IEEE Transactions on Wireless Communications, Available online 29 December 2022
DOI:10.1109/TWC.2022.3231379

Multi-Agent DRL for Resource Allocation and Cache Design in Terrestrial-Satellite Networks

Xiaonan Li, Haijun Zhang, Huan Zhou, Ning Wang, Keping Long, Saba Al-Rubaye, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—In the past few years, satellite communications have greatly affected our daily lives, and the integrated terrestrial-satellite network can combine the advantages of satellite and base stations (BSs) to provide wider coverage and lower cost. Because the resources of terrestrial-satellite network are limited, how to allocate resources of terrestrial-satellite network through effective methods has become a major challenge. This paper proposes a framework for resource allocation of terrestrial-satellite network based on non-orthogonal multiple access (NOMA). Then, a deployment method of local cache pools is given to achieve lower time delay and maximize energy efficiency in terrestrial-satellite network. In the proposed framework, we adopt a multi-agent deep deterministic policy gradient (MADDPG) method to obtain the maximum energy efficiency by user association, power control, and cache design. The MADDPG algorithm is divided into two stages, users and BSs are set as agents to complete the optimization problem in the framework. Finally, the simulation results show that the proposed method has better optimized performance compared with the traditional single-agent deep reinforcement learning algorithm and can efficiently solve the problems of resource allocation and cache design in the integrated terrestrial-satellite network.**

***Index Terms*—MADDPG, energy efficiency, resource allocation, terrestrial-satellite network, NOMA.**

I. Introduction

With rapid increase of mobile data, the scarcity of spectrum resources has brought a series of new problems and challenges in wireless communications [1]. To solve these problems and challenges, non-orthogonal multiple access (NOMA) technology based on power domain multiplexing is a significant candidate in next generation wireless communication networks [2]-[3], where NOMA can improve the total energy efficiency of the system [4]-[9].

The integrated terrestrial-satellite network that consists of BSs on the ground and satellites in the space is an important scene in the 6G system. The NOMA technology is often applied in the integrated terrestrial-satellite network [10]-[14]. It is considered as a promising scenario and worthy of research.

In the terrestrial-satellite network [16]-[19], BSs provide low-cost communication services, while satellites can be used to cover and serve users who are in underdeveloped areas. This system can achieve a wider coverage area and better service quality. However, the resources of the integrated terrestrial-satellite network are limited. One of the main challenges is how to use effective methods for resource allocation and improve the system's energy efficiency. Deploying cache pools for the BSs in the system is a promising method to improve system energy efficiency, which can reduce the time delay and support the efficient files retrieval.

Many research on the resource allocation of integrated terrestrial-satellite network have been investigated. [17] proposed a method that uses precoding to optimize resource. The [20] investigated the placement of substance and delivery problem to optimize path length. The authors in [21] investigated the problem of cross-layer design of the link scheduling, frequency assignment, and flow control in hybrid terrestrial-satellite wireless backhauling networks. In [22], the authors proposed a convex relaxation approach to achieve power and flow assignment.

[20] proposed a joint beam endowments design of cognitive satellite ground network based on NOMA. By successive convex approximation (SCA), the nonconvex maximization problem is transformed into a corresponding convex problem that is easy to solve, so as to maximize the security rate of satellite users under imperfect channel state information. [13] studied a joint optimization design of satellite ground fusion network based on NOMA, and proposes a new resource allocation scheme. On the basis of user clustering, a beamforming algorithm based on iterative penalty function is proposed. The simulation results confirm the effectiveness of this method.

Although many papers used traditional methods in the integrated terrestrial-satellite networks to allocate resource, traditional optimization methods will be difficult to solve the problem in an unstable environment. On the one hand, the environment of integrated terrestrial-satellite is unstable and

This work is supported in part by the National Key R&D Program of China (2021YFB2900801), Beijing Natural Science Foundation (L212004), and China University Industry-University-Research Collaborative Innovation Fund (2021FNA05001). This paper was presented at the IEEE Global Communications Conference, Madrid, Spain, December 2021. (Corresponding author: Huan Zhou).

Xiaonan Li, Haijun Zhang, and Keping Long are with the Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services, University of Science and Technology Beijing, Beijing, 100083, China (e-mail: haijunzhang@ieee.org).

Huan Zhou is with the College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China. (e-mail: zhouhuan117@gmail.com).

Ning Wang is with Henan Joint International Research Laboratory of Intelligent Networking and Data Analysis, School of Information Engineering, Zhengzhou University (email: ienwang@zzu.edu.cn).

Saba Al-Rubaye is with the School of Aerospace, Transport and Manufacturing, Cranfield University, Bedford MK43 0AL, U.K. (e-mail: s.alrubaye@cranfield.ac.uk).

George K. Karagiannidis is with the Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece (e-mail: geokarag@auth.gr).

the user's demand for cache files is uncertain. On the other hand, many constraints are introduced in the optimization of this scenario. Sometimes it is difficult to find an accurate mathematical model to solve these optimization problems.

To solve the above problems, deep reinforcement learning (DRL) is introduced for resource allocation and cache design in system. DRL is an effective method in solving the optimization problems under uncertainty. In [23], the authors used deep Q-network (DQN) to achieve user access. [24] proposed a cooperative multi-agent deep reinforcement learning (CM-DRL) framework to achieve the radio resources management strategy. In [25], authors used a variety of deep reinforcement learning methods to achieve power control in cognitive radio scenarios. The authors discussed the integrated terrestrial-satellite network, and used deep reinforcement learning to achieve resource optimization issues such as throughput and bandwidth in [26]. In [27], DRL was used to achieve resource allocation in a multibeam satellite system. In [28], the authors used multi-objective DRL to process cognitive satellite scenarios. In [29], the authors used DRL to achieve task scheduling.

DRL is also used in many cache design optimization problems. In [30], the authors used actor-critic frameworks in edge caching scenarios. The authors in [31] used two-layer Q network to achieve a scheme called double coded caching. In [32], the authors decomposed the joint base station and user cache optimization problem into two subproblems, then they applied value function approximation Q-learning and DQN to solve these two subproblems. In [33], the authors proposed a DRL-based algorithm, which can optimize the user association, power allocation of NOMA, deployment of unmanned aerial vehicle (UAV) and caching placement of UAVs to jointly to minimize the content delivery delay. The [34] proposed a Q-learning based caching placement and resource allocation algorithm.

The resource allocation and cache design problem in the above works for satellite scenarios are achieved by traditional DRL. The traditional DRL is the single-agent algorithm, so it cannot deal with an unstable environment when there are many agents in the scenarios. When the number of agents increases, the unstable and dynamic environment will reduce the optimization performance. At present, the research on resource allocation and cache design of integrated terrestrial-satellite network by using multi-agent reinforcement learning, is rarely investigated.

A preliminary investigation on this research problem was published in [35], and this work extends [35] in the following ways: (1) the cache design is for integrated terrestrial-satellite is now considered; (2) the users and BSs, satellites are set as agents to complete the optimization problem in the optimization framework; (3) simulation results under multiple angles are provided to verify the proposed methods. In this paper, we consider an integrated terrestrial-satellite network based on NOMA and use a multi-agent deep deterministic policy gradient method (MADDPG) to achieve user association, power control, and cache design to improve the system energy efficiency [36].

The main contributions of this paper are summarized as follows.

- We propose a cache-enabling general downlink framework for NOMA integrated terrestrial-satellite network, the users in the network are served by the BSs and satellites. The cache design is introduced into the integrated terrestrial-satellite to deploy cache equipment for BSs and satellites.
- We formulate an optimization problem to maximize the energy efficiency by dynamically optimizing the user association, power control and caching placement of BSs and satellites
- We decompose the original optimization problem of energy efficiency into two stages: resource allocation and cache design. In order to solve these two sub problems, a novel and efficient multi-agent deep reinforcement learning algorithm is used in the paper. The users and BSs, satellites are set as agents to complete the optimization problem in the optimization framework. In the framework, the user association and power control scheme based on MADDPG is first proposed. The users are set as the agents to choose the BSs or satellites and the power control factor, which has achieved the objective of optimizing resource allocation. Then, the cache design plan based on MADDPG is proposed. BSs and satellites are set as the agents to select files cache from files library to local cache pool to improve energy efficiency.
- We demonstrate the performance of the proposed MADDPG optimization framework to optimize the user association, power control and caching placement by comparing with the benchmark algorithms, the proposed algorithms in this paper achieve a good optimization performance.

