IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 15, 2020, accepted February 28, 2020, date of publication March 2, 2020, date of current version March 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977831

Forecast Based Handover in an Extensible Multi-Layer LEO Mobile Satellite System

YITAO LI[1], WUYANG ZHOU[1], (Member, IEEE), AND SHENGLI ZHOU[2], (Fellow, IEEE)
[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China
[2]Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269, USA

Corresponding author: Wuyang Zhou (wyzhou@ustc.edu.cn)

This work was supported by Project JHZK2018-203-TS-YZ.

ABSTRACT Low earth orbit mobile satellite system (LEO-MSS) is the major system to provide communication support for the regions beyond the coverage of terrestrial network systems. However, passive handover happens frequently caused by the quick movement of LEO satellites in LEO-MSS. It not only causes the waste of radio resource, but also makes it hard to guarantee the quality of service (QoS), especially for user groups in hot-spot regions. To tackle this problem, we propose an extensible multi-layer network architecture to reduce the handover rate, especially group handover rate by introducing high-altitude platforms (HAPs) and terrestrial relays (TRs) to this system. We then propose a multi-layer handover management framework and also design different handover procedures based on handover forecast for different kinds of handovers according to the proposed architecture and framework to reduce handover delay and signalling cost. Furthermore, we propose a dynamic handover optimization to reduce the dropping probability and guarantee the QoS of mobile terminals. Numerical results show that the proposed architecture reduces group handovers significantly. The proposed handover procedures also provide better performance on delay and signalling cost compared with traditional handover protocols. With the proposed dynamic handover optimization, the proposed handover procedures provide better performance on dropping probability and throughput. The proposed dynamic handover optimization has an excellent performance on dropping probability while guaranteeing the QoS of mobile terminals.

INDEX TERMS LEO-MSS, multi-layer architecture, handover rate, handover management, handover procedure, handover optimization.

I. INTRODUCTION

A. MOTIVATION

Compared with terrestrial 5G networks, low earth orbit mobile satellite system (LEO-MSS) have a prominent superiority in global coverage [1] and play a critical role in remote and non-land regions [2] and emergency communications [3]. However, due to the large coverage radius and quick movement of LEO satellites, there are still many challenges in LEO-MSS. One of the most serious challenges is the frequent passive handover caused by quick movement of LEO satellites. According to the Iridium system, handovers between satellites happen every 10 minutes and handovers between beams with 3dB radius of 200 km happen within 1 minute for all the mobile terminals (MTs). This frequent passive handover problem leads to a large dropping probability especially for user groups in hot-spot regions. Also, it causes the waste of precious radio resource because the resource can not be used until handovers finish. The larger handover delay is, the more resource will be wasted. Besides, frequent passive handovers happen together in hot-spot regions with groups of users. There have not been handover strategies and schemes to manage these group handovers efficiently, which makes it difficult to guarantee the quality of service (QoS). However, these problems in LEO satellite systems have not attracted enough attention of researchers. In this paper, we focus on the problem caused by frequent passive handovers in LEO satellite systems to reduce handover rate, dropping probability, handover delay and signalling cost.

B. RELATED WORK

Most existing works focusing on the handover problems in satellite systems devoted themselves to analyze handover models and propose handover strategies [4]–[6]. The work

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

in [4] analyzed the blocking probability using a Markov model and a service time distribution model. However, they did not analyze the throughput and dropping probability or propose efficient handover optimization to improve the performance. The work in [5] proposed a seamless handover mechanism based on the proposed software-defined satellite networking architecture. It also simulated the performance of handover latency, throughput and user quality of experience. But the work did not consider the multi-beam coverage and the handovers between beams in the same satellite. In [6], a graph-based framework supporting a variety of handover strategies in LEO satellite system was proposed. Based on the proposed framework, the handover rates of the Iridium system and Globalstar system were evaluated using three handover strategies. Nevertheless, the strategies it used were conventional and QoS of users was ignored. The work in [7] designed a provably secure and efficient access protocol and handover mechanism for Internet of Things (IoT) in space information network. But the work focused on the security properties of access and handover while we are focusing on improving the dropping probability and throughput.

These works did not solve the frequent passive group handover problem in LEO-MSS. Targeted handover protocols and efficient strategies were also lacking.

C. CONTRIBUTIONS

In this paper, we focus on solving the problem caused by frequent passive handovers in LEO satellite systems. We aim at reducing handover rate, dropping probability, handover delay and signalling cost. The contributions of this paper are summarized as follows.

- Via introducing high-altitude platforms (HAPs) and terrestrial relays (TRs) to cover hot-spot regions, we propose an extensible multi-layer network architecture to solve the group handover problems and reduce the handover rate. We analyze handover rate improvement compared with traditional LEO-MSS.
- We propose a multi-layer handover management framework according to the architecture to achieve centralized handover management efficiently. According to the framework, we also design 6 handover procedures for 6 kinds of handovers in the proposed architecture to reduce handover delay and signalling cost. Average delay and signalling cost of the designed handover procedures are analyzed and evaluated.
- We design utility functions to describe the priority of MTs and propose a dynamic handover optimization to reduce the dropping probability according to the designed utility function. Also, we propose a power optimization method to increase the system throughput and reduce dropping probability while guaranteeing the QoS of mobile terminals. To solve the problem, Lagrange dual method, Karush-Kuhn-Tucker (KKT) condition and a gradient descent algorithm are used.
- Simulations are carried out to evaluate handover rate of the proposed architecture, delay and signalling cost of the designed handover procedures, dropping probability of the proposed dynamic handover optimization and power optimization method. Numerical results show that the proposed architecture has a much lower average handover rate after introducing HAPs and TRs. The designed handover procedures provide lower delay and signalling cost than terrestrial network protocols. With the proposed dynamic handover optimization, the proposed handover procedures also provide better performance on dropping probability and throughput. In addition, the proposed dynamic handover optimization and power optimization method reduce dropping probability significantly while guaranteeing the QoS of mobile terminals.

The remainder of this paper is structured as follows. The extensible multi-layer architecture based on LEO-MSS is proposed in Section II. In Section III, forecast based handover management framework is proposed and handover procedures are designed and analyzed. In Section IV, handover optimization is proposed. In Section V, solution techniques and algorithms are introduced. In Section VI numerical results are presented and discussed. Finally, we draw the conclusions and future directions in Section VII.

II. EXTENSIBLE MULTI-LAYER LEO-MSS ARCHITECTURE AND PERFORMANCE ANALYSIS

In this section, A multi-layer architecture is proposed to solve group handover problem and the handover rate improvement will be given. We introduce the architecture in terms of node types and functions, node deployment scheme, access and mobility management mechanism, link selection scheme and frequency spectrum planning.

A typical LEO-MSS of Walker constellation (e.g., the Globalstar and Iridium) consists of 48-80 LEO satellites with multi-beam to cover the ground, one data control center (DCC) and several earth stations (ES). On this basis, HAPs and TRs are further introduced into the proposed architecture to cover hot-spot regions. Geostationary earth orbit (GEO) satellites are also introduced for relay. Typical HAPs include big unmanned aerial vehicles (UAV) [8] and airships located between 10 km and 100 km high in the air [9], [10]. HAPs can move to the destination gently according to the demands. TRs are full duplex relays with an extra directional antenna pointing to the satellites. The deployment of TRs is flexible, unlimited and low-cost because no more wired network facilities are needed. Fig. 1 and Fig. 2 show the multi-beam access scenario and the proposed extensible multi-layer architecture.

Compared with the current LEO-MSS system, we summarize the advantages of the proposed extensible multi-layer architecture as follows.

- By introducing TRs and HAPs the proposed architecture can provide communication support to small MTs without power to connect with LEO satellites directly.
- HAPs increase the system flexibility significantly because they can move purposefully.

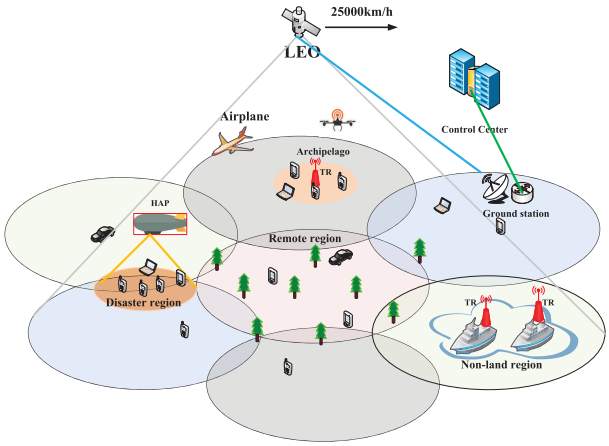


FIGURE 1. Multi-beam access scenarios.

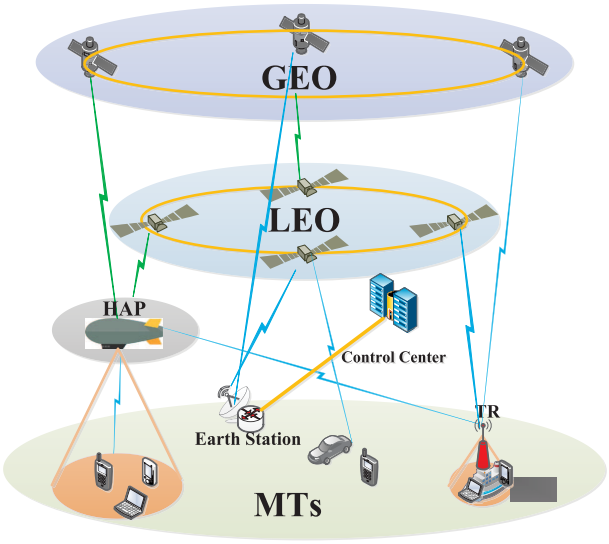


FIGURE 2. The proposed extensible multi-layer architecture.

- The capacity in the hot-spot regions can be greatly increased via HAPs and TRs.
- Via HAPs and TRs, the handover rate of the hot-spot regions can be reduced significantly. The handover signaling cost and difficulty of radio resource management can also be reduced.
- GEO satellites act as the relay to reduce the complexity of LEO routing significantly.

Note that the proposed architecture mainly focuses on the regions beyond the coverage of terrestrial 5G networks. This is different from the framework in [11], [12], addressing node identification, routing and other challenges in the future space-air-ground integrated networks, where GEO and LEO satellite networks are considered jointly with the terrestrial 5G networks. The proposed multi-layer architecture is much simpler than the space-air-ground integrated framework, and is feasible for practical implementations.

A. NODE DEPLOYMENT

In the proposed architecture, GEO satellites are uniformly deployed above the equator. Although 3 GEO satellites are able to cover the whole earth, 4 GEO satellites are deployed for more robustness. LEO satellites are deployed according to the Walker constellation [13]. TRs and HAPs are deployed according to the spatial distributions of MTs. Satellite positioning technology [14], [15] can be used for MTs to get their geographical positions. Then according to the mobility management protocols [16], their position information is reported to the system periodically. Afterwards, DCC performs the fitting of the reported position information and obtain the spatial distribution law of MTs. Finally, according to the spatial distribution law, TRs and HAPs are deployed [17]–[19].

