Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Multi-dimensional resource allocation strategy for LEO satellite communication uplinks based on deep reinforcement learning

Yu Hu[1], Feipeng Qiu[1], Fei Zheng[1*] and Jilong Zhao[1]

Abstract

In the LEO satellite communication system, the resource utilization rate is very low due to the constrained resources on satellites and the non-uniform distribution of traffics. In addition, the rapid movement of LEO satellites leads to complicated and changeable networks, which makes it difficult for traditional resource allocation strategies to improve the resource utilization rate. To solve the above problem, this paper proposes a resource allocation strategy based on deep reinforcement learning. The strategy takes the weighted sum of spectral efficiency, energy efficiency and blocking rate as the optimization objective, and constructs a joint power and channel allocation model. The strategy allocates channels and power according to the number of channels, the number of users and the type of business. In the reward decision mechanism, the maximum reward is obtained by maximizing the increment of the optimization target. However, during the optimization process, the decision always focuses on the optimal allocation for current users, and ignores QoS for new users. To avoid the situation, current service beams are integrated with high- traffic beams, and states of beams are refactored to maximize long-term benefits to improve system performance.

Simulation experiments show that in scenarios with a high number of users, the proposed resource allocation strategy reduces the blocking rate by at least 5% compared to reinforcement learning methods, effectively enhancing resource utilization.

Keywords LEO satellite communication system, Resource allocation, Deep reinforcement learning, Long-term benefits

Introduction

Recently, the LEO satellite communication system has become an integral part of the satellite communication field due to its unique advantages. These advantages include global seamless communication coverage, high communication reliability independent of geographical environment, large system capacity for massive users and multiple data services such as video calls, real-time video streaming and so on. LEO satellite communication system plays a vital role in various fields, including aviation and navigation, satellite navigation, telemedicine, smart power grids and emergency rescue [1–3].

With the dramatic increase of communication services, traditional single-beam satellite systems are no longer able to meet the communication requirements of large service capacity and high resource utilization. In response to this, multi-beam satellite systems utilize phased array antenna technology to generate multiple spot beams and employ frequency reuse techniques to enhance capacity and resource utilization. However, due to the concentrated placement of antennas

*Correspondence:
Fei Zheng
zhengfei@guet.edu.cn
[1] Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin University of Electronic Technology, Guilin 541004, China

on multi-beam satellites and the multi-coverage on the earth, each antenna receives signals from neighbor-beam and even cross-beam users on the same frequency, resulting in serious co-frequency interference between the beams. The co-frequency interference is a significant factor which restricts resource utilization.

The contradictions become more acute between the explosive growth of data services and the limited on-board resources in multi-beam satellite systems. Due to the non-uniform distribution of service requests in pace and in time, the huge business drop leads to extremely low resource utilization between beams. For high traffic beams, the scarcity of available resources results in competition between users to meet the minimum communication quality of service (QoS), and the competition ultimately reduces QoS. Conversely, for low traffic beams, a handful of resource is sufficient to meet the communication QoS, and considerable idle resources result in low resource utilization. Considering the diverse types of services and the complex satellite network environment, it is of great significance to study an efficient and intelligent resource allocation strategy. This paper focuses on uplink resource allocation in a multi-beam LEO satellite system. The main contributions of this paper are as follows:

- Taking co-frequency interference and traffic distribution into account between beams, a joint channel-power allocation strategy based on deep reinforcement learning is proposed. When the satellite is in an area with low traffic volume, the proposed method can improve resource utilization by adjusting the weights of spectral efficiency and energy efficiency, while still providing high QoS. Conversely, when the satellite is in a high-traffic area, the method can adjust the weight of the blocking rate to accommodate more users. Although this may reduce QoS, it enhances resource utilization while ensuring the minimum QoS.
- Present works focus on the QoS of current users and ignore the optimal allocation for subsequent users. Therefore, during the state reconstruction process, the interference beams and the high-traffic beams are integrated with the current serving beam, so as to maximize long-term benefits and improve the overall system performance.

The rest of this paper is organized as follows. The next section presents related work on resource allocation strategies. Section 3 introduces the uplink model of LEO satellite communications and the optimization model of resource allocation. Section 4 introduces the joint channel-power allocation strategy based on deep reinforcement learning algorithm. Section 5 provides simulation analysis. The last section is the summary of the whole paper.

Related work

In the initial stage of the development of satellite communication systems, the simplicity of the network architecture means that fixed resource allocation strategies are adequate to fulfill QoS requirements. However, with the mass terminal accessing and differentiated services, the network environment has become complex and changeable, rendering fixed resource allocation inadequate. Compared with fixed resource allocation, dynamic resource allocation can achieve higher resource utilization in such complex and dynamic network environments. Dynamic resource allocation can dynamically allocate resources such as channel, power, time and spot beams based on the distribution of traffic and beam state information. It also can deal with resources efficiently and flexibly. Consequently, dynamic resource allocation has become a research hotpot [4, 5].

Regarding dynamic resource allocation, numerous researchers have conducted extensive studies. Literature [6] proposes a channel allocation algorithm based on beam cooperation transmission. The algorithm utilizes the cooperation between beams to aggregate user signals at the receiver, thereby increasing signal energy to improve channel quality. Literature [7] considers the dynamical traffic scenario, focusing on co-frequency interference between users. The interference of channels is detected based on user location information, and then dynamical scheduling channel improves QoS. However, the complex and changeable network leads to the high complexity of the channel allocation algorithm. To mitigate the algorithm complexity, literature [8] proposes a channel allocation algorithm based on improved channel interference detection. The interference threshold is set for channels to lower complexity, and the algorithm further optimizes QoS. In literature [9], channels are dynamically reserved according to user priority, and the threshold of channel reservation is calculated by genetic algorithm. The threshold is dynamically adjusted according to the traffic distribution to reduce the handover failure rate. Literature [6–9] primarily focuses on the issue of co-channel interference and does not consider the distribution of traffic volume between beams.

Due to the different terminals between beams, the traffic is also unevenly distributed in the satellite system. With limited on-board resources, users between beams compete for resources to meet QoS, which hinders resource utilization improvement. To solve this problem, the resource allocation methods have evolved from single-resource allocation to joint resource allocation. Literature [10] proposes a joint power and channel allocation

algorithm, which allocates power and channel according to channel state information, while ensuring fairness among users. However, this approach does not consider inter-beam co-frequency interference. Literature [11] shows that co-frequency interference is the main factor to reduce communication performance. This interference affects both the uplink and downlink, limiting link capacity and system throughput. Considering the co-frequency interference between beams, literature [12, 13] investigate power and bandwidth resource allocation. In literature [12], a genetic algorithm is employed to construct a joint optimization model for power and bandwidth allocation. Literature [13] proposes an improved power and bandwidth joint allocation strategy. The strategy utilizes a sub-gradient algorithm to ensure fairness among users, so as to improve system capacity. Considering the service diversity, literature [14] proposes a random-on-demand channel allocation strategy according to the ratio between random and on-demand allocation, significantly reducing system delay and maximizing throughput. Literature [15] uses heuristic algorithms to solve frequency and beam allocation problems under resource-limited and unlimited scenarios. Aiming at minimizing the variance of supply and demand, Lagrange algorithm is used to obtain the optimal beam bandwidth allocation. Literature [10–15] proposes allocation strategies that consider traffic volume differences, but overlook the mobility of LEO satellites. The rapid movement of LEO satellites leads to a complex and dynamic network, making traditional resource allocation strategies inadequate. More efficient algorithms are needed to cope with the rapidly changing network environment.