The structure of this paper is as follows. In Section II, the system model and problem formulation are presented. In Section III, the MADDPG algorithm is introduced to solve the formulated problems. The simulation results are given in Section IV. The work is concluded in Section V.

II. System Model and Problem Formulation

A. System Model

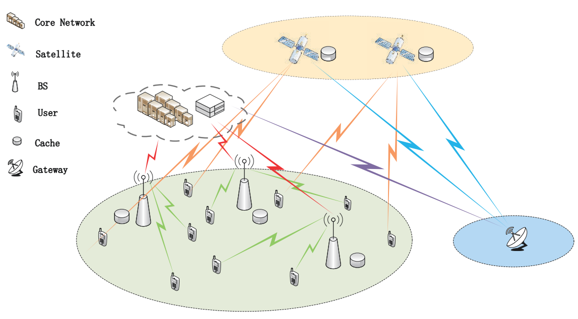

Fig. 1. The integrated terrestrial-satellite network.

Fig. 1 shows the integrated terrestrial-satellite network which consists of N Base Stations (BSs) and L low-orbit satellites. In this network, N BSs and L satellites jointly provide services for the ground users. Let B represent the set of BS, where $B = \{B_1, ...B_N\}$. The satellite set is represented by $S = \{S_1, ..., S_L\}$. The M users set is represented by $U = \{U_1, ...U_M\}$. There are K users served by ground BS, the set of them is $U_{BS} = \{U_1, ..., U_K\}$. The remaining O users are served by satellites, the set of satellite users is $U_{SA} = \{U_1, ..., U_O\}$.

NOMA scheme is implemented for the users associated with BS. In the NOMA systems, successive interference cancelation (SIC) can be used to reduce interference from other users, the superposition coding is used at the transmitter and SIC [37] technology can be used to perform user detection, correct demodulation, and interference cancellation in a certain order, which is the core concept of the NOMA. The multiple users associating with the same BS are regarded as a NOMA cluster. In the NOMA cluster, according to the best decoding order, the users with a better channel information state can reduce the interference from users who have a poor channel information state.

In each time slot, M users can only associates with one BS or one satellite in this system. Let $a_m^n(t)$ represents association situation between the mth user and nth BS, when the mth user associate with the nth BS, the value of $a_m^n(t)$ is 1, otherwise set the value of $a_m^n(t)$ to 0. In addition, $a_m^l(t)$ represents association situation between the mth user and the lth satellite, where the way to assign the value of $a_m^l(t)$ is the same to $a_m^n(t)$. The users can only associate with one BS or one satellite in the time slot.

In the system model, the NOMA is implemented for users served by BS. Then, the signal to interference plus noise ratio (SINR) of Mth user served by BS in a time slot t can be represented as

$$SINR_{Bm}(t) = \frac{a_m^n(t)|h_m^n(t)|^2 p_m(t)}{\sigma_{BI}(t) + \sigma_{BO}(t) + \sigma_S(t) + N_0}, \quad (1)$$

where $h_m^n(t)$ is the channel information state between the mth user associating with the nth BS. $p_m(t) = \alpha_m(t)\frac{p_{bmax}}{M_1}$, and $p_m(t)$ is the transmit power for user m, where $\alpha_m(t)$ is the power control factor of mth user, and one BS can serve M_1 users in network. $\sigma_{BI}(t)$ is the interference from the users in the same BS. $\sigma_{BO}(t)$ is the interference from the users in other BSs. In addition, $\sigma_S(t)$ is the interference from satellite users. N_0 is the additive white gaussian noise (AWGN) power.

To compute $\sigma_{BI}(t)$ which is caused by the users in the same cluster, we first sort the users according to the channel gain in a BS as follows $\left|h_1^n(t)\right| \geq ... \geq \left|h_m^n(t)\right| ... \geq \left|h_{M_1}^n(t)\right|$.

According to the order of channel gain, $\sigma_{BI}(t)$ is the interference from the users who have better channel condition. Therefore, the interference from the same cluster is presented as $\sigma_{BI}(t) = \sum\limits_{i=1}^{m-1} a_i^n(t)|h_i^n(t)|^2 p_i(t)$, the interference from users served by other BSs is $\sigma_{BO}(t) = \sum\limits_{j=1,j\neq n}^{N} \sum\limits_{i=1}^{M_1} a_i^j(t)|h_i^n(t)|^2 p_i(t)$, the interference from satellite users is $\sigma_S(t) = \sum\limits_{j=1}^{L} \sum\limits_{i=1}^{M_2} a_i^j(t)|g_i^n(t)|^2 p_{s,i}(t)$, where a satellite can serve M_2 users in network. $p_{s,i}(t)$ is the transmission power of satellite user.

The SINR of Mth user served by satellite is

$$SINR_{Sm}(t) = \frac{a_m^l(t)\left|g_m^l(t)\right|^2 p_{s,m}(t)}{\sigma_B(t) + \sigma_{SO}(t) + N_0}, \quad (2)$$

where $g_m^l(t)$ is the channel information state between the mth user associating with the lth satellite. And $p_{s,m}(t) = \alpha_m(t)\frac{p_{smax}}{M_2}$ is the power of satellite user. The interference from BS users and other satellite users are, respectively, $\sigma_B(t) = \sum\limits_{j=1}^{N} \sum\limits_{i=1}^{M_1} a_i^j(t)\left|h_i^l(t)\right|^2 p_{s,i}(t)$ and $\sigma_{SO}(t) = \sum\limits_{i=1,i\neq I}^{M_2} a_i^l(t)\left|g_i^l(t)\right|^2 p_{s,i}(t) + \sum\limits_{j=1,j\neq l}^{L} \sum\limits_{i=1}^{M_2} a_i^j(t)\left|g_i^l(t)\right|^2 p_{s,i}(t)$.

The energy efficiency of the mth user in the time slot t is

$$\begin{aligned} EE_m(t) = & \sum\limits_{n=1}^{N} a_m^n(t)\tfrac{\log_2(1+SINR_{Bm}(t))}{p_m(t)} \\ & + \sum\limits_{l=1}^{L} a_m^l(t)\tfrac{\log_2(1+SINR_{Sm}(t))}{p_{s,m}(t)}, \forall n \in [1, N], \forall l \in [1, L]. \end{aligned} \quad (3)$$

The cache design in the system model is described as follows. Each user individually requests files from a file library $F = \{1, ..., F\}$. Each BS and satellite is configured with a cache pool to store the files. Therefore, there are $N + L$ cache pools in the system. The size of BS cache pool is set as $N_f < F$. Each BS selects N_f files from the file library F as the combination of cache files. Each BS can store $N_f{*}s$ bit files. The size of satellite cache pool is set as $N_s < F$. Each satellite selects N_s files from the file library F as the combination of cache files. Each satellite can store $N_s{*}s$ bit files.

When the user's request arrives, the system first searches the cache files in local cache pool deployed in the BS. If the local cache pool has the files that the user needs, the transmission between the user and the local BS will occur. The file is sent back to the user from the local BS, and the power consumed is $p_{m,r}(t)$. If the BS can not meet the cache file required by local users, users will search the required files in the core network using the return link. The power consumed at this time is $p_{c,r}(t)$.

For satellite users, when the user's request arrives, if the file requested by the user has been cached by the satellite, the user can directly obtain the file from the satellite without accessing the backhaul link. If the satellite is not equipped with the file required by the user, the user's request will be forwarded to the ground gateway and access the ground gateway through the backhaul link, and download content from the core network.

There are two ways to consider the caching gain in the system, where one is the reduction of the time delay, and the other is the alleviation of power consumption. The both rewards depend on whether the files request of user is satisfied by the local cache device.

