Reinforcement Learning based Resource Management for 6G-Enabled mIoT with Hypergraph Interference Model

Jie Huang, *Member, IEEE*, Cheng Yang, Shilong Zhang, Fan Yang, *Member, IEEE*, Osama Alfarraj, Valerio Frascolla, *Member, IEEE*, Shahid Mumtaz, *Senior Member, IEEE*, Keping Yu, *Senior Member, IEEE*

***Abstract*—For the future 6G-enabled massive Internet of Things (mIoT), how to effectively manage spectrum resources to support huge data traffic under the large-scale overlapping caused by the dense deployment of massive devices is the imperative challenge. In this paper, a novel hypergraph interference model is designed, and two reinforcement learning (RL)-based resource management algorithms in the 6G-enabled mIoT are proposed to enhance the network throughput and avoid overlapping interference. Then, based on the hypergraph interference model, the resource management problem of execution network throughput maximization is theoretically formulated under large-scale overlapping interference scenarios. To handle this problem, we convert it into a Markov decision process (MDP) model and then deal with this MDP model through the advantage actor-critic (A2C)-based resource management algorithm and asynchronous advantage actor-critic (A3C)-based resource management algorithm, which aim to maximize network throughput of the spectrum resource allocation among massive devices. The simulation results verify that the proposed algorithms can not only avoid large-scale overlapping interference but also improve the network throughput.**

***Index Terms*—Internet of Things (IoT), resource management, hypergraph, Markov decision process (MDP), reinforcement Learning (RL).**

I. Introduction

A. Background and Motivation

MASSIVE Internet of Things (mIoT) is an essential component of future wireless networks, which has been envisioned to meet the significant surge in data traffic demand generated by an enormous amount of IoT devices and support various innovative applications, such as smart city, smart household, smart industrial, and vehicular communication [1]. The huge data traffic demand further spurs massive devices to the data transmission concurrently via the same spectrum and then makes the spectrum resource become more scarce [2]. Since this demand for mIoT applications will be beyond the capability of the five-generation (5G) [3], [4], ensuring simultaneous access for massive devices and handling the substantial data traffic they generate is challenging, resulting in insufficient network performance [5]. Against this backdrop, it is necessary to develop a six-generation (6G)-enabled mIoT with the device-to-device (D2D) communication [6]. Specifically, D2D communications, which enable two proximity IoT devices to transmit data directly without going through access points, e.g., remote radio heads (RRHs) [7], have emerged as a promising technique to facilitate communication among massive devices in the 6G-enabled mIoT.

The distinctive feature of the 6G-enabled mIoT is its dense deployment of devices, which poses great challenges for the traditional resource management method to fulfill the huge data traffic generated by massive devices [8]. In particular, the dense deployment of devices will inevitably generate overlapping coverage areas due to the overlapping coverage of each device's communication range [9], [10]. When there existing massive devices use the same spectrum resource in the overlapping coverage area, the overlapping interference then is generated for the 6G-enabled mIoT [11]. As the number of IoT devices in overlapping coverage areas increases, the overlapping interference will further deteriorate the network throughput of the entire 6G-enabled mIoT [12], i.e., the large-scale overlapping interference that affects the entire network is formed. For the 6G-enabled mIoT, how to achieve effective resource management and then enhance the entire network performance will be an imperative challenge under large-scale overlapping interference scenarios.

This work was supported by the Distinguished Scientist Fellowship Program (DSFP) at King Saud University, Riyadh, Saudi Arabia, in part by the National Natural Science Foundation of China under Grant (No.62301094), in part by the Chongqing Natural Science Foundation Innovation and Development Joint Fund (CSTB2023NSCQ-LMX0014), in part by Science and Technology Research Program of Chongqing Education Commission of China under Grant (KJQN202201157, KJQN202301135), in part by the 6G-SENSES project from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement (No.101139282). *(Corresponding authors: Cheng Yang, Fan Yang)*

Jie Huang, Cheng Yang, Shilong Zhang, and Fan Yang are with the School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China. (email: huangjie_cq@cqut.edu.cn, msixramYC@gmail.com, 1187923736@qq.com, yf_0220@cqut.edu.cn);

Osama Alfarraj is with the Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia. (email: oalfarraj@ksu.edu.sa);

Valerio Frascolla is with Intel Deutschland GmbH, 85579 Neubiberg, Germany. (e-mail: valerio.frascolla@intel.com);

Shahid Mumtaz is with Department of Applied Informatics, Silesian University of Technology, Akademicka 16 44-100 Gliwice, Poland, and Departement of Computer Sciences, Nottingham Trent University, Nottingham, UK. (email: dr.shahid.mumtaz@ieee.org);

Keping Yu is with the Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia, and with the Graduate School of Science and Engineering, Hosei University, Tokyo 184-8584, Japan. (email: kepingyu@ksu.edu.sa; keping.yu@ieee.org).

TABLE I
RELATED WORK

Ref.	Scenario	Main challenge	Resources	Optimization Objective	Method
[13]	IoT	Balance the network performance and service cost	Spectrum	Coverage probability	Game
[14]	IoT	Under channel imperfections to enhance communication and collaboration of IoT devices	Power Spectrum	Energy efficiency	RL
[15]	IoT	Different QoS requirements	Power Spectrum	Energy efficiency	RL
[16]	mIoT	Data traffic throughput sharply increases	Computation Spectrum	Content retrieval delays	RL Opt.
[17]	mIoT	Limited computation capabilities and energy	Power Spectrum	Energy efficiency	Opt.
[18]	mIoT	Sharp shortage of spectrum resources caused by devices' dense deployment	Spectrum	Quality of experience	RL
[19]	IoT	Co-channel interference	Spectrum	Network throughput	Opt.
[20]	IoT	Co-tier interference	Power Spectrum	Network throughput	Opt.
[21]	IoT	Cross-tier interference	Power Spectrum	Network throughput	Opt.
Proposed	mIoT	Large-scale overlapping interference	Spectrum	Network throughput	RL

B. Related Work

Resource management for IoT networks is investigated as a practical approach to obtain better network performance in the existing literature. Table I presents the relative comparison of the proposed scheme with existing schemes. For example, to balance the service cost and network performance in IoT networks, Yan *et al.* [13] investigated a joint access selection and resource management scheme, which proposed a hierarchical game framework for the access selection and bandwidth allocation problem, and then solve it though Stackelberg game theory. An imperfect channel state information (CSI)-based resource management scheme was developed in [14] to efficiently allocate spectrum and power resources, which improves the collaboration and communication between devices in IoT networks under channel imperfections. In an attempt to boost network performance and fulfill various quality of service (QoS) demands, Yang *et al.* [15] developed a resource management scheme based on reinforcement learning (RL) that focuses on energy efficiency maximization. This scheme converts the optimization problem into a Markov decision process (MDP) model and subsequently solves it using an actor-critic method. In addition to traditional IoT networks, due to the limited resources and huge data traffic characteristics of the mIoT, the resource management for massive devices and limited resources network scenario is considered as the key technology in the mIoT network [16]–[18]. The study [16] addressed the optimization problem of cooperative edge caching and spectrum resource management caused by the increased data traffic throughput resulting from the growing number of devices. The branch-and-bound methodology was used to address the cooperative edge caching aspect of this NP-hard problem, while the RL method was employed to tackle the spectrum resource management aspect. To boost the energy efficiency while simultaneously meeting the devices' maximum tolerable delay constraints, Liu *et al.* [17] formulated a joint computation and spectrum resource management problem. The authors decomposed this challenging mixed-integer non-convex problem into two individual subproblems, which were addressed individually using sequential convex programming and matching methods. To enhance the quality of experience for users under limited spectrum resources, a joint power and spectrum resource management scheme was proposed in [18]. This scheme was developed to solve an optimization problem for resource management and utilized a neural network-embedded RL algorithm to find the best solution.

Rarely do the aforementioned works [16]–[18] specifically address the impact of interference brought on by multiple devices simultaneously engaging in competition for the same spectrum resource. To alleviate the interference impact through proper resource management, the more general and compelling problem of interference has been attracting increasing research attention [19]–[21]. An interference avoidance resource management method was developed in [19] to decrease co-channel interference power and improve data transmission rates in IoT networks. The method separated the optimization problem into three individual subproblems, which can be addressed by the orthogonal deployment approach, bisection search method, and Hungarian algorithm, respectively. To minimize co-tier interference and increase the data rate, Sarma *et al.* [20] designed an efficient scheme of resource management. To maximize network throughput while accounting for the cross-tier interference that is inevitable in IoT networks, a solution including two-stage joint power control and hovering altitude was developed in [21] for the resource management problem, which mainly utilizes the Lagrange dual decomposition and concave-convex procedure method. However, most of the aforementioned works [19]–[21] rarely focus on the dense deployment of devices leading to overlapping interference problems. Furthermore, the interference avoidance and resource management schemes are typically for the traditional IoT network, which may not be suitable for the future 6G-enabled mIoT under large-scale overlapping interference scenarios. Therefore, it is an imperative challenge to obtain a resource management scheme with interference avoidance and then enhance the entire network performance for the 6G-enabled mIoT under large-scale overlapping interference scenarios.

C. Contribution

For the 6G-enabled mIoT, the large-scale overlapping interference will divide the entire network device set into multiple device subsets in which individual device interferes with each other. Specifically, the device subset represents the collective relationships among devices due to the influence of overlapping interference, which is mathematically a multi-way relationship. Since hyperedges describe a collective relationship among a set of vertices in the hypergraph model, we apply the hypergraph model to represent and analyze overlapping interference among massive devices. In this paper, to maximize the entire network throughput, we design a novel hypergraph interference model and then propose two RL-based resource management algorithms for 6G-enabled mIoT under large-scale overlapping interference scenarios. In particular, our main contribution can be summarized as follows.