Recently, the combination of AI technology and communication technology has gradually become main stream, such as intelligent medical, smart grid, smart home, and unmanned vehicles. For example, literature [16] proposes a cutting-edge deep network architecture, HighDAN for short, by embedding the adversarial learning-based DA's idea into HR-Net with Dice Loss (to reduce the effects of the class imbalance), making it largely possible to break the semantic segmentation performance bottleneck in terms of accuracy and generalization ability from cross-city studies. Among AI technologies, machine learning is the process of enabling machines to imitate human cognition and learn about the external environment. In the machine communication, interactive learning between machine and environment is used to improve communication performance [17]. As a branch of machine learning, reinforcement learning introduces a reward mechanism to achieve the goal of maximizing rewards [18]. In the heterogeneous cellular networks, literature [19] proposes a resource allocation algorithm combining game theory and reinforcement learning to reduce user power consumption. In literature [20], reinforcement learning solves the congestion control problem in satellite Internet of Things. Compared with traditional algorithms, reinforcement learning can more effectively reduce system blocking rate. For cellular networks of device-to-device (D2D), literature [21] uses reinforcement learning to obtain learning experience from the previous channel power allocation. D2D can share the channels of cellular users so as to avoid co-frequency interference [22]. Literature [23] adopts the distributed architecture and takes multiple D2D devices as agents. Literature [24] focuses on developing a novel artificial intelligence model called SpectralGPT. This model addresses challenges in processing spectral data, particularly in the context of remote sensing(RS). Literature [25] proposes a new transformer-based backbone network which is more focused on extracting spectral information, called SpectralFormer, in order to substitute for CNN- or RNN-based architectures. Without using any convolution or recurrent units, the proposed SpectralFormer can achieve state-of-the-art classification results for HS images.

It obtains the optimal power distribution scheme through a Q-learning algorithm. Literature [26] proposes a deep reinforcement learning method based on multi-agent collaboration to allocate bandwidth with low complexity. Deep reinforcement learning has more powerful performance, which makes dynamical allocation more efficient and flexible. In order to solve the multi-dimensional resource allocation in multi-beam satellite communication, literature [27] introduces a time–frequency two-dimensional resource allocation algorithm. The algorithm considers the number of users and system throughput to efficiently allocate resources. Literature [28] proposes a distributed multi-agent reinforcement learning method to improve the utilization rate of spectrum in vehicle networking scenarios. This method can efficiently allocate shared resource blocks and vehicle transmission power. It also meets the high data rate and high reliability of vehicle-to-infrastructure link. Literature [29] proposes a beam-hopping resource allocation algorithm based on deep reinforcement learning for resolving large data transmission delay. This algorithm introduces interference avoidance criterion to flexibly allocate time slots. Literature [30] proposes an approximate optimal dynamical bandwidth allocation strategy to meet time-varying traffic requirements in the multi-beam satellite communication. Currently, the existing literature on resource management mainly emphasizes immediate gains while neglecting long-term benefits. For example, whenever a new user accesses the system, the system allocates the best communication resources to achieve high QoS, which is not conducive to subsequent new user access. This paper focuses more on long-term gains. When

a new user accesses the system, the allocated resources may not be optimal, but they are more favorable for subsequent new user access, thus reducing the blocking rate.

The satellite wireless resource allocation can be regarded as a sequential decision-making problem. Deep reinforcement learning has strong environment perception ability and decision-making ability to solve this problem. In this paper, the LEO satellite is considered as the agent, each beam and each user are treated as the environmental state, and available channels and terminal transmission power are regarded as actions. The reward function is designed according to channel spectrum utilization, user energy efficiency and user blocking rate. The deep reinforcement learning algorithm is used to train the optimal joint channel power allocation strategy. State reconstruction is performed for current users to reduce the data dimensionality, so that the system can allocate channels and power for new users.

Environmental interaction model and QoS optimization model for multibeam LEO satellite systems

This section mainly introduces the multi-beam LEO satellite system model, and constructs an optimization function for spectrum, power and blocking rate.

Environmental Interaction Model

Considering the uplink of the multi-beam satellite system in Fig. 1, users have access to all frequency bands. The multi-beam LEO satellite utilizes phased array antenna technology to generate spot beams. Users are randomly distributed among different beams. The beam set is $M = \{1, 2, \ldots, S\}$, and the users in beam m are represented as $i = \{1, 2, 3 \ldots, I\}$. The beam divides the spectrum into N channels, which are represented by $n = \{1, 2, \ldots, N\}$. When the channel n in the beam m is occupied by the user, note $w_{s,n} = 1$, otherwise 0.

Considering the resource allocation over continuous time, assumes that at time t, each user occupies only one channel in its own beam. Then the channel allocation information can be represented as follows:

$$w^t = \begin{bmatrix} w_{11}^t & w_{12}^t & \cdots & w_{1N}^t \\ w_{21}^t & w_{22}^t & \cdots & w_{2N}^t \\ \vdots & \vdots & \ddots & \vdots \\ w_{S1}^t & w_{S2}^t & \cdots & w_{SN}^t \end{bmatrix} \tag{1}$$

The user is the transmitter and the satellite is the receiver in the uplink. Then the antenna receiving gain $G_R(\theta)$ can be calculated by the following formula:

$$G_R(\theta) = \mathrm{g} G_{\max} \tag{2}$$

$$G_{\max} = 2\eta\pi^2 r^2/\lambda^2 \tag{3}$$

$$\mathrm{g} = \left[\frac{J_1(u)}{2u} + 36\frac{J_3(u)}{u^3}\right]^2 \tag{4}$$

$$u = 2.07123\frac{\sin\theta}{\sin\theta_{3db}} \tag{5}$$

where $G_{\max}$ is the maximum antenna gain in the center of the satellite receiving antenna, and g is the gain factor. η is the efficiency of LEO antennas. r is the aperture of LEO antennas and λ is the carrier wavelength. $J_1(\cdot)$ and $J_3(\cdot)$ are the first and third-order Bessel functions respectively. θ is the receiving angle of the current user in its own beam. θ_{3dB} is the angle at which the received signal decreases 3 dB relative to the beam antenna gain. Unlike traditional antennas, multi-beam antennas have high receiving gain in servicing beams and low receiving gain for other beams, which can reduce the interference from users in other beams. The diagram of co-frequency interference model in uplink of LEO satellite communication is shown in Fig. 2.

When the user terminal transmits signals to the satellite, the wireless signal spreads in a spherical shape. This is known as free space path loss. L represents the free space path loss.

$$L = 92.45 + 20\lg d + 201\lg f \tag{6}$$

where d is the distance between the satellite and the ground and f is the signal frequency band. When the user terminal transmits signals to the satellite, the signal power is expressed as:

$$p_R = \frac{p_T G_R}{L} \tag{7}$$

Because the antenna sidelobe is too large, the antenna attenuation is relatively gentle, resulting in interference between adjacent beams, and residual co-frequency interference cannot be ignored. Considering the presence of co-frequency interference I and Gaussian white noise power N_0, the SINR can be expressed as:

$$\gamma = \frac{p_R}{I + N_0} \tag{8}$$

In the channel n, the transmit power of user i in beam m is $p_{i,m}^t$, $G_{i,m}^t$ is the receiving gain from current user i, $L_{i,m}^t$ is the free space loss of current user i, j is the user using channel n in other beams. On the receiver, the SINR from user i can be expressed as:

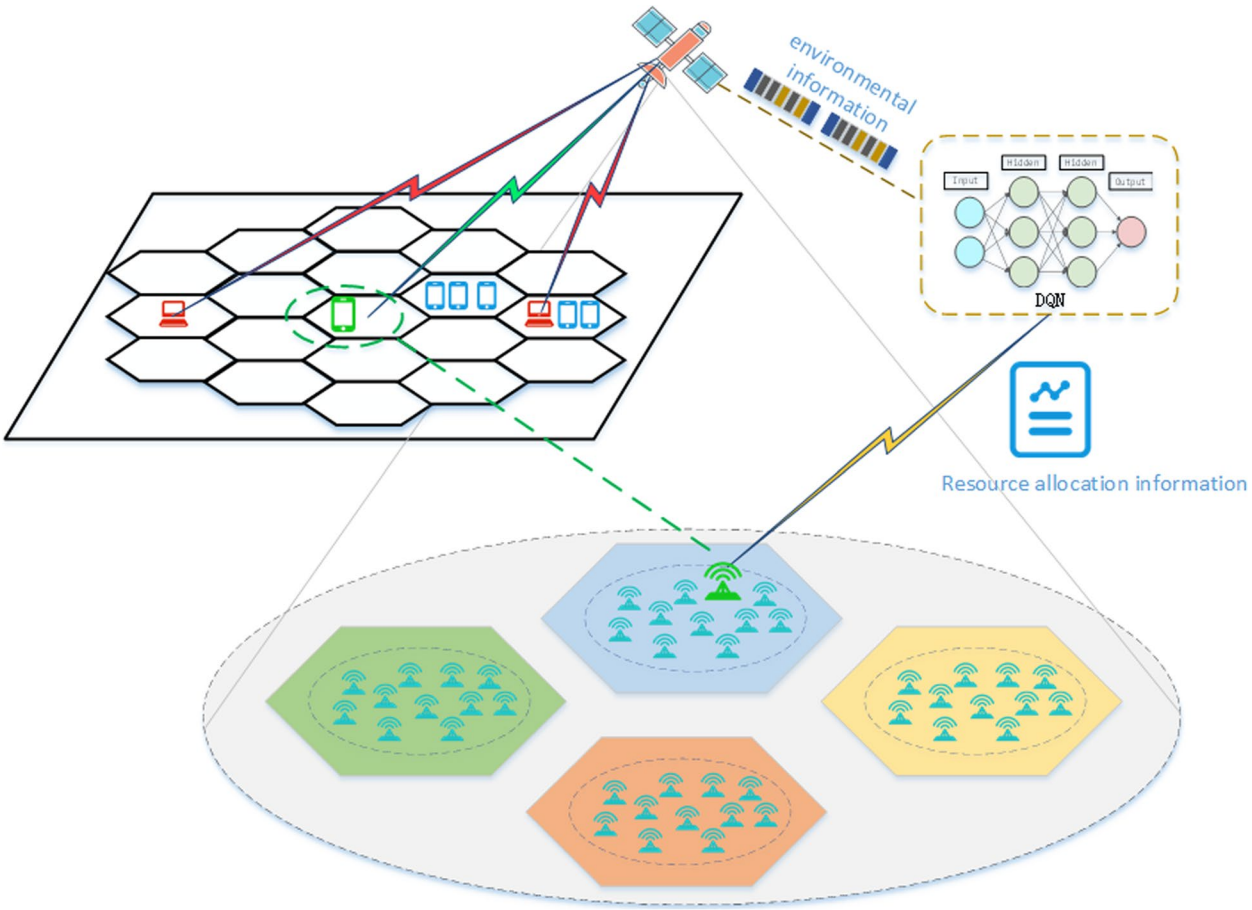

Fig. 1 Diagram of the environment interaction model of the multi-beam LEO satellite system

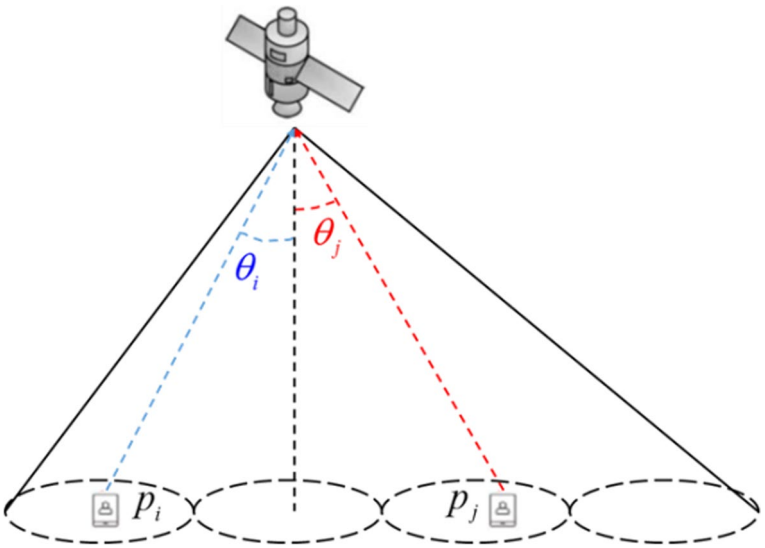

Fig. 2 Diagram of co-frequency interference model in uplink of LEO satellite communication

$$\gamma_{i,m}^{t} = \frac{p_{i,m}^{t} G_{i,m}^{t} / L_{i,m}^{t}}{\sum_{j=1, j\neq i}^{M} p_{j,m}^{t} G_{j,m}^{t} / L_{i,m}^{t} + N_0} \tag{9}$$

QoS optimization model

The communication rate from user i can be calculated by the channel model and Shannon formula:

$$R_{i,m}^{t} = B\log_2\left(1 + \gamma_{i,m}^{t}\right) \tag{10}$$

where B is the channel bandwidth. Multiple users use the same channel in a multibeam satellite with full frequency multiplexing. The channel capacity of channel n can be obtained as:

$$R_{n}^{t} = \sum_{m=1}^{M} R_{i,m}^{t} \tag{11}$$

To improve the resource utilization, the bandwidth utilization is the optimization index for channel allocation, and the bandwidth utilization is expressed as:

$$SE = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} R_{i,m}^{t}}{B} \tag{12}$$

When the channel bandwidth is constant, increasing transmit power can increase the channel capacity. However, when the channel capacity tends to saturation, the user can not improve the channel capacity by increasing the power. Using energy efficiency as the optimization index for power control, energy efficiency is expressed as:

$$EE = \sum_{m=1}^{M}\sum_{n=1}^{N} \frac{R_{i,m}^{t}}{p_{i,m}^{t}} \tag{13}$$

To meet communication requirements, a user's Signal-to-Interference-plus-Noise Ratio (SINR) must not fall below a certain threshold. Typically, this threshold is set as γ_k where k represents the user's current service type. Only when $\gamma_{i,m}^{t} \geq \gamma_k$ can users communicate normally; otherwise, users may experience dropped calls or blockage. If a new user has no available channels, or if channel allocation causes other users' SINR to fail to meet the threshold, it is also considered as blockage. Blockage for current users can be expressed as:

$$\phi_{i,m}^{t} = \left\{ \begin{array}{l} 0,\ \gamma_{i,m}^{t} \geq \gamma_k \ \ and \ \ w_m = 0 \\ 1,\ \gamma_{i,m}^{t} < \gamma_k \ \ and \ \ w_m = 1 \end{array} \right\} \tag{14}$$

At the current moment t, there are a total of U_{tot} users in the system. If the total number of users experiencing blockage in the system is $\sum_{\mathrm{m}=1}^{M}\sum_{n=1}^{N}\phi_{\mathrm{i,m}}^{\mathrm{t}}$, then the system blocking rate is:

$$VE = \frac{\sum_{\mathrm{m}=1}^{M}\sum_{n=1}^{N}\phi_{\mathrm{i,m}}^{\mathrm{t}}}{U_{tot}} \tag{15}$$

Combining the bandwidth utilization, the energy efficiency, and the blocking rate, the optimization objective function is defined as:

$$\begin{aligned} opt &\left\{ \begin{array}{l} \max \sum_{t\epsilon} SE \\ \max \sum_{t\epsilon} EE \\ \max \sum_{t\epsilon} VE \end{array} \right. \\ st &\left\{ \begin{array}{l} s1 : p_{i,m}^{t} \leq p_{\max} \\ s2 : \gamma_{i,m}^{t} \geq \gamma_k \\ s3 : w_{m,n}^{t} \in [0,1] \end{array} \right. \end{aligned} \tag{16}$$

The function means that the blocking rate should be reduced as much as possible when the resource utilization is maximized. $p_{\max}$ is the upper limit of user transmit power. The constraint $s1$ indicates the maximum transmit power of the user terminal. The constraint $s2$ indicates the SINR required for service transmission to exceed the threshold, and the constraint $s3$ indicates that each user can occupy only one channel.

