * Y$, where X and Y are the width and height of the image respectively. The values of X and Y depend on the coverage area of the satellite. Note that, matrix $k(k = 1, 2, \ldots, n)$ gives the allocation information of channel k in each beam. Values in the grid in the first n images are limited to '1' and '0'. '1' means the channel is under use and '0' means the channel is free in that grid. Apart from these n images, an extra image (i.e., matrix $n+1$) is used to show new node tasks that request for channels. In other words, this image represents the geographical distribution of new user requests. Therefore, the state representation at time step t is denoted as:

$$s_t = [\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}, \overrightarrow{R}] \tag{7}$$

where $\vec{R}$ refers to the allocation request of time step t (i.e., matrix $n+1$).

Action space: The action is a mapping from newly coming node tasks to channels to be allocated. For each coming task k, satellite allocate channel $n \in [0, N_c]$ to it, where $n = 0$ denotes that no channel is allocated to serve task k. At each time step, since the number of users is large, the size of the action space may be too large, which makes the learning to be challenging. To reduce the size of the action space, the action is divided to several mini-actions at each time step. The action set at each decision is $1, 2, 3, \ldots, N, \emptyset$. For each decision, satellite allocate channel $c_k (c_k \in [1, N_c])$ for node k in its footprint, or deny service ($\emptyset$), which is a mini-action.

Rewards: The reward is the feedback an agent gets after it takes an action. It can be obtained after each time step. The objective of this paper is to learn a power-efficient channel allocation policy to maximize the long-term reward of the satellite. Since the energy is a critical source for the satellite, the total energy consumption is the main criterion to find a successive action. To guarantee the QoS of ground users, the service blocking rate still needs to be considered in the reward function. Therefore, we split the power efficiency and service blocking rate criteria into two normalized reward function components.

First, we define a normalized reward to maintain the power consumption. The objective is to reach a minimum value of total power cost. The following function represents the power efficiency reward:

$$r_P = \alpha * \frac{\sum_{i=1}^{N_t} (P_{i,t} - P^*_{i,t})}{\sum_{i=1}^{N_t} P^*_{i,t}} \tag{8}$$

where α is a weighting factor, $P_{i,t}$ is the power set up by the agent at time step t, $P^*_{i,t}$ is the optimal power which is decided by the location of the beam. The optimal power is defined as the allocation scheme in which each node request can be satisfied in the optimal location where it consumes the minimum power, i.e., node on the ground is allocated with a channel when it moves to the nearest area to the satellite. The agent decides to allocate power to N_t nodes at time step t. The design of reward function is critical in the training stage of the agent. The better it represents the goal of the problem, the better performance it will achieve.

Second, the normalized value of the service blocking rate is used to guarantee the QoS for all users, which is represented by:

$$r_b = \beta * (\frac{\sum_{i=1}^{T} N_i}{\sum_{i=1}^{T} N^*_t}) \tag{9}$$

where β is a weighting factor of satisfactory rate, N_i is the number of served nodes that are to be out of date at time step i. In other words, the demands of the number (i.e., N_i) of nodes cannot be satisfied. N^*_t denotes the total number of nodes that are to be out of service at time step i.

Last, the overall reward r is defined as:

$$r_t = r_p + r_b \tag{10}$$

IV. IMPLEMENTATION

In this section, we focus on the procedures of utilizing our method to operate in the LEO satellite IoT network. Based on the model described above, we proposed DeepCA, a DRL solution to allocate channels.

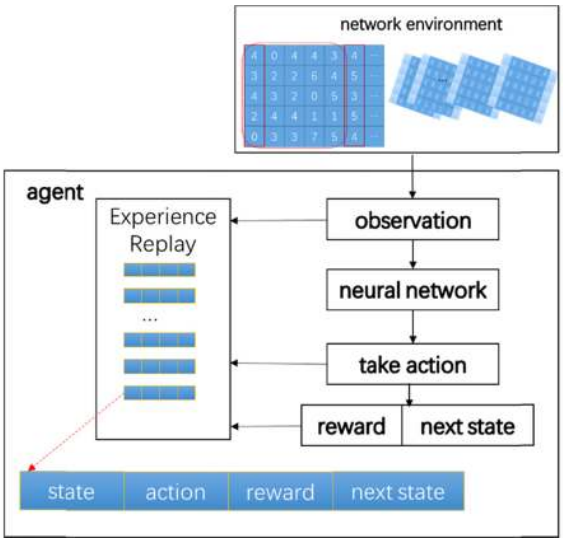


FIGURE 6. The proposed DeepCA system architecture.

