Cooperative Multilayer Edge Caching in Integrated Satellite-Terrestrial Networks

Xiangming Zhu, Chunxiao Jiang, *Senior Member, IEEE*, Linling Kuang, *Member, IEEE*, and Zhifeng Zhao, *Member, IEEE*

***Abstract*— The integrated satellite-terrestrial network is promising to provide global broadband communication service. However, the long propagation delay of satellite-terrestrial links will lead to high communication delay when users access the Internet via satellites. In this paper, we investigate the cooperative multilayer edge caching in the integrated satellite-terrestrial network to reduce the communication delay, in which the base station cache, the satellite cache, and the gateway cache cooperatively provide content service for ground users. We first propose the three-layer cooperative caching model of the network, based on which we analyze the content retrieving process and derive the cache hit probability for different caching locations. Considering limited cache sizes, we formulate the content placement problem to minimize the average content retrieving delay of users. Then, two caching strategies, the non-cooperative caching strategy and the cooperative caching strategy, are proposed with exhaustive theoretical analysis. By introducing the concept of delay reduction gains, the optimal caching strategies are obtained based on the proposed iterative algorithms. Finally, numerical results are presented to demonstrate the performance of the proposed cooperative caching architecture and the caching strategies.**

***Index Terms*— Edge caching, cooperative multilayer caching, satellite-terrestrial networks, content placement.**

I. Introduction

WITH the development of information technology, wireless communication is now indispensable for providing connectivity in modern society. From the first generation (1G) to the fourth generation (4G) wireless network, terrestrial wireless network has proved a great success for the enhancement in communication speed and quality of service (QoS). However, due to construction costs and geographic constrains, conventional terrestrial networks are poorly constructed or even absent in less-developed and sparsely populated areas [1], [2]. In fact, nearly half of the populations in the world have no access to the Internet by 2019 [3]. Extending the connectivity to the rest populations has become imperative to move forward for future communications networks. With the wide coverage ability, satellite communication networks provide a direct solution for the coverage issue. The concept of satellite Internet is recently proposed to provide global Internet access via satellite constellations [4], and has become a hotspot for both academia and industry. Various satellite constellation projects have been established to construct satellite communication networks, such as Starlink, OneWeb, and Telesat [5]. In the White Paper of the 6G wireless network, it has been proposed that the future wireless network must be able to seamlessly interface with terrestrial and satellite networks [6]. The integrated satellite-terrestrial network is the new development trend for the next generation communication network [7].

Although the integrated satellite-terrestrial network is promising to provide global broadband communication service, it also brings many new challenges due to the unique characteristics of satellite networks. Especially, the long propagation delay of satellite-terrestrial links will lead to high communication delay when users access the Internet via satellites. For MEO satellites on the orbit of 20,000 km, the two-way propagation delay will be as long as 270 ms, which is much longer than the 5G end-to-end latency requirement of 1 ms [8], and the 6G end-to-end latency requirement of 0.1 ms [6]. Thus delay optimization will be an important issue in future integrated satellite-terrestrial networks.

Currently, cloud based services are widely applied in communication networks for the advantages of high computation and storage capability at the cloud [9]. However, in satellite networks, requiring data and services from the cloud will be of high communication delay, since users can only access the cloud via satellite-terrestrial links. To avoid frequent communication with the cloud, the architecture of mobile edge computing (MEC) can be applied to place the "cloud" closer to users at the network edge [10]. By providing computing and storage services at MEC servers, the communication delay can be significantly reduced [11]. Also, with the prevalence of

Manuscript received December 21, 2020; revised April 30, 2021 and July 16, 2021; accepted September 27, 2021. Date of publication October 8, 2021; date of current version May 10, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1804800, in part by the National Natural Science Foundation of China under Grant 61922050, and in part by Zhejiang Lab under Grant 2020LB0AA01. The associate editor coordinating the review of this article and approving it for publication was S. R. Khosravirad. *(Corresponding author: Chunxiao Jiang.)*

Xiangming Zhu is with Zhejiang Lab, Hangzhou 311121, China, and also with Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China (e-mail: zhuxm@zhejianglab.com).

Chunxiao Jiang and Linling Kuang are with Tsinghua Space Center, Tsinghua University, Beijing 100084, China, and also with Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China (e-mail: jchx@tsinghua.edu.cn; kll@tsinghua.edu.cn).

Zhifeng Zhao is with Zhejiang Lab, Hangzhou 311121, China (e-mail: zhaozf@zhejianglab.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3117026.

Digital Object Identifier 10.1109/TWC.2021.3117026

social media, wireless services are extended from traditional connection-centric communications to content-centric communications [12]. Especially, in 5G and also the future 6G networks, distinct QoS guarantees are expected for delay-sensitive multimedia applications [13], [14]. By utilizing the storage capability of MEC servers, mobile edge caching is drawing much attention as a promising solution, which can help to reduce the content retrieving delay of users [15]. Since popular contents may be requested by multiple users frequently, it is of higher efficiency to cache and retrieve these popular contents from MEC servers, instead of from the remote cloud. However, compared with the increasing data traffic, the cache size of the MEC server is generally limited. Thus the caching strategies at MEC servers need to be designed elaborately to optimize the content serving capability of the network [16], [17].

In conventional terrestrial networks, contents are generally cached in BSs for mobile edge caching. However, in integrated satellite-terrestrial networks, due to the special network architecture, contents can be cached in BSs, in satellites, and also in the gateway [11]. In [18]–[20], satellites are considered to provide backhaul transmission for content retrieving from the cloud, while contents are only cached in terrestrial BSs. In [18], considering both the global content popularity and local content popularity, an off-line caching strategy was proposed to maximize the cache hit ratio. In [19], cooperative caching of BSs and users was considered with satellite backhaul transmission. The help nodes were selected based on the social relationship, and then a greedy caching strategy was proposed to minimize the content retrieving delay. In [20], the author discussed the caching problem in the satellite-terrestrial relay network. The outage probability was analyzed for two caching strategies at terrestrial relays. In [21]–[24], caching at the satellite was studied to provide content service for terrestrial users. In [21], the caching proportion of the most popular contents and general popular contents was optimized for satellite caches to maximize the cache hit ratio. In [22], a deep Q-learning approach was proposed to solve the joint networking, computing and caching problem in satellite-terrestrial networks. In [23], a distributed caching strategy of LEO satellite constellations was proposed to minimize the content retrieving delay. Inter-satellite links (ISLs) were considered to enable cooperative content retrieving from adjacent satellites. In [24], the author studied the problem of cooperative caching in multilayer satellite networks, and proposed an effective caching strategy based on Stackelberg game. In [25], the security of UAV-relayed wireless networks with caching was studied. A novel scheme was proposed based on successive convex optimization, which is effective and efficient for secure transmission in UAV relaying systems with local caching. In [26], a two-layer caching model was proposed, in which the first layer cache was the BS and the second layer cache was the satellite. The joint caching problem was then solved to minimize the bandwidth consumption.

As discussed above, in integrated satellite-terrestrial networks, the possible caching locations includes BSs, satellites, and the gateway. However, existing works mainly focus on caching in terrestrial BSs or caching in satellites separately, failing to exploit the cooperative caching architecture for improvement of the delay performance. In this paper, considering all the possible caching locations in the integrated satellite-terrestrial network, we investigate the content placement problem based on the proposed three-layer cooperative caching model. The main contributions of this paper are summarized as follows.

- We propose a three-layer cooperative caching framework for the integrated satellite-terrestrial network, in which the BS cache, the satellite cache, and the gateway cache cooperatively provide content service for ground users. Considering cooperative caching among satellites, we analyze the content retrieving process and derive the cache hit probability for different caching locations. Based on the framework, we formulate the content placement problem to minimize the average content retrieving delay of the network.
- We first propose a non-cooperative caching strategy for the content placement problem, in which each satellite acts selfishly to minimize the average content retrieving delay of its local users. We prove that the non-cooperative caching strategy of the BS is to cache the most popular contents in all BSs. Then, the cache of the satellite is divided into the duplicately caching part and the selectively caching part. By introducing the concept of delay reduction gains, the non-cooperative caching strategy of the satellite is obtained.
- We then propose a cooperative caching strategy to fully utilize the advantage of cooperation, in which each satellite acts cooperatively to minimize the average content retrieving delay of the network. The cache of the BS is divided into the duplicately caching part and the selectively caching part, while the cache of the satellite is divided into the fixedly caching part and the selectively caching part. Based on theoretical analysis, the optimal solutions of the duplicately caching part and the fixedly caching part are derived. Then, an iterative algorithm is proposed to calculate the optimal caching strategy of the entire network.

