electronics

MDPI

Article

Reinforcement Learning-Based Resource Allocation and Energy Efficiency Optimization for a Space–Air–Ground-Integrated Network

Zhiyu Chen [1], Hongxi Zhou [1], Siyuan Du [2,*], Jiayan Liu [2], Luyang Zhang [2] and Qi Liu [3]

[1] State Grid Information & Telecommunication Branch, Beijing 100761, China; zhiyu-chen@sgcc.com.cn (Z.C.); hongxizhou@sgcc.com.cn (H.Z.)
[2] School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China; liujiayan@ncepu.edu.cn (J.L.); 17755675272@163.com (L.Z.)
[3] Beijing Fibrlink Communications Co., Ltd., Beijing 100071, China; liuqi1@sgitg.sgcc.com.cn
* Correspondence: siiyuan_du@163.com

Abstract: With the construction and development of the smart grid, the power business puts higher requirements on the communication capability of the network. In order to improve the energy efficiency of the space–air–ground-integrated power three-dimensional fusion communication network, we establish an optimization problem for joint air platform (AP) flight path selection, ground power facility (GPF) association, and power control. In solving the problem, we decompose the problem into two subproblems, one is the AP flight path selection subproblem and the other is the GPF association and power control subproblem. Firstly, based on the GPF distribution and throughput weights, we model the AP flight path selection subproblem as a Markov Decision Process (MDP) and propose a multi-agent iterative optimization algorithm based on the comprehensive judgment of GPF positions and workload. Secondly, we model the GPF association and power control subproblem as a multi-agent, time-varying K-armed bandit model and propose an algorithm based on multi-agent Temporal Difference (TD) learning. Then, by alternately iterating between the two subproblems, we propose a reinforcement learning (RL)-based joint optimization algorithm. Finally, the simulation results indicate that compared to the three baseline algorithms (random path, average transmit power, and random device association), the proposed algorithm improves an overall energy efficiency of the system of 16.23%, 86.29%, and 5.11% under various conditions (including different noise power levels, GPF bandwidth, and GPF quantities), respectively.

Keywords: space–air–ground-integrated network (SAGIN); Low Earth Orbit (LEO) satellites; dynamic resource allocation; multi-agent reinforcement learning (RL); Markov Decision Process (MDP); K-armed bandit

check for updates

Citation: Chen, Z.; Zhou, H.; Du, S.; Liu, J.; Zhang, L.; Liu, Q. Reinforcement Learning-Based Resource Allocation and Energy Efficiency Optimization for a Space–Air–Ground-Integrated Network. *Electronics* **2024**, *13*, 1792. https://doi.org/10.3390/electronics13091792

Academic Editor: Christos J. Bouras

Received: 11 April 2024
Revised: 30 April 2024
Accepted: 3 May 2024
Published: 6 May 2024

1. Introduction

Against the backdrop of the widespread application of the Fifth Generation (5G) mobile communication technology, the Sixth Generation (6G) mobile communication technology has gradually become a research hotspot for scholars around the world, and the research direction in this field has demonstrated significant commercial potential, attracting extensive attention from the industrial sector, including the State Grid Corporation of China [1]. With the rapid development of new power systems and the widespread coverage of the smart grid, the flexible and random access of massive and diverse entities put higher requirements on power communication. Different business requirements arise in a variety of typical power application scenarios, many of which generate a large amount of real-time data that need to be uploaded to cloud server data centers promptly for grid regulation and control [2,3].

In the blind area of traditional network coverage or power emergency communication scenarios especially, when ground networks experience damage, faults, or lack signal coverage, the ground network data collection methods may fail to meet the data transmission requirements of GPFs. The establishment of the space–air–ground-integrated power three-dimensional converged communication network, as one of the technological visions for 6G network design, can effectively supplement the existing ground power communication systems and provide reliable support for power data transmission. It will become an effective solution for achieving secure, reliable, and flexible transmission of power business data in the future.

The SAGIN can be divided into three parts: a space-based network, an air-based network, and a ground-based network [4]. In the space-based network, satellite constellation systems led by LEO satellites [5] can achieve seamless and extensive coverage and provide cloud server data processing services for GPFs. In the air-based network, auxiliary communication airborne platforms, such as Unmanned Aerial Vehicles (UAVs), can serve as amplifiers in satellite–ground communication, completing operations related to power amplification and forwarding signals. They can also act as microbase stations for the temporary storage and processing of data.

To further enhance the application potential of the SAGIN in power emergency communication scenarios, in this paper, we propose an algorithm based on RL for the joint optimization of AP trajectories, GPF associations, and power control strategies under classification-based throughput constraints. The algorithm aims to ensure the completion of communication tasks and maximize the long-term overall system energy efficiency. The main contributions of this paper are as follows:

(1) We integrate APs with the SAGIN to categorize the communication requirements of the GPF into real-time data transmission and non-real-time data transmission. For real-time data transmission, the priority is to maximize system throughput and ensure that communication latency is low enough for the data from the GPFs to be relayed by the APs and eventually uploaded to the cloud server on the LEO satellite for processing. For non-real-time data transmission, the goal is to maximize overall system energy efficiency by storing as much data as possible in the APs. After the communication task is completed, the APs fly to the corresponding ground management facility for data unloading.

(2) To solve the optimization problem, this paper employs an RL algorithm that models the AP flight paths as an MDP and proposes a multi-agent iterative optimization algorithm based on comprehensive assessments of GPF positions and communication workload. The GPF associations and power control are modeled as a multi-agent K-armed bandit problem, and we propose an algorithm based on multi-agent Temporal Difference learning to solve this aspect. The two algorithms alternate iterations, ultimately solving for maximum overall system energy efficiency.

(3) We verify the performance of the algorithms through simulation, and the results indicate that the proposed algorithm outperforms several benchmark algorithms, and it is effective in improving long-term overall system energy efficiency.

The remainder of this paper is organized as follows. Section 2 introduces the related work in recent years. Then, Section 3 introduces the system model and problem formulation. Section 4 presents the solution approach, and Section 5 verifies the algorithm performance through simulation. Finally, Section 6 provides conclusions and future prospects.

2. Related Work

Currently, research on the SAGIN has made significant progress and produced valuable results. In terms of satellite-fused communication, reference [6] proposed using satellites as cloud servers to provide remote computing services for ground users and jointly optimized communication and computing resources to minimize overall system energy consumption. Reference [7] investigated the feasibility of seamless and efficient connections between ground communication systems and satellite networks. In terms of UAV trajectory optimization, reference [8] proposed an algorithm that utilizes the successive convex approximation (SCA) technique to optimize UAV trajectories and minimize overall system energy consumption. Reference [9] introduced an algorithm to discretize the

UAV and user movement process in a multi-UAV scenario to maximize the minimum user transmission rate. Reference [10] presented a trajectory optimization algorithm for minimizing the number of UAVs deployed in a vehicular network by solving the optimization problem using the SCA technique. Reference [11] proposed an algorithm using RL with the Q-learning algorithm to optimize the trajectory of a single UAV to minimize average communication latency.

In addition to the research directions mentioned in the above literature, another focus is on optimizing device association and power control strategies to enhance system performance and communication quality. Reference [12] proposed an algorithm based on block coordinate descent (BCD) and SCA theory for the joint optimization of device association, UAV trajectories, and power control to maximize the minimum user rate. Reference [13] described an energy consumption minimization algorithm that jointly optimized user association matching and power allocation using BCD and SCA theory. In the aforementioned studies, optimizations were conducted considering parameters such as energy consumption, throughput, minimum user rate, and average communication latency in communication systems. However, the analyses were limited to models with a single constraint and lacked consideration of specific business communication requirements in practical scenarios. For instance, in power emergency communication scenarios, different services may have varying communication throughput requirements while considering system energy consumption. Consequently, a comprehensive optimization of system efficiency considering both throughput and energy consumption becomes an effective solution. Reference [14] proposed an algorithm utilizing the genetic algorithm and simulated annealing to separately optimize device association and power selection to maximize the overall system energy efficiency. Reference [15] presented a method for power allocation and wireless backhaul bandwidth allocation under specific quality of service (QoS) constraints, aiming to maximize the system energy efficiency of downlink heterogeneous networks.

Most of the resource allocation optimization algorithms mentioned above employ traditional heuristic algorithms or SCA techniques based on BCD, requiring multiple iterations for solving, leading to a significant increase in the computational complexity of the system. Additionally, as the network scale increases, the high-speed dynamic characteristics of the SAGIN result in continuous changes in wireless environment parameters, and the adaptability of algorithms becomes a challenge. Considering the use of RL algorithms to solve the resource allocation problem as an effective solution, reference [16] proposed an algorithm utilizing RL to jointly optimize UAV paths, user device selection, and power allocation, ensuring fairness in user throughput. Reference [17] employed the Q-learning algorithm to jointly optimize UAV paths, user selections, and power allocation to maximize network capacity. Reference [18] proposed a power control scheme based on the Deep Q Network (DQN) to improve the system-level energy efficiency of two-layer 5G heterogeneous and multi-channel cells. Reference [19] proposed an innovative algorithm generalization method based on incremental reinforcement learning to enable UAVs to adjust their control strategies in dynamic environments. Reference [20] proposed a new multi-agent recurrent deterministic policy gradient algorithm to control the navigation actions of multiple UAVs. Reference [21] introduced the state-of-the-art technologies for UAV-assisted maritime communications, discussed the physical layer, resource management, cloud/edge computing, and caching UAV-assisted solutions in maritime environments and grouped them according to their performance objectives. Reference [22] presents a deep reinforcement learning-based method based on the design of a multi-head heterogeneous attention mechanism to facilitate the learning of a policy that automatically and sequentially constructs the route, while taking energy consumption into account.