1) To handle the large-scale overlapping interference for the 6G-enabled mIoT, a novel hypergraph interference model was proposed. Then, based on this hypergraph interference model, we formulate a resource management problem aimed at maximizing the network throughput within the constraints of limited resources, while avoiding overlapping interference among massive devices and ensuring data transmission.
2) To reduce the solution complexity, the resource management problem for the 6G-enabled mIoT is solved through the RL method. We reformulate the network throughput maximization problem as an MDP model and propose an advantage actor-critic (A2C)-based resource management algorithm to solve it. Specifically, the reward function of the MDP model is specially designed according to this optimization objective and constraints.
3) To speed up training and obtain higher throughput resource allocation results, we proposed an asynchronous advantage actor-critic (A3C)-based resource management algorithm combined with an asynchronous multi-threaded architecture. It is capable of avoiding overlapping interference in corresponding overlapping areas, and also preventing throughput degradation of massive devices.

The paper's remainings are organized as below. Section II presents details of the system model. Section III presents the RL-based resource management to handle the network throughput maximization problem. We then investigate the performance of the proposed two algorithms via simulation results in Section IV. Finally in Section V, the conclusions are drawn.

II. System Model

In this section, we give a concise overview of the communication model. Then, the relationship between large-scale overlapping interference and the hypergraph interference model is described. Finally, we formulate the resource management problem as a network throughput maximization optimization formulation to improve the network performance in 6G-enabled mIoT.

A. Communication Model

In Fig. 1, this paper considers a 6G-enabled mIoT supported cloud radio access network (C-RAN) framework [22], which consists of multiple RRHs, fronthaul links, a baseband unit (BBU) pool, and massive devices. This C-RAN framework deploys multiple RRHs as access points around IoT devices and then is responsible for spectrum resource management of D2D communication between massive devices [23], i.e., D2D receivers (DRs) and D2D transmitters (DTs). Wherein, we mainly focus on the spectrum resource management among massive devices in the 6G-enabled mIoT. Due to the ability of powerful centralized processors, the BBU pool is configured to optimize resource allocation [22]. The global CSI is assumed to be available at the BBU pool [24]. There are various technologies that can be used to build the fronthaul links that connect multiple RRHs to a BBU pool. Furthermore, simultaneous transmission is allowed between the BBU pool and the RRHs, as well as between the RRHs and IoT devices, and between the IoT devices, without any interference. Note that the communication range of IoT devices as DTs can overlap with each other and then form overlapping coverage areas, as shown in Fig. 1.

Assume $\mathcal{N}_{\text{DT}}$ and $\mathcal{N}_{\text{DR}}$ denote the sets of deployed single antenna IoT devices as DTs and DRs in 6G-enabled mIoT and denoted as $\mathcal{N}_{\text{DT}} = \left\{ \hat{1}, \hat{2}, \cdots, \hat{N} \right\}$ and $\mathcal{N}_{\text{DR}} = \{1, 2, \cdots, N\}$, respectively. The total spectrum resource is split into K resource blocks (RBs), which are assumed to be orthogonal and represented by $\mathcal{K} = \{1, 2, \cdots, K\}$ [25]. Let $c_{\hat{n},k} \in \{0, 1\}$ denotes whether the k-th RB is assigned for the $\hat{n}$-th DT. The 3GPP outdoor channel model [26] is the primary model that we utilize for determining the power received from a desired signal in 6G-enabled mIoT. This channel model takes into account various path losses and small-scale fading elements. At time step t, the n-th DR's received signal-to-interference-plus-noise ratio (SINR) of the desired signal from corresponding $\hat{n}$-th DT over the k-th channel is expressed by [22]

$$\gamma_n^t\left[k\right] = \frac{c_{\hat{n},k}^t p_{\hat{n}}^t\left[k\right] g_{\hat{n},n}^t\left[k\right]}{\sigma_t^2 + \sum\limits_{\tilde{n} \in \mathcal{N}_{\text{DT}}} c_{\tilde{n},k}^t p_{\tilde{n}}^t\left[k\right] g_{\tilde{n},n}^t\left[k\right]}, \tag{1}$$

where $p_{\hat{n}}^t\left[k\right]$ represents the $\hat{n}$-th DT's transmission power over the k-th RB. $p_{\tilde{n}}^t\left[k\right]$ denotes the $\tilde{n}$-th DT's transmission power over the k-th RB. $g_{\hat{n},n}^t$ and $g_{\tilde{n},n}^t$ are the desired signal channel gain and interfere signal channel gain, respectively. σ_t^2 denotes the variance of the additive white Gaussian noise (AWGN). Hence, the received achievable data rate of n-th DR is represented by

$$R_n^t = \sum_{k \in \mathcal{K}} W \log\left(1 + \gamma_n^t\left[k\right]\right), \tag{2}$$

where W indicates the assigned bandwidth for the k-th RB. Furthermore, it is necessary to take into account the minimal data rate $R_n^{\min}$ for the n-th DR, which means that $R_n^t \geq R_n^{\min}$. For the entire 6G-enabled mIoT, the network throughput can be expressed as [27]

$$R^t = \sum_{n \in \mathcal{N}_{\text{DR}}} R_n^t. \tag{3}$$

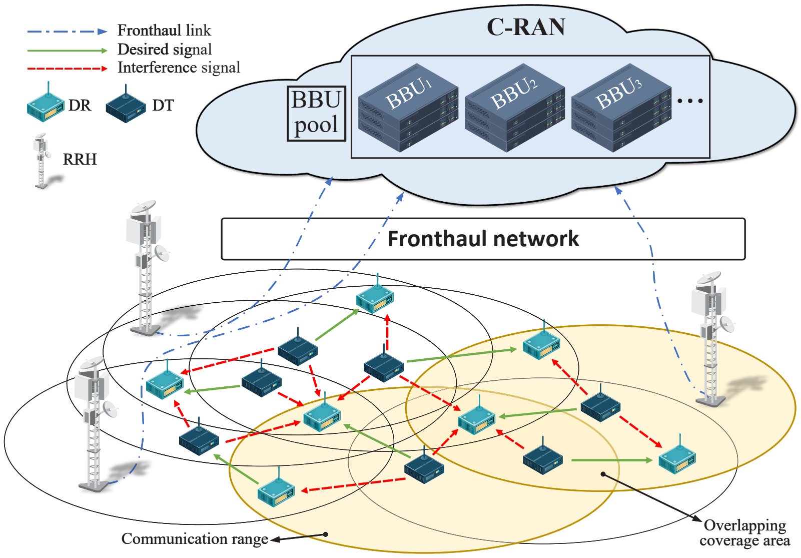

Fig. 1. Architecture of the considered 6G-enabled mIoT.

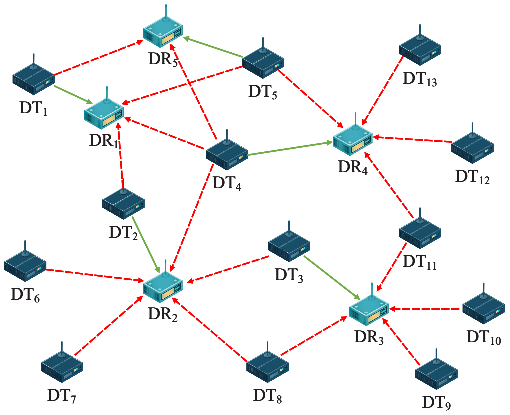

Fig. 2. Communication networks architecture.

B. Hypergraph Interference Model

In the 6G-enabled mIoT, due to the dense deployment of devices, there is large-scale overlapping coverage of DT's communication ranges, which leads to large-scale overlapping interference among DTs that reuse the same resources. Hence the SINR of DRs suffers from critical interference signal, which can cause disruptions like data failing to be received from DTs. In this context, traditional graphs can only describe the relationship between two devices and cannot establish this collective relationship among multiple devices. A hypergraph, as a promotion of a graph, is a power tool that can model the relationship among any number of IoT devices.

We denote an hypergraph at time step t by $H^t = (X^t, D^t)$, where $X^t = \{x_1^t, x_2^t, \cdots, x_n^t\}$ represents the vertex set and $D^t = \{d_1^t, d_2^t, \cdots, d_m^t\}$ the hyperedge set. To represent the hypergraph and operate the hyperedge, the definitions of incidence matrix and weak deletion of hyperedge are given.

Incidence matrix $\mathbf{H}^t$: It is a matrix with m rows and n columns, where rows represent hyperedges, columns represent vertices, and the elements in $\mathbf{H}^t$ are taken as follows

$$(x_i^t, d_j^t) = \begin{cases} 0, & x_i^t \notin d_j^t, \\ 1, & x_i^t \in d_j^t, \end{cases} \tag{4}$$

where $(x_i^t, d_j^t) = 1$ represents that vertex x_i^t is in the range of hyperedge d_j^t at time step t.

Weak deletion of a hyperedge: the weak deletion of a hyperedge d_j^t from hypergraph $H^t = (X^t, D^t)$ at time step t makes the hypergraph $H'^t = (X'^t, D'^t)$, where $D'^t = D^t \backslash \{d_j^t\}$. That is, just removing the hyperedge d_j^t will not affect the rest of the hypergraph, i.e., incidence matrix $\mathbf{H}^t$ just removes the j-th row. For any subset S^t of D^t, the $H^t \backslash S^t$ represents the hypergraph built by weakly deleting the hyperedges of S^t from H^t.