The rest of the paper is organized as follows. We first introduce the network and caching model in Section II, and then formulate the content placement problem in Section III. In Section IV, considering each satellite acts selfishly, we propose the non-cooperative multilayer caching strategy for both BSs and satellites. Then, the optimal cooperative multilayer caching strategy is proposed with detailed analysis in Section V. Numerical results are provided in Section VI to demonstrate the proposed strategies. Finally, Section VII concludes the paper.

II. System Model

In this section, we describe the mathematical model for the cooperative caching framework in the integrated satellite-terrestrial network. For improvement of the readability, the notations used in the paper are listed in Table I.

TABLE I
SUMMARY OF NOTATIONS

Notation	Description	Notation	Description
L	The number of satellites	K_l	The number of BSs served by satellite l
v_l	The satellite node in the graph	$e(v_l, v_{l'})$	The edge between satellites in the graph
F	The number of contents	f_i	The i-th popular content
p_i	The popularity distribution of content i	δ	The Zipf exponent
$c_{s,1}$	The cache size of each BS	$c_{s,2}$	The cache size of each satellite
F_1^l	The set of contents cached in BSs of satellite l	a_i^l	The caching indicator for the BS
F_2^l	The set of contents cached in satellite l	b_i^l	The caching indicator for the satellite
F_3	The set of contents cached in the gateway	$c_{s,3}$	The cache size of the gateway
τ_1	Average delay from the BS to the user	τ_2	Average delay from the satellite to the BS
τ_3	Average delay from the gateway to the satellite	τ_4	Average delay from the cloud to the gateway
τ_s	Average delay from adjacent satellites of one hop	n_h	The number of hops
L^{l,n_h}	The set of adjacent satellites of n_h hops	$N_{h,max}$	The maximum number of hops
τ_{ave}^l	The average content retrieving delay of satellite l	η^l	The weighted factor of satellite l
$\Delta\tau_{1,i}^l$	The delay reduction gain of the BS	$\Delta\tau_{2,i}^l$	The delay reduction gain of the satellite
F_1^{total}	The contents that may be cached in the BS	$F_2^{l,total}$	The contents that may be cached in satellite l
$f_{\alpha_{max}}$	The least popular content in F_1^{total}	$f_{\beta^l_{max}}$	The least popular content in $F_2^{l,total}$
$f_{\alpha_{dup}}$	The duplicately cached content for the BS	$f_{\beta^l_{fix}}$	The fixedly cached content for satellite l
$f_{\beta_{dup}}$	The duplicately cached content for the satellite		

A. Network Model

Consider an integrated satellite-terrestrial network as shown in Fig. 1, in which satellites provide backhaul transmission for terrestrial BSs in areas without connection of optical fiber, while terrestrial users access BSs for communication service. Satellites are connected to gateways via feeder links, by means of which satellites can retrieve data from the cloud. This backhaul architecture is one of the typical architectures for future satellite-terrestrial networks [27], which has been widely studied in existing works. The satellites can be either low earth orbit (LEO) satellites or medium earth orbit (MEO) satellites according to the actual network composition, which are of different coverage and different transmission delays. Considering the scenario of multiple satellites, the total number of satellites is L, and inter-satellite communication is enabled by ISLs. Abstracting satellites as nodes, and the links among satellites as edges, the satellite network can be modeled as a graph $G = (V, E)$, in which $V = \{v_1, v_2, \ldots, v_L\}$ is the set of satellite nodes, and $E = \{e(v_l, v_{l'})|, l \neq l'\}$ is the set of edges. For each edge, $e(v_l, v_{l'}) = 1$ represents that satellite l and l' are connected by the ISL, while $e(v_l, v_{l'}) = 0$ represents that there is no link between satellite l and l'. In the network, each satellite serves K_l BSs within its coverage, while BSs are isolated with each other since there is no optical fiber between BSs. The integrated satellite-terrestrial network is mainly utilized to provide communication service for areas without coverage of traditional terrestrial networks, extending the connectivity to everyone and everything. In areas of high-density populations, such as urban areas, traditional terrestrial networks are preferred for providing low cost and high-speed services.

B. Caching Model

Due to the long transmission distance of satellite links, retrieving contents from the remote cloud will be of high delay in the integrated network. Thus mobile edge caching is applied in the network to cache popular contents closer to users at the network edge. Then, the content retrieving delay can be significantly reduced by retrieving the requested contents from the edge caches. Based on the network model, we consider a three-layer cooperative caching model to minimize the content retrieving delay, including the BS cache, the satellite cache, and the gateway cache [11]. Each BS provides edge caching service for its local users, while each satellite can provide edge caching service for a large number of BSs and users within its coverage. Then, the gateway can provide edge caching service for any satellite connected with it. Since inter-satellite communication is enabled by ISLs between satellites, cooperative caching of satellites is considered in the network. The requested content can be retrieved from adjacent satellites if it is not cached in the local cache. Each satellite will maintain a cache table, which records the cached contents of its adjacent satellites. When a satellite updates its cache, it will broadcast its caching strategy to all adjacent satellites. Then, the adjacent satellites will update its cache table with the received information. Based on the cache table, each satellite can determine its caching strategy in a distributed manner.

Let $F = \{f_1, f_2, \ldots, f_F\}$ be the set of all contents, in which $\{f_1, f_2, \ldots, f_{F-1}\}$ are the $F-1$ most popular contents, f_i is the i-th popular content, and f_F is the set of the rest contents with less popularity. Without loss of generality, we assume all the contents are of the same normalized size 1. For contents of different sizes, they can be divided into units of the same size [28], [29]. In this paper, we consider the proactive caching policy, in which contents are proactively cached during off-peak times according to the content popularity [15]. Let $P = \{p_1, p_2, \ldots, p_F\}$ be the global content popularity distribution, and we have $p_1 > p_2 > \ldots > p_F, \sum_{i=1}^{F} p_i = 1$. The content popularity is found to follow

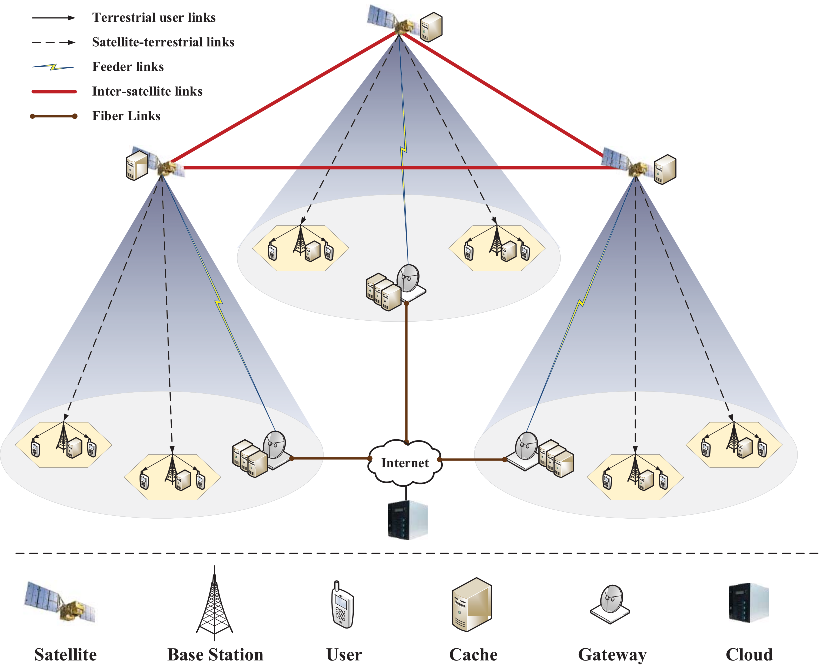


Fig. 1. The integrated satellite-terrestrial network with multilayer edge caching.

the Zipf distribution $p_i = \frac{i^{-\delta}}{\sum_{j=1}^{F} j^{-\delta}}$ for video contents in real networks [15], [30], which has been utilized in many existing works [19], [31]. In this paper, we do not restrict the content popularity distribution as any specific distribution. It can be the Zipf distribution, and can also be other distributions. Generally, the content popularity distribution changes slowly with time [32]. For example, the popularity of news video may be updated every 2-3 hours, the popularity of movies may be updated every week, and the popularity of music may be updated every month. Thus the content popularity distribution can be regarded as relatively time-nonvarying during a certain period of time. Then, the popularity distribution can be learned and predicted by the system according to statistics.