There are also some highly effective research results in further integrating the SAGIN system model. Reference [23] made a proposal to leverage LEO satellites and APs to provide edge computing services for ground users. In reference [24], on the other hand, transmitted user tasks to cloud servers on LEO satellites for processing through UAV relay were proposed. The aforementioned literature focuses on the relevant research of UAV trajectory and user resource allocation algorithms in the SAGIN, but joint optimization research specifically for emergency communication scenarios in the power industry is still limited.

3. System Model and Problem Modeling

In this section, we analyze the business requirements of GPFs in the network and construct the space–air–ground-integrated power three-dimensional fusion communication network system model applied to power emergency scenarios. Then, we propose a communication model and an energy consumption model. Finally, the optimization problem is modeled.

3.1. System Model

We consider the uplink transmission in the wireless communication system depicted in Figure 1. The system comprises a single LEO satellite, and N GPFs serve as users and M auxiliary communication airborne platforms serve as APs. In this configuration, the GPFs initiate power data transmission tasks, and the LEO satellite s, functioning as a space-based cloud server base station, replaces ground stations to receive signals.

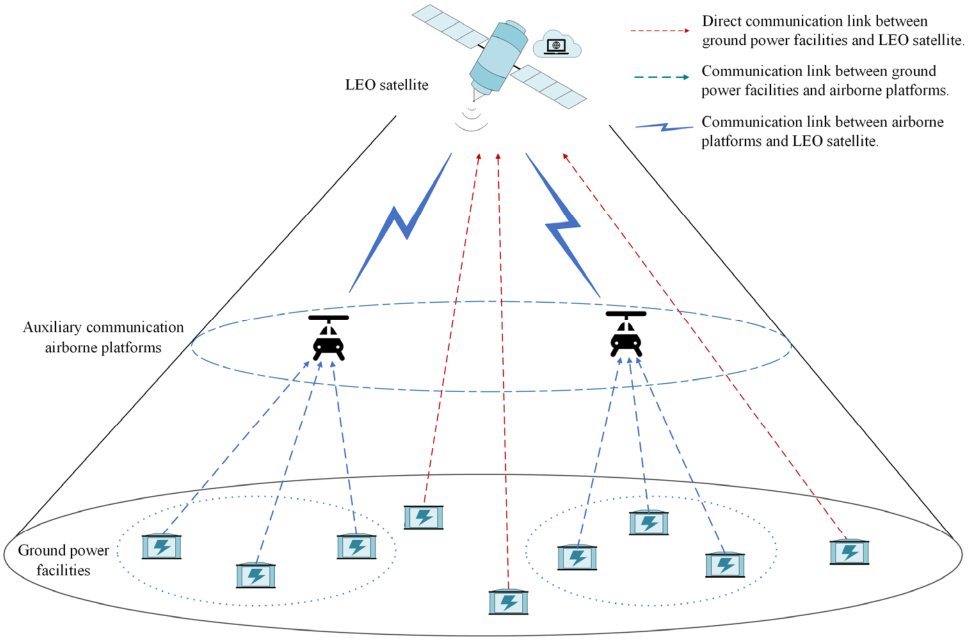
LEO satellite
Direct communication link between ground power facilities and LEO satellite.
Communication link between ground power facilities and airborne platforms.
Communication link between airborne platforms and LEO satellite.
Auxiliary communication airborne platforms
Ground power facilities

Figure 1. System model.

Some GPFs periodically generate a substantial volume of real-time power business data, which is time sensitive and needs to be immediately transmitted to the cloud server within the 5G base station for processing. However, in power emergency scenarios where public network communication facilities are damaged or malfunctioning, the 5G base station signals cannot cover the area. So, LEO satellites are employed as substitutes for ground stations to fulfill the corresponding communication services. Due to the considerable distance between LEO satellites and the Earth's surface, the direct connection speed of the GPF is relatively low, and the communication delay may not meet the needs of power businesses. To address this limitation, APs are utilized to relay and amplify certain signals, thereby enhancing communication transmission rates.

Additionally, some GPFs generate data involving non-real-time business operations that require large amounts of data transmission. Since this type of business operation is insensitive to time constraints, and considering efficiency and communication resource optimization, the data can be temporarily offloaded on APs. After completing communication tasks, the APs can fly to the nearest 5G ground station cloud server for data processing. In the system model tailored to the data transfer requirements of various business types, APs serve a dual role. They act as relay nodes for ground-to-satellite communication, facilitating signal forwarding [9]. Simultaneously, APs can function as short-term data storage devices, providing diversified services to meet the data transfer needs of GPFs [8].

N GPFs are represented by the set $\mathcal{N} = \{1, \cdots, n, \cdots, N\}$, while M APs are denoted by the set $\mathcal{M} = \{1, \cdots, m, \cdots, M\}$. The system employs a quasi-static model, dividing a coverage cycle into T time slots of length Δt, with $\mathcal{T} = \{1, \cdots, t, \cdots, T\}$. It is assumed that the positions of the GPFs are known and remain approximately constant throughout the task duration. The location of GPF n is represented as $q_n = [x_n, y_n]^T$. The APs fly at a fixed height of approximately H_a above the ground, with negligible height differences between them to prevent collisions. Within each time slot, the horizontal position of AP m is considered constant and denoted as $q_m(t) = [x_m(t), y_m(t)]^T$.

3.2. Communication Model and Energy Consumption Model

In this paper, Frequency Division Multiple Access (FDMA) technology is employed for communication to prevent interference between signals from different devices [15]. Binary variables $a_{n,m}(t)$ are used to represent the association between GPF n and AP m within each time slot. Binary variables $a_{n,s}(t)$ are used to represent the association between GPF n and LEO satellite s within each time slot. $a_{n,m}(t) = 1$ denotes that in the current time slot, GPF n chooses to upload data to AP m; otherwise, $a_{n,m}(t) = 0$. $a_{n,s}(t) = 1$ indicates that in the current time slot, GPF n opts for a direct connection to LEO satellite s for data upload; otherwise, $a_{n,s}(t) = 0$. Each device can choose only one data transmission mode in the same time slot, either connecting to the LEO satellite or a specific AP, and each device is subject to constraints as follows:

$$\sum_{m=1}^{M} a_{n,m}(t) + a_{n,s}(t) \leq 1, \ \forall t \in \mathcal{T} \tag{1}$$

According to [12], the connection between GPFs and LEO satellites, as well as APs, is characterized by Line-of-Sight (LoS) transmission links. Therefore, the channel gain between GPF n and LEO satellite s is defined as

$$h_{n,s} = g_0 \cdot (H_s)^{-\theta}, \tag{2}$$

where $g_0 = G_n^{tr} \cdot G_s^{re} \cdot \left(\frac{\lambda}{4\pi d_0}\right)^{-\theta}$ represents the channel power gain between GPF n and LEO satellite s at a unit distance $d_0 = 1m$. G_n^{tr} denotes the transmit antenna gain of GPF n, G_s^{re} denotes the receive antenna gain of LEO satellite s, $\lambda = \frac{c}{f}$ represents the carrier wavelength, c represents the speed of light, f represents the carrier frequency, and θ represents the distance attenuation factor.

Furthermore, since the distance between the GPFs and the LEO satellite is far away, the distance between LEO satellite s and GPF n is simplified to the altitude of the LEO satellite, denoted as H_s. The direct communication transmission rate between GPF n and LEO satellite s is given by

$$R_{n,s}(t) = W \cdot \log_2\left(1 + \frac{p_n(t) \cdot h_{n,s}}{\sigma^2}\right), \tag{3}$$

where W represents the allocated fixed bandwidth for communication between GPFs and LEO satellites and σ^2 denotes the variance of Additive White Gaussian Noise (AWGN). $p_n(t)$ represents the transmission power of GPF n during time slot t, which is subject to the constraint imposed by the maximum transmission power p_{max} of the power facility as follows:

$$p_n(t) \leq p_{\max}, \ \forall n \in \mathcal{N}, t \in \mathcal{T}, \tag{4}$$

when the immediate communication workload $\mu_{n,\text{immediate}}$ required by GPF n is substantial, due to the constraints on the transmission rate $p_{\max}$, the direct transmission rate $R_{n,s}$ may not be able to complete the communication task of GPF n within the coverage time.