The variable $I_m(t)$ can be used to present whether the file request of the mth BS user is satisfied by the local cache device:

$$I_m(t) = \begin{cases} 1, & \text{the requests are satisfied,} \\ 0, & \text{the requests are not satisfied.} \end{cases} \quad (4)$$

It is assumed that the popularity distribution of files follows ZipF distribution [38]. The popularity will influence the caching effect. Frequently, the popularity in our system can be follows a generalized ZipF distribution, and yield estimates for ε between 0.56 and 0.83 [39].

$$q_m = \frac{1/f^{\varepsilon}}{\sum_{f=1}^{F} 1/f^{\varepsilon}}, \forall f. \quad (5)$$

Considering the reduction of the time delay, the reward of caching deployment is given by

$$g_m(t) = I_m(t)\frac{count_m s}{T_m}, \quad (6)$$

where T_m is the time delay of downloading the content requested by mth user through the backhaul link, s is the size of file, $count_m$ is the number of the content requested by mth user, This part of file can be directly obtained from the local cache.

Similarly, considering the reduction of time delay and the transmission of files that hit the cache part, the benefit of satellite cache deployment is

$$g_{s,m}(t) = I_m(t)\frac{count_m s}{T_{s,m}}, \quad (7)$$

where $T_{s,m}$ is the time delay of downloading the content requested by mth user through the backhaul link.

The cache hit rate is defined as the proportion of users whose requests are satisfied in the system to measure the performance of the cache policy. The cache hit rate in the time slot t is

$$Hit(t) = \frac{\sum\limits_{m=1}^{M} I_m(t)}{M}. \quad (8)$$

For convenience, we use $P(t)$ to replace the total power consumption of BS users

$$P(t) = p_m(t)+(1-I_m(t))p_{c,r}(t) + I_m(t)p_{m,r}(t), \quad (9)$$

where $p_m(t)$ is the transmit power for user m, $p_{m,r}(t)$ is the data retrieval power consumption of the content requested by mth user from local BS cache device, $p_{c,r}(t)$ is the data retrieval power consumption of the content requested by mth user from core network through backhaul link.

For satellite users, we use $P_s(t)$ to replace the total power consumption of satellite users

$$P_s(t) = p_{s,m}(t)+(1-I_m(t))p_{s,c,r}(t) + I_m(t)p_{s,m,r}(t), \quad (10)$$

where $p_{s,m}(t)$ is the transmit power for user m, $p_{s,m,r}(t)$ is the data retrieval power consumption of the content requested by mth user from satellite cache device, $p_{s,c,r}(t)$ is the data retrieval power consumption of the content requested by mth user from core network through Gateway Station.

After combining base station cache with the above satellite model, the energy efficiency of the mth user in the time slot t is

$$\begin{aligned} EE_m(t) &= \sum_{n=1}^{N} a_m^n(t)\frac{\log_2(1+SINR_{Bm}(t))+g_m(t)}{P(t)} \\ &+ \sum_{l=1}^{L} a_m^l(t)\frac{\log_2(1+SINR_{Sm}(t))}{p_{s,m}(t)} \\ &= \sum_{n=1}^{N} a_m^n(t)\frac{\log_2(1+SINR_{Bm}(t))+g_m(t)}{p_m(t)+(1-I_m(t))p_{c,r}(t)+I_m(t)p_{m,r}(t)} \\ &+ \sum_{l=1}^{L} a_m^l(t)\frac{\log_2(1+SINR_{Sm}(t))}{p_{s,m}(t)}. \end{aligned} \quad (11)$$

B. Problem Formulation

The system model is introduced in last section, the optimization problem is formulated in this section. The objective of optimization problem is to maximize the total energy efficiency of all agents in the system through user association, power control and cache design.

The constraints of optimization problem are introduced as follow. Firstly, the users can only associate with one BS or one satellite in a time slot.

For each user associated to one BS or one satellite, it has its own maximum power constraint

$$p_m(t) \le \frac{p_{max}}{M_1}, \quad (12)$$

$$p_{s,m}(t) \le \frac{p_{s,max}}{M_2}, \quad (13)$$

where (12) and (13) describe the transmission power limits of BS users and satellite users respectively.

For each BS or satellite, it has the quantity of service constraints

$$\sum_{m=1}^{M_1} a_m^N(t) \le M_1, \forall m \in [1, M_1], \quad (14)$$

$$\sum_{m=1}^{M_2} a_m^L(t) \le M_2, \forall m \in [1, M_2]. \quad (15)$$

The constraint of the power control factor of the users is

$$\alpha_m(t) \in [0, 1], \forall m \in [1, M], \quad (16)$$

where (16) represents the range of user power control, the power is selected and distributed in the range.

The caching strategy used by the BS and satellite is limited by the size of the local cache capacity. The size of content requested by users is smaller than that of the local capacity. The size of local capacity is smaller than that of all file libraries.

$$count_m \le N_f \le F, \quad (17)$$

$$count_m \le N_s \le F. \quad (18)$$

The optimization problem can be formulated as

$$\max \sum_{m=1}^{M} EE_m(t). \quad (19)$$

$$
\begin{aligned}
&C1: \sum_{n=1}^{N} a_m^n(t) + \sum_{l=1}^{L} a_m^l(t) \le 1, \forall n \in [1,N], \forall l \in [1,L],\\
&C2: \sum_{m=1}^{M_1} a_m^N(t) \le M_1, \forall m \in [1,M_1],\\
&C3: \sum_{m=1}^{M_2} a_m^L(t) \le M_2, \forall m \in [1,M_2],\\
&C4: p_m(t) \le \tfrac{p_{max}}{M_1}, \forall m \in [1,M],\\
&C5: p_{s,m}(t) \le \tfrac{p_{smax}}{M_2}, \forall m \in [1,M], \forall s \in [1,S],\\
&C6: \alpha_m(t) \in (0,1], \forall m \in [1,M],\\
&C7: count_m \le N_f \le F,\\
&C8: count_m \le N_s \le F.
\end{aligned}
\tag{20}
$$

The energy efficiency optimization problem in this paper has eight constraints. $C1$ represents that the users in this system can only associate with one BS or one satellite in a time slot t. $C2$ and $C3$ the quantity of service constraints of each BS and satellite. $C4$ and $C5$ represent power limit of the mth user. $C6$ is the constraint of the power control factor. $C7$ and $C8$ is the constraint of the size of local cache capacity.

III. Multi-Agent DRL for Resource Allocation and Cache Design in Terrestrial-Satellite Network

In this section, we will introduce a MADDPG method to solve the optimization problem. The MADDPG framework will be introduced to maximize objective function and in the integrated terrestrial-satellite NOMA communication network. The optimization process contains two parts: user association and power control, and then cache design. Two algorithms based on MADDPG are proposed to solve these two problems. Different agents are selected skillfully in the two algorithms.

A. Reinforcement Learning

Reinforcement learning (RL) does not require a data set which receives reward information from the environment in each eposide, learns and then updates the parameters of the model. The agents in RL can interact with the environment and observe the reward of actions, and then learn how to change their actions to obtain higher reward. The agent is constantly making progress in a trial and error manner.

B. MADDPG Framework Formulation

In this integrated terrestrial-satellite NOMA communication network scenario, there are many agents in the environment. When the number of agents is increasing, traditional single-agent reinforcement learning will face an unstable and dynamic environment, it will lead the agent to overfit a strong policy against its competitor. The proposed MADDPG algorithm can deal with the complex multi-agent scenario which can better adapt to the complex multi-agent scenario and achieve better optimization performance.

The energy efficiency optimization problem of the integrated terrestrial-satellite network can be modeled as a Markov decision process (MDP). The MDP is composed of a state space S, an action space A, a reward space and a transition probability space. As an agent, each user can observe the environment and get the observation, then select the actions from the action space and execute them. Next, it will get a reward after executing the actions. In this paper, the agent, action, state and reward of two algorithms are defined as follows:

1) MADDPG for User Association and Power Control: The MADDPG for this research problem is presented as algorithm 1, which is the user association and power control scheme in [35]. In algorithm 1, the agent, action, state and reward are defined as follows:

Agent: Each user in the terrestrial-satellite network is considered as an agent.

Action: In the system, each agent has two actions to execute. The action space is composed of two actions, A_1=$\{A_{11}, A_{12}\}$ is the user association action and the associate situation between the agents and BSs or satellites. A_{11} is defined as: A_{11}=$\{a_1^n(t), ..., a_M^n(t)\}$. A_{12} is the power control factor, A_{12}=$\{\alpha_1(t), ..., \alpha_M(t)\}$. Firstly, each agent decides which BSs or satellites to associate with. The user association A_{11} is a discrete action, hance we need to discretize the action space A_{11}. Secondly, it determines its own power control factor.