HAP and TR are two different types of nodes introduced to cover hot-spot regions beyond the coverage of terrestrial networks. The differences are as follows. Firstly, the coverage regions of HAP are larger than those of TR. Thus, HAP is deployed preferentially over broad hot-spot regions, e.g. disaster regions, while TR is deployed preferentially on small hot-spot regions, e.g. big ships. Secondly, the flexibility of HAP is better than TR because HAP can reach almost everywhere on the earth while TR can only reach land regions or be equipped on the vehicles. Thirdly, the deployment cost of HAPs is larger than that of TRs. Finally, the available spectrum of HAPs and TRs to serve MTs is assumed to be different in this paper. TRs are deployed on the regions without enough terrestrial network facilities, so they can use the C band without considering the interference with terrestrial base stations. LEO satellites and HAPs provide much wider coverage, and hence better for them to use the Ka band to avoid the interference with terrestrial networks. For one hot-spot region, only one HAP or TR is needed. Whether to deploy HAP or TR depends on the specific application scenario.

B. ACCESS MANAGEMENT MECHANISM

The proposed architecture can provide communication service for spacecrafts, aircrafts, ships, vehicles and all kinds of hand-held terminals. The access nodes are selected according to the types of application scenarios, mobile terminals, services and position information. Specifically, only special MTs such as spacecrafts can be permitted to access to GEO satellites while ordinary MTs are not permitted to access to GEO satellites directly because the high energy consumption is beyond the scope of ordinary MTs. MTs in hot-spot regions and some specific scenarios preferentially access to HAPs and TRs while ordinary MTs in remote regions access to LEO satellites. There are two kinds of hot-spot regions. The broad hot-spot regions, including disaster regions, military bases and archipelagoes, are served by HAPs. The small hot-spot regions, including big ships, airplanes and small islands, are served by TRs. All of these access nodes are equipped with admission controller to manage the access of MTs, including authentication, load control and channel reservation.

Authentication is used to ensure the validity of MTs. Load control is to reject or postpone the access of new services and the allocation of resource for on-going services when meeting with the traffic congestion. Dedicated resources are reserved for special MTs and services to guarantee the access of high priority MTs and services.

C. MOBILITY MANAGEMENT

Mobility management is used to track, update and manage the position information. In the proposed architecture, LEO satellites and MTs have a larger range of movement than MTs in terrestrial networks, making mobility management more important. The position information of LEO satellites is updated and managed according to the orbits and time. HAPs, TRs and MTs get their own locations by satellite positioning techniques and report them to satellites periodically. The system manages and updates a table of position information of all nodes and MTs. The position information are divided into 4 levels, i.e., the GEO satellite, LEO satellite, LEO beam that cover them and the specific location information. Once registered into this system, MTs can be paged efficiently.

D. LINK SELECTION

In this architecture, links are divided into access links and routing links. The former links consists of LEO-MT links, HAP-MT links and TR-MT links, while the rest links are routing links. HAP-MT links and TR-MT links are prioritized for the MTs in hot-spot regions. LEO-MT links are selected for MTs in sparse regions. To form an integrated network, all the nodes should be connected together by routing links.

GEO satellites serve as the relay nodes in the proposed architecture.[1] The aim is to replace the complicated routing among LEO satellites, especially when the two sides of the communication are far from each other (e.g. one on the east coast of China and the other on the Atlantic). GEO satellites can connect with LEO satellites, HAPs and TRs. If the two sides of the communication are not far from each other, LEO satellites are selected as the routing nodes to avoid the large propagation delay. If the two sides of the communication are far from each other, GEO satellites are selected as the relay nodes to avoid LEO routing over a large number of hops.

E. SPECTRUM PLANNING

To provide global coverage, Ka ($17.3-20.2$ GHz and $26.5-40$ GHz) [20], [21] band is used in satellite-ground links and HAP-ground links to avoid co-channel interference (CCI) with C ($4-8$ GHz) band in terrestrial cellular networks due to its higher frequency and wider spectrum. TRs are usually deployed in the regions where terrestrial cellular networks are underdeveloped, thus TR-MT links use C band to avoid CCI with satellite-ground links. Satellite-HAP links and inter-satellite links are free-space links without atmosphere fading, where extremely high frequency (EHF, $40-300$ GHz) [22], [23] and laser [24] are used. Specifically, GEO-LEO links, GEO-HAP links, LEO-HAP links, LEO inter-orbit links, and HAP-HAP links use EHF while GEO-GEO links and LEO intra-orbit links use the laser. Detailed spectrums in different links are shown in Table 1.

[1]GEO satellites can also be used for other applications, e.g. the global positioning, meteorology, etc, to cover the increased operational cost.

TABLE 1. Spectrum in different links.

Links	Spectrum
GEO-GEO Link	Laser
LEO Intra-Orbit Inter-Satellite Link	Laser
LEO Inter-Orbit Inter-Satellite Link	EHF
GEO-LEO Link	EHF
GEO-HAP Link	EHF
LEO-HAP Link	EHF
HAP-HAP Link	EHF
LEO-MT Link	Ka Band
LEO-TR/ES Link	Ka Band
GEO-TR/ES Link	Ka Band
HAP-MT Link	Ka Band
TR-MT relay Link	C Band

HAP-ground links share the same Ka spectrum with LEO-ground links. We introduce frequency reuse schemes among LEO-ground links and HAP-ground links. Letting B_{total} and N be the total available bandwidth of LEO-ground links and the frequency reuse factor of LEO beams, respectively, n be the total reuse frequency used by the LEO beams adjacent to an HAP, the bandwidth available for the HAP is

$$B_{\text{HAP}} = \frac{N-n}{N} B_{\text{total}}. \tag{1}$$

F. HANDOVER RATE

The coverage of LEO-MSS is divided into non-land regions, remote regions and emergency regions. The target regions are fully covered by multi-beam LEO satellites in the proposed architecture, which requires seamless handovers among different beams and different satellites. In this paper, the coverage regions are divided into a whole sparse region with low-density of users and separate hot-spot regions with high-density of users. It is justified to assume that the users in hot-spot regions follow high-density uniform distributions and the users in the sparse region follow a low-density uniform distribution. Thus, the compound user distribution is expressed as

$$f(x,y) = \omega_0 f_0(x,y) + \sum_{i=1}^{I} \omega_i f_i(x,y), \tag{2}$$

where

$$\sum_{i=0}^{I} \omega_i = 1. \tag{3}$$

I is the total number of hot-spot regions, ω_0 is the percentage of MTs in the sparse region with the user distribution of

$f_0(x, y)$, and ω_i is the percentage of MTs in the ith hot-spot region with the user distribution of $f_i(x, y)$.

We now determine the handover rate, which is defined as the percentage of MTs switching from one beam to another in unit time (1 minute in the simulation) [25]. For instance, if handover happens on half of the users every minute, handover rate is 0.5/minute. The defined handover rate can be expressed as the integral of the user distribution over the altered coverage area in unit time, given by

$$R_H = \int_{\Delta S} f\,(x, y)dxdy, \tag{4}$$

where ΔS is the altered coverage area in unit time.

For the orbit at the altitude of 1500 km, the velocity of LEO satellite is about 25,000 km/h. It is more than 100 times larger than the velocity of most MTs. Therefore, we ignore the velocity of MTs served by LEO satellites. The velocity of MTs served by HAPs or TRs cannot be ignored. After GEO satellite is introduced for relay, handover rate is reduced via GEO-HAP-MT links and GEO-TR-MT links. ΔS is approximatively given by the subtraction of two circular sector regions and one prismatic region under the coverage of LEO satellites.

$$\Delta S = 2R^2 \cdot \arccos \frac{V_S \cdot t}{2R} - V_S \cdot t \cdot \sqrt{R^2 - \frac{V_S^2 t^2}{4}}, \tag{5}$$

where t and R are the unit time of this system and the radius of LEO satellite beam coverage region. V_S is the relative velocity of sub-satellite point on earth, given by

$$V_S = \frac{R_{\text{earth}} \cdot V_{\text{LEO}}}{R_{\text{earth}} + h_{\text{LEO}}}, \tag{6}$$

where V_{LEO} and R_{earth} are the velocity of LEO satellites and the radius of earth, respectively. h_{LEO} is the altitude of LEO satellites.

For the region served by TRs or HAPs, ΔS is given by

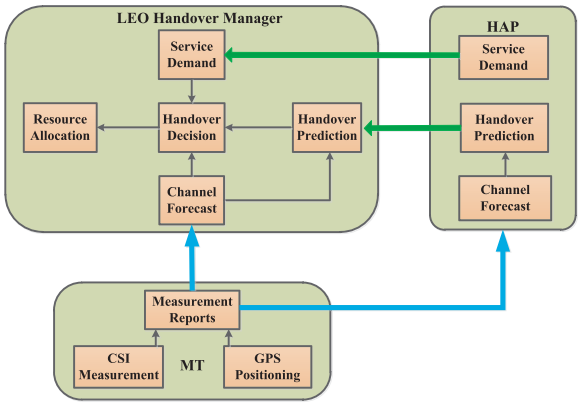
$$\Delta S = 2r^2 \cdot \arccos \frac{V_{\text{MT}} \cdot t}{2r} - V_{\text{MT}} \cdot t \cdot \sqrt{r^2 - \frac{V_{\text{MT}}^2 t^2}{4}}, \tag{7}$$

where V_{MT} and r are the average velocity of MTs and the radius of the coverage region of HAPs or TRs, respectively.

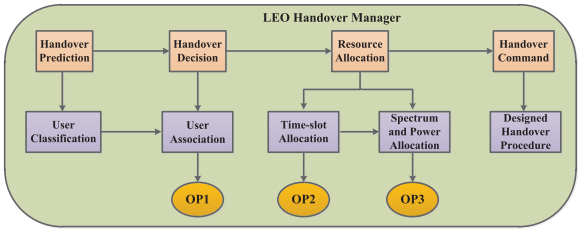
Since the velocity of LEO satellite is more than 100 times larger than that of most MTs, the average handover rate of hot-spot regions reduces more than 99% after served by GEO-TR-MT links or GEO-HAP-MT links.

III. HANDOVER MANAGEMENT AND PROCEDURE

In this section, according to the multi-layer and multi-beam architecture, we propose a handover framework to achieve centralized management of the multi-layer and multi-beam handovers. We also design handover procedures for these three types of handovers to reduce handover delay and signalling cost and compare the performance of delay and signalling cost with handover procedures in existing protocols.



(a) Handover manager



(b) Handover management process

FIGURE 3. Handover manager framework. (a) Handover manager. (b) Handover management process.