In practice, it is reasonable to assume limited cache sizes for BSs and satellites due to cost constraints. Thus only part of the $F-1$ most popular contents are cached in BSs and satellites. Let $c_{s,1} < F-1$ and $c_{s,2} < F-1$ be the cache sizes of each BS and each satellite respectively. The most popular content set is considered to be larger than the total cache size of the BS and the satellite, $F-1 > c_{s,1} + c_{s,2}$. Let $F_1^{l,k}$ be the set of contents cached in BS k of satellite l, and $a_i^{l,k} \in \{0,1\}$ be the corresponding caching indicator, while $a_i^{l,k} = 1$ indicates that content i is cached in BS k of satellite l. Similarly, let F_2^l be the set of contents cached in satellite l, and $b_i^l \in \{0,1\}$ be the corresponding caching indicator. Then the cached contents need to be carefully selected to optimize the network performance. Also, since the gateway can provide edge caching service for a large number of users, and the cache resources are relatively sufficient at ground, we assume that the gateway has relatively large cache size $c_{s,3} \geq F-1$. Then, the $F-1$ most popular contents can all be cached in the gateway. The set of cached contents in the gateway is $F_3 = \{f_1, f_2, \ldots, f_{F-1}\}$. However, the rest contents f_F with less popularity still cannot be cached in the gateway since they are of large size, and can only be retrieved from the cloud.

C. Content Retrieving Model

After the caching process during off-peak times, the network will serve the requests from users with cached contents during the peak-time. As shown in Fig. 2, the content retrieving process includes the following cases.

1. **Retrieving from the associated BS:** After receiving a content request from an user, the associated BS will first search its local cache. If the requested content is cached, the BS can directly transmit it to the user. Otherwise, the BS will send the request to the satellite.
2. **Retrieving from the associated satellite:** After receiving a content request from a BS, the associated satellite will search its local cache and also the cache table. If the requested content is cached in its local cache, the satellite will directly transmit it to the BS via the backhaul link. If the requested content is cached in its adjacent satellites, the satellite will send the request to the corresponding adjacent satellite. If the requested content cannot be found, the satellite will send the request to the gateway.
3. **Retrieving from the adjacent satellite:** After receiving a content request from an adjacent satellite, the satellite will transmit the requested content to the adjacent satellite via the ISL.
4. **Retrieving from the gateway:** After receiving a content request from a satellite, the gateway will search its local cache. If the requested content is cached, the gateway will transmit it to the satellite via the feeder link.
5. **Retrieving from the cloud:** If the requested content cannot be found in the cache, the gateway will retrieve it from the cloud, and transmit it to the satellite via the feeder link.

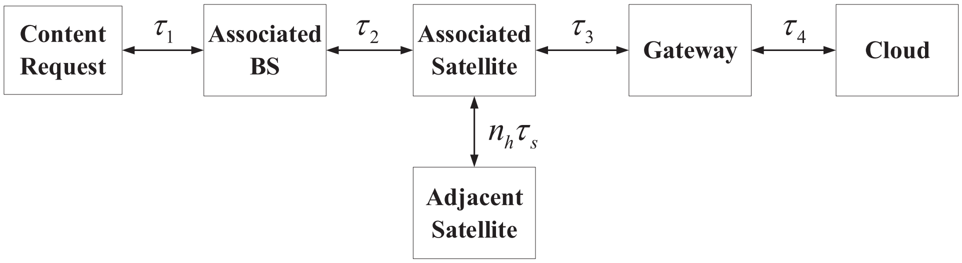


Fig. 2. The content retrieving process in the integrated satellite-terrestrial network.

In the integrated satellite-terrestrial network, satellites provide backhaul transmission for terrestrial BSs in areas without connection of optical fiber. Thus one BS can only communicate with another BS via the satellite. If BS i retrieves the requested content from another BS j, the content needs to be first transmitted from BS j to the satellite, and then transmitted back to BS i from the satellite. Generally, the transmission rate of the feeder link between the satellite and the gateway is much larger than the transmission rate of the backhaul link between the satellite and the BS. The retrieving delay from the gateway will be lower than the retrieving delay from other BSs. Thus we do not consider cooperative caching of BSs in the network.

Retrieving the requested content from different locations will be of different delays according to the transmission path. In this paper, the average content retrieving delay is defined as the total time consumed for retrieving the requested content, including the delay of sending requests and the delay of retrieving contents. Also, due to the long communication distance of satellite-terrestrial links, both the propagation delay and the transmission delay need to be considered. The propagation delay can be calculated by $\tau^p = \frac{d}{v_c}$, where d is the propagation distance and v_c is the speed of light, while the transmission delay can be calculated according to statistics or the ergodic rate [33]. Since the caching process occurs before the actual serving process, the real-time channel information cannot be acquired when determining the caching strategy. Also, in this paper, we aim to optimize the network overall service capability for all possible users during the peak-time, instead of only optimizing the experience of certain users. Thus we consider the average content retrieving delay for different caching locations in this paper. For different communication models, the average content retrieving delay may be calculated by different methods. Once the average content retrieving delay is obtained, the proposed caching strategy in this paper can be applied. For better adaptability to different communication scenarios, we do not restrict the communication model as any specific model. Similar to existing works [31], [34], we mainly focus on the cooperative caching strategy after obtaining the average content retrieving delay of different links. Let $\tau_1, \tau_2, \tau_3, \tau_4$ be the average content retrieving delay from the BS to the user, from the satellite to the BS, from the gateway to the satellite, and from the cloud to the gateway respectively. Also, let τ_s be the average content retrieving delay from adjacent satellites of one hop, and L^{l,n_h} be the set of adjacent satellites of n_h hops for satellite l. When satellite l retrieving contents from adjacent satellite l', the shortest path in graph G is selected to minimize the content retrieving delay, which can be calculated by algorithms such as the Dijkstra algorithm. Then, if the obtained shortest path covers n_h edges, satellite $l' \in L^{l,n_h}$ is considered to be the adjacent satellite of n_h hops for satellite l, and the content retrieving delay from satellite l' is $n_h\tau_s$. Obviously, the content retrieving delay from adjacent satellites should be no longer than the content retrieving delay from the gateway, or the requested content can be retrieved from the gateway instead of adjacent satellites. Thus the maximum number of hops when retrieving the requested contents from adjacent satellites is $n_h \leq N_{h,max} = \left\lfloor \frac{\tau_3}{\tau_s} \right\rfloor$.

III. Problem Formulation

In the last section, we formulated the three-layer cooperative caching model, in which the requested content can be retrieved from the BS, the satellite, the gateway, and the cloud. Since we do not consider cooperative caching of BSs in the network, the caching strategy of each BS will only influence the content retrieving delay of its local users. Then, for each satellite l, the caching strategies of all the BSs within its coverage $F_1^{l,k}$ will be the same, represented by $F_1^{l,k} = F_1^l, a_i^{l,k} = a_i^l, k \in [1, K_l]$. For an arbitrary content request from an user associated with the BS of satellite l, the cache hit probability of the BS is

$$p_{hit,1}^l = \sum_{f_i \in F_1^l} p_i = \sum_{i=1}^{F-1} a_i^l p_i. \tag{1}$$

In this case, the content retrieving delay is τ_1. If the requested content is not cached in the BS, then the cache hit probability of satellite l is

$$p_{hit,2}^l = \sum_{f_i \notin F_1^l, f_i \in F_2^l} p_i = \sum_{i=1}^{F-1} \max\{0, b_i^l - a_i^l\} p_i. \tag{2}$$

In this case, the content retrieving delay is $\tau_1 + \tau_2$. If the requested content is not cached in both the BS and the satellite, then the cache hit probability of the n_h-hop adjacent satellites is

$$\begin{aligned} p_{hit,s}^{l,n_h} &= \sum_{f_i \notin (F_1^l \cup F_2^l \ldots \cup F_s^{l,n_h-1}), f_i \in F_s^{l,n_h}} p_i \\ &= \sum_{i=1}^{F-1} \max\{0, \theta_i^{l,n_h} - \max\{a_i^l, b_i^l, \theta_i^{l,1}, \ldots, \theta_i^{l,n_h-1}\}\} p_i. \end{aligned} \tag{3}$$