In such cases, the GPFs will consider completing the communication task using the relay forwarding assistance of APs. In time slot t, the channel gain between GPF n and AP m is defined as

$$h_{n,m}(t) = g_1 \cdot (d_{n,m}(t))^{-\theta} = g_1 \cdot \left(\frac{1}{\sqrt{\|\boldsymbol{q}_m(t) - \boldsymbol{q}_n\|^2 + H_u^2}} \right)^{\theta}, \tag{5}$$

where $g_1 = G_n^{tr} \cdot G_m^{re} \cdot \left(\frac{\lambda}{4\pi d_0} \right)^{-\theta}$ represents the channel gain at a unit distance. G_m^{re} denotes the receive antenna gain of AP m. According to Shannon's formula, the transmission rate for communication between GPF n and AP m is defined as

$$R_{n,m}(t) = W \cdot \log_2 \left(1 + \frac{p_n(t) \cdot h_{n,m}(t)}{\sigma^2} \right), \tag{6}$$

and relays through AP m. According to [14], the transmission rate between GPF n and LEO satellite s is defined as

$$R_{n,m,s}(t) = W \cdot \log_2 \left(1 + \frac{\gamma_{n,m}(t) \cdot \gamma_m}{1 + \gamma_{n,m}(t) + \gamma_m} \right), \tag{7}$$

where $\gamma_{n,m}(t) = \frac{p_n(t) \cdot h_{n,m}(t)}{\sigma^2}$ represents the Signal-to-Noise Ratio (SNR) between GPF n and AP m, $\gamma_m = \frac{p_m \cdot h_m}{\sigma^2}$ is the SNR between AP m and LEO satellite s, and p_m is the fixed forwarding power of AP m. It is evident that when the AP has a sufficiently large forwarding power p_m, the transmission rate $R_{n,m,s}(t)$ of GPF n will significantly improve compared to the direct satellite transmission rate $R_{n,m}(t)$.

In conclusion, the actual data transmission rate of GPF n for immediate communication tasks in time slot t can be expressed as

$$R_{n,\text{immediate}}(t) = \sum\nolimits_{m=1}^{M} (a_{n,m}(t) \cdot R_{n,m,s}(t)) + a_{n,s}(t) \cdot R_{n,s}(t) \tag{8}$$

where the first item $\sum_{m=1}^{M} (a_{n,m}(t) \cdot R_{n,m,s}(t))$ represents the rate at which the GPF establishes a connection with an AP and $a_{n,s}(t) \cdot R_{n,s}(t)$ represents the rate at which the GPF communicates directly with the LEO satellite. In (1), we can see that the GPF can only establish one connection in a single time slot. The actual data transmission rate of GPF n during time slot t for non-immediate communication tasks is

$$R_{n,\text{non-immediate}}(t) = \sum\nolimits_{m=1}^{M} (a_{n,m}(t) \cdot R_{n,m}(t)) \tag{9}$$

The actual communication energy consumption generated by GPF n during time slot t is

$$E_n(t) = \left[\left(\sum\nolimits_{k=1}^{M} a_{n,k}(t) + a_{n,s}(t) \right) \cdot p_n(t) \right] \cdot \Delta t = p_{n,k}(t) \cdot \Delta t, \tag{10}$$

where $p_{n,k}(t)$ represents the user transmission power when associated with the kth device. If the actual transmission rate of GPF n at any time t is $R_n(t)$, according to [18], the long-term system total energy efficiency is defined as

$$\eta = \sum\nolimits_{t=1}^{T} \frac{\sum_{n=1}^{N} R_n(t) \cdot \Delta t}{\sum_{n=1}^{N} E_n(t)} = \sum\nolimits_{t=1}^{T} \frac{\sum_{n=1}^{N} R_n(t)}{\sum_{n=1}^{N} p_{n,k}(t)}. \tag{11}$$

3.3. Problem Modeling

Let the set of association factors between GPFs and LEO satellites be denoted as $\mathscr{a}_{\mathrm{s}} = (a_{n,s}(t) : \forall n \in \mathcal{N}, t \in \mathcal{T})$. The set of association factors between GPFs and APs is denoted as $\mathscr{a}_{\mathrm{m}} = (a_{n,m}(t) : \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T})$. The set of horizontal positions of APs in each time slot is denoted as $\mathcal{Q} = (q_m(t) : \forall m \in \mathcal{M}, t \in \mathcal{T})$, and the set of transmission powers of GPFs in each time slot is denoted as $\mathcal{P} = (p_n(t) : \forall n \in \mathcal{N}, t \in \mathcal{T})$. $\mathcal{N}$ represents the GPF set, $\mathcal{M}$ represents the AP set, and $\mathcal{T}$ represents the time slot set.

In this paper, we conduct joint optimization of the device association factors $\mathscr{a}_{\mathrm{s}}$ and $\mathscr{a}_{\mathrm{m}}$, AP flight trajectories $\mathcal{Q}$, and GPF transmission powers $\mathcal{P}$ to maximize the long-term system total energy efficiency η. The optimization problem is expressed as follows:

$$\begin{gathered}(\mathbf{P1}) \max_{\mathscr{a}_{\mathrm{s}},\mathscr{a}_{\mathrm{m}},\mathcal{Q},\mathcal{P}} \eta \\ \text{s.t. C1} \sim \text{C9.}\end{gathered} \tag{12}$$

C1: $a_{n,s}(t) \in \{0,1\}, \forall n \in \mathcal{N}, t \in \mathcal{T}$;
C2: $a_{n,m}(t) \in \{0,1\}, \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T}$;
C3: $\sum_{m=1}^{M} a_{n,m}(t) + a_{n,s}(t) \leq 1, \forall n \in \mathcal{N}, t \in \mathcal{T}$;
C4: $\sum_{n=1}^{N} a_{n,m}(t) \leq L_M, \forall m \in \mathcal{M}, t \in \mathcal{T}$;
C5: $\sum_{m=1}^{M} \sum_{n=1}^{N} a_{n,m}(t) + \sum_{n=1}^{N} a_{n,s}(t) \leq L_S, \forall t \in \mathcal{T}$;
C6: $0 \leq p_n(t) \leq p_{max}, \forall n \in \mathcal{N}, t \in \mathcal{T}$;
C7: $\sum_{t=1}^{T} (R_{n,\mathrm{immediate}}(t) \cdot \Delta t) \geq \mu_{n,\mathrm{immediate}}, \forall n \in \mathcal{N}$;
C8: $\sum_{t=1}^{T} (R_{n,\mathrm{non-immediate}}(t) \cdot \Delta t) \geq \mu_{n,\mathrm{non-immediate}}, \forall n \in \mathcal{N}$;
C9: $\|q_m(t+1) - q_m(t)\|^2 \leq (v_{max} \cdot \Delta t)^2, \forall m \in \mathcal{M}, t \in \mathcal{T}$.

C1 and C2 represent binary variables, indicating the association factors between GPFs and LEO satellites and APs, respectively. C3 indicates that any GPF can only choose a single AP or LEO satellite for data transmission in any time slot. C4 and C5 denote upper limits on the number of GPFs that APs and LEO satellites can associate with in any time slot. C6 represents an upper limit on the total transmission power of GPFs in any time slot. C7 indicates that each GPF needs to meet its respective minimum immediate communication workload requirement. C8 signifies that each GPF needs to satisfy its respective minimum non-immediate communication workload requirement. C9 denotes that the position variation of APs in each time slot is constrained by their maximum flight speed.

4. Reinforcement Learning-Based Joint Optimization Strategy for Flight Path and Resource Allocation

In this section, we split the problem into two subproblems and modeled them as an MDP and a multi-agent time-varying K-armed bandit model, respectively. The reinforcement learning algorithm is used to solve the subproblems and iterate alternately to find the optimal strategy of agents.

4.1. Problem Description and Decomposition

The objective of this paper is to maximize the overall energy efficiency of the system within a given operational cycle T (referred to as an episode). We optimize the position of the APs in each time slot, as well as the device association and power control strategies for GPFs. The optimization problem can be decomposed into subproblem **P2** for the optimization of the AP flight paths and the joint optimization subproblem **P3** for GPF device association and power control as follows:

$$\begin{gathered}(\mathbf{P2}) \max_{\mathcal{Q}} \eta \\ \text{s.t. C9} : ||q_m(t+1) - q_m(t)||^2 \leq (v_{max} \cdot \Delta t)^2, \ \forall m \in \mathcal{M}, t \in \mathcal{T}.\end{gathered} \tag{13}$$

$$\begin{aligned} &(\mathbf{P3})\ \max_{\mathscr{a}_{\mathrm{s}},\mathscr{a}_{\mathrm{m}},\mathcal{P}} \eta \\ &\text{s.t. C1} \sim \text{C8.} \end{aligned} \tag{14}$$

C1: $a_{n,s}(t) \in \{0,1\}, \forall n \in \mathcal{N}, t \in \mathcal{T};$
C2: $a_{n,m}(t) \in \{0,1\}, \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T};$
C3: $\sum_{m=1}^{M} a_{n,m}(t) + a_{n,s}(t) \leq 1, \forall n \in \mathcal{N}, t \in \mathcal{T};$
C4: $\sum_{n=1}^{N} a_{n,m}(t) \leq L_M, \forall m \in \mathcal{M}, t \in \mathcal{T};$
C5: $\sum_{m=1}^{M} \sum_{n=1}^{N} a_{n,m}(t) + \sum_{n=1}^{N} a_{n,s}(t) \leq L_S, \forall t \in \mathcal{T};$
C6: $0 \leq p_n(t) \leq p_{max}, \forall n \in \mathcal{N}, t \in \mathcal{T};$
C7: $\sum_{t=1}^{T} (R_{n,\text{immediate}}(t) \cdot \Delta t) \geq \mu_{n,\text{immediate}}, \forall n \in \mathcal{N};$
C8: $\sum_{t=1}^{T} (R_{n,\text{non-immediate}}(t) \cdot \Delta t) \geq \mu_{n,\text{non-immediate}}, \forall n \in \mathcal{N}.$

Subproblem **P2** assumes that the connection of devices and their transmit power are given in each time slot. The APs need to consider the communication task requirements of GPFs and the total channel gain that they can provide to GPFs to determine their movement. We propose an online path optimization algorithm based on RL to address this problem.