A. HANDOVER MANAGEMENT FRAMEWORK

The channel gains in satellite scenarios are much more deterministic than that of terrestrial 5G networks because the multi-path effect is negligible [26]. The sum of path loss and pitch angle fading of the channel gain can be calculated and predicted by the position information of MTs, TRs, HAPs and the regular orbits of LEO satellites. The sum of atmospheric fading and Rician small-scale fading of the channel gain is unchanged within one measurement and report period. Thus, the channel gains can be calculated and predicted, which is beneficial for handover management.[2]

The handover management of the proposed multi-layer architecture is shown in Fig. 3(a). All handovers in this system are managed by LEO handover managers. MTs report the channel state information (CSI) [27], [28] and position information periodically to LEO satellites or HAPs that cover them. LEO satellites and HAPs predict handovers according to the CSI and position information. Afterwards, HAPs report the handover prediction results and service demands to LEO satellites. Finally, LEO handover manager makes handover decisions and allocates resource.

Handover management process is shown in Fig. 3(b) and run at LEO satellite handover manager at every time-slot. LEO handover manager makes user classification after handover prediction. Users are divided into marginal users and central users. Then handover decisions are made to allocate

[2]The feedback of channel state information in satellite system is not timely because of the large propagation delay. In this paper, rely on predicted channels as the movements of LEO satellites are known.

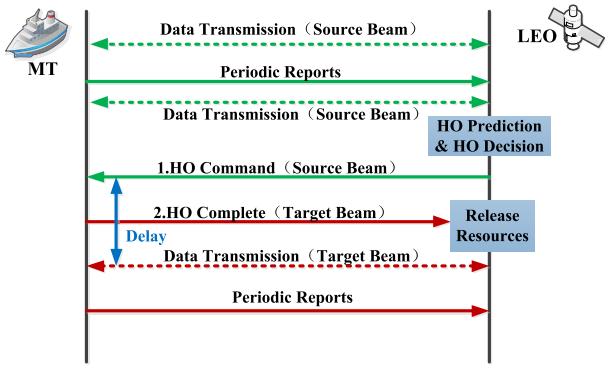


FIGURE 4. Beam handover procedure.

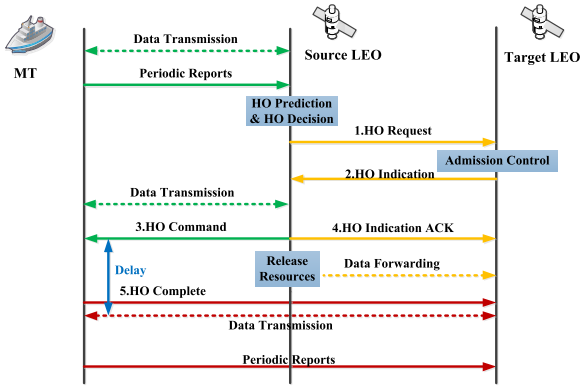


FIGURE 5. Inter-satellite handover procedure.

marginal users to the target beams. Handovers happen when the source beams and target beams are different. Afterwards, time-slot resource, spectrum resource and power are allocated to all users. Finally, the designed handover procedures are started and handover commands and signallings are sent to handover users with allocated resource.

B. HANDOVER PROCEDURE DESIGN

Handovers in the proposed architecture are divided into beam handover, inter-satellite handover and cross-layer handover. Beam handover refers to handover from one beam to another of the same LEO satellite and happens more than once a minute for all MTs. Inter-satellite handover refers to handover from one satellite to another satellite and happens more than once every 10 minutes for all MTs. Cross-layer handover refers to handover between LEO satellite and HAP/TR and happens at the edge of hot-spot regions.

According to our best knowledge, CCSDS [29] and DVB [30] are the main existing protocol standards used in the satellite network systems. However, there are no handover procedure standards in these protocols. In terrestrial network protocols, there are complicated handover procedures to guarantee the completeness and robustness of unpredictable handover scenarios due to complex channels in terrestrial 5G networks. But the handover procedure in terrestrial network protocols will cause unacceptable delay and large signalling cost if used in the proposed architecture, bringing waste of resource and difficulties to guarantee the QoS of MTs. Therefore, we design a novel handover procedure for each type of handover respectively to reduce handover delay and signalling cost. The designed handover procedures are based on positioning technique and handover prediction. Fig. 4 to Fig. 6 show the designed handover procedures for different types of handovers respectively.

In the designed handover procedure, positioning and time service technique provided by satellites are used to replace random access (RA) process. In the terrestrial network handover procedure, RA process is the key process to achieve uplink synchronization. According to [31], in satellite systems, MTs can keep the uplink synchronization with positioning accuracy in 10 meters which is already realized by existing works [32], [33].

The designed handover procedures are started at TR/LEO satellites and divided into three steps as follows.

The first step is handover (HO) prediction. Handovers are predicted according to the CSI and location of MTs, LEO satellites, TRs and HAPs. The prediction is implemented at the measurement results analysis module in LEO satellites. This module analyzes the reported CSI and positioning results of all MTs covered by this LEO satellite. The positioning results are used to estimate the path loss and pitch angle fading, while the CSI measurement results are used to calculate the sum of atmospheric fading and Rician small-scale fading. Thus, the channel gains from satellites to MTs are estimated and used as important parameters in handover prediction. The velocity of LEO satellites is more than 100 times higher than that of MTs. Thus, we consider it a reasonable assumption to ignore the change of MTs' geographical locations until the next report. Besides, the atmospheric fading and Rician small-scale fading of each link are assumed to be unchanged until the next CSI report. Therefore, the channel gains between satellites and MTs can be calculated and handovers can be forecast accurately.

The second step is handover (HO) decision. After an imprecise prediction, handovers should be decided at the accurate time-slots. Handover decisions are made according to the designed utility function in Section III.

In the proposed architecture, HAPs share the same spectrums with LEO satellites. To manage the resource and coordinate the interference unitedly and efficiently, managements are implemented at the central resource manager in LEO satellites, including handover decisions. Therefore, HAP just follows the received results calculated by the LEO satellite that covers it.

The third step is resource release. To reduce the waste of resource, LEO satellites can release the downlink (DL) resource of MTs after sending HO command. The uplink (UL) resource should not be released until receiving handover complete message.

Fig. 7 illustrates a frame structure for handover signallings. All the signallings contain the User ID, the Signalling ID and the Target ID. The User ID refers to the identity of the

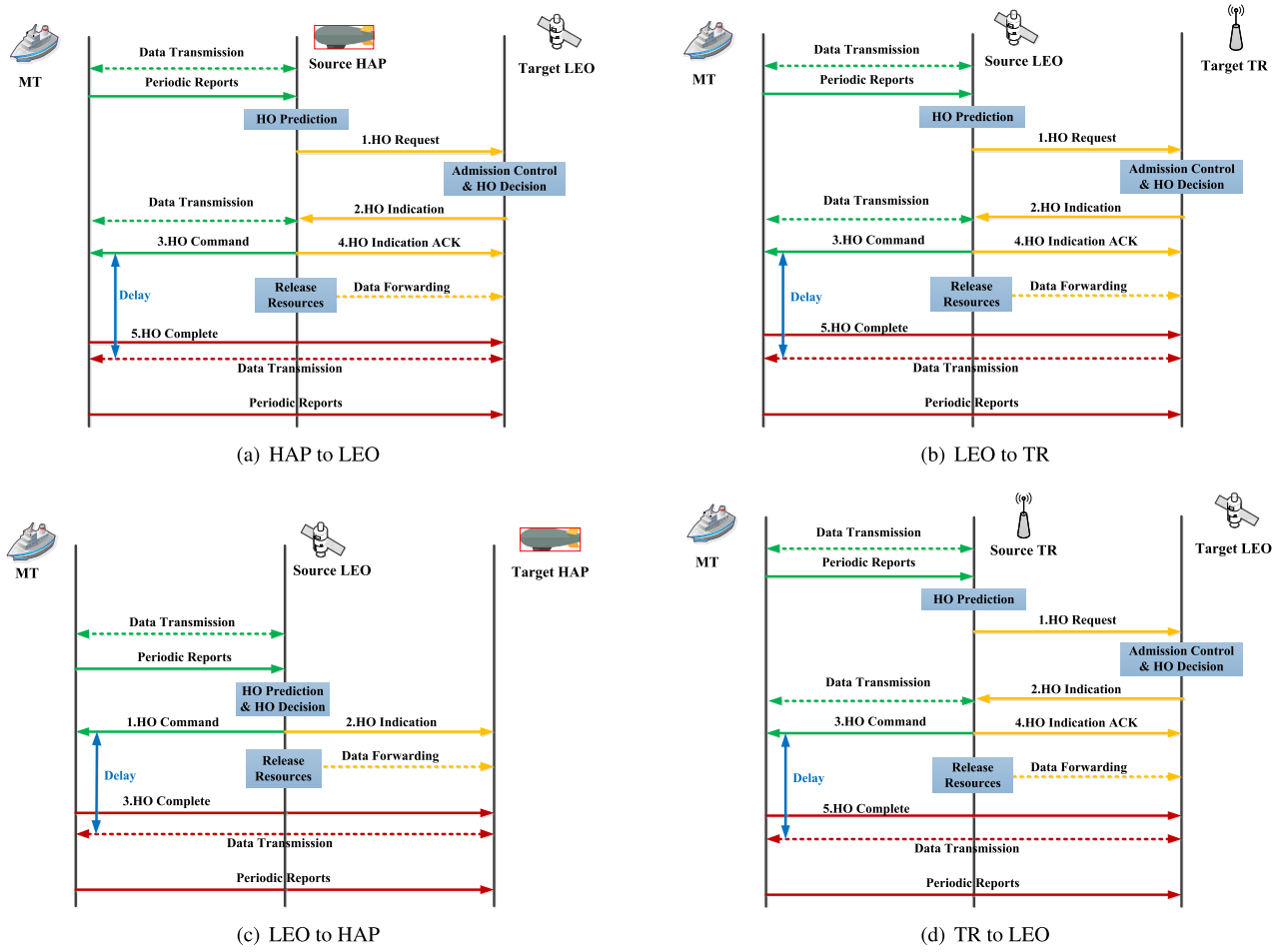


FIGURE 6. Cross-layer handover procedure.

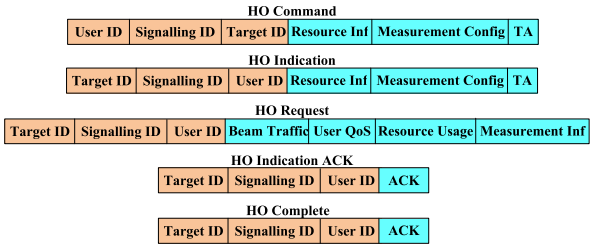


FIGURE 7. Handover signalling frame.

handover user, the Target ID refers to the identity of the target nodes and the signalling ID refers to the signalling type. HO Command and HO Indication contain the Resource Inf, Measurement Config and TA, where

- Resource Inf contains all the information of the resource allocation results, e.g. power, time-slot, bandwidth, frequency, adaptive modulation and coding.
- The Measurement Config is used to indicate the measurement configuration for the handover user, e.g. measurement period and report period.
- TA (time advance) is used for the synchronization.