Based on the relationship between vertices and hyperedges, the method of establishing hyperedge is as follows: we build the hyperedge centered on the receiver and each hyperedge contains many covered transmitters. In addition, the receiver, around which the hyperedge is constructed, falls within the communication range of all transmitters belonging to the same hyperedge.

According to the method of establishing hyperedge, Fig. 2 is modeled to the initial hypergraph model, as shown in Fig. 3, where hyperedges are built centered on $\{\mathrm{DR}_1^t, \ldots, \mathrm{DR}_5^t\}$. Based on the meaning of incidence matrix $\mathbf{H}^t$, Fig. 3 can be represented by the matrix $\mathbf{H}^t$ as shown in (5).

$$\mathbf{H}^t = \begin{array}{c} \\ d_{I_1}^t \\ d_{I_2}^t \\ d_{I_3}^t \\ d_{I_4}^t \\ d_{I_5}^t \end{array} \!\!\begin{array}{c} \begin{array}{ccccccccccccc} \mathrm{DT}_1^t & \mathrm{DT}_2^t & \mathrm{DT}_3^t & \mathrm{DT}_4^t & \mathrm{DT}_5^t & \mathrm{DT}_6^t & \mathrm{DT}_7^t & \mathrm{DT}_8^t & \mathrm{DT}_9^t & \mathrm{DT}_{10}^t & \mathrm{DT}_{11}^t & \mathrm{DT}_{12}^t & \mathrm{DT}_{13}^t \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \tag{5}$$

where columns and rows of matrix $\mathbf{H}^t$ denotes the vertices and hyperedges, respectively. $d_{I_j}^t$ represents the hyperedge established with j^{th} DR as the center.

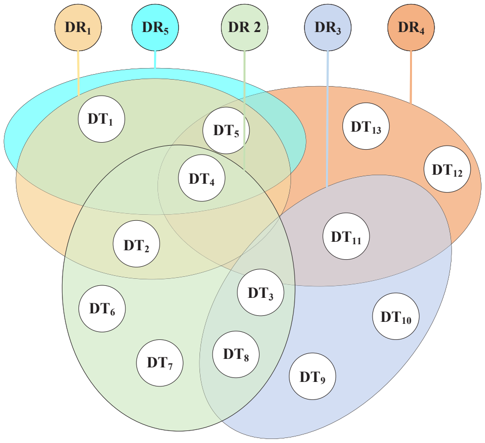

Fig. 3. Initial hypergraph interference model.

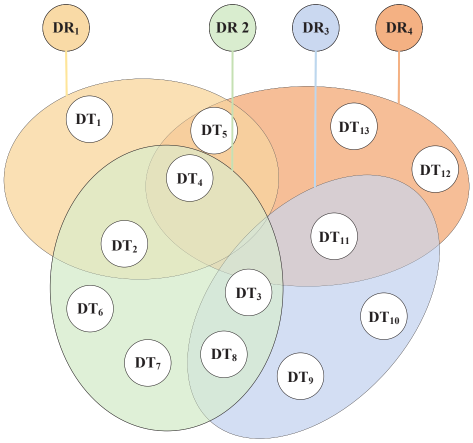

Fig. 4. Simplified hypergraph interference model.

Due to the different communication ranges of devices, in the initial hypergraph model, one hyperedge contains another hyperedge, which is a sub-hyperedge. Therefore, the initial hypergraph can be simplified by deleting sub-hyperedges, which does not change the relationship between vertices. The sub-hyperedge can be judged by the intersection of the rows. For example: in (5),

$$\begin{aligned} d_{I_1}^t &= [1\,1\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0], \\ d_{I_5}^t &= [1\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0], \\ d_{I_1}^t \cap d_{I_5}^t &= d_{I_5}^t, \end{aligned} \tag{6}$$

(5) and (6) show that hyperedge $d_{I_5}^t$ is a sub-hyperedge of $d_{I_1}^t$ at time step t. Therefore, by deleting the rows representing the sub-hyperedge, (5) is simplified to (7), i.e., Fig. 3 can be simplified to Fig. 4 through weak deletion of a hyperedge.

For the hypergraph as shown in Fig. 4, the incidence matrix of the simplified hypergraph model can be formulated as

$$\mathbf{H}_s^t = \begin{array}{c} \\ d_{I_1}^t \\ d_{I_2}^t \\ d_{I_3}^t \\ d_{I_4}^t \end{array} \!\!\begin{array}{c} \begin{array}{ccccccccccccc} \mathrm{DT}_1^t & \mathrm{DT}_2^t & \mathrm{DT}_3^t & \mathrm{DT}_4^t & \mathrm{DT}_5^t & \mathrm{DT}_6^t & \mathrm{DT}_7^t & \mathrm{DT}_8^t & \mathrm{DT}_9^t & \mathrm{DT}_{10}^t & \mathrm{DT}_{11}^t & \mathrm{DT}_{12}^t & \mathrm{DT}_{13}^t \end{array} \\ \left[\begin{array}{ccccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}. \tag{7}$$

Due to the spectrum resource not being duplicated in the same hyperedge for the resource management problem, it is the same as with the nature of the vertices coloring problem in the hypergraph. Hence, this resource management problem of the 6G-enabled mIoT can be transformed into a vertex coloring problem.

For measuring interference among DTs, we propose an interference degree matrix Φ^t based on the simplified hypergraph incidence matrix $\mathbf{H}_s^t$ at time step t to quantify the overall interference degree of mIoT, which can be defined as

$$\Phi^t = \mathbf{log}\left(\mathbf{max}\left(\mathbf{H}_s^t \mathbf{C}^t, 1\right)\right), \tag{8}$$

where $\mathbf{C}^t$ is resource allocation matrix. $\mathbf{log}(\cdot)$ denotes that all elements of the matrix are performed the log operation, e.g.,

$$\mathbf{log}\left(\left[\begin{array}{cc} 5 & 3 \\ 2 & 1 \end{array}\right]\right) = \left[\begin{array}{cc} \log 5 & \log 3 \\ \log 2 & 0 \end{array}\right]. \tag{9}$$

$\mathbf{max}(\cdot, 1)$ denotes that the maximum value is obtained by comparing each primitive of the matrix with 1, e.g.,

$$\mathbf{max}\left(\left[\begin{array}{cc} 5 & 3 \\ 2 & 0 \end{array}\right], 1\right) = \left[\begin{array}{cc} 5 & 3 \\ 2 & 1 \end{array}\right]. \tag{10}$$

In the network interference degree matrix $\Phi^t \in \{I_{d,c}\}^{|\mathcal{E}_\mathcal{H}| \times C^t}$, where $I_{d,c} > 0$ denotes that DTs belong to the d-th hyperedge is allocated the c-th RB leading to interference and $I_{d,c} \leq 0$ otherwise. Hence, the overall network interference degree can be calculated by the following formulation:

$$\varphi^t = \left\|\Phi^t\right\|_1, \tag{11}$$

where $\varphi^t = 0$ denotes that there is no interference in the mIoT and $\varphi^t \neq 0$ otherwise.

C. Problem Formulation

This paper aims to optimize spectrum resource management to increase the network throughput of the entire 6G-enabled mIoT. The presented problem is framed based on the assumption that the transmission power of IoT devices remains constant, as stated below.

$$\max_{c_{\hat{n},k}^t} R^t \tag{12a}$$

$$\text{s.t.} \varphi^t = 0, \tag{12b}$$

$$R_n^t \geq R_n^{\min}, \forall n \in \mathcal{N}_{\mathrm{DR}}, \tag{12c}$$

$$\sum_{\hat{n}} c_{\hat{n},k}^t \leq \hat{N}, \forall k \in \mathcal{K}, \tag{12d}$$

$$\sum_{k} c_{\hat{n},k}^t \leq 1, \forall \hat{n} \in \mathcal{N}_{DT}, \tag{12e}$$

$$c_{\hat{n},k}^t \in \{0, 1\}. \tag{12f}$$

In this problem, constraint (12b) guarantees that there is no overlapping interference in the entire 6G-enabled mIoT.

Constraint (12c) guarantees the minimum transmission rate requirement of DRs that receive the desired signal from DT. Constraint (12d) states that at most N DTs with different channels gain orthogonally reuse an RB. Constraint (12e) indicates that each communication link can use at most one RB. Constraint (12f) states that the DTs and RBs assignment parameters can only be integer variables 0 or 1.

In this case, the 6G-enabled mIoT can gather extensive state information. Then, it makes an all-encompassing resource management decision on all IoT devices after taking into account the current state of the environment. Nevertheless, as the network expands in size, the solution to the nonconvex problem derived from the hypergraph coloring problem becomes a computationally challenging endeavor known as NP-hard [28]. Since there is an enormous number of resource allocation results for massive devices, the resolution of this resource management problem (12) is not mathematically simple and requires significant computer resources. Consequently, the next section will focus on the efficient RL-based solution.

III. Reinforcement Learning Based Resource Management

In this section, an MDP model for 6G-enabled mIoT is specifically designed. The MDP model is employed to acquire an effective solution of the formulated optimization problem (12). Then, we proposed two actor-critic based resource management algorithms to solve the proposed MDP model and then maximize network throughput for 6G-enabled mIoT under large-scale overlapping interference scenarios. Finally, the proposed algorithms' complexity is analyzed.