Reward: In the system, each user executes actions to maximize its energy efficiency. The reward of mth user in the current time slot t is

$$reward_1(t) = EE_m(t). \tag{21}$$

State: The agent observes the change of its own energy efficiency as the state space. If the energy efficiency of mth user in time slot is higher than previous time slot, the $S_{1m}^{EE}(t)$ is set as 1. The state space of system is S_1=$\{S_{11}^{EE}(t), ..., S_{1M}^{EE}(t)\}$.

$$S_{1i}^{EE}(t) = \begin{cases} 1, & \text{if } reward_1(t) \ge reward_1(t-1) \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

where $i \in [1, M]$ is the mth user.

2) MADDPG For Cache Design: The algorithm 1 is used to optimize the resource allocation in the network. Then algorithm 2 is used to optimize the cache design of the BSs and satellites. Therefore, the agents, actions, states and rewards of the two algorithms are slightly different.

Agent: The BSs or satellites selects which files from the files library, and the BS or satellite is considered as an agent in Algorithm 2.

Action: In every time slot, each BS or satellite selects which files from the files library. A_2=$\{A_{21}\}$. The files in the local cache pool is the combinations of the file libraries.

Reward: In the system, each BS or satellite executes actions to maximize its energy efficiency. The number of agents is $N + L$. The total energy efficiency of users served by nth BS or satellite is set as reward. The reward of nth BS in the current time slot t is

$$reward_2(t) = EE_n(t) = \left\{ \begin{array}{ll} \sum\limits_{m=1}^{K} EE_m(t), & n \in [1,N], \\ \sum\limits_{m=1}^{O} EE_m(t), & n \in [N+1, N+L] \end{array} \right. \tag{23}$$

$\boldsymbol{State}$: The agent in algorithm 2 is BS or satellite. Therefore if the reward of nth BS or satellite in this current time slot is higher than the previous time slot, the $S_{2n}^{EE}(t)$ is set to 1. The state space of system is S_2=$\{S_{21}^{EE}(t), ..., S_{2N}^{EE}(t)\}$. The $S_{2i}^{EE}(t)$ is set as

$$S_{2i}^{EE}(t) = \left\{ \begin{array}{ll} 1, & \text{if } reward_2(t) \geq reward_2(t-1), \\ 0, & \text{otherwise.} \end{array} \right. \tag{24}$$

MADDPG algorithm can obtain the actions executed by other agents to reduce instability. The transition probability is presented as follow

$$\begin{array}{l} P(s^{'}|s, a_1, ...a_N, \chi_1,...,\chi_M) = P(s^{'}|s, a_1, ...a_N) \\ = P(s^{'}|s, a_1, ...a_N, \chi_1^{'},...,\chi_M^{'}). \end{array} \tag{25}$$

As shown by the state transition probability in (25) , when the policy of agents is dynamically changed and updated, the environment is still stable. In the formula (25), $a_i, \forall i \in [1, M]$ is the action of the agent, s is the state. There are M agents in the network, and the policy values of all M agents in the system are set as χ= $\{\chi_1$,...,$\chi_M\}$. The policy of each agent has its corresponding parameter value ϖ= $\{\varpi_1$,...,$\varpi_M\}$.

MADDPG is presented to solve the integrated terrestrial-satellite NOMA communication network optimization problem. In the MADDPG algorithm, each agent aims to obtain the maximum return by optimizing its policy. The gradient of the objective function can be solved by the following equation.

$$\nabla_{\varpi_i} J(\chi_i) = E_{x,a \sim D}[\nabla_{\varpi_i} \chi_i(a_i|o_i) \nabla_{a_i} Q_i^{\chi}(x, a_1, ..., a_N)], \tag{26}$$

where x, a are respectively the observation space and action space of M agents, D is the replay memory.

The $actor$ network and $critic$ network play different roles in the MADDPG algorithm. The $actor$ network will select actions according to policy. The actions A_1 or A_2 are selected according to the strategy value, which the action space is continuous. The $critic$ network evaluates the actions that will be executed. The way to evaluate the actions is to update the Q function. As is shown in the (26), the Q function to evaluate the actions is $Q_i^{\chi}(x, a_1, ..., a_N)$. The $actor$ and $critic$ network update in different methods in the MADDPG algorithm. The $actor$ network updates the policy network for selecting actions through gradient descent in formula (26) [40]. The $critic$ network update the Q function that evaluates the actions selected by $actor$ network to minimize the loss function $L(\varpi_i)$ below

$$y = r_i + rQ_i^{u^{'}}(x^{'}, a_1^{'}, ..., a_N^{'})|_{a_j^{'}=u_j^{'}(o_j)}. \tag{27}$$

C. Algorithm Description

We describe The algorithm 1 and algorithm 2 in this section.

In this section, the MADDPG algorithm for system resource allocation is presented as algorithm 1 [35]. The MADDPG algorithm for system cache design problem is presented as algorithm 2.

Algorithm 1 MADDPG algorithm for terrestrial-satellite network resource allocation problem

```
Input: The parameters of deep neural networks and the replay memory.
for $episode = 1$ to $Ep$ do
    Initialize the observation of the terrestrial-satellite network, including user association and power control.
    for $agent = 1$ to $N$ do
        for $step = 1$ to $St$ do
            Each BS or satellite gets the observation state $S_1$=$\{S_{11}^{EE}(t), ..., S_{1N}^{EE}(t)\}$.
            Each BS or satellite selects the user association and power control from $A_1$=$\{A_{11}, A_{12}\}$.
            Each BS or satellite observes reward via (21).
        end for
    end for
    Sample a random batch from replay memory.
    Each agent update $actor$ network and $critic$ network.
    Update the parameters of the target network.
end for
Output: The parameters of the trained deep neural networks and user association and power control.
```

Algorithm 2 MADDPG algorithm for terrestrial-satellite network cache design problem

```
Input: The parameters of deep neural networks and the optimized user association and power control.
for $episode = 1$ to $Ep$ do
    Use the optimization result of algorithm 1 to initialize the integrated terrestrial-satellite network scenario.
    for $agent = 1$ to $N$ do
        for $step = 1$ to $St$ do
            Each BS or satellite gets the observation state $S_2$=$\{S_{21}^{EE}(t), ..., S_{2N}^{EE}(t)\}$.
            Each BS or satellite selects the cache files from the file library $A_2$=$\{A_{21}\}$ based on the optimization result of algorithm 1.
            Each BS or satellite observes reward referring to (23) and next state.
        end for
        Each agent update $actor$ network and $critic$ network. Update the parameters of the target network.
    end for
end for
Output: The optimization of cache design and renewed energy efficiency.
```

Firstly, the algorithm initializes the parameters of neural networks. At the same time, initializes the replay memory. The $actor$ network selects behavior based on the probability, the $critic$ network evaluates the behavior selected by the $actor$ network. And the $actor$ changes the probability based on the evaluation of the $critic$ network. Secondly, the algorithm gives the initial state of the agents in the iterative process of the MADDPG algorithm for the terrestrial-satellite network. Next, for each step in an episode, each agent observes its new state which deeps the energy efficiency compare with it in last moment. Then agent selects action based on exploration and policy. After each agent has executed the action, it obtains the reward of this action and gets the new state. Finally, store the above values in the replay memory.

IV. Simulation Results

A. Simulation Environment

In this section, we set the experimental environment such as experimental parameters and the hyperparameters of algorithm 1 and algorithm 2. Some simulation results are given to present the convergence performance of the MADDPG framework and the result compared with the traditional deep reinforcement learning algorithm.

The network consists of 32 agents, 6 BSs, and 2 satellites. We set $M_1 = M_2 = 4$. The channel of BS is assumed to be Rayleigh channel. The parameters of satellites in this paper are defined according to [41]. The carrier frequency is set at the S band. The maximum transmit power of BSs is set to 31 dBm, and the maximum transmit power of satellites is set to 43 dBm.