HO Request contains the Beam Traffic, User QoS, Resource Usage and Measurement Inf, where

- The beam traffic contains the information of the traffic of this LEO beam, HAP beam or TR beam.
- The User QoS contains all the QoS demands of the handover user.
- Resource Usage refers to the resource usage and utilization of this LEO beam, HAP beam or TR beam.
- The Measurement Inf contains the measurement configuration in the source beam of the handover user.

HO Indication ACK and HO Complete contain the ACK (acknowledgement character) for the handshake. The handover signaling is generated at the handover manager and transmitted as short messages via the control channel.

- **Signalling Cost**

There are only two signallings (handover command and handover complete) between the MT and LEO satellite during beam handover. The signalling cost is the smallest among all types of handovers.

Compared with beam handover, there is one additional signalling during cross-layer handover from LEO to HAP. The additional signalling is handover indication from LEO

satellite to the target HAP. No interaction is needed after the handover decision is made in the source LEO satellite.

Compared with beam handover, during inter-satellite handover, cross-layer handover from LEO satellite to TR, cross-layer handover from TR to LEO satellite and cross-layer handover from HAP to LEO satellite, another three signallings are needed. These three signallings are used to achieve interaction between the source nodes and the target nodes.

- **Signalling Delay**

According to the designed procedures, downlink is broken off when handover command is sent and rebuilt after receiving handover complete. Thus, downlink delay is calculated as the sum of 2 propagation delay and 2 processing delay. Uplink is broken off after receiving handover command and rebuilt after sending handover complete. Thus, uplink delay is calculated as 1 processing delay.

We ignore the propagation delay from HAP and TR to MTs. The propagation delay of cross-layer handover depends on the relative position of HAP/TR and LEO satellite.

IV. HANDOVER OPTIMIZATION

Handover decision requires a comprehensive consideration of multiple factors to make more accurate decision. We propose a utility function including user priority, minimum rate requirement, delay requirement, channel gain and the data traffic of beams as the basis of handover decision. Then time-frequency resource and transmission power are allocated to minimize the dropping probability.

The probability of handover between TRs and LEO satellites is very small and the handover problem between TRs and LEO satellites has little difference with that of terrestrial networks. Therefore, we consider handovers among beams, between LEO satellites, and between HAP and LEO satellite in this section. According to the existing work [34], the quasi-static characteristic of HAPs will bring great challenges to handovers if HAPs are equipped with multiple beams. Thus, HAPs are equipped with directional single beam to cover the hot-spot regions in this paper.

As mentioned in Section III-B, HAP does not need to implement the strategies and algorithms, it just follows the results calculated by the LEO satellite that covers it. We consider the HAP as a special beam for simplicity. And we do not consider the cooperative transmission for MTs provided by different beams or nodes at the same time.

A. CHANNEL MODEL

The channel model mainly consists of the pitch angle fading, path loss, atmospheric fading and Rician small-scale fading [35]. The pitch angle fading is given by [36]

$$G_H(\psi) = Ap_{\text{eff}} \cdot \cos(\psi)^{\eta} \frac{32 \log 2}{2\left(2 \arccos\left(\sqrt[\eta]{0.5}\right)\right)^2}, \quad (8)$$

where Ap_{eff} is the antenna aperture efficiency (assumed to be unity), η is the roll-off factor of the antenna. ψ is the pitch angle. For MT j located in (x_j, y_j) covered by beam i, the pitch angle can be calculated approximatively as

$$\psi_{i,j} = 2 \arctan \frac{\sqrt{(x_j - o_{i,1})^2 + (y_j - o_{i,1})^2}}{2h}, \quad (9)$$

where h is the altitude of satellites or HAPs, $(o_{i,1}, o_{i,2})$ is the center position of the coverage region of beam i. The atmospheric fading is given by

$$A(d) = 10^{\left(\frac{3d\chi}{10h}\right)}, \quad (10)$$

where χ is the attenuation through the clouds and rain in dB/km, d is the propagation distance between HAPs or satellites and the MT, given approximatively by

$$d = \sqrt{h^2 + (x_j - o_{s,1})^2 + (y_j - o_{s,2})^2}, \quad (11)$$

where $(o_{s,1}, o_{s,2})$ is the position right below satellite or HAP.

The specific expression of the channel model is expressed as

$$G = \left(\frac{c_{\text{light}}}{4\pi d f_c}\right)^2 \cdot G_H(\psi) \cdot A(d) \cdot \varphi, \quad (12)$$

where φ is the Rician small-scale fading. c_{light} and f_c are the speed of light and the carrier frequency.

The received power is given by

$$P_{\text{rx}} = P_{\text{tx}} \cdot G_{\text{tx}} \cdot G \cdot G_{\text{rx}}, \quad (13)$$

where P_{tx} is the transmission power, G_{tx} and G_{rx} are the antenna gains of the transmitter and receiver, respectively.

B. UTILITY DESIGN

The handover optimization is proposed to reduce the dropping probability and increase the throughput. Dropping probability in this paper refers to the probability of on-going services which are broken off. Services are broken off if the packets are not transmitted during their maximum time delay. Thus, dropping probability is calculated as the ratio of broken services number and the total service number, expressed as

$$P_d = \frac{N_{\text{drop}}}{N_{\text{total}}}, \quad (14)$$

where N_{drop} and N_{total} are the numbers of broken services and total services, respectively.

The utility function is used in user association and time-slot allocation, and designed according to the following observations.

- To reduce the dropping probability, on-going services should have a higher priority than new services when allocating resource.
- As the channel gain rises, the channel gain utility should also rise. to make use of the predicted channel gain, the channel gain at increasing tendency should have a little larger utility than that at the decreasing tendency when the channel gain is the same.
- The packet with a larger delay than the upper delay limit should be dropped. The delay utility should decrease

rapidly as delay increases and approaches the upper delay limit [37].
- The data traffics of beams should be considered when determining the handover. MTs should be preferentially served by the unoccupied beams.
- To reduce the dropping probability, services whose rate demand is lower than the surplus capacity of beams should be served preferentially.

Therefore, for every time-slot t, the traffic utility function $U_{\mathrm{tr}}^{i}(t)$ is defined as

$$U_{\mathrm{tr}}^{i}(t)=\exp\left(-R_{\mathrm{tr}}^{i}(t)\right), \tag{15}$$

where $R_{\mathrm{tr}}^{i}(t)$ is the ratio of traffic to the average capacity of beam i.

The rate demand utility function $U_{\mathrm{r}}^{i,j}(t)$ is defined as

$$U_{\mathrm{r}}^{i,j}(t)=\frac{1}{1+\left(\exp\left(20\left(r_{i,j}(t)-1.2\right)\right)\right)^{p_{\mathrm{r}}}} \tag{16}$$

and shown in Fig. 8(a), where $r_{i,j}(t)$ is the ratio of of rate demand of MT j to surplus capacity of beam i and p_{r} is the speed of descent.

The channel gain utility function $U_{g}^{i,j}(t)$ is defined as

$$U_{\mathrm{g}}^{i,j}(t)=G_{i,j}(t)\cdot\exp\left(p_{\mathrm{g}}\left(G_{i,j}(t-\Delta t)-G_{i,j}(t)\right)\right) \tag{17}$$

and shown in Fig. 8(b), where $G_{i,j}(t)$ and p_{g} are the channel gain between MT j and beam i and the weight of the predicted channel gain.

The delay utility function $U_{d}^{j}(t)$ is defined as

$$U_{\mathrm{d}}^{j}(t)=1-\exp\left(\frac{t-d_{j}^{\max}(t)}{p_{\mathrm{d}}}\right) \tag{18}$$

and shown in Fig. 8(c), where p_{d} and $d_{j}^{\max}(t)$ are the speed of descent and the upper delay limit, respectively.

C. PROBLEM FORMULATION

According to Fig. 3(b), we divide the handover problem into user association, time slot, frequency resource and power allocation. User association is made first via comparing the designed user association utility function of MTs in different beams. Then MTs with better time-slot allocation utility are allocated time-slot preferentially in each beam. Finally, spectrum and transmission power are optimized to maximize throughput and minimize dropping probability.[3] The power optimization of downlinks are more difficult than that of uplinks. Thus we consider the downlink in this subsection. Important notations used in this paper are shown in Table 2.

[3]These three steps are run by the LEO satellite handover manager at every time-slot. The results of utility functions at the current time-slot t are affected by the results of power allocation at the previous time slot. Due to the processing and propagation delay, the problem of time-slot t is calculated at an early time-slot t_0 based on the predicted information.

TABLE 2. Important notations.

K	total bandwidth
I	total beam including HAP
J	total MTs
k	the k^{th} spectrum resource block (RB)
i	the i^{th} beam
j	the j^{th} MT
J_i	total MTs served by beam i
K_i	total bandwidth of beam i
j_i	the j^{th} MT served by beam i
$j_{i,k}$	MT j served by beam i with RB k

1) USER ASSOCIATION

The user association is the first step. Handovers happen when users are associated from one beam to another. The designed user association utility function is expressed as

$$U_{1}^{i,j}(t)=\left[U_{\mathrm{tr}}^{i}(t)\right]^{\omega_1}\left[U_{\mathrm{g}}^{i,j}(t)\right]^{\omega_2}, \tag{19}$$

where ω_1 and ω_2 are the weights of U_{tr}^{i} and $U_{\mathrm{g}}^{i,j}$, respectively. The user association problem is formulated as OP1.

$$\begin{aligned}\text{OP1}\quad &\max\sum_{i=1}^{I}\sum_{j=1}^{J}a_{i,j,t}U_{1}^{i,j}(t)\\ &\text{s.t. C1}:\ \sum_{i=1}^{I}a_{i,j,t}\le 1,\quad \forall j,t\\ &\text{C2}: a_{i,j,t}\in\{0,1\},\quad \forall i,j,t,\end{aligned} \tag{20}$$

where I is the total number of beams including beams of HAPs, and J is the total number of MTs. $a_{i,j,t}=1$ refers to associate the j^{th} MT to the i^{th} beam at time-slot t. C1 is the constraint that each MT can only be associated to one beam at the same time. C2 is the constraint of the value of $a_{i,j,t}$. OP1 is a zero-one integer linear programming problem. The simplex algorithm [38] is a low-complexity algorithm to solve it at every time slot.