The proposed DeepCA architecture is shown in Figure 6, which takes the specifically designed structure of training samples to implicitly learn and predict the future. It is based on the example scenario in Figure 4. We depict our training algorithm as follows.

First, at time step t, the agent (satellite) gets the observation, which includes the bandwidth resource requests of all nodes and the current state of channel allocation.

Once obtaining the observation, the agent takes power allocation action based on the state through neural network. The parameters in neural network represent an allocation policy $\pi(a_t, s_t)$. As is illustrated in prior section, for simplicity, we treat the power and demand per beam as discrete variables. Therefore, the state and action space are also discontinuous. This is impractical for agent to train a model using neural network, so we normalize the state space as input to neural networks.

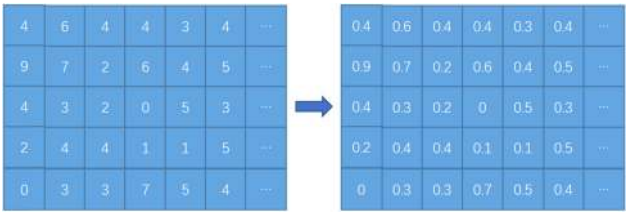


FIGURE 7. State space normalization example.

As is shown in Figure 7, we give an example of the normalization of the state space. Suppose that the satellite can serve no more than 10 nodes in each beam. The matrix on the left is the initial number of node requests in a footprint. The process of normalization is that each of the matrix numbers is divided by the maximum capability (i.e. 10). After doing this operation, the value of all the elements is between 0 and 1, which can be seen on the right matrix. Then the matrix can be put into the neural networks as a continuous variable.

After allocating power to beams, the agent can get a reward r_t and observe the current state of channels and bandwidth demands of nodes (i.e., state s_{t+1}). The tuple of (s_t, a_t, r_t, s_{t+1}) is called an experience. To fully utilize the experience, we use experience replay to use these experiences more than once.

The map of node requests is presented in Figure 8, satellite moves one column of nodes at each time step. In this example, the area whose number of node requests are 4, 3, 4, 2, 0 respectively (i.e., the left column) will be out of the scope at the next time step and the area whose number of node requests are 4, 5, 3, 5, 4 (i.e., the right column) will come into the footprint of satellite.

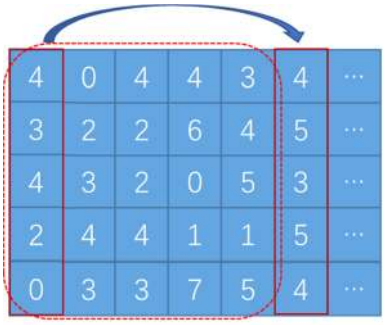


FIGURE 8. Node requests on the ground.

V. EVALUATION

A. ENVIRONMENT MODEL

A comprehensive evaluation is conducted to evaluate the performance of our proposed DeepCA scheme. The simulation parameters used in the experiment are listed in TABLE 2.

TABLE 2. Simulation parameters.

Symbol	Quantity
total power P_{tot}*/dBW*	23
Number of beams	25
Number of channels	16
Total satellite power P_{tot}	23
Maximum transmission power of one beam/dBW	20
Transmission power of each channel/dBW	0.35
discount factor γ	0.99

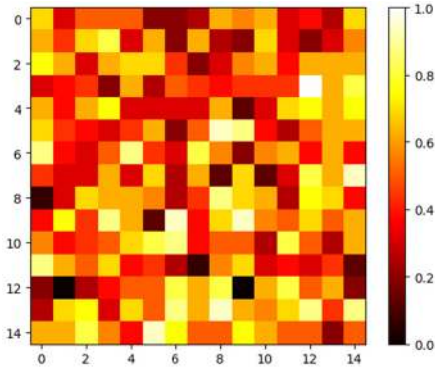


FIGURE 9. Distribution of bandwidth demands.