In this case, the content retrieving delay is $\tau_1 + \tau_2 + n_h\tau_s$. Also, F_s^{l,n_h} is the set of contents cached in the n_h-hop adjacent satellites of satellite l, and $\theta_i^{l,n_h} \in \{0,1\}$ is the auxiliary caching indicator derived from $\mathbf{b} = [b_1^1,\ldots,b_{F-1}^1,\ldots,b_1^L,\ldots,b_{F-1}^L]$, where $\theta_i^{l,n_h} = 1$ indicates that content i is cached in the n_h hop adjacent satellites of satellite l. If the requested content still cannot be found, then the cache hit probability of the gateway is

$$\begin{aligned} p_{hit,3}^l &= \sum_{f_i \notin (F_1^l \cup F_2^l \cup F_s^{l,1} \ldots \cup F_s^{l,N_{h,max}}), f_i \in F_3} p_i \\ &= \sum_{i=1}^{F-1} (1 - \max\{a_i^l, b_i^l, \theta_i^{l,1}, \ldots, \theta_i^{l,N_{h,max}}\}) p_i. \end{aligned} \tag{4}$$

In this case, the content retrieving delay is $\tau_1+\tau_2+\tau_3$. Since the $F-1$ most popular contents are all cached in the gateway, the content will be retrieved from the cloud only when an user requests the rest contents f_F with less popularity. The cache hit probability of the cloud is

$$p_{hit,4}^l = p_{hit,4} = p_F. \tag{5}$$

In this case, the content retrieving delay from the cloud is $\tau_1+\tau_2+\tau_3+\tau_4$. The average content retrieving delay of users associated with BSs of satellite l can then be calculated by

$$\begin{aligned} \tau_{ave}^l = \tau_1 &+ (1 - p_{hit,1}^l)\tau_2 + \sum_{n_h=1}^{N_{h,max}} p_{hit,s}^{l,n_h} n_h \tau_s \\ &+ (1 - p_{hit,1}^l - p_{hit,2}^l - \sum_{n_h=1}^{N_{h,max}} p_{hit,s}^{l,n_h})\tau_3 + p_F\tau_4. \end{aligned} \tag{6}$$

For simplification, we use the average content retrieving delay of satellite l to refer to τ_{ave}^l.

Once the caching strategy is determined, the average content retrieving delay can be calculated based on (6). With limited cache sizes, not all the popular contents can be cached in BSs and satellites. Thus the cached contents need to be carefully selected to minimize the average content retrieving delay of the entire network. Let η^l be the weighted factor that reflects the load of the satellite, satisfying $\sum_{l=1}^{L} \eta^l = 1$. Then, we formulate the content placement problem as

$$\begin{aligned} &\min_{\mathbf{a},\mathbf{b}} \sum_{l=1}^{L} \eta^l \tau_{ave}^l \\ &s.t. \ \sum_{i=1}^{F-1} a_i^l \le c_{s,1}, \forall l \in [1,L], \\ &\qquad \sum_{i=1}^{F-1} b_i^l \le c_{s,2}, \forall l \in [1,L], \end{aligned} \tag{7}$$

where the optimization variables $\mathbf{a} = [a_1^1,\ldots,a_{F-1}^1,\ldots,a_1^L,\ldots,a_{F-1}^L]$, $\mathbf{b} = [b_1^1,\ldots,b_{F-1}^1,\ldots,b_1^L,\ldots,b_{F-1}^L]$ are the content caching indicators of all BSs and satellites. Due to the coupling of the caching strategies of different satellites, the proposed content placement problem is of high complexity that cannot be directly solved. In the following sections, we propose two caching strategies by analyzing the properties of the optimal solution. To reduce the content retrieving delay of users, retrieving contents from adjacent satellites is enabled by ISLs in both the two proposed caching strategies. The difference of the two strategies is the optimization target of each satellite, in which we consider each satellite acts selfishly and cooperatively, respectively.

IV. Non-Cooperative Multilayer Caching Strategy

In this section, we first propose a non-cooperative multilayer caching strategy, in which each satellite acts selfishly to minimize the average content retrieving delay of its local users. When determining the caching strategy, each satellite will not consider the content retrieving delay of its adjacent satellites.

A. BS Caching Strategy Analysis

Lemma 1: In the non-cooperative caching strategy, the contents cached in any satellite l do not overlap with the contents cached in the BSs within its coverage, represented by $F_1^l \cap F_2^l = \emptyset, \forall l \in$ [1], $[L]$.

Proof: For any satellite l, if there exists a certain content f_m cached in both the BS and the satellite, $f_m \in F_1^l \cap F_2^l$, we can replace the content f_m in the satellite by any other content f_n that has not been cached, $f_n \notin F_1^l \cup F_2^l$. Then, for users of BSs associated with satellite l, the retrieving delay of content f_n will be reduced from $\tau_n^l(1) \ge \tau_1 + \tau_2 + \tau_s$ to $\tau_n^l(2) = \tau_1 + \tau_2$, while the retrieving delays of other contents stay the same. The average content retrieving delay of satellite l will be reduced. Thus the contents cached in any satellite l do not overlap with the contents cached in the BSs within its coverage. ■

Lemma 2: In the non-cooperative caching strategy, the satellite does not cache the $c_{s,1}$ most popular contents, represented by $\forall f_i \in F_2^l, i > c_{s,1}$.

Proof: Assume that satellite l caches one of the $c_{s,1}$ most popular contents, $f_m \in F_2^l, m \le c_{s,1}$. Then, based on Lemma 1, content f_m will not be cached in the BSs associated with satellite l. Thus there will exist a certain content f_n with popularity ranking lower than $c_{s,1}$ cached in the BSs, $f_n \in F_1^l, n > c_{s,1}$. We exchange the caching locations of content f_m and f_n. Then, for users of BSs associated with satellite l, the variation of the average content retrieving delay can be calculated as $\eta^l(p_n - p_m)\tau_2$. Since $p_m > p_n$, the average content retrieving delay will be reduced. Thus the satellite does not cache the $c_{s,1}$ most popular contents. ■

Based on Lemma 2, the $c_{s,1}$ most popular contents are not cached in all satellites. Thus the $c_{s,1}$ most popular contents cannot be retrieved from the associated satellite or adjacent satellites. To minimize the average content retrieving delay, each BS should cache the $c_{s,1}$ most popular contents. We have Theorem 1 as follows.

Theorem 1: In the non-cooperative caching strategy, each BS caches the $c_{s,1}$ most popular contents, represented by $F_1^l = \{f_1, f_2, \ldots, f_{c_{s,1}}\}, \forall l \in [1,L]$.

Since Theorem 1 is derived without any additional assumption, it is optimal for minimizing the average content retrieving delay of each satellite. We can find that in the non-cooperative caching strategy, the caches of all BSs in the network will be the same, $F_1^l = F_1 = \{f_1, f_2, \dots, f_{c_{s,1}}\}$.

B. Satellite Caching Strategy Analysis

1) The Delay Reduction Gain: The set of contents that may be cached in satellite l can be represented by $F_2^{l,total} = \{f_{c_{s,1}+1}, f_{c_{s,1}+2}, \dots, f_{F-1}\}$. Assume there is no content in the cache of satellite l at the beginning, in which case all the contents in $F_2^{l,total}$ can only be retrieved from adjacent satellites or the gateway. Then, by adding content $f_i \in F_2^{l,total}$ into the cache, the retrieving delay of content f_i will be reduced. The delay reduction gain $\Delta\tau_{2,i}^l$ obtained from caching content f_i can be calculated as follows.

- If content f_i is not cached in all the adjacent satellites of no more than $N_{h,max}$ hops, the retrieving delay will be reduced from $\tau_i^l(1) = \tau_1 + \tau_2 + \tau_3$ to $\tau_i^l(2) = \tau_1 + \tau_2$. The delay reduction gain is $\Delta\tau_{2,i}^l = \eta^l p_i \tau_3$.
- If content f_i is not cached in all the adjacent satellites of less than n_h hops, but is cached in the adjacent satellite of n_h hops, the retrieving delay will be reduced from $\tau_i^l(1) = \tau_1 + \tau_2 + n_h\tau_s$ to $\tau_i^l(2) = \tau_1 + \tau_2$. The delay reduction gain is $\Delta\tau_{2,i}^l = \eta^l p_i n_h \tau_s$.