Resource allocation strategy for LEO satellite communication uplinks based on deep reinforcement learning

This chapter focuses on the joint channel power allocation strategy to improve resource utilization and reduce blockage. In deep reinforcement learning, the strategy can be mapped as the satellite intelligences to maximize the benefit for each user. The overall framework of the algorithm is shown in Fig. 3.

The satellite is defined as the intelligent agent. The beams are defined as the environment. And the gain function is associated with the resource allocation problem. The satellite senses the new users and obtains the optimal resource allocation strategy according to the service information, the channel allocation matrix and the traffic distribution. The algorithm complexity is reduced by state reconfiguration, and the decision performance is improved by experience replay pool and Q-network training.

State definition

State space

The state space contains the main information of the external environment. The resource allocation needs to obtain the current user traffic distribution, the state of channel resource occupation and the service information.

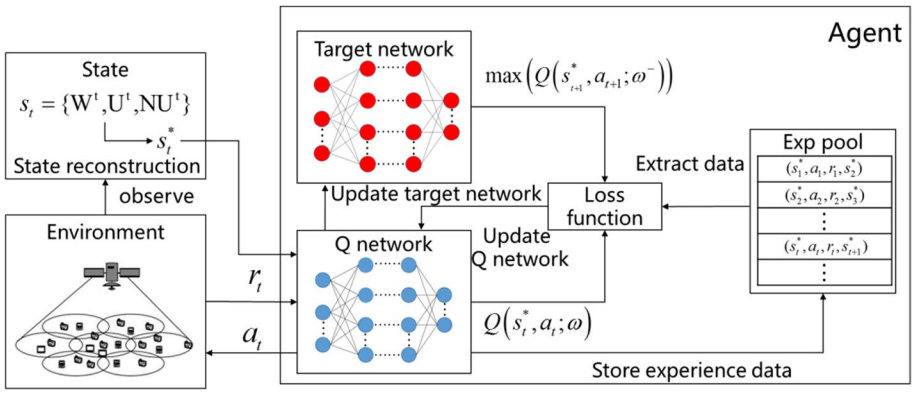

Fig. 3 The overall framework of resource allocation algorithm based on deep reinforcement learning

Therefore, the state space S_t contains the channel allocation matrix W^t, the user traffic distribution matrix U^t, and the new user service information NU^t. The state space S_t is expressed as:

$$\mathrm{S_t} = \left\{\mathrm{W^t}, \mathrm{U^t}, \mathrm{NU^t}\right\} \tag{17}$$

where NU^t represents the servic information of the new user. NU^t contains the beam of new user and the threshold of SINR. U^t represents the number of users in each beam. The current state is considered as a terminal state, when all users have been allocated channels and power, or when no resources are available. And the system proceeds to the next training round.

Action space

The intelligent agent selects an appropriate action based on the current state. Therefore, the action space is defined as follows:

$$a_t = \{m, p\} \tag{18}$$

where m is the selected channel number and p is the selected user transmission power. The transmit power can be divided into multiple power domains. At time t, the intelligent agent inputs the environmental state information s into the deep Q-network, when a new user appears in the beam. Then the deep Q-network selects the free channel and transmission power for the user. When a_t={0,0}, the new user is not allocated a channel and power. The intelligent agent aims to maximize long-term rewards for the new user. When resources are allocated to a_t new user, it may result in other users being unable to transmit their services properly. Therefore, the scenario of not allocating resources should also be considered.

Reward function

The intelligence agents maximize the accumulated reward through strategies learning. The optimization goal is to improve the resource utilization and reduce the blocking rate. We can associate the reward function with the optimization index. After the three indexes are processed by the Ψ normalization function, the weighted sum can be expressed as:

$$Z = a_1\Psi(SE) + a_2\Psi(EE) + a_3\Psi(1 - VE) \tag{19}$$

where a_1, a_2 and a_3 are the weighted values of spectral efficiency, energy efficiency, and blocking rate, respectively. The reward function is defined as:

$$r_t = \begin{cases} 1, \Delta Z > 0 \\ 0, \Delta Z \leq 0 \end{cases} \tag{20}$$

where ΔZ is the increment of the function, $\Delta Z = Z_{t+1} - Z_t$. When $\Delta Z > 0$, the new action will be rewarded and when $\Delta Z \leq 0$, the new action will not obtain rewards.

Analysis of state reconfiguration

The algorithm complexity is too large if all the state information inputs to the deep Q-network for training. Therefore, it is effective to reconstruct the state space. The elements are only used for new users in the state space. The beams mutually influence each other, and co-frequency interference originates from the surrounding two concentric beams. However, it is disadvantageous to only consider the surrounding two concentric beams for long-term benefits.

As shown in Fig. 4, beam a and beam b have two available channels w_1 and w_2 at the current moment. If we only consider the surrounding two concentric beams, it

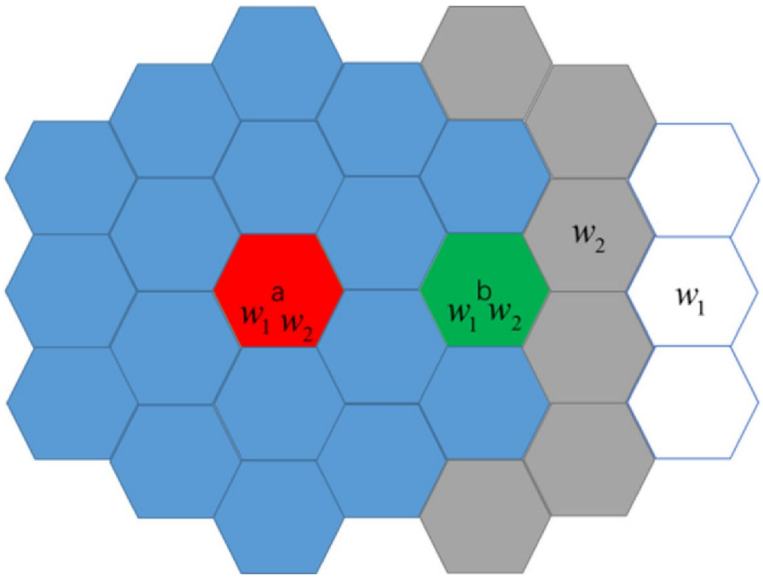

Fig. 4 Schematic diagram of state reconstruction analysis

would assume that both channels in beam b are available. In this case, channel w_1 would be the optimal choice in beam a. After the system allocating channel w_1 to the new user of beam a, the subsequent new users in beam b will be blocked by strong co-channel interference when accessing the channel w_1. However, if we consider the surrounding three concentric beams, we can allocate channel w_2 to the new users in beam a. The co-frequency interference for new users in beam b is relatively weaker. Therefore, the reconstructed state space can be represented as s^* according to the beam about new user and the surrounding three concentric beams.

Q-network training and updating

Compared to the reinforcement learning, neural network reinforcement learning can efficiently process high-dimensional state data and action data [31, 32]. There is a correlation between states and actions, and neighboring states or actions can influence each other.

Deep reinforcement learning introduces the experience replay mechanism, which reduces the correlation between data. It makes deep Q-networks easy to converge and the training update process more stable. Deep reinforcement learning introduces the target Q-network to reduce the correlation between the Q-value increase and the target Q-value through an error function, thereby improving algorithm stability.