Distribution of bandwidth demands is plotted in Figure 9. We simulate the generation of node tasks with Gaussian distribution, whose pattern is unknown for satellite.

We compare our method with two traditional satellite channel allocation methods: random power allocation (RPA) algorithm and the greedy algorithm (GA). Random power allocation is a naïve channel allocation scheme where all the channels are randomly allocated to each beam while avoiding CCI.

GA is a class of algorithms. In this scenario, it aims at either of the two targets: optimizing the power consuming of satellite or maximizing the satisfaction rate of all nodes, which is called greedy-1 and greedy-2 algorithm respectively. Specifically, with the given demands at a time slot, agent that focuses on demand satisfactory always allocates communication power to those nodes that are to be out of date priorly, while for agent that targets at the necessity of reducing power, minimizing power has a higher priority than satisfying all customers at one time step.

In the experiment setting, we train and test DeepCA in different demand densities to improve its ability under various ground situation, while keeping total power, number of beams, number of channels and some other parameters of satellite unchanged. We set the maximum training time steps to 3 million in practice. In the training phase, one episode comprises the process of moving of 1000 blocks.

We train our learning model on a computer with an Intel Core i7-8550U CPU and a Nvidia GeForce MX150 GPU. Based on the learning curve, 3 million time steps are used for training in total.

B. EVALUATION RESULTS

This part focuses on different analyses that account for the performance of each algorithm using models presented in the previous section. We consider the cumulated power consumption and satisfaction rate as performance metrics. The satisfaction rate is defined as the ratio of nodes that have been allocated channels to sum of nodes.

Figure 10 shows the reward-episode curve in training phase. We can see that the mean episode reward increases rapidly during the first one hundred episodes, and then gets very close to zero.

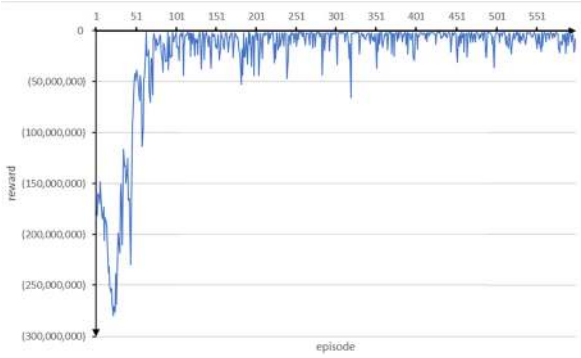


FIGURE 10. Reward in training phase.

Due to the variety of demand distribution, agent may come across some situations where its model does not address well, so there are still some uncertainties of the reward it gets. That is the reason why fluctuation exists after many episodes. However, it does no significant matter to the performance of the model for it can deal with most of the demand distributions.

1) POWER CONSUMPTION

Results of the four schemes are shown in Figure 11. The data shows advantages of our DRL-based algorithm over the other algorithms. The reason for this is that agent has learned how to allocate channels with the aim of minimizing power consumption in a long term instead of a short period of time.

The power consumption rate that reflects the degree of power efficiency is calculated as follows:

$$R_p = \frac{\sum_{t=1}^{T} P_t - P_o}{P_o} \tag{11}$$

where P_t is actual power consumption at time step t, $\sum_{t=1}^{T} P_t$ is the total power consumption in one episode. P_o denotes optimal power consumption in one episode, which is decided by the demand distribution and the location where they are served the most energy-efficiently.

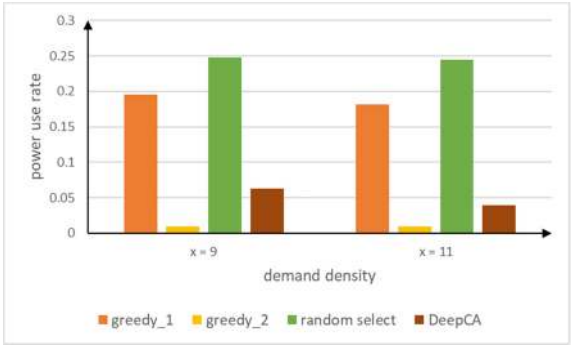


FIGURE 11. Power consumption under different demand distributions.