2) The Least Popular Content: Since the cache size of the satellite is $c_{s,2}$, if a satellite caches a certain content f_n with popularity ranking lower than $c_{s,1} + c_{s,2}$, $f_n \in F_2^l, n > c_{s,1} + c_{s,2}$, there will exist a certain content f_m that is not cached in the satellite, $f_m \notin F_2^l, c_{s,1} < m \leq c_{s,1} + c_{s,2}$. The maximum delay reduction gain of content f_n needs to be higher than the minimum delay reduction gain of content f_m, represented by $\max(\Delta\tau_{2,n}^l) = \eta^l p_n \tau_3 \geq \min(\Delta\tau_{2,m}^l) = \eta^l p_{c_{s,1}+c_{s,2}} \tau_s$. Otherwise, caching content f_m instead of content f_n can always reduce the average content retrieving delay. Thus the least popular content $f_{\beta_{max}^l}$ that may be cached in satellite l can be calculated as

$$\beta_{max}^l = \beta_{max} = \max\{i | p_i \tau_3 \geq p_{c_{s,1}+c_{s,2}} \tau_s\}. \tag{8}$$

For the Zipf distribution, the least popular content can be obtained as

$$\beta_{max}^l = \beta_{max} = \left\lfloor (\frac{\tau_3}{\tau_s})^{\frac{1}{\delta}} (c_{s,1} + c_{s,2}) \right\rfloor. \tag{9}$$

Then, the set of contents that may be cached in satellite l can be represented by $F_2^{l,total} = F_2^{total} = \{f_{c_{s,1}+1}, f_{c_{s,1}+2}, \dots, f_{c_{s,1}+c_{s,2}}, \dots, f_{\beta_{max}}\}$.

3) The Duplicately Cached Contents: For satellite l, in order to minimize the average content retrieving delay of its local users, it is obvious that satellite l should cache the contents with the largest delay reduction gain $\Delta\tau_{2,i}^l$. Based on the calculation above, we can find that the delay reduction gain of each content is linear to its popularity. Caching contents of high popularity will always bring high delay reduction gains, since they are requested more frequently. Consequently, failing to cache these most popular contents in the local cache of the satellite may lead to higher average retrieving delay in all cases. Thus we divide the set F_2^{total} into two sets, $F_{2,0}^{part}$ and $F_{2,0+}^{part}$. The set of contents $F_{2,0}^{part} = \{f_{c_{s,1}+1}, f_{c_{s,1}+2}, \dots, f_{\beta_{dup}}\}$ will be cached in all satellites duplicately, whether these contents are cached in adjacent satellites or not. The contents in $F_{2,0+}^{part} = \{f_{\beta_{dup}+1}, f_{\beta_{dup}+2}, \dots, f_{\beta_{max}}\}$ will be cached selectively according to the caches of adjacent satellites. Obviously, $F_{2,0}^{part}$ cannot be larger than the cache size of the satellite, $\beta_{dup} \leq c_{s,1} + c_{s,2}$. For calculating the set of contents $F_{2,0}^{part}$, we have Theorem 2 as follows.

Theorem 2: In the non-cooperative caching strategy, all satellites cache the set of contents $F_{2,0}^{part}$ duplicately, represented by $F_{2,0}^{part} = \{f_{c_{s,1}+1}, f_{c_{s,1}+2}, \dots, f_{\beta_{dup}}\}, \beta_{dup} = \max\{i | p_i \tau_s \geq p_{c_{s,1}+c_{s,2}+1} \tau_3, c_{s,1} < i \leq c_{s,1} + c_{s,2}\}$.

Proof: Assume that a certain content in the set $F_{2,0}^{part}$ is not cached in satellite l, $f_m \in F_{2,0}^{part}, f_m \notin F_2^l$. The minimum delay reduction gain of content f_m is obtained as $\Delta\tau_{2,m}^l = \eta^l p_m \tau_s$, when content f_m is cached in adjacent satellites of one hop.

Then, let f_n be the least popular content in F_2^l. Since the cache size of the satellite is $c_{s,2}$ and $m \leq c_{s,1} + c_{s,2}$, we have $n \geq c_{s,1} + c_{s,2} + 1$, $f_n \in F_{2,0+}^{part}$. The maximum delay reduction gain of content f_n is obtained as $\Delta\tau_{2,n}^l = \eta^l p_n \tau_3 \leq \eta^l p_{c_{s,1}+c_{s,2}+1} \tau_3$, when content f_n is not cached in all adjacent satellites.

For any content $f_m \in F_{2,0}^{part}$, we have $p_m \tau_s \geq p_{c_{s,1}+c_{s,2}+1} \tau_3$. Then replacing the content f_n in satellite l by content f_m can reduce the average content retrieving delay of satellite l. Content f_m will be cached in the satellite instead of content f_n. Thus all satellites cache the set of contents $F_{2,0}^{part}$ duplicately. ■

Since Theorem 2 is derived without any additional assumption, it is optimal for minimizing the average content retrieving delay of each satellite. For the Zipf distribution, the set of contents $F_{2,0}^{part}$ can be obtained as

$$\beta_{dup} = \left\lfloor (\frac{\tau_s}{\tau_3})^{\frac{1}{\delta}} (c_{s,1} + c_{s,2} + 1) \right\rfloor. \tag{10}$$

If $\beta_{dup} \leq c_{s,1}$, it means that there is no solution for β_{dup}, and the set of $F_{2,0}^{part}$ is empty, $F_{2,0}^{part} = \emptyset$. All the contents $F_2^{total} = \{f_{c_{s,1}+1}, f_{c_{s,1}+2}, \dots, f_{\beta_{max}}\}$ are cached selectively. If $\beta_{dup} = c_{s,1} + c_{s,2}$, it means that all satellites cache the same set of contents $F_{2,0}^{part} = \{f_{c_{s,1}+1}, f_{c_{s,1}+2}, \dots, f_{c_{s,1}+c_{s,2}}\}$.

4) The Non-Cooperative Caching Strategy: In Theorem 1, we have obtained the optimal solution for the caching strategy of the BS, while Theorem 2 gives the optimal solution for the duplicately cached contents in the satellite. Then, considering both the duplicately cached contents and the selectively cached contents, we propose the non-cooperative caching strategy of the satellite as Algorithm 1. At the beginning, we initialize the cache of each satellite as the most popular contents in F_2^{total}. Then, we iteratively calculate the delay reduction gain of satellites. Each satellite will keep caching the set of contents $F_{2,0}^{part}$, while the rest cache is updated by the contents with the largest delay reduction gains in $F_{2,0+}^{part}$. The caching strategies will be updated until the results converge, and we then obtain the non-cooperative caching strategies of all satellites. To accelerate convergence, each satellite will not update the cache if it deteriorates the average content

Algorithm 1 Non-Cooperative Caching Strategy of the Satellite

```
Initialize the cache of each BS as $F_1 = \{f_1, f_2, \dots, f_{c_{s,1}}\}$
Initialize the cache of each satellite as the most popular contents
Initialize the set of adjacent satellites for each satellite by calculating the shortest path in graph $G$
Calculate $\beta_{max}$ according to (8), and calculate $\beta_{dup}$ according to Theorem 2
repeat
  for $l = 1$ to $L$ do
    for $i = \beta_{dup} + 1$ to $\beta_{max}$ do
      Calculate the delay reduction gain $\Delta\tau^l_{2,i}$ according to the caches of adjacent satellites
    end for
    Update the cache of satellite $l$ by $F^l_2 = \{F^{part}_{2,0}, F^{l,part'}_{2,0+}\}$. $F^{l,part'}_{2,0+}$ is the set of contents with the largest delay reduction gains in $F^{part}_{2,0+}$
  end for
until The results converge
```

retrieving delay of the network. Then, the result obtained in each iteration is monotonically nonincreasing. Also, it is obvious that the network has a lower bound as the minimum average content retrieving delay. Based on the fact that a monotonic sequence converges if and only if it is bounded, the proposed Algorithm 1 will finally converge.

In each iteration, the computational complexity for calculating the delay reduction gain of satellites is $O(L^2F)$, and the computational complexity for calculating the set of contents with the largest delay reduction gains is $O(LF\log_2 F)$. Thus the total computational complexity of Algorithm 1 is $O\big(ILF(L + \log_2 F)\big)$, where I is the number of iterations. In Section VI, it is shown that Algorithm 1 converges fast within only several iterations.

V. Cooperative Multilayer Caching Strategy

In this section, we propose a cooperative multilayer caching strategy, in which each satellite acts cooperatively to minimize the average content retrieving delay of the network. Since the contents cached in the satellite can be retrieved by its local users and also the users of its adjacent satellites, each satellite needs to consider the content retrieving delays of all possible users when determining the caching strategy.

A. BS Caching Strategy Analysis

1) The Delay Reduction Gain: Similarly, for the BS associated with satellite l, we calculate the delay reduction gain $\Delta\tau^l_{1,i}$ obtained from caching content f_i as follows.