In Q-Learning, the value function $Q(s_t, a_t)$ is stored in a Q-value table. In deep reinforcement learning, the value function (Q-value) is parameterized as a function $Q(s_t, a_t)$ and mapped from the state space to action Q-values using a deep Q-network.

$$\mathrm{Q}(s_t, a_t; \omega) \approx \mathrm{Q}(s_t, a_t) \tag{21}$$

Each value function $Q(s_t, a_t)$ corresponds to a network parameter ω, where ω represents the weight value of the neural network. The intelligent agent selects the action a_t according to the reconstructed state s_t^*. After the action is applied to the environment, the environment provides feedback a reward r_t and the next state s_{t+1}^* to the agent.

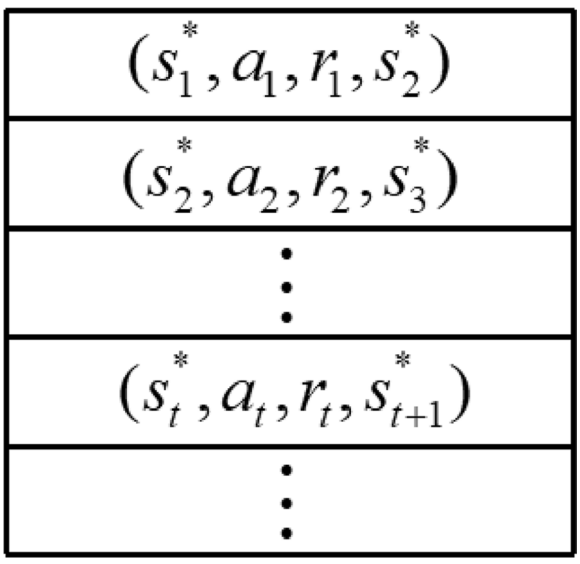

Fig. 5 Schematic diagram of the experience replay pool architecture

Experience data $(s_t^*, a_t, r_t, s_{t+1}^*)$ is extracted from the target Q-network and is stored in an experience replay pool, which is as illustrated in Fig. 5.

Updating the value function $Q(s_t, a_t)$ is equivalent to updating the network parameters. The updating formula is as follows:

$$\begin{aligned} Q'(s_t^*, a_t) = Q(s_{t+1}^*, a_{t+1}) + \\ \alpha(r_t + \lambda \max(Q(s_{t+1}^*, a_{t+1}) - Q(s_{t+1}^*, a_{t+1}))) \end{aligned} \tag{22}$$

where α is the learning rate and λ is the discount factor for long-term benefit. During the training process of deep reinforcement learning, the error between the two Q-networks is calculated using an error function. And the Q network parameters are updated according to the error in reverse.

In order to approximate the action-value function, the error function needs to approach 0. The error function is defined as follows:

$$Loss(\omega) = E\left[\left(\begin{array}{c} r_t + \gamma \max(Q'(s_{t+1}^*, a_{t+1}, \omega^-)) \\ -Q(s_t^*, a_t, \omega) \end{array}\right)^2\right] \tag{23}$$

Similar to the error back propagation algorithm, the current Q-network passes the error calculation results backward and updates the parameters ω through the gradient descent method. The update formula is as follows:

$$\omega_{t+1} = \omega_t + \alpha[r + \lambda \max Q(s_{t+1}^*, a_{t+1}; \omega^-) - Q(s_{t+1}^*, a_t; \omega_t)] \nabla Q(s_t^*, a_t; \omega_t)] \tag{24}$$

To prevent the satellite intelligences from falling into local optimum, the actions are selected by the $\varepsilon-$ greedy algorithm. It selects unexecuted actions according to the probability ε and selects existing actions according to the probability $1-\varepsilon$. In addition, the Q-network parameters ω are updated at each step. The Q-network assigns the parameter ω to the parameter ω^- of the target Q-network at each interval of certain steps.

Analysis of algorithm complexity

The neural network in the proposed strategy includes convolutional and fully connected layers. The complexity of the algorithm can be calculated by evaluating the time complexity of these layers. The time complexity of convolutional layers is:

$$O\left(\sum_{\nu=1}^{V} K_\nu^2 H_\nu^2 C_{\nu-1} C_\nu\right) \tag{25}$$

Here, V represents the number of convolutional layers, K_ν the size of the convolution kernel in layer ν, H_ν the output data dimension of layer ν, C_ν the number of output channels in layer ν, which is equivalent to the number of convolution kernels, and $C_{\nu-1}$ the number of input channels. The time complexity of fully connected layers is:

$$O\left(\sum_{\nu=1}^{V'} 2X_\nu Y_\nu\right) \tag{26}$$

In this equation, V' indicates the number of fully connected layers, X_ν the input to layer ν of the fully connected layers, and Y_ν the output of the fully connected layers. The total complexity of the algorithm is the sum of the complexities of these individual layers:

$$O\left(\sum_{\nu=1}^{V} K_\nu^2 H_\nu^2 C_{\nu-1} C_\nu + \sum_{\nu=1}^{V'} 2X_\nu Y_\nu\right) \tag{27}$$

Regarding state reconstruction, the Q-learning algorithm considers three beam layers, while the DQN algorithm takes into account four beam layers. DQN algorithm has a higher data dimensionality and involves a deep network, leading to a higher complexity compared to the Q-learning algorithm. However, the complexity of the DQN algorithm decreases as the training process converges, making it adaptable to the high mobility environments of LEO satellites.

Resource allocation algorithm

In each time slot, new users randomly appear in the system, and the deep reinforcement learning algorithm allocates channels and power to these new users. The algorithm process is as follows: The scene parameters are first initialized, and then the state space is constructed to allocate resources as the action space. In each training process, starting from the first state, the action is randomly selected. The action is executed and rewarded, the training goes to the next state, and the next state is reconstructed. After that, the experience pool is played back. The network parameters are updated, and the above steps are repeated. When the training reaches the last state, or when no available resource can be allocated, the training ends and goes to the next round of training. The DQN-based joint channel power allocation algorithm is shown in Table 1.

Simulation analysis

In our scenario, 37 spot beams are set up, and 200 users randomly appear according to Poisson distribution. The comparison algorithms are deep reinforcement learning algorithm and Q-Learning algorithm. The weights

Table 1 Resource allocation algorithm

DQN-based joint channel power allocation algorithm
1 Initialize scene parameters and algorithm parameters
2 Obtain channel assignment information, user distribution information, and new user service information
3 **for** episode = 1:max_ episode
4 Initialize state space s_t
5 State Reconfiguration s_t^*
6 **for** t = 1,2,3……,T -1
7 Select action by ε — greedy algorithm
8 Execute the action a_t, get the reward valuer_t, and observe the next state s_{t+1}
9 Reconstruct s_{t+1} as s_{t+1}^* and put experience data $\left(s_t^*, a_t, r_t, s_{t+1}^*\right)$ into the replay experience pool
10 Randomly selected sample data from the replay experience pool
11 Calculate the error function
12 Updating Q-network parameters using gradient descent ω
13 Update the target Q network parameters ω^-
14 **end**
15 **end**
16 Get deep reinforcement learning network parameters
17 Output the channel and power assigned to each new user