As shown in Figure 11, the power consumption rate of greedy-2 almost equals to zero. Which means that, it always selects the most energy-efficient area to serve at each time step. It minimizes the power consumption on one hand. On the other hand, it causes a low satisfaction rate for it neglects the nodes to be out of date, results of which can be seen in the next part. We also see that the power use of DeepCA decreases 78.41% and 83.96% compared with greedy_1 algorithm and random select algorithm respectively when average demand desity is 9. And when demand density increases, the decrease amount is 67.86% and 74.60%.

2) SATISFACTION RATE

In our experiment, the satisfaction rate is reflected through blocking rate, and sum of both equals to 1. It is surprising that the service blocking rate of DeepCA in Figure 12 is lower than that of the other algorithms, even greedy_1 algorithm which aims at maximizing the satisfaction rate. The superiority of DeepCA is also evident over the other two schemes. As a matter of fact, this result reveals that the goal

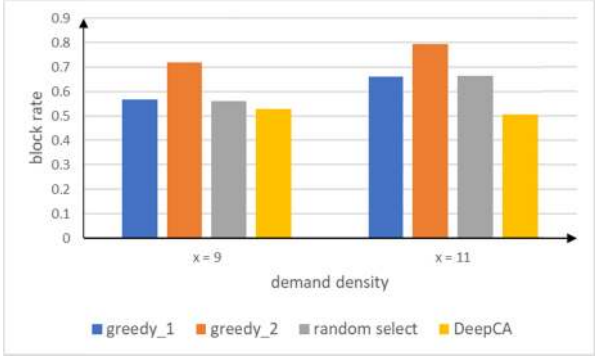


FIGURE 12. Demand blocking rate under different demand distributions.

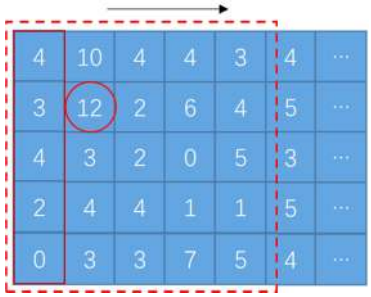


FIGURE 13. One of the scenarios where greedy algorithm performs bad.

of an agent is to minimize power consumption while keeping satisfaction rate in a long period instead of achieving optimal satisfaction rate or power consumption scheme at one time step.

This phenomenon can be explained by the foresight ability of the model. As is shown in Figure 13, when the number of channels are very limited, considering the lasting effect of current channel allocation decision, DeepCA model may deny some service requests that are becoming invalid at the next time step, while satisfying requests in the other grids which has high density of nodes. However, for greedy_1 algorithm, it will choose the nodes that are to be invalid firstly. Following this policy, at the next time step, the grid that has 12 units of requests will probably not be satisfied due to power and channel restrictions.

As the most critical objective metric in this paper, satisfaction rate is the most excellent performance of DeepCA among those channel allocation algorithms. As is inferred before, due to the advantage of model-free learning method, agent can learn how to take action to maximize cumulative reward in a long term. For channel allocation in LEO satellite system, which aims at minimize power consumption while securing coordinative satisfaction rate with the other algorithms or better, it can learn the latent pattern of the state-action pair that achieve the best rewards.

VI. CONCLUSION

In this paper, we proposed DeepCA, a novel approach for dynamic channel allocation in SIoT. We introduced a new sliding block scheme to facilitate the modeling of dynamic feature of the LEO satellite, and formulated the dynamic channel allocation problem in SIoT as an MDP. A deep reinforcement learning based algorithm is proposed for optimal channel allocation. To accelerate the learning process of DeepCA, we utilized the image form to represent the requests of users to reduce the input size, and carefully divided an action into multiple mini-actions to reduce the size of the action set. Simulation results showed that DeepCA consistently outperforms classic channel allocation algorithms. For future research work, we hope to address the energy saving problem by designing a novel battery model, which takes the battery load and constraints into consideration.