- If content f_i is cached in satellite l, the delay reduction gain is $\Delta\tau^l_{1,i} = \eta^l p_i \tau_2$.
- If content f_i is not cached in satellite l and all the adjacent satellites less than n_h hops, but is cached in the adjacent satellite of n_h hops, the delay reduction gain is $\Delta\tau^l_{1,i} = \eta^l p_i(\tau_2 + n_h\tau_s)$.
- If content f_i is not cached in satellite l and all the adjacent satellites of no more than $N_{h,max}$ hops, the delay reduction gain is $\Delta\tau^l_{1,i} = \eta^l p_i(\tau_2 + \tau_3)$.

2) The Least Popular Content: Since the cache size of the BS is $c_{s,1}$, if the BS associated with satellite l caches a certain content f_n with popularity ranking lower than $c_{s,1}$, $f_n \in F^l_1, n > c_{s,1}$, one of the $c_{s,1}$ most popular contents will not be cached in the BS, $f_m \notin F^l_1, m \le c_{s,1}$. The maximum delay reduction gain of content f_n needs to be larger than the minimum delay reduction gain of content f_m, represented by $\max(\Delta\tau^l_{1,n}) = \eta^l p_n(\tau_2 + \tau_3) \ge \min(\Delta\tau^l_{1,m}) = \eta^l p_{c_{s,1}}\tau_2$. Otherwise, caching content f_m instead of content f_n can always reduce the average content retrieving delay. Thus the least popular content $f_{\alpha^l_{max}}$ that may be cached by the BS can be calculated as

$$\alpha^l_{max} = \alpha_{max} = \max\{i | p_i(\tau_2 + \tau_3) \ge p_{c_{s,1}}\tau_2\}. \tag{11}$$

Also, for the Zipf distribution, the least popular content can be obtained as

$$\alpha^l_{max} = \alpha_{max} = \left\lfloor (1 + \frac{\tau_3}{\tau_2})^{\frac{1}{\delta}} c_{s,1} \right\rfloor. \tag{12}$$

Then, the set of contents that may be cached in the BS can be represented by $F^{l,total}_1 = F^{total}_1 = \{f_1, f_2, \dots, f_{c_{s,1}}, \dots, f_{\alpha_{max}}\}$.

3) The Duplicately Cached Contents: In the cooperative caching strategy, the contents cached in the BS may overlap with the contents cached in the associated satellite. Each BS will cache $c_{s,1}$ contents with the largest delay reduction gains. Similarly, we divide the set F^{total}_1 into two sets, $F^{part}_{1,0}$ and $F^{part}_{1,0+}$. The set of contents $F^{part}_{1,0} = \{f_1, f_2, \dots, f_{\alpha_{dup}}\}$ will be cached duplicately in all BSs, while contents in $F^{part}_{1,0+} = \{f_{\alpha_{dup}+1}, f_{\alpha_{dup}+2}, \dots, f_{\alpha_{max}}\}$ are cached selectively. Obviously, $F^{part}_{1,0}$ cannot be larger than the cache size of the BS, $\alpha_{dup} \le c_{s,1}$. For calculating the set of contents $F^{part}_{1,0}$, we have Theorem 3 as follows.

Theorem 3: In the cooperative caching strategy, all BSs cache the set of contents $F^{part}_{1,0}$ duplicately, represented by $F^{part}_{1,0} = \{f_1, f_2, \dots, f_{\alpha_{dup}}\}, \alpha_{dup} = \max\{i | p_i(\tau_2 + \tau_s) \ge p_{c_{s,1}+1}(\tau_2 + \tau_3), i \le c_{s,1}\}$.

Proof: Assume that a certain content in the set $F^{part}_{1,0}$ is not cached in the BS associated with satellite l, $f_m \in F^{part}_{1,0}, f_m \notin F^l_1$. Then, let f_n be the least popular content in F^l_1. Since the cache size of the BS is $c_{s,1}$ and $m \le c_{s,1}$, we have $n \ge c_{s,1}+1$, $f_n \in F^{part}_{1,0+}$.

Case 1 (Content f_m Is Not Cached in Satellite l): The minimum delay reduction gain of content f_m is obtained as $\Delta\tau^l_{1,m} = \eta^l p_m(\tau_2 + \tau_s)$, when content f_m is cached in adjacent satellites of one hop. The maximum delay reduction gain of content f_n is obtained as $\Delta\tau^l_{1,n} = \eta^l p_n(\tau_2 + \tau_3) \le \eta^l p_{c_{s,1}+1}(\tau_2 + \tau_3)$, when content f_n is not cached in all adjacent satellites. For any content $f_m \in F^{part}_{1,0}$, we have $p_m(\tau_2 + \tau_s) \ge p_{c_{s,1}+1}(\tau_2 + \tau_3)$. Then, replacing the content f_n in the BS by content f_m can reduce the average content retrieving delay of the network.

Case 2 (Content f_m Is Cached in Satellite l, and Is Not Retrieved by Any Adjacent Satellites): We replace the content f_m in satellite l by f_n, and replace the content f_n in the BS by f_m. Then, the variation of the average content retrieving delay of satellite l can be calculated as $\eta^l(p_n - p_m)\tau_2 < 0$. The average content retrieving delay of satellite l is reduced.

Since content f_m is not retrieved by any adjacent satellites, whether to cache content f_m in satellite l will not influence the content retrieving delay of adjacent satellites. Also, by caching content f_n in satellite l, the adjacent satellites of satellite l can retrieve content f_n from satellite l, while the original paths for retrieving content f_n still exist. The content retrieving delay of adjacent satellites for content f_n will be nonincreasing. Considering satellite l and all adjacent satellites, the average content retrieving delay of the network is reduced.

Case 3 (Content f_m Is Cached in Satellite l, and Is Retrieved by Adjacent Satellites):

For any adjacent satellite l' that retrieves content f_m from satellite l, content f_m will not be cached in satellite l' and the BSs associated with it. Similarly, let $f_{n'}$ be the least popular content in $F_1^{l'}$. Then we have $f_{n'} \in F_1^{l'}, n' \geq c_{s,1}+1$, $f_{n'} \in F_{1,0+}^{part}$. The original average content retrieving delay of satellite l' for these contents can be calculated as $\tau^{l'}(1) \geq \eta^{l'}[p_m(\tau_1 + \tau_2 + \tau_s) + p_n\tau_n^{l'}(1) + p_{n'}\tau_1]$.

We replace the content f_m in satellite l by f_n, replace the content f_n in the BS associated with satellite l by f_m, and replace the content $f_{n'}$ in the BS associated with satellite l' by f_m. Then, the variation of the average content retrieving delay of satellite l can be calculated as $\eta^l(p_n - p_m)\tau_2 < 0$. The average content retrieving delay of satellite l is reduced.

The average content retrieving delay of satellite l' is reduced to $\tau^{l'}(2) \leq \eta^{l'}[p_m\tau_1 + p_n\tau_n^{l'}(2) + p_{n'}(\tau_1 + \tau_2 + \tau_3)]$. Then, the variation of the average content retrieving delay of satellite l' can be calculated as $\tau^{l'}(2) - \tau^{l'}(1) \leq \eta^{l'}[p_{n'}(\tau_2 + \tau_3) - p_m(\tau_2 + \tau_s)] + \eta^{l'}p_n[\tau_n^{l'}(2) - \tau_n^{l'}(1)]$. Similarly, the content retrieving delay of satellite l' for content f_n is nonincreasing, $\tau_n^{l'}(2) \leq \tau_n^{l'}(1)$. Also, for any content $f_m \in F_{1,0}^{part}$, we have $p_m(\tau_2 + \tau_s) \geq p_{c_{s,1}+1}(\tau_2 + \tau_3) \geq p_{n'}(\tau_2 + \tau_3)$. Then, the average content retrieving delay of satellite l' is reduced. Considering satellite l and all adjacent satellites l', caching content f_m in the BS can reduce the average content retrieving delay of the network.

Based on the above analysis of the three cases, content f_m will be cached in the BS instead of content f_n. Thus all BSs cache the set of contents $F_{1,0}^{part}$ duplicately. ■

Since Theorem 3 is derived without any additional assumption, it is optimal for minimizing the average content retrieving delay of the network. For the Zipf distribution, the set of contents $F_{1,0}^{part}$ can be obtained as

$$\alpha_{dup} = \left\lfloor \left(\frac{\tau_2 + \tau_s}{\tau_2 + \tau_3}\right)^{\frac{1}{\delta}} (c_{s,1} + 1) \right\rfloor. \tag{13}$$

If $\alpha_{dup} = c_{s,1}$, it means that all BSs cache the same set of contents $F_{1,0}^{part} = \{f_1, f_2, \ldots, f_{c_{s,1}}\}$.