Table 2 Simulation parameters

Simulation Parameters	**Value**
Satellite Altitude	780 km
Number of Beams	37
Number of channels	16
The Number of Users	200
User Maximum Transmit Power	20dbW
Business Minimum SINR	3db
Individual Channel Bandwidth	1 MHz
Free Space Loss	212 dB
Transmitting Antenna Gain	40 dB
Receiving Antenna Gain	50 dB
Convolution Kernel 1	6
Convolution Kernel 2	2
Output Dimension 1	3
Output Dimension 2	2
Output Channels 1	1*16
Output Channels 2	16*32
Fully Connected Layer 1	128*128
Fully Connected Layer 2	128*16
Experience Pool Capacity	10000
Discount Factor	0.9
Learning Rate	0.001
Exploration rate	0.01 ~ 1

are set to (1/3, 1/3, 1/3) and (1/4, 1/4, 1/2), corresponding to spectral efficiency, energy efficiency, and blocking rate, respectively. When the number of users in the system is low, the weights for spectral efficiency and energy efficiency can be appropriately increased. When the number of users is high, the weight for the blocking rate can be increased to ensure more users can access the system. The deep reinforcement learning algorithm considers the four-layer beam and the Q-Learning algorithm considers the three-layer beam. Although reinforcement learning algorithms have excellent computational performance, they are not adept at handling high-dimensional data, especially in complex and highly mobile scenarios. Therefore, to achieve a better comparative effect, the experimental design involves restructuring the environmental state into three layers of beams to reduce data dimensionality and enhance algorithm performance. In the early stages when the number of users is not high, the results obtained from the two approaches will be very close. It is only when the number of users is sufficiently high that significant differences in results between the two approaches will emerge. Factors such as the discount factor, learning rate, and exploration rate can influence the convergence performance of the algorithm. To ensure convergence, the learning rate is set to 0.01, the discount factor to 0.9, and the initial exploration rate to 1, which is then gradually reduced to 0.01 as training progresses. We show the simulation parameters in Table 2.

As shown in Fig. 6, the blocking rate increases as the number of users increases. It can be observed that the blocking rate starts to rise significantly when the number of users reaches 125. When the number of users reaches 200, the blocking rate of the Q-learning algorithm is about 20% when the weight value is 1/3, and the blocking rate of the DQN algorithm is reduced to about 15%. When the blocking rate weight value is 1/2, the blocking rate of the DQN algorithm is about 12%. In this case, users prioritize improving the system's co-frequency interference by reducing power instead of pursuing high data rates. It leads to the improved channel quality, allowing more users to access the system and reducing congestion.

As shown in Fig. 7, the spectral efficiency gradually increases as the number of users increases. When the number of users reaches 100, the spectral efficiency of Q-learning algorithm is higher than that of DQN algorithm. However, after reaching 100, the rate of increase in spectral efficiency slows down. When the number of users reaches 125, more users can transmit services normally in the DQN algorithm compared with the Q-learning algorithm, and the spectral efficiency is relatively higher, approximately 268 Mbps/MHz.. When the frequency efficiency weight value is 1/4, the system requires users to reduce the power in order to pursue a lower

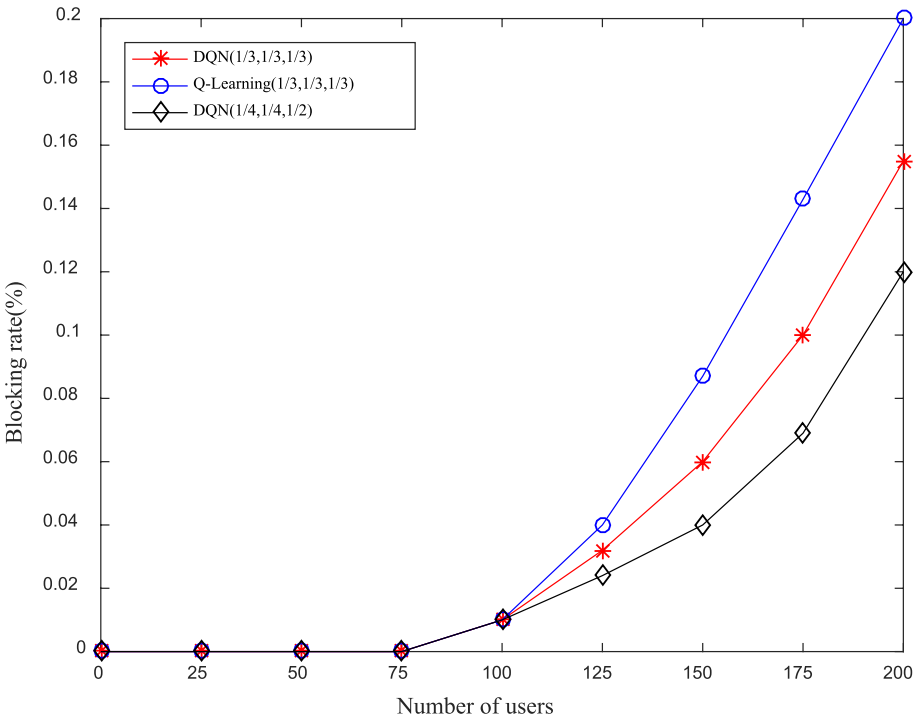

Fig. 6 The relationship between the blocking rate and the number of users

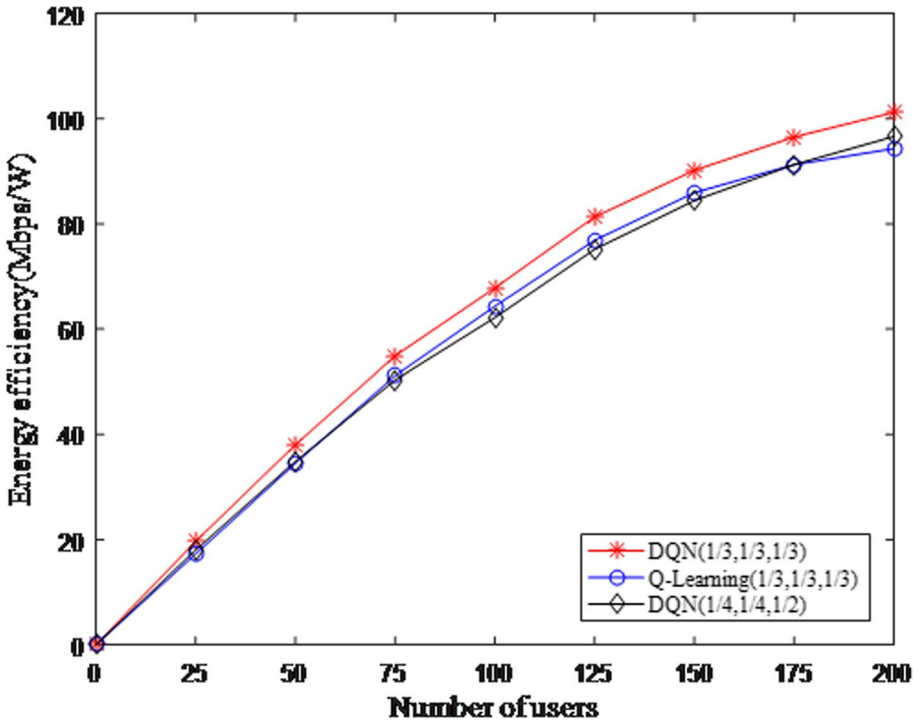

Fig. 7 Variation of channel frequency efficiency with the number of users

blocking rate. And reducing the transmission power will result in lower user rates with constant channel bandwidth. Compared to the 1/3 weight Q-learning algorithm, the frequency efficiency of the 1/4 weight DQN algorithm is lower when the number of early-stage users is not large. As the number of users increases, along with the increase in blocked users, the frequency efficiency of the DQN algorithm increases more significantly than that of Q-learning. When the number of users reaches 200, the frequency efficiency of the 1/4 weight DQN is about 350 Mbps/MHz, higher than the 345 Mbps/MHz of the 1/3 weight Q-learning algorithm.