REFERENCES

[1] N. K. Giang, J. Im, D. Kim, M. Jung, and W. Kastner, "Integrating the EPCIS and building automation system into the Internet of Things: A lightweight and interoperable approach," *J. Wireless Mobile Netw., Ubiquitous Comput., Dependable Appl.*, vol. 6, no. 1, pp. 56–73, 2015.

[2] B. Pokrić, S. Krčo, D. Drajić, M. Pokrić, V. Rajs, Ž. Mihajlović, P. Knežević, and D. Jovanović, "Augmented reality enabled IoT services for environmental monitoring utilising serious gaming concept," *J. Wireless Mobile Netw., Ubiquitous Comput., Dependable Appl.* vol. 6, no. 1, pp. 37–55, 2015.

[3] I. V. Kotenko, I. Saenko, and A. Branitskiy, "Applying big data processing and machine learning methods for mobile Internet of Things security monitoring," *J. Internet Services Inf. Secur.*, vol. 8, no. 3, pp. 54–63, Aug. 2018.

[4] J. A. Fraire, S. Céspedes, and N. Accettura, "Direct-to-satellite IoT—A survey of the state of the art and future research perspectives," in *Proc. Int. Conf. Ad-Hoc Netw. Wireless*. Cham, Switzerland: Springer, 2019, pp. 241–258.

[5] H. S. Dhillon, H. Huang, and H. Viswanathan, "Wide-area wireless communication challenges for the Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 168–174, Feb. 2017.

[6] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 60–67, Oct. 2016.

[7] H. Huang, S. Guo, W. Liang, K. Wang, and A. Y. Zomaya, "Green data-collection from geo-distributed IoT networks through low-earth-orbit satellites," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 806–816, Sep. 2019.

[8] W. Liu, F. Tian, and Z. Jiang, "Beam-hopping based resource allocation algorithm in LEO satellite network," in *Proc. Int. Conf. Space Inf. Netw.* Singapore: Springer, 2018, pp. 113–123.

[9] Z. Qu, G. Zhang, H. Cao, and J. Xie, "LEO satellite constellation for Internet of Things," *IEEE Access*, vol. 5, pp. 18391–18401, 2017.

[10] P. Angeletti, D. F. Prim, and R. Rinaldo, "Beam hopping in multi-beam broadband satellite systems: System performance and payload architecture analysis," in *Proc. 24th AIAA Int. Commun. Satell. Syst. Conf.*, Jun. 2006, p. 5376.

[11] C. C. White, *Markov Decision Processes*. New York, NY, USA: Springer, 2001.

[12] H. Huang, S. Guo, W. Liang, and K. Wang, "Online green data gathering from geo-distributed IoT networks via LEO satellites," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[13] A. C. Fu, E. Modiano, and J. N. Tsitsiklis, "Optimal energy allocation and admission control for communications satellites," *IEEE/ACM Trans. Netw.*, vol. 11, no. 3, pp. 488–500, Jun. 2003.

[14] Q. Liu, X. Wang, B. Han, X. Wang, and X. Zhou, "Access delay of cognitive radio networks based on asynchronous channel-hopping rendezvous and CSMA/CA MAC," *IEEE Trans. Veh. Technol.*, vol. 64, no. 3, pp. 1105–1119, Mar. 2015.

[15] Y. Kawamoto, H. Nishiyama, Z. M. Fadlullah, and N. Kato, "Effective data collection via satellite-routed sensor system (SRSS) to realize global-scaled Internet of Things," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3645–3654, Oct. 2013.

[16] M. Anteur, V. Deslandes, N. Thomas, and A.-L. Beylot, "Ultra narrow band technique for low power wide area communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[17] T. Deng, J. Zhu, and Z. Nie, "An adaptive MAC protocol for SDCS system based on LoRa technology," in *Proc. 2nd Int. Conf. Automat., Mech. Control Comput. Eng. (AMCCE)*, 2017.