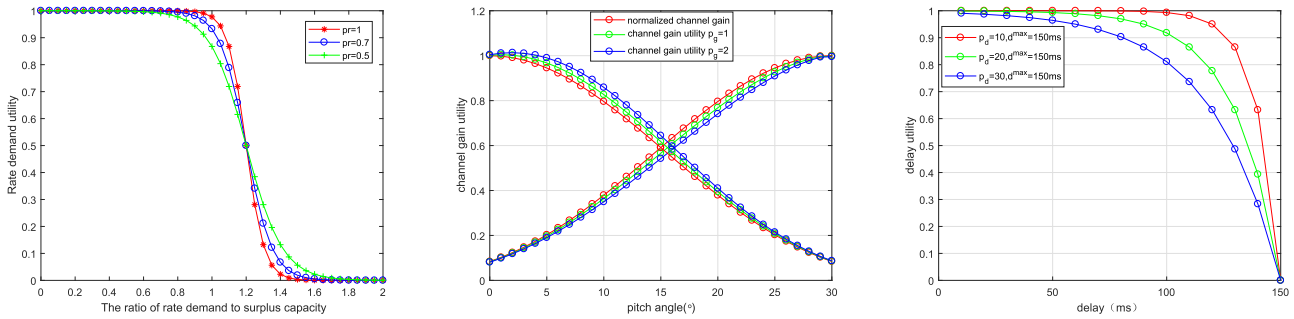
2) TIME-SLOT ALLOCATION

After user association, the j^{th} MT served by beam i is represented as j_i. We design a time-slot allocation utility function based on the predicted information to allocate the optimal time-slot resource for all MTs.

The designed time-slot allocation utility function is expressed as

$$\begin{aligned}U_{2}^{i,j}(t)=&\left[U_{\mathrm{s}}^{j}(t)\right]^{\omega_3}\left[U_{\mathrm{g}}^{i,j}(t)\right]^{\omega_4}\left[U_{\mathrm{r}}^{i,j}(t)\right]^{\omega_5}\\ &\cdot\left[\left(\max\left\{U_{\mathrm{d}}^{j}(t),\varepsilon_{\mathrm{d}}\right\}\right)\right]^{-\omega_6},\end{aligned} \tag{21}$$

where $U_{\mathrm{s}}^{j}(t)$ is the priority of the service of MT j, ω_3 to ω_6 are the weights of each utility. ε_{d} is the threshold of the delay utility function, making the tradeoff between the fairness and throughput.



(a) Rate demand utility (b) Channel gain utility (c) Delay utility

FIGURE 8. **Utility functions. (a) Rate demand utility. (b) Channel gain utility. (c) Delay utility.**

The matrix $\boldsymbol{B}_t = [b_{i,j_i,t}]$ is defined as the allocation results at time-slot t. $b_{i,j_i,t}$ is the allocation results of MT j_i in beam i at t. $b_{i,j_i,t} = 1$ refers to allocate the timeslot to MT j_i in beam i at t. The optimal time-slot allocation problem is formulated as OP2.

$$\begin{aligned} \text{OP2} \quad & \max \sum_{i=1}^{I} \sum_{j_i=1}^{J_i} b_{i,j_i,t} U_2^{i,j_i}(t) \\ & \text{s.t. C1} : \sum_{j_i=1}^{J_i} b_{i,j_i,t} \le K_i, \quad \forall i, t \\ & \text{C2} : b_{i,j_i,t} \in \{0,\ 1\}, \quad \forall i, j_i, t, \end{aligned} \tag{22}$$

where J_i is the total number of MTs associated to beam i and K_i is the total bandwidth of beam i. C1 is the constraint of total bandwidth. C2 is the constraint of the value of $b_{i,j_i,t}$. OP2 is also a zero-one integer linear programming problem. The simplex algorithm [38] is a low-complexity algorithm to solve it at every time slot.

3) SPECTRUM AND POWER ALLOCATION

Minimize the dropping probability is equivalent to maximize the system throughput when allocating spectrums and powers. According to our previous work [25], the best spectrum reuse scheme is four color reuse among multiple beams. HAP reuse the spectrum of LEO satellites. The spectrums and powers of LEO beams should be allocated before those of HAPs. LEO satellites allocate spectrums and powers of HAPs according to the spectrum and power allocation results of LEO beams. According to the previous research results of our team [39], the interference of HAP to MTs served by LEO beams can be ignored most of the time. Thus, the spectrum and power allocation of this system is divided into two steps. The first step is to allocate the spectrum and power of multi-beam LEO satellites. The second step is to allocate the spectrum and power of HAPs according to the interference of LEO beams.

The rate for the MT $j_{i,k}$ at t is given by

$$R_{j_{i,k},t} = B\log_2\left(1 + \frac{s_{j_i,k,t} P_{i,k,t} G_{i,j_i,t}}{\sum\limits_{\substack{i'=1 \\ i' \ne i}}^{I} \sum\limits_{j_{i'}=1}^{J_{i'}} s_{j_{i'},k,t} P_{i',k,t} G_{i',j_i,t} + \sigma^2}\right), \tag{23}$$

where B is the bandwidth of each RB. $P_{i,k,t}$ and $G_{i,j_{i,k},t}$ are the transmission power of beam i with RB k and the channel gain between beam i and MT $j_{i,k}$ at t, respectively. $s_{j_i,k,t} = 1$ refers to allocate spectrum k to MT j in beam i at t and $s_{j_i,k,t} = 0$ refers not to allocate spectrum k to MT j in beam i at t.

The system throughput at t is given by

$$\text{Throughput}_t = \sum_{i=1}^{I} \sum_{k=1}^{K} R_{j_{i,k},t}. \tag{24}$$

The spectrum and power allocation for MTs in this system is formulated as OP3

$$\begin{aligned} \text{OP3} \quad & \max \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{j_i=1}^{J_i} s_{j_i,k,t} R_{j_{i,k},t} \\ & \text{s.t.} \\ & \text{C1} : \sum_{j_i=1}^{J_i} s_{j_i,k,t} \le 1, \quad \forall k, t \\ & \text{C2} : \sum_{k=1}^{K} s_{j_i,k,t} \le 1, \quad \forall j_i, t \\ & \text{C3} : 0 \le s_{j_i,k,t} \le 1, \quad \forall j_i, k, t \\ & \text{C4} : R_{\min}^{j_{i,k},t} \le R_{j_{i,k},t} \le R_{\max}^{j_{i,k},t}, \quad \forall j_{i,k}, t \\ & \text{C5} : \sum_{k=1}^{K} P_{i,k,t} \le P_{\max}^{i}, \quad \forall i, t \\ & \text{C6} : 0 \le P_{i,k,t} \le P_{\max}, \quad \forall i, k, t, \end{aligned} \tag{25}$$

where $R_{\min}^{j_{i,k},t}$ and $R_{\max}^{j_{i,k},t}$ are the minimal and maximal rate demands of MT j_i in RB k at t. $P_{\max}^{i}$ and $P_{\max}$ are the upper power limits of beam i and every RB.

V. SOLUTION TECHNIQUES AND ALGORITHMS

OP1 and OP2 are zero-one integer linear programming problems. The simplex algorithm [38] is a low-complexity algorithm to solve them at every time slot. In this section, we will focus on how to obtain the solutions of OP3.

A. SOLUTION TECHNIQUE

For simplicity we omit index t in this section because the solution method is not relevant the time-slot series. We solve OP3 by dividing it into spectrum allocation problem OP4 and power optimization problem OP5.

The spectrum allocation for MTs in this system is formulated as OP4.

$$\begin{aligned}\text{OP4}\quad \max\ & \sum_{k=1}^{K} s_{j_i,k}\overline{R}_{j_i,k}\\ \text{s.t. C1}:\ & \sum_{k=1}^{K} s_{j_i,k} \le 1, \quad \forall j_i\\ \text{C2}:\ & \sum_{j_i=1}^{J_i} s_{j_i,k} \le 1, \quad \forall k\\ \text{C3}:\ & s_{j_i,k} \in \{0,\ 1\}, \quad \forall j_i, k, \end{aligned} \tag{26}$$

where $\overline{R}_{j_i,k}$ is the expected average rate of j_i in beam i with spectrum resource k. Letting $\overline{P}$ be the expected average transmission power for equal power transmission, $\overline{R}_{j_i,k}$ is expressed as

$$\overline{R}_{j_i,k} = B\log_2\left(1+\frac{\overline{P}_{i,k}G_{i,j_i}}{\sum\limits_{i'=1,i'\neq i}^{I}\overline{P}_{i',k}G_{i',j_i}+\sigma^2}\right). \tag{27}$$

OP4 is also a zero-one integer linear programming problem. The simplex algorithm [38] is a low-complexity algorithm to solve it at every time slot.

The power optimization problem is formulated as OP5.

$$\begin{aligned}\text{OP5}\quad \max\ & \sum_{i=1}^{I}\sum_{k=1}^{K} R_{j_i,k}\\ \text{s.t. C1}:\ & R_{\min}^{j_{i,k}} \le R_{j_i,k} \le R_{\max}^{j_{i,k}}, \quad \forall j_{i,k}\\ \text{C2}:\ & \sum_{k=1}^{K} P_{i,k} \le P_{\max}^{i}, \quad \forall i\\ \text{C3}:\ & 0 \le P_{i,k} \le P_{\max}, \quad \forall i, k, \end{aligned} \tag{28}$$

OP5 is a non-convex problem. According to [40], OP5 is with strong duality if I and K are big enough. We can find the optimal results of OP5 by using the Lagrangian dual method.

The Lagrangian function of OP5 is given by

$$\begin{aligned}L(P_{i,k},\lambda^{\min},\lambda^{\max},\mu) = & \sum_{i=1}^{I}\sum_{k=1}^{K} R_{j_i,k}\\ & + \sum_{i=1}^{I}\sum_{k=1}^{K}\lambda_{i,k}^{\min}\left(R_{j_i,k}-R_{\min}^{j_{i,k}}\right)\\ & + \sum_{i=1}^{I}\sum_{k=1}^{K}\lambda_{i,k}^{\max}\left(R_{\max}^{j_{i,k}}-R_{j_i,k}\right)\\ & + \sum_{i=1}^{I}\mu_i\left(P_{\max}^{i}-\sum_{k=1}^{K}P_{i,k}\right), \end{aligned} \tag{29}$$

where $(\lambda^{\min},\lambda^{\max}) = \{(\lambda_{i,k}^{\min},\lambda_{i,k}^{\max}), k\in K, i\in I\} \ge 0$ and $\mu = \{\mu_i, i\in I\} \ge 0$ are Lagrangian dual variables. The lagrangian equation of OP5 is expressed as

$$h(\lambda^{\min},\lambda^{\max},\mu) = \sup_{P_{i,k}\ge 0} L(P_{i,k},\lambda^{\min},\lambda^{\max},\mu). \tag{30}$$

The dual problem of OP5 is expressed as

$$\min_{(\lambda^{\min},\lambda^{\max},\mu)\ge 0} h(\lambda^{\min},\lambda^{\max},\mu). \tag{31}$$