A. MDP model

In this 6G-enabled mIoT environment, it is often for the state transition probabilities and expected rewards of all states are frequently unknown. Consequently, we deal with the spectrum resource management problem (12) in 6G-enabled mIoT by employing an RL framework. This requires obtaining a stochastic optimal policy by interacting with this environment. The proposed MDP, designed specifically for the 6G-enabled mIoT, consists of the following essential elements: state space, action space, reward function, policy and value function.

1) State Space: The proposed MDP model setup involves an RL agent that observes the 6G-enabled mIoT environment in discrete time. As for the centralized scheme, the BBU pool as an RL agent should know all information about whole IoT devices, e.g., association state, transmit power and overlapping interference.

(1) $\mathbf{H}^t$: The incidence matrix of the hypergraph interference model.
(2) φ^t: The interference degree of the entire 6G-enabled mIoT.
(3) $\mathbf{P}^t$: The set of DTs' transmission power.
(4) $\mathbf{C}^{t-1}$: The resource allocation matrix.
(5) R^t: The network throughput of the entire 6G-enabled mIoT.

Hence, in the 6G-enabled mIoT environment, the state s_t at time step t can be formulated as

$$s_t = \left\{\mathbf{H}^t, \varphi^t, \mathbf{P}^t, \mathbf{C}^{t-1}, R^t\right\} \tag{13}$$

and the state space for the 6G-Enabled mIoT environment can be formulated as $\mathcal{S}$. The 6G-enabled mIoT environment is assumed to transition from state s_t to next state s_{t+1} by the RL agent taking an action in the MDP model.

2) Action Space: In this 6G-Enabled mIoT environment, the action space can be denoted as $\mathcal{A}$. During resource management of the 6G-enabled mIoT, the BBU pool as an RL agent makes decisions for the communication request from IoT devices. Hence, the performed action is a resource allocation matrix defined as $a_t \in \{0,1\}^{\hat{N}\times K}$ at time step t, which can be expressed as

$$a_t = \begin{bmatrix} c_{\hat{1},1}^t & \cdots & c_{\hat{N},1}^t \\ \vdots & \ddots & \vdots \\ c_{\hat{1},K}^t & \cdots & c_{\hat{N},K}^t \end{bmatrix}. \tag{14}$$

In addition, the sum of all elements in a matrix column does not exceed the DTs' number $\hat{N}$ to meet the constraint (12d). The sum of all elements in a resource allocation matrix row does not exceed 1 to meet the constraint (12e).

3) Reward Function: The design of the reward function is crucial for resource management with avoidance interference as it influences the convergence performance and network performance of learning algorithms. The reward function would be used to evaluate the value of the state space and action space. As mentioned in Section II-C, the network throughput, as the agent's optimal goal, will be maximized in the learning process. Moreover, for this proposed MDP model, the design of the reward function must satisfy constraints (12b) and (12c) to improve the network throughput of the entire 6G-enabled mIoT. Hence, at time step t, the reward function r_t includes the network throughput R^t and the overlapping interference penalty, which can be defined as

$$r_t = \begin{cases} R^t, & \text{if (12b) and (12c) are satisfied,} \\ -\varphi^t, & \text{otherwise.} \end{cases} \tag{15}$$

Moreover, by utilizing the proposed MDP model as its foundation, the RL algorithm can efficiently address sequential decision problems of resource management, which involve selecting the maximum cumulative reward through a series of states. The cumulative discounted reward, denoted as $G_t = \sum_{\tau=0}^{T-t} \lambda^\tau r_{t+\tau}$, is computed by summing the rewards $r_{t+\tau}$ multiplied by the discount factor λ^τ for each time step τ from 0 to $T-t$, where T represents the total number of time steps. The reward discount factor, denoted by $\lambda \in [0,1]$, quantifies the impact of future rewards on their present value [29]. This hyperparameter of the RL algorithm can be modified.

4) Policy: In the proposed MDP model, the policy refers to a probability for selection actions that aim to optimize long-term performance. To thoroughly explore the complete set of possible actions, the proposed MDP model adopts a stochastic policy $\pi\left(a_t|\, s_t\right) = \Pr\left(a_t|\, s_t\right)$. The probability of executing an action a_t in state s_t is represented by this policy $\pi\left(a_t|\, s_t\right)$ [29]. Therefore, the process of choosing an action can be formulated as

$$a_t \sim \pi\left(\cdot|\, s_t\right). \tag{16}$$

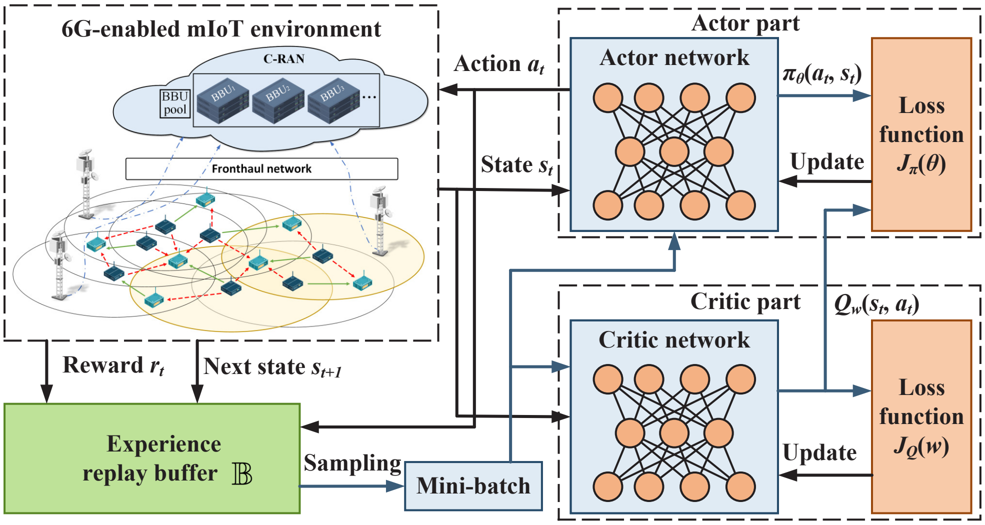

Fig. 5. A2C-based resource management framework.

The above formula represents the action a_t obtained by sampling based on all actions' probability distribution in the state s_t.

5) Value Function: The MDP model under consideration classifies the value functions into two distinct categories: the state-value function, denoted as $V\left(s_t\right)$, and the action-value function, also known as the Q-value, denoted as $Q\left(s_t, a_t\right)$. $V\left(s_t\right)$ represents the value for the current state s_t from an expectation perspective under policy π, and can be represented as

$$V\left(s_t\right) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\left[\sum_{\tau=0}^{T-t} \lambda^\tau r_{t+\tau} \middle| s_t\right], \tag{17}$$

where $\mathbb{E}\left[\cdot\right]$ denotes the expectation operation. The action-value function $Q\left(s_t, a_t\right)$ estimates the expected cumulative discounted rewards over time when commencing in the current state s_t and taking action a_t while the action selection follows the policy π. The action-value function $Q\left(s_t, a_t\right)$ is formulated as

$$Q\left(s_t, a_t\right) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\left[\sum_{\tau=0}^{T-t} \lambda^\tau r_{t+\tau} \middle| s_t, a_t\right], \tag{18}$$

where $V\left(s_t\right) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\left[Q\left(s_t, a_t\right)\right]$. Considering the Markov property, the action-value function through the Bellman equation can be rewritten as [30]

$$Q\left(s_t, a_t\right) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\left[r_t + \lambda\gamma \sum_{a \in \mathcal{A}} \pi\left(a|\, s_{t+1}\right) Q\left(s_{t+1}, a\right)\right]. \tag{19}$$

B. Actor-Critic Based Resource Management Method

In this paper, we propose two actor-critic based resource management algorithms to solve the above MDP model for 6G-enabled mIoT under large-scale overlapping interference scenarios.

1) Advantage Actor-Critic based Resource Management Algorithm: Fig. 5 shows the proposed A2C-based resource management framework. Its critic network can provide a value function to evaluate resource allocation results generated by the actor network. We use a parameterized function of θ to express the actor network, where θ is the stochastic policy's parameters $\pi_\theta\left(a_t|\, s_t\right)$. The parameter of action-value function $Q_w\left(s_t, a_t\right)$ is denoted by the function parameterized function w, which defines the critic network. To enhance the accuracy of the action-value function, the critic network is used to estimate the long-term reward for the state-action pair. In the proposed A2C-based resource management algorithm, the parameter update of the critic network is defined as

$$w \leftarrow w - \eta_c \nabla_w J_Q\left(w\right), \tag{20}$$

where η_c is the critic network's learning rate. With the experience replay mechanism, the critic network's loss function $J_Q\left(w\right)$ can be defined as [31]

$$J_Q\left(w\right) = \mathbb{E}_{\mathbb{B}}\left[y_t - Q_w\left(s_t, a_t\right)\right]^2, \tag{21}$$

where y_t is the target value, which can be expressed as [32]

$$y_t = r_t + \gamma \sum_{a \in \mathcal{A}} \pi_\theta\left(a|\, s_{t+1}\right) Q_w\left(s_{t+1}, a\right). \tag{22}$$

$\mathbb{B}$ is a experience replay buffer. The experience replay mechanism allows the RL agent to update the parameters of the critic network in an efficient manner. To accomplish this manner, the agent can make use of a random mini-batch sampling from $\mathbb{B}$, represented as $\left\{s_t, a_t, r_t, s_{t+1}\right\} \sim \mathbb{B}$. A discrepancy between the estimated value $Q_w\left(s_t, a_t\right)$ and target value y_t is known to be the temporal-difference error. Updating the critic network's parameters can be accomplished by taking the average value of a mini-batch of size I that is sampled from the experience replay buffer $\mathbb{B}$. Hence, the loss function's gradient $\nabla_w J_Q(w)$ is expressed by

$$\nabla_w J_Q(w) = \frac{1}{I}\sum_{i=1}^{I} \nabla_w \left(Q_w\left(s_t^i, a_t^i\right)\right)\left[y_t^i - Q_w\left(s_t^i, a_t^i\right)\right], \tag{23}$$

where index i referring to the i-th sample.