The cache related simulation environment are as follows, the size of the file library F is 40, the size of BS cache device N_f is 3, and the size of satellite cache capacity N_s is also set to 3. The number of files required by the users $count_m$ is set to 1, and the file content size s is set to 2 bits.

About the power of the transmission files. The data retrieval powers consumption of the content requested by users from local BS cache device $p_{m,r}(t)$ is set to 13 dBm. The data retrieval powers consumption of the content requested by users from core network through backhaul link $p_{c,r}(t)$ are set to 26 dBm. The power consumed by the user to request files from the satellite cache $p_{s,m,r}(t)$ is set to 17 dBm. The power consumed by the satellite to access the ground gateway through backhaul link and download content from the core network $p_{s,c,r}(t)$ is set to 30 dBm. Moreover, the cache design of Algorithm 2 optimizes the result of Algorithm 1.

Here are the hyperparameters of the model in this simulation. The optimizer is AdamOptimizer and activation function is $ReLU$. The learning rate of neural networks is $alr = clr =$ 0.001. The discount factor is 0.95. The batch size is set as 10. The total episode of the experiment is set to 1000. In each episode, the agent needs to complete 100 steps.

B. Simulation Results

When the number of agents is set to 24, 32 and 40 respectively, optimization effect of algorithm 1 is presented in Fig. 2. When the number of agents is 24, the parameters in the network are set as $M = 24$, $N = 6$, $S = 2$, $M_1 = M_2 = 3$. To test the convergence of algorithm 1 with the different number of agents in a more complex environment, we increase the M_1 and M_2 from 3 to 5. It can be seen that the three curves in Fig. 2 roughly reach the maximum reward of around 700 episodes, so the good convergence performance can be obtained in all the three cases. As can be see from the Fig. 2, when there are many agents, the terrestrial-satellite network becomes more complex, the algorithm 1 can still converge well. The observation demonstrates that algorithm 1 can optimize objective function in the system.

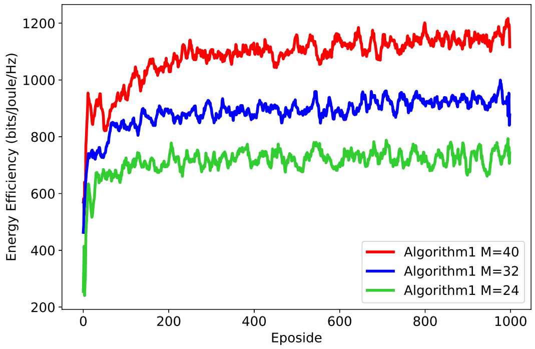

Fig. 2. Convergence of MADDPG in different numbers of agents.

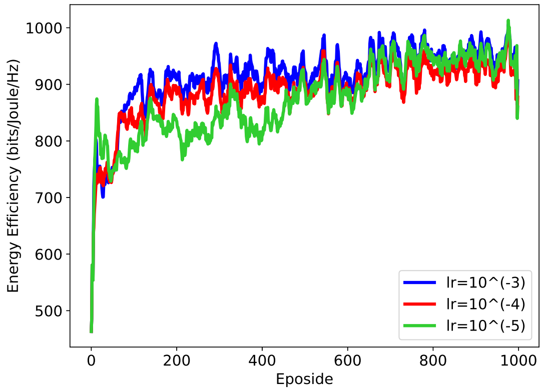

Fig. 3. Convergence of MADDPG in different learning rate.

Fig. 3 presents the results of algorithm 1 when using different learning rate. We set $M = 32, N = 6, S = 2, M_1 = M_2 = 4$ in the network. As the algorithm adopting different learning rates, the speed of convergence is slightly different. When the learning rate is relatively large, the convergence point arrives faster. Besides, the curves of three different learning rates converge to a similar height and get a similar reward.

The two subgraphs of Fig. 4 respectively show the convergence of BSs and the satellites. We use the same experimental settings as the two figures above. The two parts converge at roughly the same speed and they both converge very quickly.

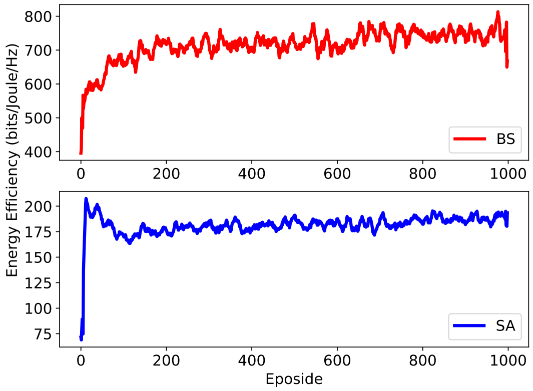

Fig. 4. Energy efficiency convergence of BS users and satellite users.

Fig. 4 presents the energy efficiency of users served by BSs is higher than that of the satellite users. The total energy efficiency of BSs users converges to about 750 bits/Joule/Hz, and the total energy efficiency of satellites users converges to about 175 bits/Joule/Hz. The reason is that the users served by BSs have better channel conditions than the satellite users.

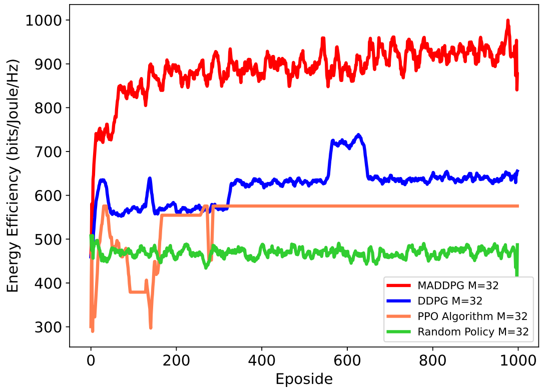

Fig. 5. Comparison of algorithm 1 and benchmark algorithms.

In order to verify the optimization performance of algorithm 1 proposed in this paper, the integrated satellite network resource optimization algorithm1 based on MADDPG proposed in this paper and three benchmark algorithms are listed as follows:
(1) Deep Deterministic Policy Gradient (DDPG) algorithm: DDPG algorithm is introduced to compare with algorithm 1 in Fig. 5, as a traditional deep reinforcement learning algorithm, DDPG algorithm is the baseline algorithm among Policy Gradient (PG) algorithms. (2) Proximal Policy Optimization (PPO) algorithm: PPO algorithm is a new policy gradient algorithm. (3) Genetic algorithm (GA): GA is a algorithm to search the optimal solution by simulating the natural evolution process. (4) Random Policy: In each episode, the user selects a random action value to determine the actions of user collaboration and power control.

In Fig. 5, the curve shows the comparison of convergence processes of different algorithms when the number of users is $M = 32$. As can be seen in the Fig. 5, the rewards of the proposed algorithm 1, DDPG algorithm and PPO algorithm can reach convergence. For the same number of users, the energy efficiency of algorithm1 is the highest among the three algorithms, which can get the best resource optimization performance and better optimize the objective function of the system. The energy efficiency optimized by DDPG algorithm converges to 625 bits/Joule/Hz. PPO algorithm starts from 450bits/Joule/Hz and converges to 580 bits/Joule/Hz after about 400 episodes. The curve of random policy oscillates between 400 and 500 bits/Joule/Hz, which can not converge well like other algorithms. Compared with the proposed algorithm 1, the effect of the optimization objective function of other algorithms is poor and the stability is not strong.

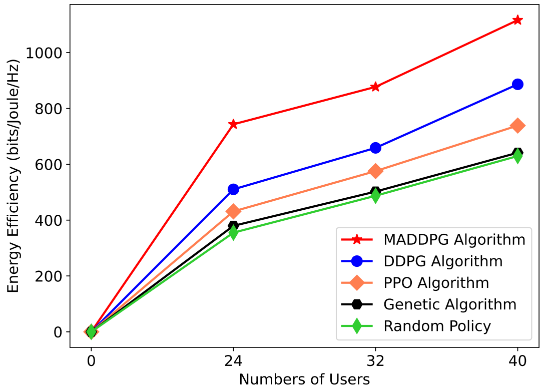

Fig. 6. The energy efficiency of different algorithms varies with the number of users in the system.