Subproblem **P3** assumes that the positions of the APs are given in each time slot. GPFs need to choose an appropriate device association and provide transmit power for data transmission within each time slot. From the constraints of the subproblem, it can be observed that each GPF can only choose one device for association in each time slot, and each device has a maximum connection limit, leading to the issue of connection competition. After determining the associated device, GPFs further optimize their transmit power to achieve maximum throughput. We propose a joint optimization sub-algorithm for device association and power control based on RL to address this problem.

4.2. Aerial Platform Flight Path Optimization Algorithm

In this paper, each AP makes independent decisions in each time slot based on the value of each state provided by environmental feedback. The environment changes based on the decisions made and the rewards obtained in each state, eventually forming a finite MDP within an episode. Therefore, the MDP can be employed to model this subproblem.

In this model, each AP is considered as an intelligent agent. The system can be described using a quadruple $\Phi = \{\mathcal{M}, \mathcal{S}, \mathscr{a}, \mathcal{R}\}$, where $\mathcal{M}$ represents the number of agents, $\mathcal{S}$ represents the states in which each agent is located, $\mathscr{a}$ represents the actions chosen by the agents in those states, and $\mathcal{R}$ represents the rewards obtained from the actions of the agents in those states.

Based on the four elements mentioned above, the definition of the agent's state space, action space, and reward function is as follows:

(1) State Space: The entire two-dimensional coordinates within the communication task range of all APs are discretely represented. The obtained set of two-dimensional coordinates constitutes the state space of the agents. The position $q_m(t) = [x_m(t), y_m(t)]^T$ of each agent in each time slot represents the agent's state S_m.

(2) Action Space: we simplify the Computational complexity in this algorithm by considering the agent's maximum speed constraint $(v_{max}\Delta t)$ and selecting a set of equidistant fixed displacement values λ_k, $k = 1, \cdots, K$. By applying these fixed displacement values to both the x and y coordinates of the two-dimensional state positions, a combination of K^2 attainable target positions in one time slot is generated. The selection of all fixed displacement values for the x and y axes constitutes the action space of the agent.

(3) Reward Function: The reward function $R_m(a_m|s_m)$ for AP m is defined as the weighted sum of the channel gains of various GPFs at the target position based on their communication workload. This sum is then added to a function of the distances between the AP and other APs and multiplied by a certain proportionality factor ρ. The expression is as follows:

$$R_m(A_m|S_m) = \sum_{i=1}^{N} \gamma_i h_{i,m}(t) + \rho \cdot \log_2\left(1 + \sum_{j=1, j \neq m}^{M-1} d_{j,m}\right) \tag{15}$$

where γ_i serves as the weighting factor for the workload of GPFs. $h_{i,m}(t)$ represents the channel gain between GPF i and AP m. $d_{j,m}$ denotes the distance between AP m and other APs. The purpose of the second term is to prevent prolonged path overlap between different APs. Based on the proposed model, we present a comprehensive sub-iteration multi-agent iterative AP path optimization algorithm that considers both the positions and workload of GPFs. The specific algorithm procedure is illustrated in Algorithm 1.

Initially, the maximum training episodes N_{epi} are initialized, and the initial state $s_m(0)$ is given for each AP (Step 1). In each time slot within an episode loop (Step 2), the algorithm calculates the reward function for each reachable state of each AP (Step 2a). According to a greedy strategy, the APs move and update their states $s_m(t)$ (Step 2b). In the overall algorithm flow, the weights of the reward function are updated based on the remaining communication workload for GPFs derived from Algorithm 2 (Step 2c). After completing one training episode, the primary algorithm parameters are initialized, and the next training episode is executed (Step 3). This process is repeated until all training episodes are completed (Step 4).

Algorithm 1: Aerial Platform Path Optimization Algorithm Based on Comprehensive Evaluation of GPF Position and Workload

```
(1) Initialization: Maximum training episodes $N_{\text{epi}}$, for any AP $m$, given its initial position state $s_m(0)$.
(2) For each iteration in episode $N_{\text{epi}}$, $t = 1, 2, \ldots, T$, instruct each AP to independently execute the following steps:
    (a) Based on the positions of each GPF, remaining workload, and the current state of the AP in this time slot, update the reward function for each reachable target state of AP $m$ using the following expression:
        $R_m(A_m|S_m) \leftarrow \sum_{i=1}^{N} \gamma_i h_{i,m}(t) + \rho \cdot \log_2\left(1 + \sum_{j=1,j\neq m}^{M-1} d_{j,m},\right)$
    calculate the reward function for the reachable target positions of AP m in the current time slot;
    (b) According to a greedy strategy, each AP will move towards the target state with the highest reward and update its state $s_m(t)$;
    (c) Conduct the iteration of Algorithm 2 for each GPF, update the weight $\gamma_i$ for the remaining workload of each GPF, return to step (a), and proceed to the next iteration until completion.
(3) Set $N_{\text{epi}} = N_{\text{epi}} - 1$.
(4) Repeat steps (1) to (3) until $N_{\text{epi}} = 0$.
```

4.3. Device Association and Power Control Algorithm

In this paper, when GPFs make action selections, the moving strategy of APs is not considered. Only the current position of APs at the current time slot is taken as the environmental state. Therefore, the decision-making process for device association and power control of GPFs can be modeled as a time-varying K-armed bandit model with multiple intelligent agent states. The action selection of the intelligent agent is considered not to affect changes in the environmental state.

This model regards GPFs as intelligent agents, and the system can be represented by a quadruple $\Phi = \{\mathcal{M}, \mathcal{S}, \mathcal{a}, \mathcal{R}\}$. The definition of the agent's state space, action space, and reward function is as follows:

(1) State Space: To simplify the algorithm complexity, in the proposed algorithm, the distances between APs and intelligent agents are discretized into L state distributions based on the provided channel gain. The distances between each AP and intelligent agent constitute a state space with L^M elements.

(2) Action Space: Defined as the device association and power control schemes that agents can choose. Agents determine their action choices based on the current state matrix and the policy π specified by the algorithm at the current time.

(3) Reward Function: The reward function for the agent n at time slot t is defined as the instantaneous throughput and instantaneous energy consumption as follows:

$$R_n(s(t), a(t)) = R_n(t) \cdot \Delta t, \tag{16}$$

$$P_n(s(t), a(t)) = E_n(t), \tag{17}$$

where $R_n(t)$ represents the transmission rate of the agent n in time slot t and $E_n(t)$ represents the actual communication energy consumption generated by the agent n in time slot t.

Let $Q_n^k(s,a)$ represent the estimated value of the action chosen by intelligent the agent n in state s after $k-1$ selections, which is expressed as follows:

$$Q_n^k(s,a) = \frac{R_n^1 + R_n^2 + \cdots + R_n^{k-1}}{k-1}, \tag{18}$$

$$Q_n^{k+1}(s,a) = Q_n^k(s,a) + \frac{1}{k}\Big(R_n^k - Q_n^k(s,a)\Big) \tag{19}$$

Let $P_n^k(s,a)$ represent the estimated cost of the agent n choosing action a in state s after $k-1$ selections, which is expressed as follows:

$$P_n^k(s,a) = \frac{P_n^1 + P_n^2 + \cdots + P_n^{k-1}}{k-1}, \tag{20}$$

$$P_n^{k+1}(s,a) = P_n^k(s,a) + \frac{1}{k}\Big(P_n^k - P_n^k(s,a)\Big), \tag{21}$$

then, the efficiency estimation function for the agent n choosing action a in state s for the k-th time can be expressed as

$$\eta_n^k(s,a) = \frac{Q_n^k(s,a)}{P_n^k(s,a)} = \frac{R_n^1 + R_n^2 + \cdots + R_n^{k-1}}{P_n^1 + P_n^2 + \cdots + P_n^{k-1}}. \tag{22}$$

The policy for the action selection of the agent n is represented as

$$A_n^t = \underset{a}{\operatorname{argmax}}\Big[\eta_n^k(s,a)\Big]. \tag{23}$$

This signifies that the agents always choose the strategy that maximizes the current estimated energy efficiency. However, to balance exploration and exploitation, there is a probability $\varepsilon \in (0,1)$ for exploration to discover better actions. The specific policy is expressed as follows:

$$\pi_n^\varepsilon(A_n^t|s) = \begin{cases} 1-\varepsilon, \ A_n^t = \underset{a}{\operatorname{argmax}}\Big[\eta_n^k(s,a)\Big] \\ \qquad \varepsilon, otherwise \end{cases}, \tag{24}$$

and to enhance the algorithm's overall performance and expedite the convergence speed of iterations, ε can be set as a linear function related to the iteration episodes N_{epi}, as shown below:

$$\varepsilon = 0.5 - \rho_{\text{epi}} \cdot N_{\text{epi}}, \tag{25}$$

where ε represents the exploration rate and ρ_{epi} denotes the rate at which the exploration rate decreases with N_{epi}. When $\varepsilon = 0$, the training process will conclude, yielding the corresponding training results.