In the proposed algorithm, the actor part is used to take charge of policy evaluation. The actor part implements the policy gradient method to generate the parameterized policy. The actor–critic method's goal is to obtain the optimal policy of action for maximizing the expectation function $J_\pi(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)]$, or long-term reward from the standpoint of expectations, where $r(\tau) = \sum_{t=0}^{T} \lambda^t r_t$ represents the cumulative discount reward with restricted step. τ is the sampling trajectory. Hence, the update of the actor network's parameters θ can be defined as

$$\theta \leftarrow \theta - \eta_a \nabla J_\pi(\theta), \tag{24}$$

where η_a is an actor network's learning rate. For the actor network, the specific derivation of the policy gradient can be expressed as

$$\nabla_\theta J_\pi(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \left(\log \pi_\theta\left(a_t \middle| s_t\right)\right) A_w\left(s_t, a_t\right)\right], \tag{25}$$

where $A_w(s_t, a_t)$ denotes the advantage function. The advantage function $A_w(s_t, a_t)$ can significantly decrease the variance and boost the accuracy of the function approximation in the critic when utilized in the gradient calculation. The advantage function $A_w(s_t, a_t)$ can be defined as

$$A_w(s_t, a_t) = Q_w(s_t, a_t) - \sum_{a\in A} \pi_\theta\left(a \middle| s_t\right) Q_w(s_t, a). \tag{26}$$

By using the experience replay buffer $\mathbb{B}$, combining (25) and (26), the gradient of actor network's parameters $\nabla_\theta J_\pi(\theta)$ can be approximated as

$$\nabla_\theta J_\pi(\theta) = \frac{1}{I}\sum_{i=1}^{I}\left[\nabla_\theta \log \pi_\theta\left(a_t^i \middle| s_t^i\right) A_\omega\left(s_t^i, a_t^i\right)\right]. \tag{27}$$

The proposed A2C-based resource management algorithm utilizes two distinct deep neural networks, each with unique parameters, to depict the actor-critic networks. Concurrently, we update the parameters of both the actor network and critic network sequentially and simultaneously. The proposed A2C-based resource management algorithm, outlined in Algorithm 1, is formed by the combination of the actor network and critic network. Moreover, actor-critic methods in effectively learning parameterized stochastic policies and exhibiting favorable convergence properties have been confirmed in [33].

2) Asynchronous Advantage Actor-Critic based Resource Management Algorithm: To enhance network performance and minimize the learning process time, we propose the A3C-based resource management algorithm. The proposed A3C-based resource management algorithm employs an asynchronous multi-threaded architecture for improving system management performance. This A3C-based resource management algorithm is composed of a global actor-critic network and multiple workers as shown in Fig. 6. Each worker has its own local actor-critic networks, enabling them to interact independently with the environment. The parameters of the global actor-critic networks are shared between all thread-specific workers that are able to select an action depending on the current state in order to get a reward and progress to the next state of the environment. With the asynchronous multi-threaded architecture, the A3C-based resource management algorithm can train the actor network and critic network reliably.

Algorithm 1 A2C-based resource management algorithm.

```
Initialization:
    The variables of the environment,
    experience replay buffer $\mathbb{B}$,
    actor network's parameters $\theta$,
    critic network's parameters $w$;
for episode = 1 to $E_{\max}$ do
    Reset the 6G-enabled mIoT environment's state $s_0$;
    for time = 1 to $T$ do
        Agent executes the action $a_t$ according to $\pi_\theta(\cdot| s_t)$;
        Calculate the reward $r_t$ and obtain next state $s_{t+1}$;
        Put the tuple $\{s_t, a_t, r_t, s_{t+1}\}$ into replay buffer $\mathbb{B}$;
        Random sample a subset of $I$ tuples from $\mathbb{B}$;
        Network parameters updating:
         $\theta \leftarrow \theta - \eta_a \nabla_\theta J_\pi(\theta)$,
         $w \leftarrow w - \eta_c \nabla_w J_Q(w)$;
    end for
end for
return The parameters of actor-critic networks $\theta$ and $w$.
```

Subsequently, the global actor-critic network parameters are updated using the accumulated gradient, which are expressed as

$$\theta \leftarrow \theta - \eta_a d\theta, w \leftarrow w - \eta_c dw, \tag{28}$$

where $d\theta$ is the global actor network's accumulated gradient. dw is the global critic network's accumulated gradient. In the A3C-based resource management algorithm, the critic network's accumulated gradient for each worker is given by

$$dw \leftarrow dw + \left(y_t - Q_{w'}(s_t, a_t)\right)\nabla_{w'} Q_{w'}(s_t, a_t), \tag{29}$$

where w' denotes the critic network's parameters of the thread-specific worker. Moreover, to deal with the challenge represented by the actor-critic method in achieving a trade-off between exploration and exploitation, we adopt taking advantage of the entropy function $H(\pi_{\theta'}(s_i))$ to motivate exploration during training while avoiding premature convergence. In the A3C-based resource management framework, the actor network's parameter update for each worker is given by

$$d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi_{\theta'}\left(a_t \middle| s_t\right) A_{w'}(s_t, a_t) + \delta \nabla_{\theta'} H\left(\pi_{\theta'}(s_t)\right), \tag{30}$$

where θ' represents the actor network's parameters of the specific worker. δ denotes the intensity of the entropy regularization, which can adjust the trade-off between exploration and exploitation [34]. The proposed A3C-based resource management algorithm for the resource management of a 6G-enabled mIoT is presented in Algorithm 2, which is a centralized

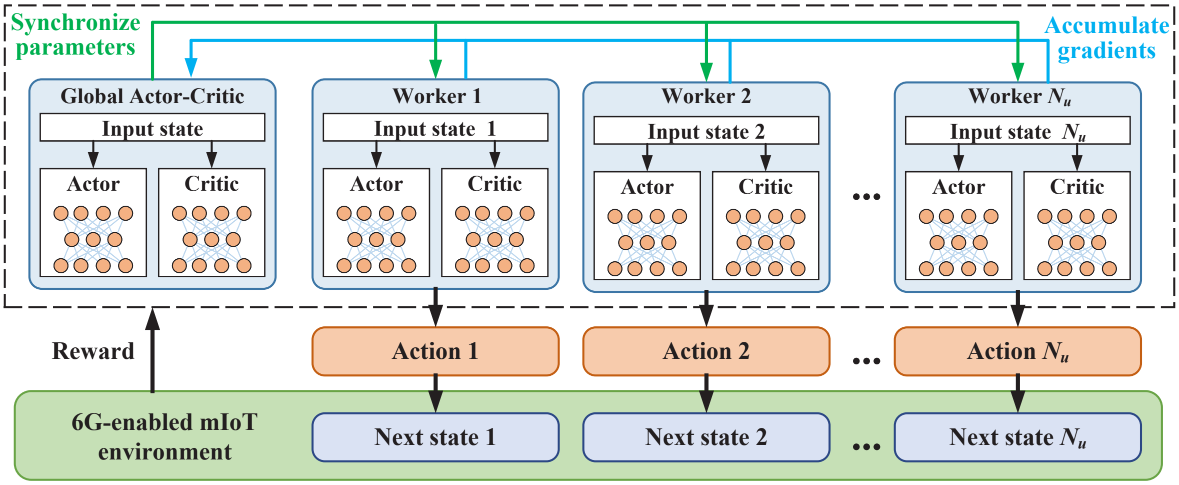

Fig. 6. A3C-based resource management framework.

framework and observations made for the entire 6G-enabled mIoT.

Algorithm 2 A3C-based resource management algorithm.

```
Initialization:
    The variables of the environment,
    the parameters of thread-specific networks $\theta'$ and $w'$,
    the parameters of global actor-critic networks $\theta$ and $w$;
for episode = 1 to $E_{\max}$ do
    for each thread-specific worker parallel do
        Initialize the accumulated gradients $dw$ and $d\theta$;
        Synchronize the parameters of worker's networks;
        Obtain the current environment state $s_t$;
        for $t$ =1 to $T$ do
            Execute action $a_t$ getting from $\pi_{\theta'}\left(\cdot|\, s_t\right)$;
            Calculate the reward $r_t$ and obtain state $s_{t+1}$;
        end for
        if $s_t$ is not terminal state then
            $r = \sum_{a\in\mathcal{A}} \pi_{\theta'}\left(a|\, s_t\right) Q_{w'}\left(s_t, a\right)$;
        else
            $r = 0$;
        end if
        for $i \in \{t-1, t-2, \cdots, 1\}$ do
            Obtain accumulate gradient for $\theta'$ with (30);
            Obtain accumulate gradient for $w'$ with (29);
        end for
        Asynchronous update the parameters $w$ and $\theta$;
        episode $\leftarrow$ episode $+ 1$;
    end for
end for
return Global networks' parameters $\theta$ and $w$.
```

C. Complexity Analysis

Under the assumption that the actor network and critic network individually consist of N_{a} and N_{c} full connected layers. In each episode, the neural networks' computational complexity is formulated as $O\left(\sum_{i=0}^{N_a-1} L_a^i L_a^{i+1} + \sum_{j=0}^{N_c-1} L_c^j L_c^{j+1}\right)$ [35], where L_c^j and L_a^i indicate the total amount of neurons at layer j in the critic network and the total amount of neurons at layer i in the actor network, respectively. Hence, the Algorithm 1 complexity can be formulated by $O\left(E_{\max} T \left(\sum_{n=0}^{N_a-1} L_a^n L_a^{n+1} + \sum_{n=0}^{N_c-1} L_c^n L_c^{n+1}\right)\right)$, where $E_{\max}$ is the number of episode as training steps. T denotes the total number of times in each episode. In addition, the Algorithm 2 complexity can be formulated by $O\left(E_{\max} T \left(\sum_{n=0}^{N_a-1} L_a^n L_a^{n+1} + \sum_{n=0}^{N_c-1} L_c^n L_c^{n+1}\right) \Big/ N_u\right)$ [36], where N_u is the thread-specific workers' number in this algorithm.