In Fig. 6, the total energy efficiency of different algorithms varies with the number of users per BS and satellite. As can be seen in the Fig. 6, the total energy efficiency of the system will increase with the increase of the number of users of each BS and satellite. When the number of users of each BS and satellite increases from 3 to 5, the proposed algorithm 1 can achieve higher energy efficiency than other benchmark algorithms. When the number of users of each BS and satellite is 5, the total energy efficiency of the system can reach 1180 bits/Joule/Hz. It can be seen that the total energy efficiency of DDPG algorithm, PPO algorithm, GA algorithm and random policy is much lower than that of the algorithm1 proposed in this paper, which fully shows the performance of the proposed algorithm 1.

Fig. 7 presents the convergence of algorithm 2 in different local capacity and file library. Although the number of training episode is set as 1000, the convergence speed of algorithm 2 is fast. It reach convergence in about 60 episodes. In Fig. 7, four cases of $N_f = 3, F = 40$,$N_f = 4, F = 40$,$N_f = 3, F = 50$ and $N_f = 4, F = 50$ are compared.

We compare the convergence of algorithm 2 when the size of BS and satellite local capacity and size of files library are different. In the case of four different size of capacities,

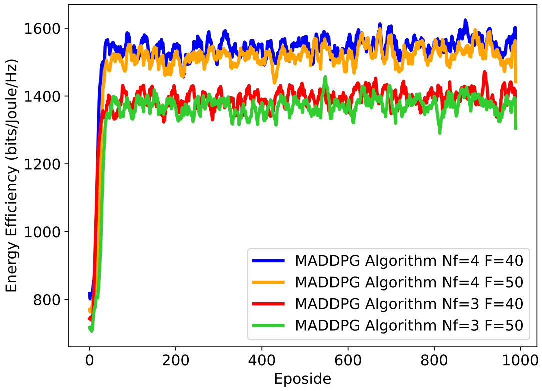

Fig. 7. The convergence of cache reward algorithm 2 in different local capacity and file library.

algorithm 2 can get good convergence performance. In the early episode of train, the cache reward is less than 1090 bits/Joule/Hz. This is because poor local cache deployment will cause additional power consumption. With the number of episodes increase, the cache rewards gradually increase. After training about 50 episodes, the cache rewards converge. In the framework, when the local cache capacity N_f is the same, the larger the file library F, the smaller the cache reward. Because when the cache file library F increases, it will be more difficult to hit the file in the algorithm 2 framework. On the contrary, for the same cache file library F, the local cache capacity N_f becomes larger and the cache reward becomes larger.

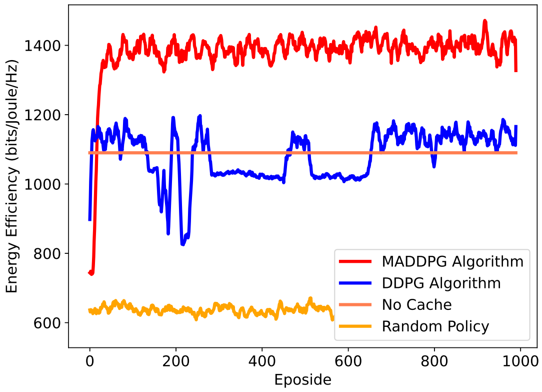

Fig. 8. Comparison of algorithm 2 and benchmark algorithms.

In Fig. 8 and Fig. 9, the curve shows the comparison of the energy efficiency convergence process and cache hit rate convergence process of different cache optimization algorithms when $N_f = 3, F = 40$. As can be seen in the Fig. 8, the energy efficiency of the proposed algorithm 2 and DDPG algorithm can reach convergence. For the same cache capacity and file library size, the energy efficiency of algorithm 2 is the highest among the four algorithms, which can get the best cache optimization effect and better optimize the objective function of the system. DDPG algorithm starts training from 900 bits/Joule/Hz and converges to 1100 bits/Joule/Hz after about 700 episodes. The curve of random policy oscillates between 600 bits/Joule/Hz and 650 bits/Joule/Hz, which can not converge well like other algorithms. Compared with the proposed algorithm 2, the effect of the optimization objective function of other algorithms is poor and the stability is not strong. As can be seen in Fig. 9, the cache hit rate of algorithm 2 starts to rise from 0.13 and finally converges to 0.33. The cache hit rate of DDPG algorithm can not converge well. From the perspective of cache hit rate, the proposed algorithm 2 can get better results.

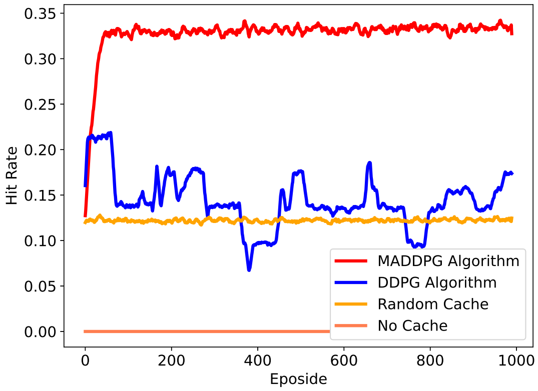

Fig. 9. Comparison of MADDPG and DDPG.

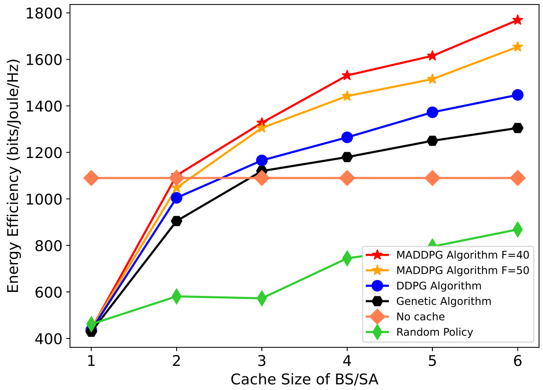

Fig. 10. Cache reward with different number of local cache capacity.

Fig. 10 and Fig. 11 respectively show the curves of energy efficiency of different algorithms with the cache capacity of each BS and satellite and the curves of cache hit rate with the cache capacity of each BS and satellite. As shown in Fig. 8, when the cache capacity is 1, the energy efficiency of MADDPG algorithm is less than that of uncached strategy. This is because the local cache capacity is too small, it is difficult for the algorithm to cache files suitable for users, so it is difficult to get good results. As the cache capacity increases from 1 to 6, the energy efficiency of MADDPG algorithm increases gradually. The proposed MADDPG algorithm can

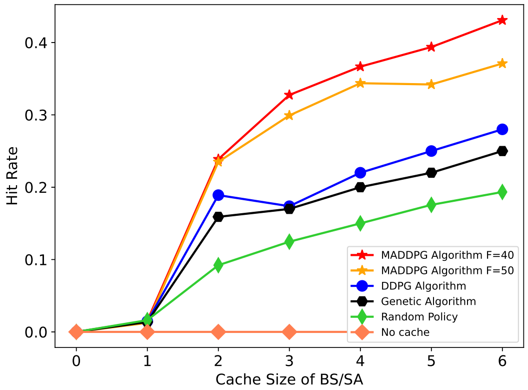

Fig. 11. Cache hit rate with different number of local cache capacity.

achieve higher energy efficiency than other algorithms, and the gap with other algorithms increases gradually. The trend of cache hit rate in Fig. 10 is roughly the same as that in Fig. 9. It can be seen from the figure that the cache hit rate of the proposed MADDPG algorithm is higher than that of other algorithms. Fig. 9 and Fig. 10 further illustrate the relationship between cache reward, cache hit rate, cache capacity and file library. In a certain range, when the local cache capacity is the same, the larger the file library, the smaller the cache reward and cache hit rate. Because when the file library increases, it is more difficult to hit the required files. On the contrary, for the same file library, the larger the local cache capacity, the greater the cache reward and cache hit rate.