Based on the aforementioned model, we propose a device association and power control algorithm based on the multi-agent K-armed bandit. The specific algorithm procedure is illustrated in Algorithm 2.

Firstly, we initialize the maximum training episodes N_{epi}, exploration parameter ε, state-action value function $Q_n^1(s,a)$, and energy consumption estimation function $P_n^1(s,a)$, along with the positions and initial workload $\mu_n = (\mu_{n,\text{immediate}}, \mu_{n,\text{non-immediate}})$ for each GPF (Step 1). In each time slot within an episode loop (Step 2), it is determined whether GPFs have immediate data tasks based on μ_n. If immediate tasks exist, execute policy $\pi_{n,\text{immediate}}$ (Step 2a), which means that GPFs will upload real-time data to the nearest AP at maximum power. If the optional APs have reached the maximum number of connections, they will be directly connected to the LEO satellite. If only non-immediate data tasks remain, select action $a_n(t)$ based on policy π_n^{ϵ} (Step 2b). Then, calculate the instantaneous reward $R_n(s(t),a(t))$ and instantaneous cost $P_n(s(t),a(t))$ based on action $a_n(t)$, update the remaining workload, and use the remaining workload as a state change for the iteration of Algorithm 1 in the next step (Step 2c). Update the state-action value function $Q_n^k(s,a)$ and energy cost estimation function $P_n^k(s,a)$ using $R_n(s(t),a(t))$ and $P_n(s(t),a(t))$, and compute the energy efficiency estimation function $\eta_n^k(s,a)$ for the current time t (Step 2d). After completing one training episode, initialize the main algorithm parameters and execute the next training episode (Step 3). Repeat the above process until all training episodes are completed (Step 4).

Algorithm 2: Algorithm for Device Association and Power Control Based on Multi-Agent K-Armed Bandit

```
(1) Initialization: Maximum training episodes $N_{\text{epi}}$, exploration parameter $\varepsilon$, state-action value function $Q_n^1(s,a)=0$, and energy cost estimation function $P_n^1(s,a)=0$.
(2) For GPF $n$, given its position and initial workload $\mu_n = (\mu_{n,\text{immediate}}, \mu_{n,\text{non-immediate}})$.
(3) For each iteration in episode $N_{\text{epi}}$, $t=1,2,\ldots,T$, prioritize GPFs based on the type and quantity of the remaining workload. Then, instruct each GPF to independently execute the following steps:
    (a) Check if immediate data tasks still exist; if yes, execute policy $\pi_{n,\text{immediate}}$;
    (b) If only non-immediate data tasks remain, select action $a_n(t)$, based on policy $\pi_n^{\epsilon}$;
    (c) Based on the executed action $a_n(t)$, obtain instantaneous reward $R_n(s(t),a(t))$, and instantaneous cost $P_n(s(t),a(t))$, update the remaining workload, and use the remaining workload as a state change for the iteration of Algorithm 1 in the next step;
    (d) Update the state-action value function $Q_n^k(s,a)$ and energy cost estimation function $P_n^k(s,a)$ based on the instantaneous reward and cost, and update the energy efficiency estimation function $\eta_n^k(s,a)$ for the current time.
(4) Set $N_{\text{epi}} = N_{\text{epi}} - 1$.
(5) Repeat steps (2) to (4) until $N_{\text{epi}} = 0$.
```

4.4. Overall Algorithm Flow

Combining Algorithm 1 and Algorithm 2, we propose a joint optimization strategy for flight path and resource allocation based on RL, as shown in Algorithm 3. In each time slot, the algorithm first runs Algorithm 1 based on the remaining throughput of GPFs to obtain the current coordinates of APs for that time slot (Step 2b). Subsequently, each GPF obtains the state of the current time slot. Based on the trained energy efficiency estimation matrix, Algorithm 2 is executed to select the optimal resource allocation strategy (Step 2c). The specific flow is illustrated in Figure 2.

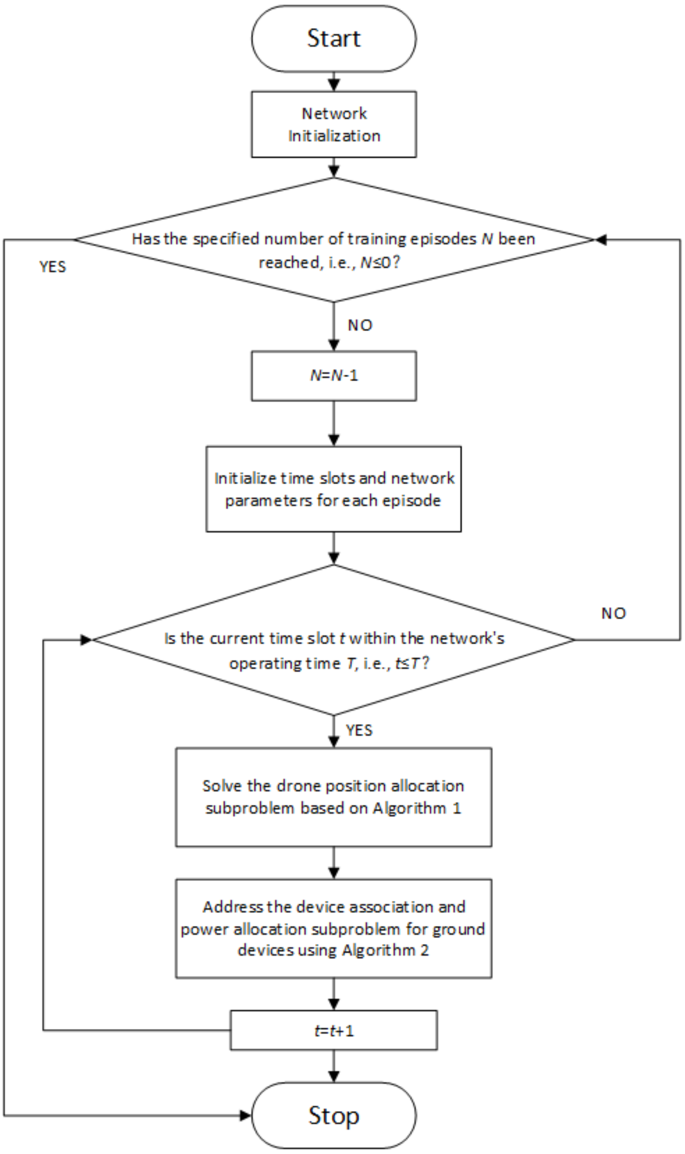
Start
Network Initialization
Has the specified number of training episodes *N* been reached, i.e., *N*≤0?
YES
NO
N=*N*-1
Initialize time slots and network parameters for each episode
Is the current time slot *t* within the network's operating time *T*, i.e., *t*≤*T*?
NO
YES
Solve the drone position allocation subproblem based on Algorithm 1
Address the device association and power allocation subproblem for ground devices using Algorithm 2
t=*t*+1
Stop

Figure 2. Reinforcement learning-based joint optimization strategy for flight path and resource allocation—process flowchart.

The concept of the proposed algorithm draws on the Monte Carlo method and the single-step Temporal Difference method. It updates the value at a single moment and selects actions through a soft greedy strategy to balance exploration and optimization. It only requires a certain sampling sequence, that is, from the environment. Using the sequence of states, actions, and benefits sampled, we can iterate to the optimal strategy.

Algorithm 3: A Joint Optimization Strategy for Flight Path and Resource Allocation Based on RL

```
(1) Input: GPFs set $\mathcal{N}$, APs set $\mathcal{M}$, each algorithm parameter, Environmental state information,
GPFs' workload $\mu$, the upper limit of transmission power of GPFs $p_{max}$.
Output: Optimal Flight Path strategy $\mathcal{Q}$ and optimal device association and power control
strategy $\mathscr{a}_{\mathrm{s}}, \mathscr{a}_{\mathrm{m}}, \mathcal{P}$.
(2) for each training episodes $N_{\mathrm{epi}}$ do
    a. for each time slot $t$ do
    b. Based the GPFs' workload and position obtained from Algorithm 2 in time slot $t$, decide the
       Flight Path strategy $\mathcal{Q}(t)$ according to Algorithm 1.
    c. Based on Flight Path strategy $\mathcal{Q}(t)$ obtained from Algorithm 1 in time slot $t$, the device
       association and power control strategy $\mathscr{a}_{\mathrm{s}}, \mathscr{a}_{\mathrm{m}}, \mathcal{P}$ is decided using Algorithm 2.
    d. Update Environmental state information.
    e. $t = t + 1$.
    f. end for
    g. until $t > T$.

(3) $N_{\mathrm{epi}} = N_{\mathrm{epi}} - 1$.
(4) end for
(5) until $N_{\mathrm{epi}} \leq 0$.
(6) Return $\mathcal{Q}, \mathscr{a}_{\mathrm{s}}, \mathscr{a}_{\mathrm{m}}, \mathcal{P}$.
```

5. Simulation Results

The main simulation parameters are listed in Table 1 [14,16,18]. The simulation environment in this paper is based on MATLAB and is employed for computational simulation to validate the effectiveness of the proposed algorithm. It is assumed that in a square area with an extent denoted as 500 m × 500 m, GPFs are randomly distributed. Each GPF is characterized by specific real-time and non-real-time communication throughput requirements. Three APs are assumed to fly within this area at an altitude denoted as $H_u = 100$ m, aiming to establish LoS communication with GPFs. Additionally, an LEO satellite serves as a cloud server data center, processing real-time power business data from GPFs. Specifically, the APs depart from a predefined starting position and, within a specified time denoted as T, fulfill the communication requirements of all GPFs. The specific simulation parameters are detailed in Table 1.