IV. Simulations Results

In this section, all simulation experiments were executed using Python 3.9.13, Pytorch 2.0.1, and NetworkX 3.2 implemented on a Dell Server with two Nvidia GeForce RTX 3080Ti GPUs, an Intel® Xeon® Gold 6242R CPU and 64GB memory. In our simulations, we assume that each IoT device experiences independent Rayleigh fast fading, and the noise variance on each user is the same. We consider the pathloss model as the model in 3GPP TR 38.901 UMi scenario [26]. The simulation results are presented to validate the network performance of the 6G-enabled mIoT. The performance of the proposed two algorithms (labeled as the proposed-A3C-based-algorithm and proposed-A2C-based-algorithm) is compared with two other algorithms, i.e., the proximal policy optimization (PPO)-based resource management algorithm [37] (labeled as the PPO-based-algorithm) and random-based resource management algorithm (labeled as the random-based-algorithm). Furthermore, the simulations contain various other parameters, which are outlined in Table II.

A. Convergence performance

Fig. 7 displays the summary statistics of the cumulative rewards achieved by the proposed-A2C-based-algorithm at four

TABLE II
SIMULATION PARAMETERS [17]

Parameters	Values
Total number of time	10
Transmission power	23 dBm
Time duration	1 s
Antenna height	1.5 m
Number of RBs	40
Antenna gain	3 dBi
RB bandwidth	1 MHz
Receiver noise figure	9 dB
Noise power spectral density	−174 dBm/Hz
Pathloss model	UMi scenario pathloss model
Discount factor	0.98
Fast fading	Rayleigh fading

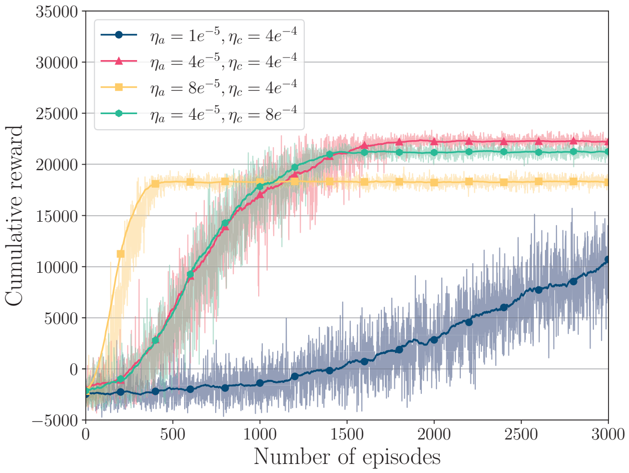

Fig. 7. Convergence of the proposed-A2C-based-algorithm under different learning rates.

different learning rates. The experiment was conducted with a total of 500 IoT devices and a communication range radius of 200 m. From this figure, we can see that the cumulative reward is improving as the number of episodes increases when the agent learns the 6G-enabled mIoT environment. When the critic learning rate is fixed as $4e^{-4}$, increasing the actor learning rate can speed up the convergence of the proposed-A2C-based-algorithm to a certain extent, but the system hardly continues to explore higher cumulative reward. As shown in Fig. 7, the convergence performance of the proposed-A2C-based-algorithm is fastest when its actor learning rate is $8e^{-5}$. It converges when the needed number of episodes is less than 500, but its convergence result value is less than the result of the actor learning rate being $4e^{-5}$. And when the learning rate is $1e^{-5}$, it cannot converge at the end of the entire training process. In addition, when the actor learning rate is fixed as $4e^{-4}$, increasing the critic learning rate as $8e^{-4}$ does not make it faster convergence, and the obtained convergence results are reduced. Hence, the proposed-A2C-based-algorithm's convergence performance is demonstrated and has a higher reward when actor learning rate $\eta_a = 4e^{-5}$ and critic learning rate $\eta_c = 4e^{-4}$.

In Fig. 8, it shows the cumulative reward of different algorithms as the number of episodes increases where the number of IoT devices is 500 and the radius of the IoT devices' communication range is 200 m. We set the learning rates of the PPO-based-algorithm, proposed-A2C-based-algorithm and proposed-A3C-based-algorithm to be the same, i.e., their actor learning rate $\eta_{\mathrm{a}} = 4e^{-5}$ and critic learning rate $\eta_{\mathrm{c}} = 4e^{-4}$ according to the result of Fig. 8. Since the random-based-algorithm cannot learn from the environment, it cannot increase the cumulative reward as the number of episodes increases. Compared with the PPO-based-algorithm, since the proposed two algorithms adopt a stochastic policy mechanism to sample action, they can explore higher reward results and converge faster in Fig. 8. In addition, the proposed-A3C-based-algorithm adopts an asynchronous multi-threaded architecture that can explore more reward results through multiple workers and the parallel architecture can speed up the convergence speed than the proposed-A2C-based-algorithm.

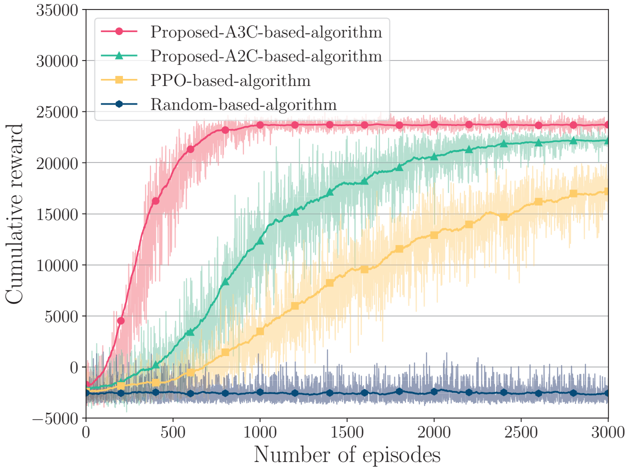

Fig. 8. Convergence performance under different algorithms.

B. Network performance

To demonstrate the advantages of the proposed two algorithms, Fig. 9 compares the network throughput of the 6G-enabled mIoT system employing various algorithms, where the x-axis indicates the five different numbers of IoT devices and the radius of IoT devices' communication range is 200 m. As shown in Fig. 9, the network throughput of the random-based algorithm, PPO-based algorithm, and the proposed two algorithms increases as the number of IoT devices increases Since the presence of overlapping areas in the 6G-enabled mIoT, the random-based algorithm can not avoid the overlapping interference generating and then results in reduced network throughput. The PPO-based-algorithm can effectively manage resources for the entire 6G-enabled mIoT, which reduces the overlapping interference when the spectrum resources are limited and then can enhance the network throughput. When the interference increases due to the increased number of IoT devices, the network performance of the PPO-based-algorithm is gradually better than that of the random-based-algorithm. Through the incentive feedback mechanism based on the hypergraph interference model deployed in the RL, the network throughput of the proposed two algorithms can be maximized. The proposed-A2C-based-algorithm and

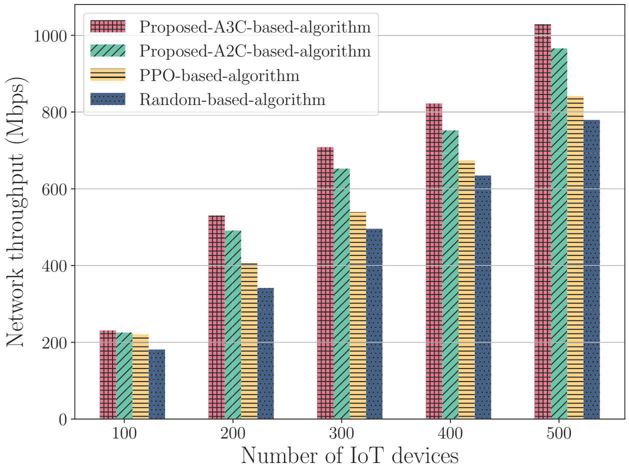

Fig. 9. Network throughput under different numbers of IoT devices.

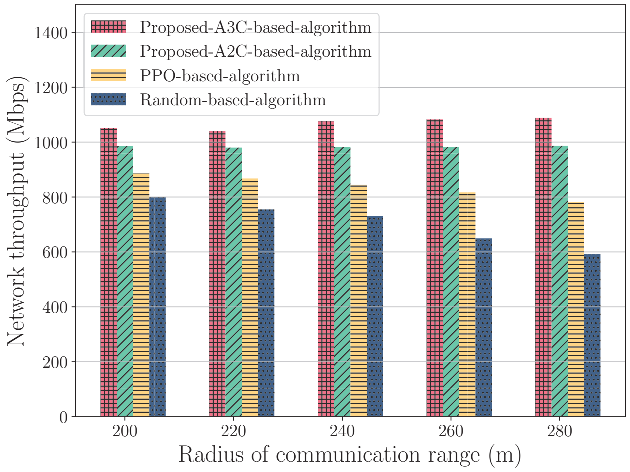

Fig. 11. Network throughput under different communication ranges.