V. Conclusion

In this paper, we propose a resource allocation and cache design scheme based on multi-agent deep reinforcement learning in an integrated terrestrial-satellite NOMA network. The objective is to maximize the energy efficiency of the system. We adopt a MADDPG algorithm to achieve user association, power control and cache design to improve the total energy efficiency of the system. The multi-agent deep reinforcement learning algorithm in this paper is divided into two stages, users and BSs are cleverly set as agents to complete the optimization problem in the framework. First, the users are set as the agents to choose the BSs or satellites and the power control factor. Then, the cache design scheme based on MADDPG is proposed. The BSs and satellites are set as the agents to select files cache from files library to local cache pool to improve energy efficiency. According to the simulation results, the proposed framework has good effectiveness and potential in solving the problem. Compared with the traditional single-agent deep reinforcement learning algorithm DDPG and other benchmark algorithms, it has a better optimization performance. In the future research, MADDPG algorithm will be used to deal with the optimal resource allocation problem of multi-layer satellite network model. More detailed power consumption will be investigated in future work.

References

[1] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, “Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges,” *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C. I and Z. Wang, “Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends,” *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74—81, Sep. 2015.

[3] Z. Ding, R. Schober, and H. V. Poor, “A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, June. 2016.

[4] H. Zhang et al., “Energy efficient dynamic resource optimization in NOMA system,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5671–5683, Sep. 2018.

[5] A. A. Nasir, H. D. Tuan, T. Q. Duong and M. Debbah, “NOMA throughput and energy efficiency in energy harvesting enabled networks,” *IEEE Trans. Wireless Commun.*, vol. 67, no. 9, pp. 6499–6511, Sep. 2019.

[6] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong and V. C. M. Leung, “Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2074–2085, Sep. 2020.

[7] H. Zhang, M. Feng, K. Long, G. K. Karagiannidis, V. C. M. Leung and H. V. Poor, “Energy efficient resource management in SWIPT enabled heterogeneous networks with NOMA,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 835–845, Feb. 2020.

[8] Y. Xu, Z. Qin, G. Gui, H. Gacanin, H. Sari and F. Adachi, ”Energy efficiency maximization in NOMA enabled backscatter communications with QoS guarantee,” in *IEEE Wireless. Commun. Lett.*, vol. 10, no. 2, pp. 353–357, Feb. 2021.

[9] M. Zeng, W. Hao, O. A. Dobre and H. V. Poor, ”Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA,” in *IEEE Trans. on Veh. Technol.*, vol. 68, no. 3, pp. 3000–3004, Mar. 2019.

[10] W. Lu, K. An, T. Liang and X. Yan, “Robust beamforming in multibeam satellite systems with non-orthogonal multiple access,” *IEEE Wireless Commun. Lett.*, to be published.

[11] X. Zhu, C. Jiang, L. Kuang, N. Ge and J. Lu, “Non-orthogonal multiple access based integrated terrestrial-satellite networks,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2253–2267, Oct. 2017.

[12] M. Jia, Q. Gao, Q. Guo, X. Gu and X. Shen, “Power multiplexing NOMA and bandwidth compression for satellite-terrestrial networks,” *IEEE Trans. on Veh. Technol.*, vol. 68, no. 11, pp. 11107–11117, Nov. 2019.

[13] Z. Gao, A. Liu, C. Han and X. Liang, ”Sum rate maximization of massive MIMO NOMA in LEO satellite communication system,” in *IEEE Wireless. Commun. Lett.*, to be published.

[14] A. Wang, L. Lei, E. Lagunas, A. I. Pérez-Neira, S. Chatzinotas and B. Ottersten, “NOMA-enabled multi-beam satellite systems: joint optimization to overcome offered-requested Data Mismatches,” *IEEE Trans. on Veh. Technol.*, vol. 70, no. 1, pp. 900–913, Jan. 2021.

[15] H. Li, S. Zhao, Y. Li and C. Peng, ”Sum Secrecy Rate Maximization in NOMA-Based Cognitive Satellite-Terrestrial Network,” in IEEE Wireless Communications Letters, *IEEE Wireless. Commun. Lett.*, vol. 10, no. 10, pp. 2230–2234, Oct. 2021.

[16] S. Fu, J. Gao and L. Zhao, “Integrated resource management for terrestrial-satellite systems,” *IEEE Trans. on Veh. Technol.*, vol. 69, no. 3, pp. 3256–3266, Mar. 2020.

[17] B. Deng, C. Jiang, J. Yan, N. Ge, S. Guo and S. Zhao, “Joint multigroup precoding and resource allocation in integrated terrestrial-satellite networks,” *IEEE Trans. on Veh. Technol.*, vol. 68, no. 8, pp. 8075–8090, Aug. 2019.

[18] B. Di, H. Zhang, L. Song, Y. Li and G. Y. Li, “Ultra-dense LEO: integrating terrestrial-satellite networks into 5G and beyond for data offloading,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.

[19] Y. Zhang, L. Yin, C. Jiang and Y. Qian, ”Joint beamforming design and resource allocation for terrestrial-satellite cooperation system,” *IEEE Trans. Wireless Commun.*, vol. 68, no. 2, pp. 778–791, Feb. 2020.

[20] Z. Ji, S. Wu, C. Jiang and W. Wang, ”Popularity-driven content placement and multi-hop delivery for terrestrial-satellite networks,” *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2574–2578, Nov. 2020.

[21] M. Shaat, A. I. Perez-Neira, G. Femenias and F. Riera-Palou, ”Joint frequency assignment and flow control for hybrid terrestrial-satellite backhauling networks,” *ISWCS.*, Bologna, 2017, pp. 293–298.

[22] E. Lagunas, L. Lei, S. Chatzinotas and B. Ottersten, “Power and flow assignment for 5G integrated terrestrial-satellite backhaul networks,” *WCNC.*,2019, pp. 1–6.

[23] Y. Cao, S. -Y. Lien and Y. -C. Liang, “Deep reinforcement Learning for multi-user access control in non-terrestrial networks,” *IEEE Trans. Wireless Commun.*, vol. 69, no. 3, pp. 1605–1619, Mar. 2021.

[24] X. Liao et al., ”Distributed intelligence: a verification for multi-agent DRL-based multibeam satellite resource allocation,” *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2785–2789, Dec. 2020.

[25] H. Zhang, N. Yang, W. Huangfu, K. Long and V. C. M. Leung, “Power control based on deep reinforcement learning for spectrum sharing,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4209–4219, Jun. 2020.

[26] P. V. R. Ferreira et al., “Reinforcement learning for satellite communications: from LEO to deep space operations,” *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 70–75, May. 2019.

[27] X. Hu, S. Liu, R. Chen, W. Wang and C. Wang, “A deep reinforcement learning-based framework for dynamic resource allocation in multibeam satellite systems,” *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1612–1615, Aug. 2018.

[28] P. V. R. Ferreira et al., “Multiobjective reinforcement learning for cognitive satellite communications using deep neural network ensembles,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 1030–1041, May. 2018.

[29] C. Zhou et al., “Deep reinforcement learning for delay-oriented IoT task scheduling in space-air-ground integrated network,” *IEEE Trans. Wireless Commun.*, to be published.

[30] C. Zhong, M. C. Gursoy and S. Velipasalar, “Deep reinforcement learning-based edge caching in wireless networks,” *IEEE Trans. Cognit. Commun. Networking.*, vol. 6, no. 1, pp. 48–61, Mar. 2020.

[31] Z. Zhang, H. Chen, M. Hua, C. Li, Y. Huang and L. Yang, “Double coded caching in ultra dense networks: caching and multicast scheduling via deep reinforcement learning,” *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1071–1086, Feb. 2020.

[32] Y. Qian, R. Wang, J. Wu, B. Tan and H. Ren, “Reinforcement learning-based optimal computing and caching in mobile edge network,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2343–2355, Oct. 2020.

[33] T. Zhang, Z. Wang, Y. Liu, W. Xu, and A. Nallanathan, “Joint resource, deployment and caching optimization for AR applications in dynamic UAV NOMA networks,” *IEEE Trans. Wireless Commun.*,to be published.

[34] T. Zhang, Z. Wang, Y. Liu, W. Xu and A. Nallanathan, “Caching placement and resource allocation for cache-enabling UAV NOMA networks,” *IEEE Trans. on Veh. Technol.*, vol. 69, no. 11, pp. 12897–12911, Nov. 2020.

[35] X. Li, H. Zhang, W. Li, and K. Long, “Multi-Agent DRL for user association and power control in Terrestrial-Satellite network,” *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–5.

[36] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, “Multiagent actor-critic for mixed cooperative-competitive environments,” *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6379-6390.