B. Satellite Caching Strategy Analysis

1) The Delay Reduction Gain: Since all BSs duplicately cache the set of contents $F_{1,0}^{part}$, the contents in $F_{1,0}^{part}$ will not be cached in the satellite. Then, the set of contents that may be cached in satellite l can be represented by $F_2^{l,total} = \{f_{\alpha_{dup}+1}, f_{\alpha_{dup}+2}, \ldots, f_{F-1}\}$. In the cooperative caching strategy, each satellite needs to consider the content retrieving delays of all possible users, including its local users and the users of its adjacent satellites. We define the delay reduction gain obtained from caching content f_i as $\Delta\tau_{2,i}^l = \sum_{n_h=0}^{N_{h,max}} \Delta\tau_{2,i}^{l,n_h}$, in which $\Delta\tau_{2,i}^{l,0}$ represents the delay reduction gain obtained from the local users of satellite l, while $\Delta\tau_{2,i}^{l,n_h} = \sum_{l' \in L^{l,n_h}} \Delta\tau_{2,i}^{l,n_h,l'}$ represents the delay reduction gain obtained from users of its adjacent satellites of n_h hops. The delay reduction gain $\Delta\tau_{2,i}^{l,0}$ can be calculated as follows.

- If content f_i is cached in the BSs associated with satellite l, the delay reduction gain is $\Delta\tau_{2,i}^l = 0$.
- If content f_i is not cached in the BSs and all the adjacent satellites of less than n_h hops, but is cached in the adjacent satellite of n_h hops, the delay reduction gain is $\Delta\tau_{2,i}^{l,0} = \eta^l p_i n_h \tau_s$.
- If content f_i is not cached in the BSs and all the adjacent satellites of no more than $N_{h,max}$ hops, the delay reduction gain is $\Delta\tau_{2,i}^{l,0} = \eta^l p_i \tau_3$.

Then, for each adjacent satellite l' of n_h hops, the delay reduction gain $\Delta\tau_{2,i}^{l,n_h,l'}$ can be calculated as follows.

- If content f_i is cached in satellite l', the BSs associated with satellite l', or other adjacent satellites of no more than n_h hops for satellite l', the delay reduction gain is $\Delta\tau_{2,i}^{l,n_h,l'} = 0$.
- If content f_i not cached in satellite l' and the BSs, but is cached in other adjacent satellites of $n_h' > n_h$ hops for satellite l', the delay reduction gain is $\Delta\tau_{2,i}^{l,n_h,l'} = \eta^{l'} p_i (n_h' - n_h)\tau_s$.
- If content f_i is not cached in satellite l', the BSs, and all other adjacent satellites of no more than $N_{h,max}$ hops for satellite l', the delay reduction gain is $\Delta\tau_{2,i}^{l,n_h,l'} = \eta^{l'} p_i (\tau_3 - n_h\tau_s)$.

2) The Least Popular Content: We have proved that the least popular content that may be cached in the BS is content $f_{\alpha_{max}}$. Then, the set of contents that can only be cached in the satellite can be represented by $F_2^{only} = \{f_{\alpha_{max}+1}, f_{\alpha_{max}+2}, \ldots, f_{F-1}\}$. If satellite l caches a certain content f_n with popularity ranking lower than $\alpha_{max} + c_{s,2}$, $f_n \in F_2^l, n > \alpha_{max} + c_{s,2}$, there will exist a certain content f_m that is not cached in the satellite, $f_m \notin F_2^l, \alpha_{max} < m \leq \alpha_{max} + c_{s,2}$. The maximum delay reduction gain of content f_n needs to be higher than the minimum delay reduction gain of content f_m, represented by $\max(\Delta\tau_{2,n}^l) = p_n[\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)] \geq \min(\Delta\tau_{2,m}^l) = \eta^l p_{\alpha_{max}+c_{s,2}}\tau_s$. Otherwise, caching content f_m instead of content f_n can always reduce the average content retrieving delay. Due to the impact of weighted factors, the least popular content $f_{\beta_{max}^l}$ is different for each satellite, which can be calculated as

$$\beta^l_{max} = \max\{i| \ \ p_i[\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)] \geq \eta^l p_{\alpha_{max}+c_{s,2}}\tau_s\}. \tag{14}$$

For the Zipf distribution, the least popular content can be obtained as

$$\beta^l_{max} = \left\lfloor \left(\frac{\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)}{\eta^l\tau_s} \right)^{\frac{1}{\delta}} (\alpha_{max}+c_{s,2}) \right\rfloor. \tag{15}$$

The set of contents that may be cached in satellite l can be represented by $F_2^{l,total} = \{f_{\alpha_{dup}+1}, f_{\alpha_{dup}+2}, \ldots, f_{\beta^l_{max}}\}$.

3) The Fixedly Cached Contents: Since the set of contents $F_2^{l,total}$ is different for different satellites, we divide the set $F_2^{l,total}$ into two sets, $F_{2,0}^{l,part}$ and $F_{2,0+}^{l,part}$. The set of contents $F_{2,0}^{l,part}$ will be cached in satellite l fixedly, whether these contents are cached in adjacent satellites or not, while contents in $F_{2,0+}^{l,part}$ are cached selectively. For calculating the set of contents $F_{2,0}^{l,part}$, we have Theorem 4 as follows.

Theorem 4: In the cooperative caching strategy, satellite l caches the set of contents $F_{2,0}^{l,part}$ fixedly, represented by $F_{2,0}^{l,part} = \{f_{\alpha_{max}+1}, f_{\alpha_{max}+2}, \ldots, f_{\beta^l_{fix}}\}, \beta^l_{fix} = \max\{i|\eta^l p_i \tau_s \geq p_{\alpha_{dup}+c_{s,2}+1} \ [\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'} \ (\tau_3 - n_h\tau_s)], \alpha_{max} < i \leq \alpha_{dup} + c_{s,2}\}$.

Proof: Assume that a certain content in the set $F_{2,0}^{l,part}$ is not cached in satellite l, $f_m \in F_{2,0}^{l,part}, f_m \notin F_2^l$. Since $m > \alpha_{max}$, content f_m will not be cached in the BS associated with satellite l. The minimum delay reduction gain of content f_m is obtained as $\Delta\tau^l_{2,m} = \eta^l p_m \tau_s$, when content f_m is cached in adjacent satellites of one hop and all adjacent satellites will not retrieve content f_m from satellite l.

Then, let f_n be the least popular content in F_2^l. The set of contents that may be cached in satellite l is $F_2^{l,total} = \{f_{\alpha_{dup}+1}, f_{\alpha_{dup}+2}, \ldots, f_{\beta^l_{max}}\}$. Since the cache size of the satellite is $c_{s,2}$ and $m \leq \alpha_{dup} + c_{s,2}$, we have $n \geq \alpha_{dup} + c_{s,2} + 1$, $f_n \in F_{2,0+}^{l,part}$. The maximum delay reduction gain of content f_n is obtained as $\Delta\tau^l_{2,n} = p_n[\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)] \leq p_{\alpha_{dup}+c_{s,2}+1}[\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)]$, when content f_n is not cached in all adjacent satellites and all adjacent satellites will retrieve content f_n from satellite l.

For any content $f_m \in F_{2,0}^{l,part}$, we have $\eta^l p_m \tau_s \geq p_{\alpha_{dup}+c_{s,2}+1}[\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)]$. Then, replacing the content f_n in satellite l by content f_m can reduce the average content retrieving delay of the network. Content f_m will be cached in satellite l instead of content f_n. Thus satellite l caches the set of contents $F_{2,0}^{l,part}$ fixedly. ■

Since Theorem 4 is derived without any additional assumption, it is optimal for minimizing the average content retrieving delay of the network. For the Zipf distribution, the set of contents $F_{2,0}^{l,part}$ can be obtained as

$$\beta^l_{fix} = \left\lfloor \left(\frac{\eta^l\tau_s}{\eta^l\tau_3 + \sum_{n_h=1}^{N_{h,max}} \sum_{l' \in L^{l,n_h}} \eta^{l'}(\tau_3 - n_h\tau_s)} \right)^{\frac{1}{\delta}} (\alpha_{dup}+c_{s,2}+1) \right\rfloor. \tag{16}$$

If $\beta^l_{fix} \leq \alpha_{max}$, it means that there is no solution for β^l_{fix}, and the set of $F_{2,0}^{l,part}$ is empty, $F_{2,0}^{l,part} = \emptyset$. All the contents $F_2^{l,total} = \{f_{\alpha_{dup}+1}, f_{\alpha_{dup}+2}, \ldots, f_{\beta^l_{max}}\}$ are cached selectively.