Figure 8 illustrates the comparison of cumulative energy efficiency for the two algorithms under different weight values. The energy efficiency of the DQN algorithm is consistently higher than that of the Q-learning algorithm for a weight of 1/3. The DQN algorithm with a weight value of 1/3 shows a stronger preference for energy efficiency compared to a weight value of 1/4. When the number of users reaches 125, the energy efficiencies of the DQN algorithms with weights of 1/3 and 1/4 are approximately 82.5 Mbps/W and 75.6 Mbps/W, respectively, while the energy efficiency of the reinforcement learning algorithm with a weight of 1/3 is about 77.8 Mbps/W. In situations with low co-frequency interference, users can achieve high rates without the need for high transmission power. However, in scenarios with strong co-frequency interference, higher transmission power is required to achieve the same rate. Therefore, as the number of users increases from 125 to 200, due to the stronger co-channel interference within the system, the energy efficiency of new users decreases, and the growth in cumulative energy efficiency is somewhat slowed.

From Fig. 9, it can be observed that before reaching 125, the Q-learning algorithm has higher power consumption compared to the DQN algorithm. When the number of users is within the 50–75 range, the power consumed by the Q-learning algorithm is even about 20W more than that consumed by the DQN algorithm. When the number of users is in the range of 150–200, the interference level, within the satellite system, is more severe in the Q-learning algorithm than in the DQN algorithm. New users need to transmit high power to meet the service minimum requirements. High power will cause strong interference to other users, and the probability of blocking will be higher.

When the number of users reaches 200, although the DQN algorithm with a blocking rate weight of 1/2 consumes up to 638W, there are more users in the system who can transmit business normally.

To better highlight the performance of the proposed algorithm, the comparative experiment design involves state reconstruction to reduce data dimensionality, thus enhancing the performance of the reinforcement learning

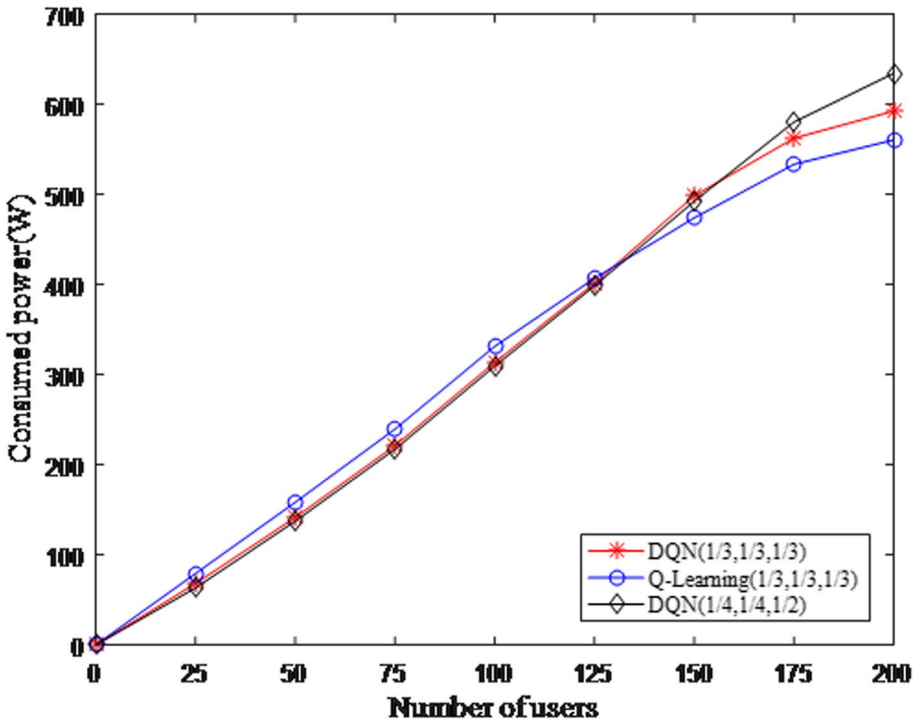

Fig. 8 The relationship between user energy efficiency and the number of users

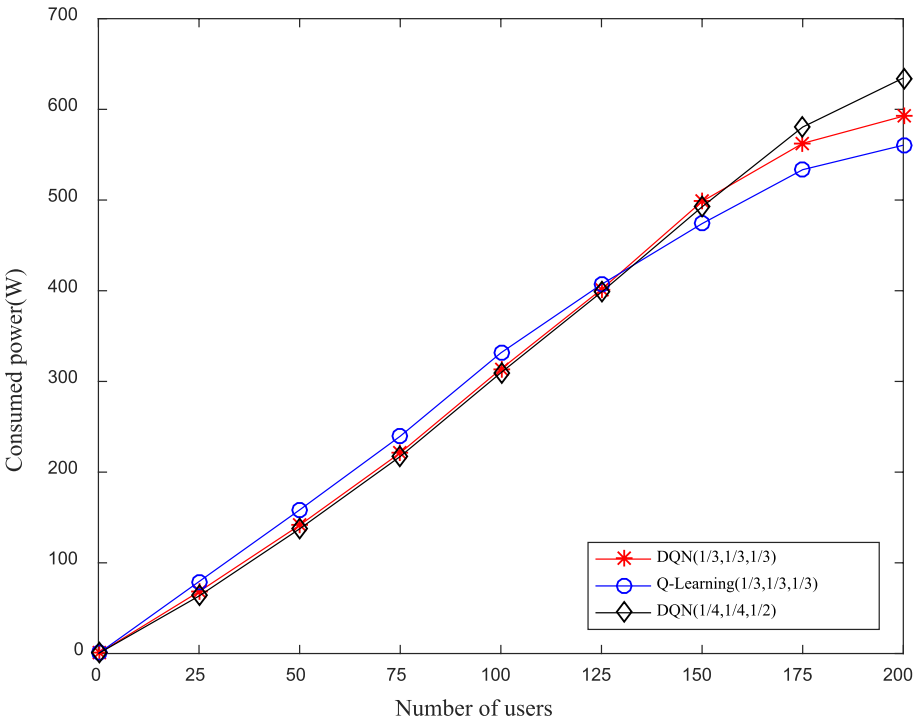

Fig. 9 Power consumption system capacity versus number of users

algorithm. The reinforcement learning considers co-channel interference within three layers of beams, while the proposed method considers not only the co-channel interference but also the traffic volume in the fourth layer of beams. The simulation shows that when the number of users is low, the results of the two methods are not significantly different. However, as the number of users gradually increases, the proposed method achieves better results. This is because when the number of users is low and there are sufficient resources available, the decision-making difference between the two methods is not significant. In scenarios with a higher number of users, the proposed method takes into account the traffic volume in addition to what is considered by reinforcement learning, thus being more conducive to maximizing long-term benefits. A lower blocking rate means that the system accommodates more users, thereby relatively improving resource utilization.

Conclusion

This paper addresses the issue of low resource utilization caused by limited onboard resources and uneven distribution of user traffic. It proposes a channel and power allocation strategy based on deep reinforcement learning. The LEO satellite is considered as the intelligent agent. The spot beams are treated as the system environment. The state information of channel allocation, the user requests, and user traffic is regarded as the environment state. Through the interaction between the satellite and beams, the system allocates appropriate channel and power to users, aiming to improve the resource utilization and reduce user blocking. In the reward mechanism, the maximum reward is obtained by maximizing a weighted sum of spectrum efficiency, energy efficiency, and blocking probability increment. Moreover, state integration is performed by merging beams with high user traffic and the current service beam to avoid bias towards current users while neglecting subsequent new users, in order to maximize long-term benefits. To validate the performance of the proposed approach, simulation experiments are designed to evaluate the effectiveness of the above strategy. The simulation results demonstrate that the resource allocation algorithm based on deep reinforcement learning can achieve lower user blocking rates and higher resource utilization as the number of users in the system gradually increases, compared to reinforcement learning. Due to the rapid movement of LEO satellites, which leads to a complex and dynamic network, the joint allocation strategy proposed in this paper can adapt to the complex and dynamic LEO satellite network system. When there are not many users in the system, resource utilization can be improved by increasing spectral efficiency and energy efficiency. When there are more users, the weight of the blocking

rate can be increased to accommodate as many users as possible, thus improving resource utilization. However, in the joint allocation strategy proposed in this paper, each user occupies a channel and can only use the resources of their own beam. When there are fewer users in the system, idle channels can be allocated to users within the beam to improve resource utilization. Additionally, users can use channels from adjacent beams for transmission, thereby enhancing signal strength and reducing transmission power. Therefore, subsequent work could explore inter-beam cooperative transmission strategies to achieve higher resource utilization.