$L(P_{i,k},\lambda^{\min},\lambda^{\max},\mu)$ is a convex function when $P_{i,k}\ge 0$. The optimal power should satisfy the constraint of KKT condition

$$\frac{\partial L}{\partial P_{i,k}} = 0, \tag{32}$$

According to Equation. 32, we get

$$\sum_{i'=1}^{I} P_{i',k}G_{i',j_{i,k}} = \frac{\left(1+\lambda_{i,k}^{\min}-\lambda_{i,k}^{\max}\right)BG_{i,j_{i,k}}\log_2 e}{\mu_i} - \sigma^2. \tag{33}$$

Letting

$$c_{i,k} = \frac{\left(1+\lambda_{i,k}^{\min}-\lambda_{i,k}^{\max}\right)BG_{i,j_{i,k}}\log_2 e}{\mu_i} - \sigma^2, \tag{34}$$

$$\begin{aligned}\boldsymbol{C}_k &= \left[c_{1,k}c_{2,k}\cdots c_{I,k}\right]\\ \boldsymbol{P}_k &= \left[P_{1,k}P_{2,k}\cdots P_{I,k}\right]\\ \boldsymbol{G}_k &= \begin{bmatrix} G_{1,j_{1,k}} & G_{1,j_{2,k}} & \cdots & G_{1,j_{I,k}}\\ G_{2,j_{1,k}} & G_{2,j_{2,k}} & \cdots & G_{2,j_{I,k}}\\ \vdots & \vdots & & \vdots\\ G_{I,j_{1,k}} & G_{I,j_{2,k}} & \cdots & G_{I,j_{I,k}}\end{bmatrix}, \end{aligned} \tag{35}$$

The optimal transmission powers can be expressed as

$$\boldsymbol{P}_k^{*} = \boldsymbol{C}_k\boldsymbol{G}_k^{-1}\forall k, \tag{36}$$

where $\boldsymbol{G}_k^{-1}$ is the inverse matrix of $\boldsymbol{G}_k$.

The Lagrangian equation is expressed as

$$\begin{aligned}&h(\lambda^{\min},\lambda^{\max},\mu)\\ &\quad = \sup_{P_{i,k}\ge 0} L(P_{i,k},\lambda^{\min},\lambda^{\max},\mu)\end{aligned}$$

$$= \sup_{P_{i,k}\geq 0} \sum_{i=1}^{I}\sum_{k=1}^{K} \lambda_{i,k}^{\min}\left(R_{j_{i,k}}^{*} - R_{\min}^{j_{i,k}}\right) + \sum_{i=1}^{I}\sum_{k=1}^{K} \lambda_{i,k}^{\max}\left(R_{\max}^{j_{i,k}} - R_{j_{i,k}}^{*}\right) + \sum_{i=1}^{I}\sum_{k=1}^{K} R_{j_{i,k}}^{*} + \sum_{i=1}^{I} \mu_i\left(P_{\max}^{i} - \sum_{k=1}^{K} P_{i,k}^{*}\right), \quad (37)$$

where $R_{j_{i,k}}^{*}$ is the optimal rate for the MT j_i with RB k. $h(\lambda^{\min}, \lambda^{\max}, \mu)$ is a convex function, but its gradient may not exist at some points. We use sub-gradient method [41] to find the results. Let

$$\lambda_{i,k}^{\min}(l+1) = \left[\lambda_{i,k}^{\min}(l) - \beta_1^{i,k}(l)\left(R_{j_{i,k}} - R_{\min}^{j_{i,k}}\right)\right]^{+}$$
$$\lambda_{i,k}^{\max}(l+1) = \left[\lambda_{i,k}^{\max}(l) - \beta_2^{i,k}(l)\left(R_{\max}^{j_{i,k}} - R_{j_{i,k}}\right)\right]^{+}$$
$$\mu_i(l+1) = \left[\mu_i(l) - \beta_3^{i}(l)\left(P_{\max}^{i} - \sum_{k=1}^{K} P_{i,k}\right)\right]^{+}, \quad (38)$$

where $(x)^{+} \equiv \max\{0, x\}$, $l \in \{1, 2, \ldots, l_{\max}\}$ is the iteration index. $\beta_m^{i,k}(l) = \beta_m^{i,k}(1)/l$ is the step length, where $m = 1, 2, 3$.

Algorithm 1 Algorithm for OP5

Initialize $l_{\max}$, threshold ε and step length $\beta_m^{i,k}(1)$, where $m = 1, 2, 3$.
Initialize Lagrangian $\lambda_{i,k}^{\min}(1)$, $\lambda_{i,k}^{\max}(1)$ and $\mu_i(1)$.
Initialize $l = 1$, $n = 0$ and $h(0)$.
while $n = 0$ **do**
  for $i = 1 : I$ **do**
    for $i = 1 : K$ **do**
      Update $P_{i,k}^{*}(l)$
    end for
  end for
  Calculate $h(l)$.
  if $|h(l) - h(l-1)| \geq \varepsilon$ and $l \leq l_{\max}$ **then**
    for $i = 1 : I$ **do**
      for $i = 1 : K$ **do**
        Update $\beta_m^{i,k}(l+1)$, Lagrangian $\lambda_{i,k}^{\min}(1+1)$, $\lambda_{i,k}^{\max}(1+1)$ and $\mu_i(1+1)$.
      end for
    end for
    $l = l + 1$.
  else
    $n = 1$
  end if
end while
Output optimal power $P_{i,k}^{*}$.

B. PROPOSED ALGORITHMS

Since the dual problem in equation (31) is convex and has the same optimal result as OP5, we adopt the sub-gradient algorithm [41] to find the solution based on the KKT condition. The algorithm is shown in Algorithm 1 and run in LEO satellites with the complexity of $O(LIK)$, where L is the maximum iterations in the algorithm. The algorithm iterates along the sub-gradient and reduces the step size gradually to converge to the optimal result. Once the difference between the objective functions of the latest two iterations is smaller than a predefined threshold ε, the iteration stops.

VI. NUMERICAL RESULTS AND DISCUSSION

In this section, we carry out simulations to evaluate handover rate improvement of the proposed multi-layer architecture compared with traditional LEO-MSS. Then we evaluate handover delay and signalling cost of the designed handover procedures compared with handover procedure in terrestrial network protocols [42]. Afterwards, we evaluate the performance of the proposed dynamic handover optimization and power optimization method. The simulations are implemented by MATLAB R2017a and the simulation results will be presented and discussed as follows.

A. SYSTEM PARAMETERS

The important system parameters are shown in Table 3. The altitude of LEO satellite with multi-beam is about 1500 km, and the ground-track speed V_{LEO} is approximately 25000 km/h. The factor of LEO satellite antenna η is 20, which means the LEO beam coverage radius of 3 dB is about 200 km. The channel model of HAP-MT link is the same as LEO-MT link, as shown in equation 12. However, HAP is equipped with single beam instead of multi-beam and the antenna parameters and the path loss constants of HAP are different from those of LEO satellites. The altitude of HAP is 20 km, and the antenna factor η of HAP is 1, whose beam coverage radius of 3 dB is about 20 km.

To validate the handover rate improvement of the proposed multi-layer architecture, as in the previous analysis, we focus on the case that GEO satellite is connected directly to HAPs or TRs for routing. In actual scenarios, the probability of this case is P_{GEO}. Here, we set $P_{\text{GEO}} = 0.5$. And the average velocity of MTs in hot-spot regions is $V_{\text{MT}} = 5$ m/s.

To validate handover delay of the designed handover procedures, we set the average processing delay to be 4 ms per handshake [42]. There are 19 beams per LEO satellite.

To evaluate the proposed dynamic handover optimization and power control method, we consider a rectangular region covered by 24 LEO beams and one HAP beam for simplicity. The frequency reuse factor $N = 4$. The total bandwidth of LEO-ground downlink is 240 MHz. The bandwidth of each RB $B = 6$ MHz. The length of one timeslot $t_s = 10$ ms. For LEO satellites, $P_{\max}^{i} = 3$ dBW and $P_{\max} = 10$ dBW. For HAPs, $P_{\max}^{i} = -20$ dBW and $P_{\max} = -10$ dBW. Noise power spectral density is -173 dBm/Hz and Rician small-scale fading $\varphi = 20$ dB. The carrier frequency of Ka

TABLE 3. Important parameters.

Parameters	Value
Altitude of LEO Satellite	1500 km
$V_{\mathbf{LEO}}$	25000 km/h
LEO Beam Coverage Radius of 3 dB	200 km
Altitude of HAP	20-100 km
Average Radius of Small Hot-Spot Regions	2 km
Average Radius of Large Hot-Spot Regions	20 km
LEO Satellite Antenna η	20
HAP Antenna η	1
TR Antenna η	20
Receiving Gain of TR $G_{\mathbf{reTR}}$	3 dB
Receiving Gain of MT $G_{\mathbf{reMT}}$	0 dB
Noise Power Spectral Density	-173 dBm/Hz
Carrier Frequency of Ka Band	30 GHz
$B_{\mathbf{total}}$	240 MHz
Bandwidth of Each RB B	6 MHz
Timeslot t_s	10 ms
Average Velocity of MTs in Hot-Spot Regions V_r	5 m/s

band is 30 GHz. There are 100 MTs generated in the hot-spot region with a radius of 20km. The rest 900 MTs are generated in the whole rectangular region with length 2800km and width 1000km. The user density of the remote region is 37.5 per beam and the user density of the hot-spot region is 250 times larger than that of the remote region. 20% of the MTs are very important MTs, whose user priority utility is 1, and the user priority utility of the rest MTs is 0.1. The maximum delay tolerances of services are 300 ms, 500 ms, and 1 s, respectively. There are 200 services on-going at the first time-slot. New services arrival processes are modeled as Poisson distributed with average arrival rate λ. Packets of ongoing services arrive every 100 ms on average and the packet sizes are normally distributed with the means of 500k bits [37]. The duration time of services follows the exponential distribution with mean $\theta = 1$ minute. The maximal rate demands of MTs are the total data waiting to be transmitted. The minimum rate demands of MTs are 80k bits per slot. MTs will not be allocated resource until the total data waiting to be transmitted are larger than 80k bits. The weight of the utility is different according to the data traffic. $\omega_1 = 0.25$ and $\omega_2 = 0.75$ when serving under-load data traffic, $\omega_1 = 0.5$ and $\omega_2 = 0.5$ when serving full-load data traffic. $\omega_3 = 0.375$, $\omega_4 = 0.25$, $\omega_5 = 0.125$ and $\omega_6 = 0.25$ when serving under-load data traffic, $\omega_3 = 0.4375$, $\omega_4 = 0.1875$, $\omega_5 = 0.1875$ and $\omega_6 = 0.1875$ when serving full-load data traffic.