Table 1. System model simulation parameters.

Parameter	Value	Parameter	Value
N	10	W	50 kHz
M	3	p_{max}	41 dBm
H_u	100 m	T	100 s
v_{max}	10 m/s	H_s	250 km
g_0	1.4×10^{-4}	L_M	3
σ^2	−110 dBm	L_s	10
p_m	47 dBm	Δt	1 s
$\mu_{n,\mathrm{immediate}}$	3.5 $\sim$ 6.5 Mbit	$\mu_{n,\mathrm{non-immediate}}$	55 $\sim$ 67 Mbit

The number of algorithm iterations N_{epi}, discretization state parameters K and L, and distance scaling factors ρ in Tables 2 and 3 are reasonably set, taking into account factors such as the size of the simulation environment equipment, computer computing power, and simulation time.

Table 2. Algorithm 1 simulation parameters.

Parameter	Value
N_{epi}	300
K	17
ρ	6×10^{-9}

Table 3. Algorithm 2 simulation parameters.

Parameter	Value
N_{epi}	300
L	5
ρ_{epi}	1.67×10^{-3}

Figure 3 depicts the flight paths of three APs in the system when accommodating 10 GPFs. It can be observed that the APs tend to move towards regions with denser GPF positions. Additionally, the three APs predominantly serve distinct groups of GPFs. As the remaining throughput requirements of GPFs change, the APs continuously adjust their positions. While meeting the demands of the majority of GPFs, efforts are made to consider GPFs with a higher remaining throughput. This indicates that through optimized flight paths, APs can enhance service provision for GPFs, consequently improving the overall system energy efficiency.

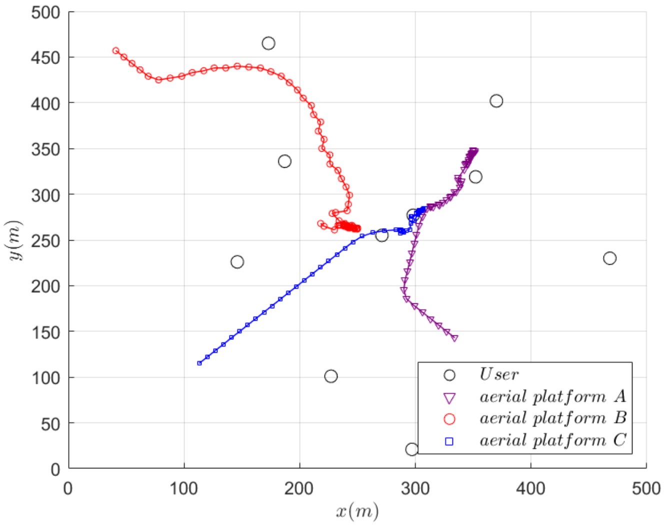

Figure 3. The positions of GPFs and the flight paths of APs in the simulated scenario.

In Figure 4, the variation of the overall system energy efficiency with the number of training episodes is compared under different exploration rates. To visually represent this variation more intuitively, Figure 5 averages the obtained actual energy efficiency values every 10 episodes to obtain the corresponding fluctuation curve. It can be observed in the graph that when the exploration rate is relatively high, the algorithm converges to a lower maximum energy efficiency value. However, the exploration of possible action strategies is more thorough. As the exploration rate decreases, the convergence speed of the algorithm decreases, but it can converge to higher energy efficiency values. The proposed algorithm incorporates a decreasing exploration rate with an increasing number of training episodes. It initially sacrifices energy efficiency to enhance convergence speed, and after a sufficiently thorough exploration of the policy, it converges to the globally optimal energy efficiency value, aiming to improve the algorithm's performance.

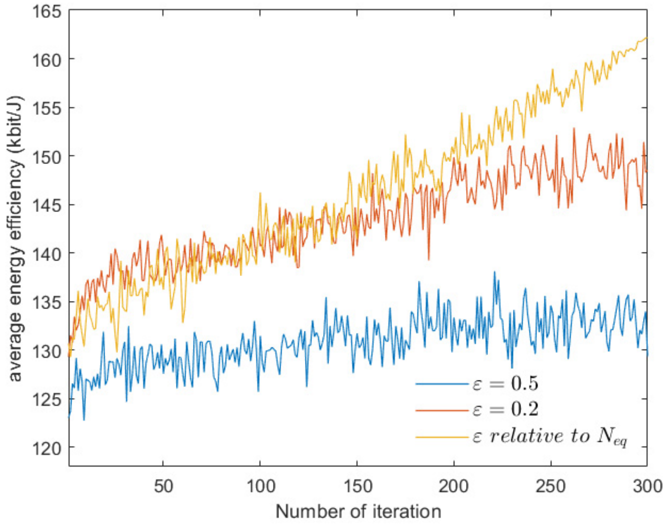

Figure 4. Actual energy efficiency variations during training under different exploration rate strategies.

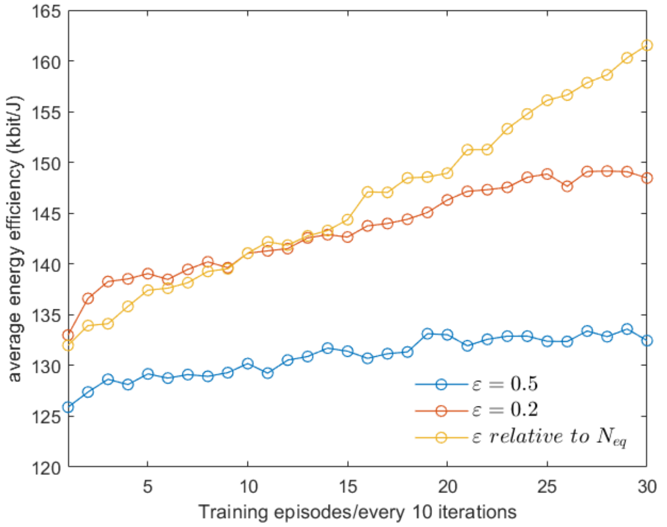

Figure 5. Average energy efficiency variations during training under different exploration rate strategies.

To illustrate the effectiveness of the proposed algorithm in this paper, the algorithm is compared with the following three algorithms:

(1) Random Path Algorithm (RP algorithm): Within the state space defined in Algorithm 1, the movement of each AP is set to move towards a random GPF in each time slot. To facilitate comparison, the device association and power control strategy of Algorithm 2 remain unchanged.

(2) Random Device Selection Algorithm (RS algorithm): All GPFs are randomly associated with available APs. The path selection for APs and power control strategy for GPFs remain consistent with the proposed algorithm.

(3) Equal Transmit Power Algorithm (EP algorithm): All GPFs set their transmission power to the same fixed value, ensuring that each GPF meets its throughput requirement. The path selection for APs and device association strategy for GPFs remain consistent with the proposed algorithm.

Figure 6 depicts the curve of total system energy efficiency against noise power variations. The energy efficiency of all four algorithms decreases with the increase in noise power. Higher noise power leads to a significant reduction in the transmission rates of GPFs, and when the noise power is excessive, the system energy efficiency decreases substantially, as GPFs must increase their maximum transmit power to meet throughput requirements. It is evident that, under all noise power conditions, the performance of the proposed algorithm is superior to the other three algorithms. The curve of the energy efficiency of the proposed algorithm changing with noise power is improved by 17.23%, 162.87%, and 4.45%, respectively, compared with the RP algorithm, EP algorithm, and RS algorithm.

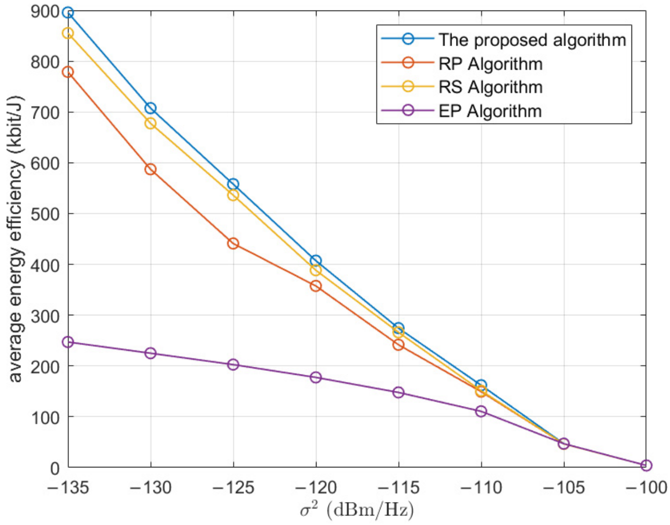

Figure 6. Total system energy efficiency with respect to variations in noise power.