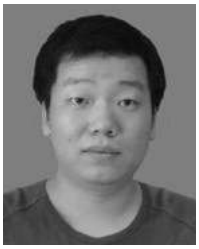
[18] Y. Yang, M. Xu, D. Wang, and Y. Wang, "Towards energy-efficient routing in satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3869–3886, Dec. 2016.

[19] Z.-Y. Ai, F. Song, and X. Wang., "Combining SDN and ICN for network survivability improvement," *J. Internet Services Inf. Secur.*, vol. 8, no. 1, pp. 18–30, 2018.

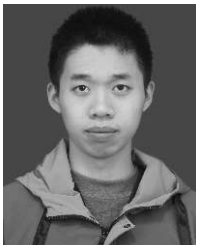
[20] A. Detti, A. Caponi, and N. Blefari-Melazzi, "Exploitation of information centric networking principles in satellite networks," in *Proc. IEEE 1st AESS Eur. Conf. Satell. Telecommun. (ESTEL)*, Oct. 2012, pp. 1–6.

[21] I. F. Akyildiz, H. Uzunalio lu, and M. D. Bender, "Handover management in low earth orbit (LEO) satellite networks," *Mobile Netw. Appl.*, vol. 4, no. 4, pp. 301–310, 1999.

[22] L. Chen, J. Lingys, K. Chen, and F. Liu, "AuTO: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization," in *Proc. Conf. ACM Special Interest Group Data Commun. (SIGCOMM)*, 2018, pp. 191–205.
[23] Z. Zhang, L. Ma, K. Poularakis, K. K. Leung, J. Tucker, and A. Swami, "MACS: Deep reinforcement learning based SDN controller synchronization policy design," in *Proc. IEEE 27th Int. Conf. Netw. Protocols (ICNP)*, Oct. 2019, pp. 1–11.
[24] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. 15th ACM Workshop Hot Topics Netw. (HotNets)*, 2016, pp. 50–56.



BAOKANG ZHAO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the National University of Defense Technology, Changsha, Hunan, China, in 2002, 2004, and 2009, respectively, all in computer science. He is currently an Associate Professor with the National University of Defense Technology. His research interests include high-performance networking, data center networking, space-air-integrated networking, Internet of Things, as well as cloud and distributed computing.



JIAHAO LIU received the B.S. degree in network engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is currently pursuing the master's degree in cyberspace security. His research interests include satellite communications, and machine learning techniques applied in computer networks.

ZILING WEI received the B.S. and M.S. degrees in computer science from the National University of Defense Technology (NUDT), Changsha, China, in 2012 and 2014, respectively, and the Ph.D. degree in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2019. He is currently an Assistant Professor with NUDT. His research interests include network security, wireless communications, and edge computing.

ILSUN YOU (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Dankook University, Seoul, South Korea, in 1997 and 2002, respectively, and the second Ph.D. degree from Kyushu University, Japan, in 2012. From 1997 to 2004, he was at the THINmultimedia Inc., Internet Security Company, Ltd., and Hanjo Engineering Company, Ltd., as a Research Engineer. He is currently an Associate Professor with the Department of Information Security Engineering, Soonchunhyang University. Especially, he has focused on 4G/5G security, security for wireless networks and mobile internet, IoT security, and so forth, while publishing more than 180 papers in these areas. He is a Fellow of the IET. He has served or is currently serving as the General Chair or a Program Chair of international conferences and workshops such as WISA'19-20, MobiSec'16-19, AsiaARES'13-15, MIST'09-17, MobiWorld'08-17, and so forth. He is the EiC of the *Journal of Wireless Mobile Networks*, *Ubiquitous Computing*, and *Dependable Applications* (JoWUA). He is in the Editorial Board for *Information Sciences* (INS), the *Journal of Network and Computer Applications* (JNCA), IEEE Access, the *Intelligent Automation & Soft Computing* (AutoSoft), the *International Journal of Ad Hoc and Ubiquitous Computing* (IJAHUC), *Computing and Informatics* (CAI), and the *Journal of High Speed Networks* (JHSN).