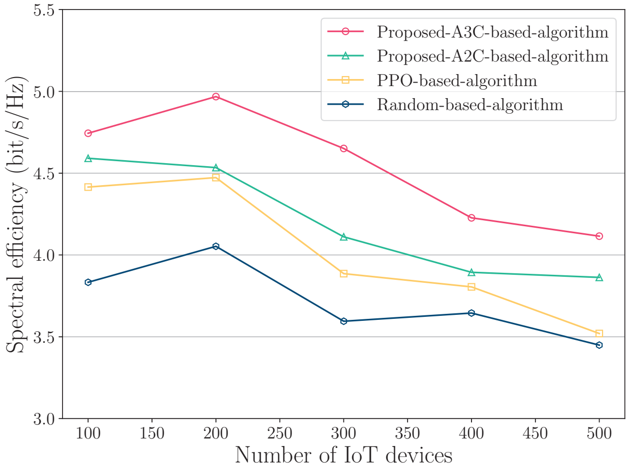

Fig. 10. Spectral efficiency under different numbers of IoT devices.

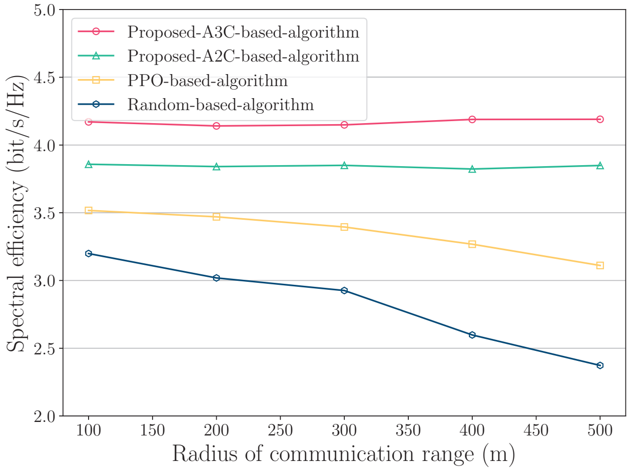

Fig. 12. Spectral efficiency under different communication ranges.

proposed-A2C-based-algorithm can allocate RBs dynamically according to the environment of the 6G-enabled mIoT, whereas the proposed-A3C-based-algorithm has the highest network throughput of more than 1000 Mbps when the numbers of IoT devices is 500.

Fig. 10 illustrates the relation between the spectral efficiency and the number of IoT devices where the radius of IoT devices' communication range is 200 m. Results indicate that with the increase in the number of IoT devices, the spectral efficiency of all algorithms is reducing. Due to the large-scale overlapping area among IoT devices' communication range and the overlapping interference caused by massive devices in the random-based-algorithm, the spectral efficiency is lower than that of other algorithms. The PPO-based-algorithm has more effective resource management due to its learning ability, and so the spectral efficiency of the PPO-based-algorithm is higher than the random-based-algorithm. Compared with the PPO-based-algorithm, the proposed-A2C-based-algorithm and proposed-A3C-based-algorithm can dynamically manage spectrum resources under limited resources in the 6G-enabled mIoT, due to their used randomness policy which can explore higher spectral efficiency. In addition, the proposed-A3C-based-algorithm adopting the asynchronous multi-threaded architecture can further explore more resource allocation strategies to obtain higher spectral efficiency than the proposed-A2C-based-algorithm. Compared with the other algorithms, when the number of IoT devices is 200, the spectral efficiency of the proposed-A3C-based-algorithm is achieved by close to 5.0 bit/s/Hz.

As shown in Fig. 11, the x-axis indicates five different radius of IoT devices' communication range to the IoT device in a 6G-enabled mIoT where the number of IoT devices is 500. In Fig. 11, an increase in the radius of IoT devices' communication range leads to an overall reduction in network throughput for both the random-based algorithm and PPO-based algorithm. This is because a gradual expansion in the communication range will further lead to an increase in the overlapping area and cause more serious large-scale overlapping interference for the 6G-enabled mIoT. The random-based algorithm is ineffective in managing the growth in overlapping interference, leading to a progressive drop in network throughput. As a result, its performance has consistently been poorer than other algorithms. Compared with the PPO-based-algorithm, the two proposed algorithms can effectively avoid the occurrence of overlapping interference and ensure stable network throughput of the system due to adopting a hypergraph interference model. As the communication range expands, the network throughput of the two proposed algorithms is always higher than the PPO-based-algorithm.

Fig. 12 illustrates the radius of IoT devices' communication range versus the spectral efficiency where the number of IoT devices is 500. As the communication range expands, the spectral efficiency of the random-based-algorithm and PPO-based-algorithm are on a downward trend. From this figure, due to PPO-based-algorithm learning the 6G-enabled mIoT environment, we can see that the spectral efficiency of the PPO-based-algorithm and the proposed algorithms are improving more than the random-based-algorithm. By adopting the stochastic policy mechanism, the proposed algorithms' spectral efficiency can obtain the resource allocation result with higher spectral efficiency. Compared with the PPO-based-algorithm, the proposed algorithms can overcome the impact of overlapping interference caused by the expanded communication range by adopting a hypergraph interference model to design a reward function, which more effectively avoids the occurrence of overlapping interference. Furthermore, compared with the proposed-A2C-based-algorithm, the proposed-A3C-based-algorithm can further explore more resource allocation results by using multiple workers due to asynchronous multi-threaded architecture and then obtain the resource allocation result with higher spectral efficiency.

V. Conclusion

In this paper, we designed the novel hypergraph interference model and proposed RL-based resource management algorithms to solve the resource management problem of the 6G-enabled mIoT under the large-scale overlapping interference scenario. Considering the characteristics of overlapping interference, the relationship between overlapping coverage area and overlapping interference is analyzed, and the interference degree of the entire network is calculated through the hypergraph interference model. Then, to solve the problem, we build an MDP model based on a hypergraph interference model for the 6G-enabled mIoT. Finally, we propose the A2C-based resource management algorithm and A3C-based resource management algorithm to solve the MDP model, where the A3C-based resource management algorithm uses an asynchronous multi-threaded architecture to improve learning speed and obtain higher network performance. Simulation results verify the correctness of theoretical results and show that our proposed algorithms outperform other algorithms. The work of this paper provides a reference for the study of the resource management problem with large-scale overlapping interference.

References

[1] B. Qian, H. Zhou, T. Ma, K. Yu, Q. Yu, and X. Shen, "Multi-operator spectrum sharing for massive iot coexisting in 5g/b5g wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 881–895, 2021.

[2] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek, "Analyzing random access collisions in massive iot networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6853–6870, 2018.

[3] H. S. Jang, B. C. Jung, T. Q. S. Quek, and D. K. Sung, "Resource-hopping-based grant-free multiple access for 6g-enabled massive iot networks," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15 349–15 360, 2021.

[4] F. Yang, C. Yang, J. Huang, K. Yu, S. Garg, and M. Alrashoud, "Hypergraph-based resource-efficient collaborative reinforcement learning for b5g massive iot," *IEEE open j. Commun. Soc.*, vol. 4, pp. 2439–2450, 2023.

[5] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Enabling massive iot toward 6G: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11 891–11 915, 2021.

[6] I. N. A. Ramatryana, G. B. Satrya, and S. Y. Shin, "Adaptive traffic load in irsa-noma prioritizing emergency devices for 6g enabled massive iot," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2713–2717, 2021.

[7] Y. Han, L. Liu, L. Duan, and R. Zhang, "Towards reliable uav swarm communication in d2d-enhanced cellular networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1567–1581, 2021.

[8] T.-P. Chu and S. Rappaport, "Overlapping coverage with reuse partitioning in cellular communication systems," *IEEE Trans. Veh. Technol.*, vol. 46, no. 1, pp. 41–54, 1997.

[9] R. Borralho, A. Quddus, A. Mohamed, P. Vieira, and R. Tafazolli, "Coverage and data rate analysis for a novel cell-sweeping-based ran deployment," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[10] J. Huang, S. Zhang, F. Yang, T. Yu, L. V. N. Prasad, M. Guduri, and K. Yu, "Hypergraph-based interference avoidance resource management in customer-centric communication for intelligent cyber-physical transportation systems," *IEEE Trans. Consum. Electron.*, pp. 1–1, 2023.

[11] C.-C. Lai, L.-C. Wang, and Z. Han, "The coverage overlapping problem of serving arbitrary crowds in 3d drone cellular networks," *IEEE Trans. Mob. Comput.*, vol. 21, no. 3, pp. 1124–1141, 2022.

[12] R. M. Radaydeh, "Distributed d2d resource allocation for reducing interference in ultra-dense networks with generic imperfection," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023, pp. 0029–0034.

[13] S. Yan, M. Peng, and X. Cao, "A game theory approach for joint access selection and resource allocation in uav assisted iot communication networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1663–1674, 2019.

[14] A. Kaur, K. Kumar, A. Prakash, and R. Tripathi, "Imperfect csi-based resource management in cognitive iot networks: A deep recurrent reinforcement learning framework," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 9, no. 5, pp. 1271–1281, 2023.

[15] H. Yang, W.-D. Zhong, C. Chen, A. Alphones, and X. Xie, "Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive internet of things," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5677–5689, 2020.