Figure 7 illustrates the curve of total system energy efficiency with respect to variations in bandwidth allocation for GPFs. The total system energy efficiency increases with the increase in channel bandwidth allocated to GPFs. Higher channel bandwidth provides GPFs with higher transmission rates, leading to an overall improvement in system energy efficiency. It can be observed that, under all scenarios of GPF channel bandwidth allocation, the performance of the proposed algorithm is superior to the other three algorithms. The curve of the energy efficiency of the proposed algorithm changing with channel bandwidth for ground devices is improved by 16.63%, 116.42%, and 4.96%, respectively, compared with the RP algorithm, EP algorithm, and RS algorithm.

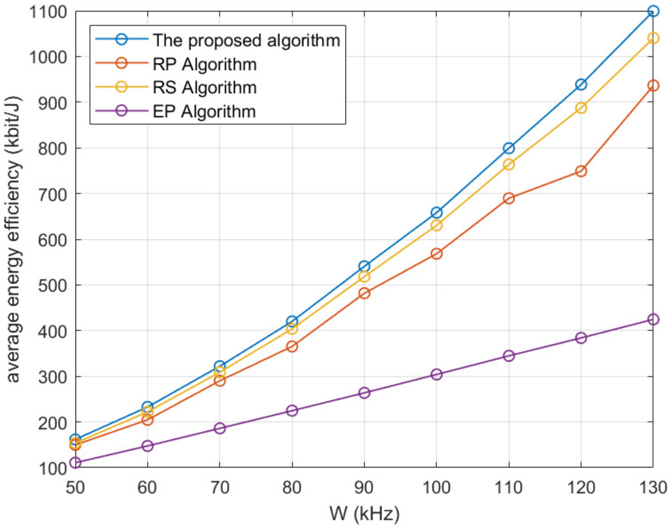

Figure 7. Total system energy efficiency with respect to variations in channel bandwidth for ground devices.

Figure 8 depicts the variation curve of total system energy efficiency with respect to the throughput load rate of APs. Constrained by their flying altitude, APs have a limited maximum throughput they can provide to GPFs in each time slot. The throughput load rate is defined as the ratio of the total GPF throughput demand to the maximum throughput that the APs can provide. In the graph, it can be observed that when the load rate is below 100%, the system's total energy efficiency remains relatively stable. However, when the load rate exceeds 100%, it shows a decreasing trend and may exhibit a sharp decline under excessively high load rates. This is due to the inability of APs to meet the throughput demands of GPFs, forcing an increase in the transmission power level of GPFs and resulting in a decrease in the overall system energy efficiency. This can be improved by adjusting the altitude of APs or allocating a larger bandwidth to each GPF. Under any load rate condition, the performance of the proposed algorithm is superior to the other three algorithms. The curve of the energy efficiency of the proposed algorithm changing with throughput load rate is improved by 7.92%, 34.71%, and 6.79%, respectively, compared with the RP algorithm, EP algorithm, and RS algorithm.

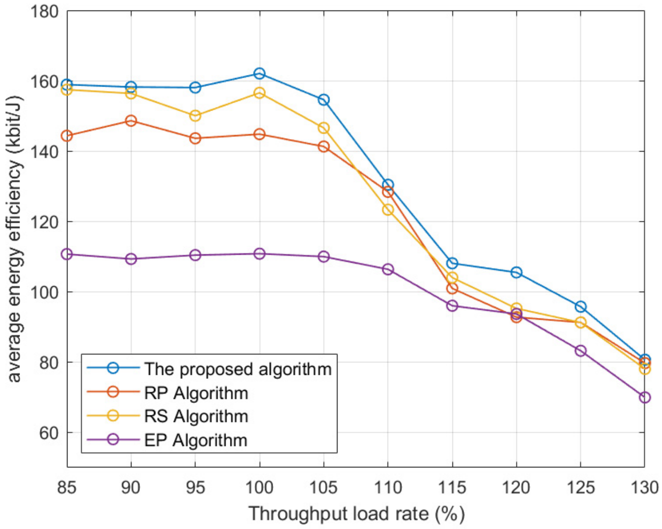

Figure 8. Total system energy efficiency with respect to throughput load rate.

Figure 9 illustrates the variation curve of total system energy efficiency with respect to the number of GPFs. In the graph, the number of GPFs starts from six and incrementally increases up to thirteen. When the number of GPFs is low or too high, the performance of each algorithm decreases. This may be attributed to a mismatch between the scaling factor ρ in Algorithm 1 and the number of GPFs. Validation confirms that adjusting the scaling factor ρ can improve algorithm performance. The algorithm achieves its maximum performance when the number of GPFs equals the sum of the maximum connectable devices for each AP. As the number of GPFs gradually increases, the total system energy efficiency shows a decreasing trend. This is due to intensified competition for connections among GPFs in each time slot as the number of GPFs increases, leading to some GPFs being unable to establish connections with APs. Improvement can be achieved by increasing the number of APs or enhancing the number of connectable devices for each AP. It is evident that, under most scenarios of device numbers, the performance of the proposed algorithm is superior to the other three algorithms. The curve of the energy efficiency of the proposed algorithm changing with the number of ground devices is improved by 7.13%, 31.15%, and 4.23%, respectively, compared with the RP algorithm, EP algorithm, and RS algorithm.

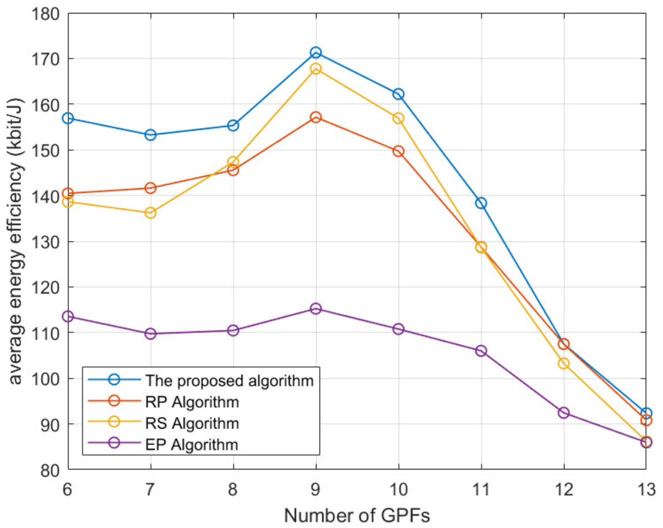

Figure 9. Total system energy efficiency with respect to the number of ground devices.

In the table below, we list the quantitative indicators of the proposed algorithm's improvement in energy efficiency. It can be seen that the proposed algorithm has a significant improvement compared with the EP algorithm, but the optimization effect is not obvious compared with the other two algorithms. This may be because the number of APs in the simulation scenario is small and the users are relatively close to each other, so the effect of equipment selection and path optimization is not obvious enough. However, in future work, we will study the larger number and distribution of users, and it is expected that more ideal results can be obtained.

In addition to the analysis of simulation performance, we also analyze the limitations and computational complexity of the algorithm. First of all, in terms of algorithm limitations, compared with the RP algorithm, the proposed algorithm considers the path optimization of the AP and will perform better in scenarios where the AP needs to move in a large range. Compared with the AP algorithm, the proposed algorithm can optimize the transmit power of GPFs, adapt to the different communication needs of each GPF, and greatly improve the energy efficiency of the system. Compared with the RS algorithm, the proposed algorithm can optimize the device selection of the GPF and ensure that the system can perform better in scenarios where a large number of GPFs are connected to the network. The proposed algorithm breaks through the limitations of GPF's large-scale distribution, multi-service communication, and large-scale access. Therefore, the proposed algorithm has strong applicability. However, when dealing with larger-scale state space and action space problems in the future, the proposed algorithm needs to perform a certain discretization of the state space and action space, and the training time is long, and sensitive to parameter selection. This is also the limitation of the proposed algorithm.

In this article, Algorithm 1 is used to optimize the flight path of APs, and its computational complexity is $O(KTM)$, where K represents the number of iterations, T represents the number of task slots, and M represents the number of APs. Algorithm 2 is used to optimize the device selection and power control of GPFs, and its computational complexity is $O(KTN)$, where N represents the number of GPFs. The computational complexity of the overall algorithm flow is $O(KT(M+N))$. As shown in Table 4, except for the slightly lower complexity of the RP algorithm, the complexity of the other two benchmark algorithms is consistent with the proposed algorithm.

Table 4. Quantitative indicator of the energy efficiency increases in the proposed algorithm compared with each algorithm.