Authors' contributions
Y.H found the shortcomings of previous studies and gave the instructive opinion. F.Q investigated the background and wrote the main manuscript text. F.Z developed the concept and supervised the entire work. J.Z validated and checked this work. All authors reviewed the manuscript.

Funding
This research was supported by Dean Project of Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education under Grant No. CRKL200107.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
The research has consent for Ethical Approval.

Competing interests
The authors declare no competing interests.

Received: 2 November 2023 Accepted: 28 February 2024
Published online: 08 March 2024

References

1. Ye N, Jihong Yu, Wang A, Zhang R (2022) Help from space: grant-free massive access for satellite-based IoT in the 6G era [J]. Digital Communications and Networks 8(2):215–224
2. Wang F, Li G, Wang Y, Rafique W, Khosravi MR, Liu G, Liu Y, Qi L (2022) Privacy-aware traffic flow prediction based on multi-party sensor data with zero trust in Smart City [J]. ACM Trans Internet Technol. https://doi.org/10.1145/3511904
3. Yang Y, Yang X, Heidari M, Srivastava G, Khosravi MR, Qi L (2022) ASTREAM: data-stream-driven scalable anomaly detection with accuracy guarantee in IIoT environment [J]. IEEE Trans Netw Sci Eng. https://doi.org/10.1109/TNSE.2022.3157730
4. Li G, Zijie Hong Yu, Pang YX, Huang Z (2022) Resource allocation for sum-rate maximization in NOMA-based generalized spatial modulation [J]. Digital Communications and Networks 8(6):1077–1084
5. Xie H, Yongjun Xu (2022) Robust Resource Allocation for NOMA-assisted Heterogeneous Networks [J]. Digital Communications and Networks 8(2):208–214
6. Hang L, Zhe Z, Zhen G, et al (2014) Dynamic Channel Assignment Scheme with Cooperative Beam Forming for Multi-beam mobile satellite networks [C]. 6th International Conference on Wireless Communications and Signal Processing (WCSP), IEEE, 1–5
7. Umehira M (2012) Centralized Dynamic Channel Assignment Schemes for Multi-beam Mobile Satellite Communications Systems [C]. AIAA International Communications Satellite System Conference (ICSSC), 24–27
8. Umehira M, Fujita S, Zhen G, et al (2014) Dynamic Channel Assignment Based on Interference Measurement with Threshold for Multi-beam Mobile Satellite Networks [C]. Communications
9. Chang R, He Y, Cui G, et al (2016) An allocation scheme between random access and DAMA channels for satellite networks [C]. IEEE International Conference on Communication Systems (ICCS), IEEE, 1–5
10. Choi JP, Chan VWS (2005) Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks [J]. IEEE Trans Wireless Commun 4(6):2983–2993
11. Lutz E (2015) Co-channel interference in high-throughput multi-beam satellite systems [C]. 2015 IEEE International Conference on Communications (ICC), IEEE, 885–891
12. Wang L, Zheng J, He C et al (2021) Resource allocation in high throughput multi-beam communication satellite systems [J]. Chin Space Sci Technol 41(05):85–94
13. Shi Y, Zhang BN, Guo DX et al (2018) Joint Power and Bandwidth Allocation Algorithm with Inter-beam Interference for Multi-beam Satellite [J]. Comput Eng 21:103–106
14. Zhou J, Ye X, Pan Y et al (2015) Dynamic channel reservation scheme based on priorities in LEO satellite systems [J]. J Syst Eng Electron 26(1):1–9
15. Zuo P, Peng T, Linghu W et al (2018) resource allocation for cognitive satellite communications downlink [J]. IEEE Access 6:75192–75205
16. Hong D, Zhang B, Li H et al (2023) Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. Remote Sens Environ 299:113856
17. Zengjing C, Wang Lu, Chengzhi X (2023) Efficient dynamic channel assignment through laser chaos: a multiuser parallel processing learning algorithm [J]. Sci Rep 13(1):1353
18. Jiang C, Zhang H, Ren Y et al (2017) Machine learning paradigms for next-generation wireless networks [J]. IEEE Wirel Commun 24(2):98–105
19. Chen X, Zhang H, Tao C, et al (2013) Improving Energy Efficiency in Green Femtocell Networks: A Hierarchical Reinforcement Learning Framework [C]. IEEE International Conference on Communications (ICC), IEEE, 2241–2245
20. Wang Z, Zhang J, X Zhang, et al (2019) Reinforcement Learning Based Congestion Control in Satellite Internet of Things [C]. 11th International Conference on Wireless Communications and Signal Processing (WCSP), IEEE1–6
21. Zhi Y, Tian J, Deng X, Qiao J, Dianjie Lu (2022) Deep reinforcement learning-based resource allocation for D2D communications in heterogeneous cellular networks [J]. Digital Communications and Networks 8(5):834–842
22. Liu X, Zheng J, Zhang M, Li Y, Wang R, He Y (2021) A novel D2D-MEC method for Rnhanced computation capability in cellular networks [J]. Sci Rep 11(1):16918
23. Qiu Y, Ji Z, Zhu Y, et al (2018) Joint Mode Selection and Power Adaptation for D2D Communication with Reinforcement Learning [C]. 15th International Symposium on Wireless Communication Systems (ISWCS), 1–6
24. Hong D, Zhang B, Li X et al (2023) SpectralGPT: Spectral Foundation Model
25. Hong D, Han Z, Yao J et al (2022) SpectralFormer: rethinking hyperspectral image classification with transformers. IEEE Trans Geosci Remote Sensing 60:1–15. https://doi.org/10.1109/TGRS.2021.3130716
26. Hu X, Liao X, Liu Z et al (2020) Multi-agent deep reinforcement learning-based flexible satellite payload for mobile terminals [J]. IEEE Trans Veh Technol 9:9849–9865
27. He Y, Sheng B, Yin H, Yan D, Zhang Y (2022) Multi-objective deep reinforcement learning based time-frequency resource allocation for multi-beam satellite communications [J]. China Communications 19(1):77–91
28. J. Li, J. Zhao, X. Sun (2021) Deep Reinforcement Learning Based Wireless Resource Allocation for V2X Communications [C]. 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), Changsha, China, 1–5
29. Y. Han, C. Zhang, G. Zhang (2021) Dynamic Beam Hopping Resource Allocation Algorithm Based on Deep Reinforcement Learning in Multi-Beam Satellite Systems [C]. 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 68–73
30. S. Ma, X. Hu, X. Liao, W. Wang (2021) Deep Reinforcement Learning for Dynamic Bandwidth Allocation in Multi-Beam Satellite Systems [C]. 2021

IEEE 6th International Conference on Computer and Communication Systems (ICCCS), Chengdu, China, 955–959
31. Xiaolong Xu, Jiang Q, Zhang P, Cao X, Khosravi MR, Alex LT, Qi L, Dou W (2022) Game theory for distributed IoV task offloading with fuzzy neural network in edge computing [J]. IEEE Trans Fuzzy Syst 30(11):4593–4604
32. Jia Y, Liu B, Dou W, Xiaolong Xu, Zhou X, Qi L, Yan Z (2022) CroApp: A CNN-based resource optimization approach in edge computing environment [J]. IEEE Trans Industr Inf 18(9):6300–6307

Publisher's Note