[37] K. Higuchi and Y. Kishiyama, “Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink,” in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Las Vegas, NV, 2013, pp. 1–5.

[38] X. Liu, H. Zhang, K. Long, A. Nallanathan, and V. C. M. Leung, “Energy efficient user association, resource allocation and caching deployment in fog radio access networks,” *IEEE Trans. on Veh. Technol.*, vol. 71, no. 2, pp. 1846–1856, Feb. 2022.

[39] Y. Jin, Y. Wen and C. Westphal, “Optimal transcoding and caching for adaptive streaming in media cloud: an analytical approach,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1914–1925, Dec. 2015.

[40] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath, “Deep reinforcement learning: a brief survey,” *IEEE Signal Process Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[41] D. Christopoulos, S. Chatzinotas and B. Ottersten, “Multicast multigroup precoding and user scheduling for frame-based satellite communications,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4695–4707, Sept. 2015.

XiaoNan Li received the M.S. degree from the School of Computer and Communication Engineering, University of Science and Technology Beijing, China, in 2022. His research interests include satellite communications and deep reinforcement learning.

Haijun Zhang (M’13, SM’17) is currently a Full Professor and Associate Dean at University of Science and Technology Beijing, China. He was a Postdoctoral Research Fellow in Department of Electrical and Computer Engineering, the University of British Columbia (UBC), Canada. He serves/served as Track Co-Chair of WCNC 2020, Symposium Chair of Globecom’19, TPC Co-Chair of INFOCOM 2018 Workshop on Integrating Edge Computing, Caching, and Offloading in Next Generation Networks, and General Co-Chair of GameNets’16. He serves as an Editor of IEEE Transactions on Communications, IEEE Transactions on Network Science and Engineering, and IEEE Transactions on Vehicular Technology. He received the IEEE CSIM Technical Committee Best Journal Paper Award in 2018, IEEE ComSoc Young Author Best Paper Award in 2017, and IEEE ComSoc Asia-Pacific Best Young Researcher Award in 2019.

Huan Zhou (M’14) received his Ph.D. degree from the Department of Control Science and Engineering at Zhejiang University. He was a visiting scholarat the Temple University from Nov. 2012 to May. 2013, and a CSC supported postdoc fellow at the University of British Columbia from Nov. 2016 to Nov. 2017. Currently, he is a full professor at the College of Computer and Information Technology, China Three Gorges University. He was a Lead Guest Editor of Pervasive and Mobile Computing, TPC Chair of EAI BDTA 2020, Local Arrangement Chair of I-SPAN 2018, Special Session Chair of the 3rd International Conference on Internet of Vehicles (IOV 2016), and TPC member of IEEE Globecom, ICC, ICCCN, etc. He has published more than 50 research papers in some international journals and conferences, including IEEE JSAC, TPDS, TWC and so on. His research interests include Opportunistic Mobile Networks, VANETs, Mobile Data Offloading, and Mobile Edge Computing.He receives the Best Paper Award of I-SPAN 2014 and I-SPAN 2018, and is currently serving as an associate editor for IEEE ACCESS and EURASIP Journal on Wireless Communications and Networking.

Ning Wang (M’13) received the B.E. degree in communication engineering from Tianjin University, China, in 2004, the M.A.Sc. degree in electrical engineering from The University of British Columbia, Canada, in 2010, and the Ph.D. degree in electrical engineering from the University of Victoria, Canada, in 2013. In 2013, he was on the Finalist of the Governor General’s Gold Medal for outstanding graduating doctoral student with the University of Victoria. From 2004 to 2008, he was with the China Information Technology Design and Consulting Institute as a Mobile Communication System Engineer, specializing in planning and design of large-scale commercial mobile communication networks, network traffic analysis, and radio network optimization. From 2013 to 2015, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia. Since 2015, he has been with the School of Information Engineering, Zhengzhou University, Zhengzhou, China, where he is currently a Professor. He also holds adjunct appointments with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada. His research interests include resource allocation and security designs of future cellular networks, channel modeling for wireless communications, statistical signal processing, and cooperative wireless communications. He was on the technical program committees of international conferences, including the IEEE GLOBECOM, IEEE ICC, IEEE WCNC, and CyberC.

Keping Long (SM'06) received the M.S. and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, in 1995 and 1998, respectively. From September 1998 to August 2000, he was a Postdoctoral Research Fellow at the National Laboratory of Switching Technology and Telecommunication Networks, Beijing University of Posts and Telecommunications (BUPT), China. From September 2000 to June 2001, he was an Associate Professor at BUPT. From July 2001 to November 2002, he was a Research Fellow with the ARC Special Research Centre for Ultra Broadband Information Networks (CUBIN), University of Melbourne, Australia. He is currently a professor and Dean at the School of Computer and Communication Engineering, University of Science and Technology Beijing. He has published more than 200 papers, 20 keynote speeches, and invited talks at international and local conferences. His research interests are optical Internet technology, new generation network technology, wireless information networks, value added services, and secure technology of networks. Dr. Long has been a TPC or ISC member of COIN 2003/04/05/06/07/08/09/10, IEEE IWCN2010, ICON2004/06, APOC2004/06/08, Co-Chair of the organization Committee for IWCMC2006,TPC Chair of COIN 2005/08, and TPC Co-Chair of COIN 2008/10. He was awarded by the National Science Fund for Distinguished Young Scholars of China in 2007 and selected as the Chang Jiang Scholars Program Professor of China in 2008. He is a member of the Editorial Committees of Sciences in China Series F and China Communications.

Saba Al-Rubaye (IEEE Senior Member) reader in Autonomous and Connected Systems and Head of Advanced Connectivity & System Integration research group in the School of Aerospace, Transport and Manufacturing at Cranfield. She has more than 17 years of professional experience in industry and academia with demonstrated track record of launching innovative solutions in design, testing, consultation, leadership, and program development. Dr Al-Rubaye has been involved in several projects sponsored by Innovation UK, Research England, EPSRC, and Department for Transport (DfT) as well as Canada-NSERC, and USA Government. She is participating in developing industry standards by being an active research group member of IEEE P1932.1 standard of License/Unlicensed Interoperability and IEEE P1920.2 Standard for Vehicle-to-Vehicle Communications for Unmanned Aircraft Systems (UAS). Dr Al-Rubaye has published many papers in prestige IEEE journals and conferences and a recipient of the best technical paper award twice published in IEEE Vehicular Technology in 2011 and 2015, respectively. Dr Al-Rubaye has contributed as a general co-chair and organized a cutting-edge workshop for 6G Communication Networks at IEEEICC2020 Conference in Ireland and was a keynote speaker for Industry day in IEEE ICCVE2019 in Austria. Her main research interests are UAS/Aircraft communications, 5G/6G Networks, Advance Air Mobility, Safety and Security of Autonomous Vehicle. She is a Chartered Engineer (CEng), member of IET, Life Senior member of IEEE and certified Unmanned Aircraft System (UAS) Pilot.

George K. Karagiannidis (IEEE Fellow) is currently Professor in the Electrical & Computer Engineering Dept. of Aristotle University of Thessaloniki, Greece and Head of Wireless Communications & Information Processing (WCIP) Group. His research interests are in the areas of Wireless Communications Systems and Networks, Signal processing, Optical Wireless Communications, Wireless Power Transfer and Applications and Communications & Signal Processing for Biomedical Engineering. Dr. Karagiannidis was in the past Editor in several IEEE journals and from 2012 to 2015 he was the Editor-in Chief of IEEE Communications Letters. From September 2018 to June 2022 he served as Associate Editor-in Chief of IEEE Open Journal of Communications Society. Currently, he is in the Steering Committee of IEEE Transactions on Cognitive Communications & Networks. Recently, he received two prestigious awards: The 2021 IEEE Communications Society Radio Communications (RCC) Committee Technical Recognition Award, for his Outstanding Contributions to Wireless Systems and the 2018 Signal Processing and Communications Electronics (SPCE) Technical Recognition Award of the IEEE ComSoc for his Outstanding Technical Contributions to Signal Processing for Communications. Dr. Karagiannidis is one of the highly-cited authors across all areas of Electrical Engineering, recognized from Clarivate Analytics as Web-of-Science Highly-Cited Researcher in the seven consecutive years 2015-2021.