Dependent Variable	RP Algorithm	EP Algorithm	RS Algorithm
noise power	17.23%	162.87%	4.45%
channel bandwidth	16.63%	116.42%	4.96%
throughput load rate	7.92%	34.71%	6.79%
the number of ground devices	7.13%	31.15%	4.23%
average value	12.23%	86.29%	5.11%
computational complexity	$O(KTN)$	$O(KT(M+N))$	$O(KT(M+N))$

6. Conclusions

In this paper, we study the energy efficiency optimization problem of the space–air–ground-integrated power three-dimensional converged communication network in emergency communication scenarios. On the basis of the system model, we establish the optimization problems of joint AP flight path selection, GPF association, and power control. The AP flight path is modeled as an MDP, and the AP path optimization algorithm is proposed based on a comprehensive evaluation of GPF location and workload. The GPF association and power control subproblem is modeled as a multi-agent, time-varying K-arm bandit model. We propose a device association and power control algorithm based on Temporal Difference learning to solve this problem. Combining these two algorithms, we propose a joint optimization strategy for flight path and resource allocation based on RL. Through the collaborative iteration of the two algorithms, the optimization of AP flight paths, GPF association, and power control can be achieved, which can improve the overall energy efficiency of the system while meeting certain throughput requirements. Finally, through computer simulations and comparisons with three other benchmark algorithms, the proposed algorithm's effectiveness in enhancing the overall system energy efficiency is validated. In practical scenarios, adjusting the scaling factor of the AP can adapt to different numbers of GPFs to achieve optimal system performance under current throughput requirements.

In future work, we will improve the settings of some parameters in the algorithm, integrate deep learning content to further optimize the algorithm iteration process, train neural networks as a framework for strategy selection, and conduct simulations in a larger-scale environment compared with other classic algorithms in the field of reinforcement learning to conduct a more in-depth analysis of the simulation results.

Author Contributions: Conceptualization, Z.C., H.Z., S.D. and J.L.; methodology, Z.C., S.D., J.L. and L.Z.; software, S.D. and J.L.; validation, H.Z., S.D. and L.Z.; formal analysis, J.L.; investigation, Z.C.; resources, Z.C., H.Z. and J.L.; data curation, J.L. and L.Z.; writing—original draft preparation, S.D. and J.L.; writing—review and editing, Q.L., Z.C. and L.Z.; visualization, H.Z.; supervision, Q.L. and J.L.; project administration, Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Foundation of the State Grid Corporation of China under Grant 5700-202352266A-1-1-ZN.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Zhiyu Chen and Hongxi Zhou was employed by the company State Grid Information & Telecommunication Branch, and author Qi Liu was employed by Beijing Fibrlink Communications Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GPF	Ground Power Facility
AP	Aerial Platform
MDP	Markov Decision Process
SAGIN	Space–Air–Ground-Integrated Network
LEO	Low Earth Orbit
RL	Reinforcement Learning
5G	Fifth Generation Mobile Communications Technology
6G	Sixth Generation Mobile Communications Technology
UAV	Unmanned Aerial Vehicle
SCA	Sequential Convex Approximation
QoS	Quality of Service
FDMA	Frequency Division Multiple Access
LoS	Line of Sight
AWGN	Additive White Gaussian Noise
SNR	Signal-to-Noise Ratio

References

1. Siasos, G.; Tousoulis, D.; Oikonomou, E.; Zaromitidou, M.; Verveniotis, A.; Plastiras, A.; Kioufis, S.; Maniatis, K.; Miliou, A.; Siasou, Z.; et al. On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 905–974.
2. Wang, J.; Jiang, C.; Wei, Z.; Pan, C.; Zhang, H.; Ren, Y. Joint UAV Hovering Altitude and Power Control for Space-Air-Ground IoT Networks. *IEEE Int. Things J.* **2019**, *6*, 1741–1753. [CrossRef]
3. Bedi, G.; Venayagamoorthy, G.K.; Singh, R.; Brooks, R.R.; Wang, K.C. Review of Internet of Things (IoT) in Electric Power and Energy Systems. *IEEE Int. Things J.* **2018**, *5*, 847–870. [CrossRef]
4. Hu, Y.; Chen, M.; Saad, W. Joint Access and Backhaul Resource Management in Satellite-Drone Networks: A Competitive Market Approach. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3908–3923. [CrossRef]
5. Cui, H.; Zhang, J.; Geng, Y.; Xiao, Z.; Sun, T.; Zhang, N.; Liu, J.; Wu, Q.; Cao, X. Space-air-ground integrated network (SAGIN) for 6G: Requirements, architecture and challenges. *China Commun.* **2022**, *19*, 90–108. [CrossRef]
6. Song, Z.; Hao, Y.; Liu, Y.; Sun, X. Energy-Efficient Multiaccess Edge Computing for Terrestrial-Satellite Internet of Things. *IEEE Int. Things J.* **2021**, *8*, 14202–14218. [CrossRef]
7. Kapovits, A.; Corici, M.; Gheorghe-Pop, I.; Gavras, A.; Burkhardt, F.; Schlichter, T.; Covaci, S. Satellite communications integration with terrestrial networks. *China Commun.* **2018**, *15*, 22–38. [CrossRef]
8. Li, B.; Na, Z.; Liu, R.; Lin, B. Energy Consumption Minimization of Rotary-Wing UAVs for Data Distribution. *IEEE Commun. Lett.* **2023**, *27*, 1819–1823. [CrossRef]
9. Liu, L.; Zhang, S.; Zhang, R. CoMP in the Sky: UAV Placement and Movement Optimization for Multi-User Communications. *IEEE Trans. Commun.* **2019**, *67*, 5645–5658. [CrossRef]
10. Samir, M.; Sharafeddine, S.; Assi, C.; Nguyen, T.M.; Ghrayeb, A. Trajectory Planning and Resource Allocation of Multiple UAVs for Data Delivery in Vehicular Networks. *IEEE Netw. Lett.* **2019**, *3*, 107–110. [CrossRef]
11. Zhang, G.; Yan, Y.; Cui, M.; Chen, W.; Zhang, J. Online optimization design of flight routes for UAV base stations. *J. Electron. Inf. Technol.* **2021**, *43*, 3605–3611.
12. Wu, Q.; Zeng, Y.; Zhang, R. Joint Trajectory and Communication Design for Multi-UAV Enabled Wireless Networks. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 2109–2121. [CrossRef]
13. Cui, G.; Xu, Y.; Zhang, S.; Wang, W. Secure data offloading strategy for multi-UAV wireless networks based on minimum energy consumption. *J. Commun.* **2021**, *42*, 51–62.
14. Meng, S.; Zhu, S.; Wang, Z.; Zhang, R.; Han, J.; Liu, J.; Sun, H.; Qin, P.; Zhao, X. JDAPCOO: Resource Scheduling and Energy Efficiency Optimization in 5G and Satellite Converged Networks for Power Transmission and Distribution Scenarios. *Sensors* **2022**, *22*, 7085. [CrossRef] [PubMed]
15. Zhang, H.; Liu, H.; Cheng, J.; Leung, V.C.M. Downlink Energy Efficiency of Power Allocation and Wireless Backhaul Bandwidth Allocation in Heterogeneous Small Cell Networks. *IEEE Trans. Commun.* **2018**, *66*, 1705–1716. [CrossRef]
16. Arani, A.; Hu, P.; Zhu, Y. Fairness-Aware Link Optimization for Space-Terrestrial Integrated Networks: A Reinforcement Learning Framework. *IEEE Access* **2021**, *9*, 77624–77636. [CrossRef]
17. Cui, J.; Liu, Y.; Nallanathan, A. Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 729–743. [CrossRef]
18. Giannopoulos, A.; Spantideas, S.; Kapsalis, N.; Karkazis, P.; Trakadas, P. Deep Reinforcement Learning for Energy-Efficient Multi-Channel Transmissions in 5G Cognitive HetNets: Centralized, Decentralized and Transfer Learning Based Solutions. *IEEE Access* **2021**, *9*, 129358–129374. [CrossRef]

19. Ma, B.; Liu, Z.; Dang, Q.; Zhao, W.; Wang, J.; Cheng, Y.; Yuan, Z. Deep Reinforcement Learning of UAV Tracking Control Under Wind Disturbances Environments. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–13. [CrossRef]
20. Xue, Y.; Chen, W. Multi-Agent Deep Reinforcement Learning for UAVs Navigation in Unknown Complex Environment. *IEEE Trans. Intell. Veh.* **2024**, *9*, 2290–2303. [CrossRef]
21. Nomikos, N.; Gkonis, P.K.; Bithas, P.S.; Trakadas, P. A Survey on UAV-Aided Maritime Communications: Deployment Considerations, Applications, and Future Challenges. *IEEE Open J. Commun. Soc.* **2023**, *4*, 56–78. [CrossRef]
22. Fan, M.; Wu, Y.; Liao, T.; Cao, Z.; Guo, H.; Sartoretti, G.; Wu, G. Deep Reinforcement Learning for UAV Routing in the Presence of Multiple Charging Stations. *IEEE Trans. Veh. Technol.* **2023**, *72*, 5732–5746. [CrossRef]
23. Ding, C.; Wang, J.-B.; Zhang, H.; Lin, M.; Li, G.Y. Joint Optimization of Transmission and Computation Resources for Satellite and High Altitude Platform Assisted Edge Computing. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 1362–1377. [CrossRef]
24. Wang, Y.; Li, Z.; Chen, Y.; Liu, M.; Lyu, X.; Hou, X.; Wang, J. Joint Resource Allocation and UAV Trajectory Optimization for Space–Air–Ground Internet of Remote Things Networks. *IEEE Syst. J.* **2021**, *15*, 4745–4755. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.