[16] F. Zhang, G. Han, L. Liu, M. Martínez-García, and Y. Peng, "Joint optimization of cooperative edge caching and radio resource allocation in 5g-enabled massive iot networks," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 14 156–14 170, 2021.

[17] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient mec in noma-enabled massive iot networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1015–1027, 2021.

[18] J. Zhao, S. Xu, and D. Li, "Qoe driven resource allocation in massive iot: A deep reinforcement learning approach," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[19] S. A. Gbadamosi, G. P. Hancke, and A. M. Abu-Mahfouz, "Interference avoidance resource allocation for d2d-enabled 5g narrowband internet of things," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22 752–22 764, 2022.

[20] S. S. Sarma, R. Hazra, and A. Mukherjee, "Symbiosis between d2d communication and industrial iot for industry 5.0 in 5g mm-wave cellular network: An interference management approach," *IEEE Trans. Ind. Inf.*, vol. 18, no. 8, pp. 5527–5536, 2022.

[21] J. Wang, C. Jiang, Z. Wei, C. Pan, H. Zhang, and Y. Ren, "Joint uav hovering altitude and power control for space-air-ground iot networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1741–1753, 2019.

[22] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang, "Energy-efficient resource allocation for d2d communications underlaying cloud-ran-based lte-a networks," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 428–438, 2016.

[23] Z. Gao, M. Ke, Y. Mei, L. Qiao, S. Chen, D. W. K. Ng, and H. V. Poor, "Compressive sensing-based grant-free massive access for 6g massive communication," *IEEE Internet Things J.*, pp. 1–1, 2023.

[24] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, “Joint precoding and rrh selection for user-centric green mimo c-ran,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2891–2906, 2017.

[25] A. Ebrahim and E. Alsusa, “Interference and resource management through sleep mode selection in heterogeneous networks,” *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 257–269, 2017.

[26] 3GPP, “Study on channel model for frequencies from 0.5 to 100 ghz,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.901, Jun. 2018, version 15.0.0.

[27] M. Elnourani, S. Deshmukh, and B. Beferull-Lozano, “Distributed resource allocation in underlay multicast d2d communications,” *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3409–3422, 2021.

[28] H. Liang, S. Zhou, X. Liu, F. Zheng, X. Hong, X. Zhou, and L. Zhao, “A dynamic resource allocation model based on smdp and drl algorithm for truck platoon in vehicle network,” *IEEE Internet Things J.*, vol. 9, no. 12, pp. 10 295–10 305, 2022.

[29] C. Shang, Y. Sun, H. Luo, and M. Guizani, “Computation offloading and resource allocation in noma–mec: A deep reinforcement learning approach,” *IEEE Internet Things J.*, vol. 10, no. 17, pp. 15 464–15 476, 2023.

[30] F. Tang, Y. Zhou, and N. Kato, “Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5g hetnet,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2773–2782, 2020.

[31] Y. Cao, H. Wang, D. Li, and G. Zhang, “Smart online charging algorithm for electric vehicles via customized actor–critic learning,” *IEEE Internet Things J.*, vol. 9, no. 1, pp. 684–694, 2022.

[32] H. Dong, H. Dong, Z. Ding, S. Zhang, and Chang, *Deep Reinforcement Learning*. Springer, 2020.

[33] Y. Wei, F. R. Yu, M. Song, and Z. Han, “Joint optimization of caching, computing, and radio resources for fog-enabled iot using natural actor–critic deep reinforcement learning,” *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2061–2073, 2019.

[34] J. Du, W. Cheng, G. Lu, H. Cao, X. Chu, Z. Zhang, and J. Wang, “Resource pricing and allocation in mec enabled blockchain systems: An a3c deep reinforcement learning approach,” *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 33–44, 2022.

[35] L. Li, L. Tang, Q. Liu, Y. Wang, X. He, and Q. Chen, “Handoff control and resource allocation for ran slicing in iot based on dtn: An improved algorithm based on actor–critic framework,” *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13 370–13 384, 2023.

[36] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y.-C. Liang, “Intelligent resource scheduling for 5g radio access network slicing,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7691–7703, 2019.

[37] X. He, Y. Mao, Y. Liu, P. Ping, Y. Hong, and H. Hu, “Channel assignment and power allocation for throughput improvement with ppo in b5g heterogeneous edge networks,” *Digit. Commun. Netw.*, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864823000536

Jie Huang (Member, IEEE) received the B.E. degree in communication engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2011 and Ph.D. degree from Chongqing University, Chongqing, China in 2017. He is currently a lecture at Chongqing University of Technology. His interests include wireless communication systems, cognitive radio networks and the next-generation mobile communication.

Cheng Yang (Graduate Student Member, IEEE) received the B.E. degree in electronic engineering from Chongqing University of Technology, Chongqing, China, in 2022. He is currently pursuing a master’s degree at the School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing, China. His research interests include wireless communication resource allocation and next-generation mobile communication.

Shilong Zhang (Graduate Student Member, IEEE) received the B.E. degree from the Department of Electronic Engineering, Hunan Institute of Technology, China, in 2020. He is currently pursuing a master’s degree at the School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing, China. His interests include wireless communication resource allocation and next-generation mobile communication.

Fan Yang (Member, IEEE) received the B.E, M.S and Ph.D. degree from Chongqing University, Chongqing, China in 2006, 2010 and 2018 respectively. He is currently an associate Professor with the School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing, China. He served many Telecom companies and institutes in China. His research interests include adaptive transmission in wireless communication systems, mobile broadband wireless networks, the next generation mobile communication hypergraph theory and resource management.

Osama Alfarraj received the master’s and Ph.D. degrees in information and communication technology from Griffith University, in 2008 and 2013, respectively. He is currently a Professor of computer sciences with King Saudi University, Riyadh, Saudi Arabia. His current research interests include eSystems (eGov, eHealth, and ecommerce), cloud computing, and big data. For two years, he has served as a Consultant and a member for the Saudi National Team for Measuring E-Government, Saudi Arabia.

Valerio Frascolla (Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering. He is currently the Director of Research and Innovation with Intel, Munich. He has expertise in wireless systems architecture and protocols design, requirements management, standards bodies attendance, and project/program management. He is a mentor and a coach. He has been working in different roles with Ancona University, Comneon, Infineon, as a reviewer for the European Commission and an Evaluator of the Portuguese and the Romanian Science Foundations. He is in the advisory board of six research projects and contributed to other 17. He is the author of more than 70 publications. His main research interests include 5G and beyond system design, with a focus on spectrum management, AI, and edge technologies. He is a member of the Board of Directors of the BDVA Association. He serves as a reviewer for more than 30 journals. He has participated in the TPC of more than 80 conferences and has a track record as an organizer of special sessions, workshops, and panels at main international venues. He serves as the chair for several workgroups in European associations.

Shahid Mumtaz (Senior Member, IEEE) is a Professor with Nottingham Trent University (NTU), U.K. He authorizes four technical books, 12 book chapters, and 300+ technical papers (200+ IEEE Journals/transactions, 100+ conferences, two IEEE best paper awards) in mobile communications. Most of his publication is in the field of Wireless Communication. He is an IET Fellow, Founder, and EiC of IET “Journal of Quantum Communication,” Vice-Chair: Europe/Africa Region- IEEE ComSoc: Green Communications & Computing Society. He is a Scientific Expert and Evaluator for various research funding agencies. In 2012, he was awarded an “Alain Bensoussan fellowship.” China awarded him the young scientist fellowship in 2017.

Keping Yu (Senior Member, IEEE) received the M.E. and Ph.D. degrees from the Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan, in 2012 and 2016, respectively. He was a Research Associate, Junior Researcher, Researcher with the Global Information and Telecommunication Institute, Waseda University, from 2015 to 2019, 2019 to 2020, 2020 to 2022, respectively. He is currently an Associate Professor, the Vice Director of Institute of Integrated Science and Technology, and the Director of the Network Intelligence and Security Laboratory (YU Lab), Hosei University, Japan. He is also a Fellow of the Distinguished Scientist Fellowship Program at King Saud University, Saudi Arabia.

Dr. Yu has hosted and participated in more than ten projects, is involved in many standardization activities organized by ITU-T and ICNRG of IRTF, and has contributed to ITU-T Standards Y.3071 and Supplement 35. He has been a Highly Cited Researcher identified by Clarivate™ (2023) and the World's Top 2% Scientists identified by Stanford University (2022, 2023) . He received the Best Symposium Award from IWCMC 2023, the IEEE Outstanding Leadership Award from IEEE BigDataSE 2021, the Best Paper Award from IEEE Consumer Electronics Magazine Award 2022 (1st Place Winner), IEEE ICFTIC 2021, ITU Kaleidoscope 2020, the Student Presentation Award from JSST 2014. He has authored more than 200 peer-review research papers and books, including over 80 IEEE/ACM Transactions papers. He is an Associate Editor of IEEE Open Journal of Vehicular Technology, Journal of Intelligent Manufacturing, Journal of Circuits, Systems and Computers, and IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. He has been a Guest Editor for IEEE Transactions on Computational Social Systems, IEEE Journal of Biomedical and Health Informatics, and Renewable & Sustainable Energy Reviews. He served as general co-chair and publicity co-chair of the IEEE VTC2020-Spring 1st EBTSRA workshop, general co-chair of IEEE ICCC2020 2nd EBTSRA workshop, general co-chair of IEEE TrustCom2021 3nd EBTSRA workshop, session chair of IEEE ICCC2020, ITU Kaleidoscope 2016.