B. COMPARISON BETWEEN GEO RELAY AND LEO ROUTING

We here provide a numerical example to compare GEO relay and LEO routing on the normalized total delay, routing complexity and the hop number using the constellation model of the Iridium. Since the topology and the antenna pointing of LEO inter-satellite links changes quickly due to the quick

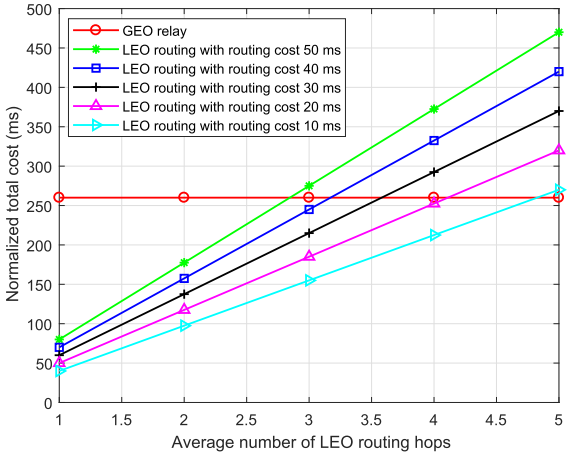


FIGURE 9. Normalized total cost comparison.

movement of LEO satellites, the average cost of routing algorithm is much larger than that of terrestrial networks.

Letting N_{h} be the hop number of LEO routing, T_{G}, T_{L} and T_{LL} be the average propagation delay of GEO-MT links, LEO-MT links and LEO-LEO inter-satellite links, respectively, T_{pr}, C_{pr} and C_{r} be the processing delay, the processing cost and the dynamic routing cost of each hop, respectively, the normalized total costs of GEO relay and LEO routing are

$$C_{\mathrm{GEO}} = 2T_{\mathrm{G}} + T_{\mathrm{pr}} + C_{\mathrm{pr}}, \tag{39}$$

$$C_{\mathrm{LEO}} = 2T_{\mathrm{L}} + (N_{\mathrm{h}} - 1)T_{\mathrm{LL}} + N_{\mathrm{h}}(T_{\mathrm{pr}} + C_{\mathrm{pr}} + C_{\mathrm{r}}). \tag{40}$$

Letting $T_{\mathrm{G}} = 120$ ms, $T_{\mathrm{L}} = 5$ ms, $T_{\mathrm{LL}} = 25$ ms, $T_{\mathrm{pr}} = 10$ ms (one time-slot) and $C_{\mathrm{pr}} = 10$ ms, we get the normalized total costs with different routing hop number and routing cost as Fig. 9. The results demonstrate that GEO relay satellites are suitable to replace the LEO routing with the hop number of more than 3.

C. RESULTS AND DISCUSSION OF THE PROPOSED ARCHITECTURE

Fig. 10 shows that after introducing HAPs and TRs to cover hot-spot regions and GEO satellite as relay, the proposed extensible multi-layer architecture reduces the average handover rate dramatically, because in hot-spot regions, handovers are not passive any more and the low-speed movement of MTs becomes the main reason causing handovers. As the percentage of MTs in hot-spot regions increases, the average handover rate of the whole system decreases linearly. 0.5 per minute is the smallest handover rate in our simulations because $(1 - P_{\mathrm{GEO}}) = 0.5$.

D. RESULTS AND DISCUSSION OF THE PROPOSED PROCEDURE

Fig. 11 shows the signalling cost of the proposed procedures and terrestrial network procedure. The signalling costs of the proposed procedures are smaller than those of the terrestrial network procedure after deleting RA process. The signalling

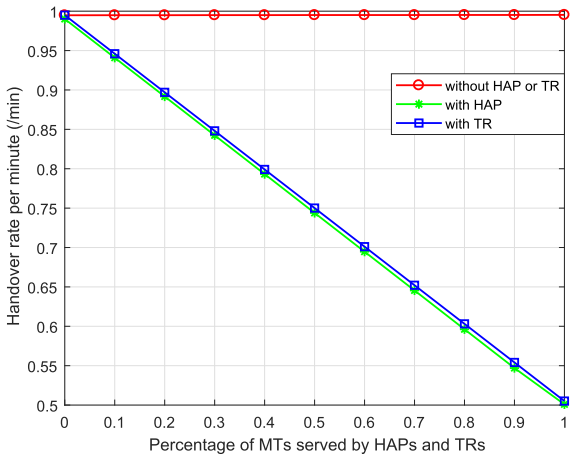


FIGURE 10. Average handover rate.

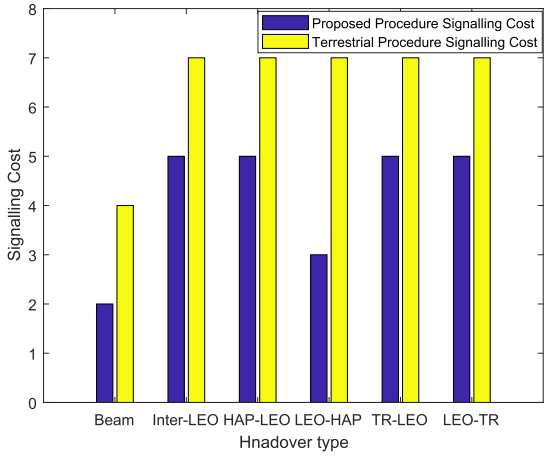


FIGURE 11. Signalling of the proposed and terrestrial network procedures.

costs during beam handover is the smallest since no signalling is required between beams of the same LEO satellite. The proposed procedure has a greatest improvement at cross-layer handover from LEO satellite to HAP. The reason is that HAP follows the command of LEO satellite in resource reuse during handovers and the signallings for interaction are deleted. The signalling costs of the other four types of handovers are the same.

The delays of the proposed procedures and terrestrial network procedure are shown in Fig. 12. Two times of processing delay and propagation delay are reduced during all types of handovers by removing the RA procedure. No matter in the downlink or uplink, during all types of handovers, the delays of the proposed procedures are smaller than those of the terrestrial network procedure. There are one more processing delay and two more propagation delay in the in the downlink than in the uplink during all types of handovers. Delay and delay improvement of different types of handovers are different because the propagation delays of different types of handovers are different.

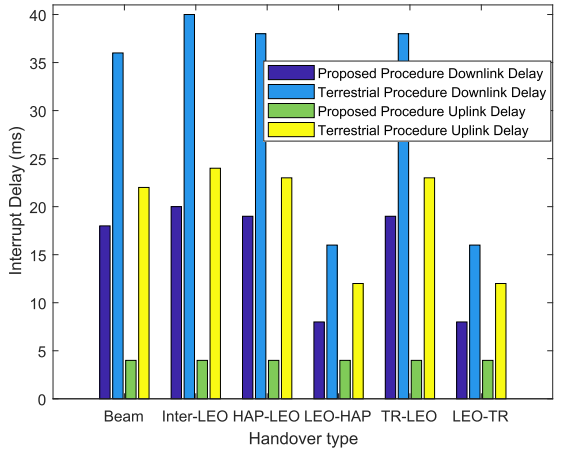


FIGURE 12. Delay of the proposed and terrestrial network procedures.

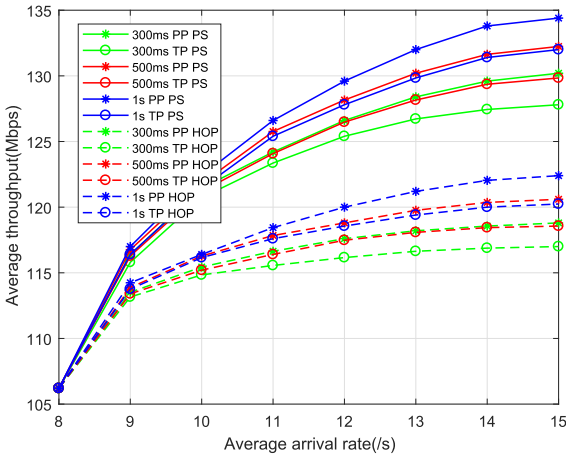


FIGURE 13. Average throughput per beam.

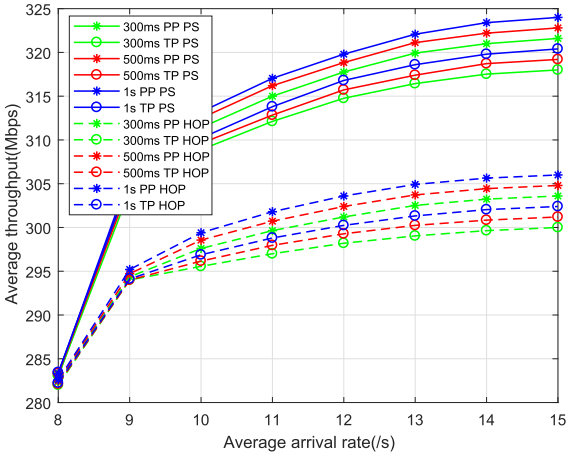


FIGURE 14. Average throughput of HAP.

E. RESULTS AND DISCUSSION OF THE PROPOSED OPTIMIZATION

A total of 100 Monte-Carlo simulations are carried, with each simulation lasting 5 minutes and containing 30000 timeslots. We compare the dropping probability and throughput of the proposed handover optimization and the handover

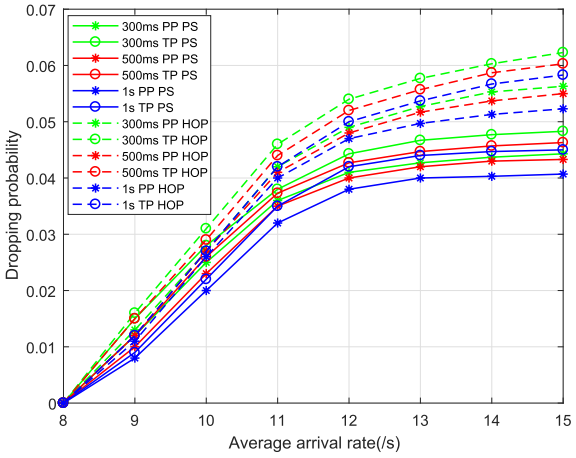


FIGURE 15. Average dropping probability.

protection (HOP) scheme in [43] with different arrival rate and maximum delay tolerance in the proposed architecture. We also compare the dropping probability and throughput of designed procedures and terrestrial network protocols. Finally we get the following average results in Figs. 13 −15.

Fig. 13 and Fig. 14 demonstrate that compared with HOP scheme, the proposed strategy (PS) provides more than 10% throughput improvement while guaranteeing the dropping probability due to the power optimization. Compared with terrestrial protocols (TP), the proposed procedures (PP) provide better throughput because the resource during handover delay cannot be used immediately. The shorter the handover delay is, the less resource will be wasted. The longer the maximum delay tolerance of service is, the larger the throughput is, especially for MTs covered by LEO satellites.

Fig. 15 demonstrates that the proposed handover optimization provides smaller dropping probability compared with HOP scheme. Similar to the throughput, the proposed procedures and longer maximum delay tolerance of service contribute to smaller dropping probability. As the arrival rate increases, the dropping probability firstly rises and then almost remains stable because on-going services have a higher priority to be transmitted than new services when traffic rises, sacrificing the blocking probability of new services.

VII. CONCLUSION AND FUTURE DIRECTION

In this paper, we focused on the problem of frequent passive handovers in LEO-MSS. Via introducing high-altitude platforms (HAPs) and terrestrial relays (TRs), we proposed an extensible multi-layer network architecture to reduce handover rates in hot-spot regions and solve the problem of group handovers. We also proposed a multi-layer handover framework according to the architecture to achieve centralized handover management efficiently. Based on the proposed architecture and handover framework, we designed different handover procedures for different types of handovers to reduce handover delay and signalling cost. Afterwards, we proposed a dynamic handover optimization to reduce the dropping probability while guaranteeing the QoS of mobile terminals. Lagrange dual method and Karush-Kuhn-Tucker (KKT) conditions are used to find the optimal solutions. Also, we presented corresponding gradient descent algorithm to get the results. Numerical results show that the proposed architecture reduces group handovers significantly. The proposed handover procedures provide better performance on delay and signalling cost compared with traditional handover protocols. With the proposed dynamic handover optimization, the proposed handover procedures also provide better performance on dropping probability and throughput. In addition, the proposed dynamic handover optimization has an excellent performance on dropping probability while guaranteeing the QoS of mobile terminals.

The proposed extensible multi-layer architecture based on LEO-MSS can be used for future space-air-ground communication networks to provide better communication services. Introducing HAPs and TRs is a method to solve the problem of group handovers in LEO satellite systems. The handover procedures we designed can be used in future space-air-ground networks and provide better performance on delay, signalling cost, dropping probability and throughput. The proposed dynamic handover optimization yields an excellent performance on dropping probability while guaranteeing the QoS of mobile terminals. And the handovers in space-air-ground networks can be further optimized via introducing uplink handover strategies.

ACKNOWLEDGMENT

This article was presented in part at the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-FALL).

REFERENCES

[1] H. Yao, L. Wang, X. Wang, Z. Lu, and Y. Liu, "The space-terrestrial integrated network: An overview," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 178–185, Sep. 2018.
[2] C. Sacchi, K. Bhasin, N. Kadowaki, and F. Vong, "Technologies and applications of future satellite networking [guest editorial]," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 154–155, May 2015.
[3] T. Pecorella, L. S. Ronga, F. Chiti, S. Jayousi, and L. Franck, "Emergency satellite communications: Research and standardization activities," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 170–177, May 2015.
[4] R. Musumpuka, T. M. Walingo, and J. M. Smith, "Performance analysis of correlated handover service in LEO mobile satellite systems," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2213–2216, Nov. 2016.
[5] B. Yang, Y. Wu, X. Chu, and G. Song, "Seamless handover in software-defined satellite networking," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1768–1771, Sep. 2016.
[6] Z. Wu, F. Jin, J. Luo, Y. Fu, J. Shan, and G. Hu, "A graph-based satellite handover framework for LEO satellite communication networks," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1547–1550, Aug. 2016.
[7] K. Xue, W. Meng, S. Li, D. S. L. Wei, H. Zhou, and N. Yu, "A secure and efficient access and handover authentication protocol for Internet of Things in space information networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5485–5499, Jun. 2019.
[8] W. Feng, J. Wang, Y. Chen, X. Wang, N. Ge, and J. Lu, "UAV-aided MIMO communications for 5G Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1731–1740, 2018.
[9] A. Mohammed, A. Mehmood, F.-N. Pavlidou, and M. Mohorcic, "The role of high-altitude platforms (HAPs) in the global wireless connectivity," *Proc. IEEE*, vol. 99, no. 11, pp. 1939–1953, Nov. 2011.
[10] E. Cianca, R. Prasad, M. De Sanctis, A. De Luise, M. Antonini, D. Teotino, and M. Ruggieri, "Integrated satellite-HAP systems," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. supl.33–supl.39, Dec. 2005.

[11] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. S. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101–109, Jul. 2017.

[12] J. Zhang, X. Zhang, M. A. Imran, B. Evans, Y. Zhang, and W. Wang, "Energy efficient hybrid satellite terrestrial 5G networks with software defined features," *J. Commun. Netw.*, vol. 19, no. 2, pp. 147–161, Apr. 2017.

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

[14] M. Kirkko-Jaakkola, J. Parviainen, J. Collin, and J. Takala, "Improving TTFF by two-satellite GNSS positioning," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 4, pp. 3660–3670, Oct. 2012.

[15] X. Luo, S. Li, and H. Xu, "Results of real-time kinematic positioning based on real GPS L5 data," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1193–1197, Aug. 2016.

[16] A. Z. M. Shahriar, M. Atiquzzaman, and S. Rahman, "Mobility management protocols for next-generation all-IP satellite networks," *IEEE Wireless Commun.*, vol. 15, no. 2, pp. 46–54, Apr. 2008.

[17] X. Wang, X. Gao, and R. Zong, "Energy-efficient deployment of airships for high altitude platforms: A deterministic annealing approach," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2011, pp. 1–6.

[18] X. Wang, "Deployment of high altitude platforms in heterogeneous wireless sensor network via MRF-MAP and potential games," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 1446–1451.

[19] T. Tang, T. Hong, H. Hong, S. Ji, S. Mumtaz, and M. Cheriet, "An improved UAV-PHD filter-based trajectory tracking algorithm for multi-UAVs in future 5G IoT scenarios," *Electronics*, vol. 8, no. 10, p. 1188, Oct. 2019.

[20] S. Mumtaz, J. Rodriguez, and I. Otung, "Smart resource allocation scheme for 5G terrestrial networks," in *Proc. 35th AIAA Int. Commun. Satell. Syst. Conf.*, 2017, p. 5443.

[21] V. S. Kumar and D. G. Kurup, "A new broadband magic tee design for Ka-band satellite communications," *IEEE Microw. Wireless Compon. Lett.*, vol. 29, no. 2, pp. 92–94, Feb. 2019.

[22] E. Cianca, T. Rossi, A. Yahalom, Y. Pinhasi, J. Farserotu, and C. Sacchi, "EHF for satellite communications: The new broadband frontier," *Proc. IEEE*, vol. 99, no. 11, pp. 1858–1881, Nov. 2011.

[23] M. W. Atwood, G. N. Marcoux, and W. P. Craig, "Demonstration of two-way extremely high frequency (EHF) satellite communication (SATCOM) using submarine-survivable phased arrays," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2008, pp. 1–7.

[24] Z. Sodnik, B. Furch, and H. Lutz, "Free-space laser communication activities in europe: SILEX and beyond," in *Proc. 19th Annu. Meeting IEEE Lasers Electro-Optics Soc. (LEOS)*, Oct. 2006, pp. 78–79.

[25] Y. Li, H. Wei, L. Li, Y. Han, J. Zhou, and W. Zhou, "An extensible multi-layer architecture model based on LEO-MSS and performance analysis," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–6.

[26] G. E. Corazza and F. Vatalaro, "A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 738–742, 3rd Quart., 1994.

[27] W. Wang, A. Liu, Q. Zhang, L. You, X. Gao, and G. Zheng, "Robust multigroup multicast transmission for frame-based multi-beam satellite systems," *IEEE Access*, vol. 6, pp. 46074–46083, 2018.

[28] W. Feng, Y. Chen, N. Ge, and J. Lu, "Optimal energy-efficient power allocation for distributed antenna systems with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7759–7763, Sep. 2016.

[29] Accessed: Sep. 28, 2017. [Online]. Available: https://standards.nasa.gov/ccsds-standards

[30] Accessed: Aug. 2019. [Online]. Available: https://www.dvb.org/standards

[31] K. Ohya, S. Kameda, H. Oguma, A. Taira, N. Suematsu, T. Takagi, and K. Tsubouchi, "Experimental evaluation of timing synchronization accuracy for QZSS short message synchronized SS-CDMA communication," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.

[32] J. A. del Peral-Rosado, R. Raulefs, J. A. Lopez-Salcedo, and G. Seco-Granados, "Survey of cellular mobile radio localization methods: From 1G to 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1124–1148, 2nd Quart., 2018.

[33] C. Laoudias, A. Moreira, S. Kim, S. Lee, L. Wirola, and C. Fischione, "A survey of enabling technologies for network localization, tracking, and navigation," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3607–3644, 4th Quart., 2018.

[34] Y. Albagory, M. Nofal, and A. Ghoneim, "Handover performance of unstable-yaw stratospheric high-altitude stations," *Wireless Pers. Commun.*, vol. 84, no. 4, pp. 2651–2663, May 2015.

[35] A. Ibrahim and A. S. Alfa, "Using lagrangian relaxation for radio resource allocation in high altitude platforms," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5823–5835, Oct. 2015.

[36] J. Thornton, D. Grace, M. H. Capstick, and T. C. Tozer, "Optimizing an array of antennas for cellular coverage from a high altitude platform," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 484–492, May 2003.

[37] L. Li, N. Deng, W. Ren, B. Kou, W. Zhou, and S. Yu, "Multi-service resource allocation in future network with wireless virtualization," *IEEE Access*, vol. 6, pp. 53854–53868, 2018.

[38] K. G. Murty, *Linear Programming*. Hoboken, NJ, USA: Wiley, 1983.

[39] S. Wang, Y. Li, Q. Wang, M. Su, and W. Zhou, "Dynamic downlink resource allocation based on imperfect estimation in LEO-HAP cognitive system," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.

[40] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[41] Y. Li, N. Deng, and W. Zhou, "A hierarchical approach to resource allocation in extensible multi-layer LEO-MSS," *IEEE Access*, vol. 8, pp. 18522–18537, 2020.

[42] Accessed: Apr. 9, 2019. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/25_series/25.912/

[43] K. Adachi, J. Joung, Y. Zhou, and S. Sun, "A distributed resource reservation scheme for handover failure reduction," *IEEE Wireless Commun. Lett.*, vol. 4, no. 5, pp. 537–540, Oct. 2015.

YITAO LI received the B.S. degree in electronic engineering from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include low earth orbit mobile satellite network, geostationary earth orbit relay satellite network, space-air-ground integrated network, 5G network, radio resource management, handover management, and RRC and MAC protocols of wireless access network.

WUYANG ZHOU (Member, IEEE) received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2000. He is currently a Professor with the Department of Electronic Engineering and Information Science, USTC. His research interests include space-air-ground integrated network, 5G network, interference cancellation in satellite network, intelligent computing, and positioning technologies.

SHENGLI ZHOU (Fellow, IEEE) received the B.S. and M.Sc. degrees in electrical engineering and information science from the University of Science and Technology of China, Hefei, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2002.

He is currently a Professor with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. His general research interests are in the areas of wireless communications and signal processing. He received the 2007 ONR Young Investigator Award and the 2007 Presidential Early Career Award for Scientists and Engineers. He was an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2005 to 2007, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, from 2008 to 2010, and the IEEE JOURNAL OF OCEANIC ENGINEERING, from 2010 to 2016.