C. Cooperative Caching Strategy of the Network

For both BSs and satellites, it is optimal to cache the contents with the largest delay reduction gains. However, the delay reduction gain of each BS is determined by its associated satellite and the adjacent satellites, while the delay reduction gain of each satellite is determined by the BSs associated with it and the adjacent satellites. In Theorem 3, we have obtained the optimal solution for the duplicately cached contents in the BS, while Theorem 4 gives the optimal solution for the fixedly cached contents in each satellite. Then, based on the analysis of the caching strategies above, we propose the cooperative caching strategy of the network as Algorithm 2. At the beginning, we initialize the cache of each BS and satellite as the most popular contents. Then, we iteratively calculate the delay reduction gain of the BSs and satellites. Each BS will keep caching the set of contents $F_{1,0}^{part}$, while the rest cache is updated by the contents with the largest delay reduction gains in $F_{1,0+}^{part}$. Also, satellite l will keep caching the set of contents $F_{2,0}^{l,part}$, while the rest cache is updated by the contents with the largest delay reduction gains in $F_{2,0+}^{l,part}$. The caching strategies will be updated until the results converge, and we then obtain the cooperative caching strategies of the network. In each iteration, let τ_{ave} and τ'_{ave} be the average content retrieving delay of the network before and after updating the cache of the BS associated with satellite l. We have $\tau'_{ave} \leq \tau_{ave}$, since the average content retrieving delay of satellite l is reduced, while the average content retrieving delay of other satellites stays the same. Similarly, let τ'_{ave} and τ^*_{ave} be the average content retrieving delay of the network before and after updating the cache of satellite l. We have $\tau^*_{ave} \leq \tau'_{ave}$, since the delay reduction gain of the satellite is calculated by considering the average content retrieving delay of the network. Thus, the result obtained in each iteration is monotonically nonincreasing. Also, the network has a lower bound as the minimum average content retrieving delay. Based on the fact that a monotonic sequence converges if and only if it is bounded, the proposed Algorithm 2 will finally converge.

Algorithm 2 Cooperative Caching Strategy of the Network

```
Initialize the cache of each BS and satellite by the most popular contents
Initialize the set of adjacent satellites for each satellite by calculating the shortest path in graph $G$
Calculate $\alpha_{max}$ according to (11), and calculate $\alpha_{dup}$ according to Theorem 3
Calculate $\beta^l_{max}$ according to (14), and calculate $\beta^l_{fix}$ according to Theorem 4
repeat
  for $l = 1$ to $L$ do
    for $i = \alpha_{dup} + 1$ to $\alpha_{max}$ do
      Calculate the delay reduction gain of the BS $\Delta\tau^l_{1,i}$ according to the caches of satellite $l$ and the adjacent satellites
    end for
    Update the cache of the BS by $F^l_1 = \{F^{part}_{1,0}, F^{l,part'}_{1,0+}\}$. $F^{l,part'}_{1,0+}$ is the set of contents with the largest delay reduction gains in $F^{part}_{1,0+}$
    for $i = \beta^l_{fix} + 1$ to $\beta^l_{max}$ do
      Calculate the delay reduction gain of the satellite $\Delta\tau^l_{2,i}$ according to the caches of the BS and the adjacent satellites
    end for
    Update the cache of satellite $l$ by $F^l_2 = \{F^{l,part}_{2,0}, F^{l,part'}_{2,0+}\}$. $F^{l,part'}_{2,0+}$ is the set of contents with the largest delay reduction gains in $F^{l,part}_{2,0+}$
  end for
until The results converge
```

In each iteration, the computational complexity for calculating the delay reduction gain of BSs is $O(L^2F)$, the computational complexity for calculating the set of contents with the largest delay reduction gains of BSs is $O(LF\log_2 F)$, the computational complexity for calculating the delay reduction gain of satellites is $O(L^3F)$, and the computational complexity for calculating the set of contents with the largest delay reduction gains of satellites is $O(LF\log_2 F)$. Thus the total computational complexity of Algorithm 2 is $O\big(ILF(L^2+\log_2 F)\big)$, where I is the number of iterations. In Section VI, it is shown that Algorithm 2 converges fast within only several iterations.

VI. Simulation Results

In this section, by using the simulator of MATLAB, numerical results are presented to demonstrate the performance of the proposed three-layer cooperative caching architecture and the caching strategies.

A. Simulation Setup

The default settings of the system are $F = 1{,}000$ contents, $c_{s,1} = 100$, $c_{s,2} = 100$, $L = 100$ satellites, $\tau_1 = 20$ ms, $\tau_2 = 200$ ms, $\tau_3 = 180$ ms, $\tau_4 = 500$ ms, and $\tau_s = 40$ ms. The content popularity is considered to follow the Zipf distribution with the Zipf exponent $\delta = 0.5$. The satellite network is created by polar constellations. Each satellite is connected with four adjacent satellites by ISLs [35], including two adjacent satellites in the same orbit and two satellites in adjacent orbits. When retrieving contents from adjacent satellites, the shortest path is selected by the Dijkstra algorithm. The loads of satellites are generated by poisson distribution with poisson parameter λ_l, in which we set $\frac{\max \lambda_l}{\min \lambda_l} = 5$. The caching strategies implemented for comparison are listed as follows:

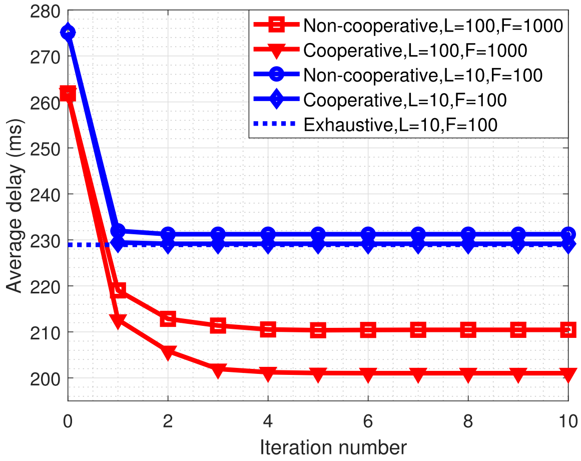


Fig. 3. The convergence process of proposed caching strategies.

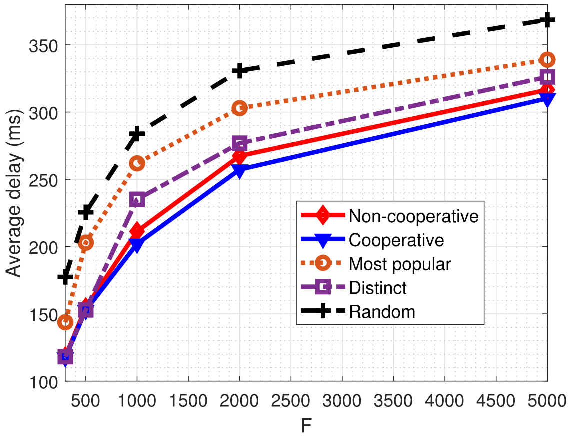


Fig. 4. Performance comparison for different content numbers.

1. **Non-cooperative**. This is the non-cooperative caching strategy proposed in Section IV.
2. **Cooperative**. This is the cooperative caching strategy proposed in Section V.
3. **Most popular**. This is the benchmark strategy for content placement problems, in which the most popular contents are cached in a greedy manner.
4. **Distinct**. This strategy is derived from the horizontal cooperative caching strategy proposed in [34]. While the most popular contents are duplicately cached, the less popular contents are cached distinctly in all satellites for increasing the diversity.
5. **Random**. All contents are cached randomly.

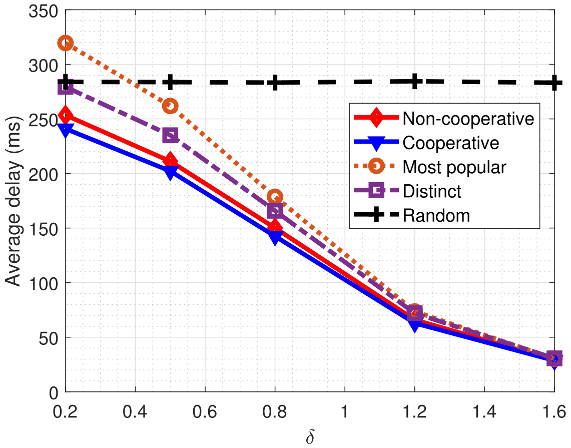


Fig. 5. Performance comparison for different Zipf exponents.

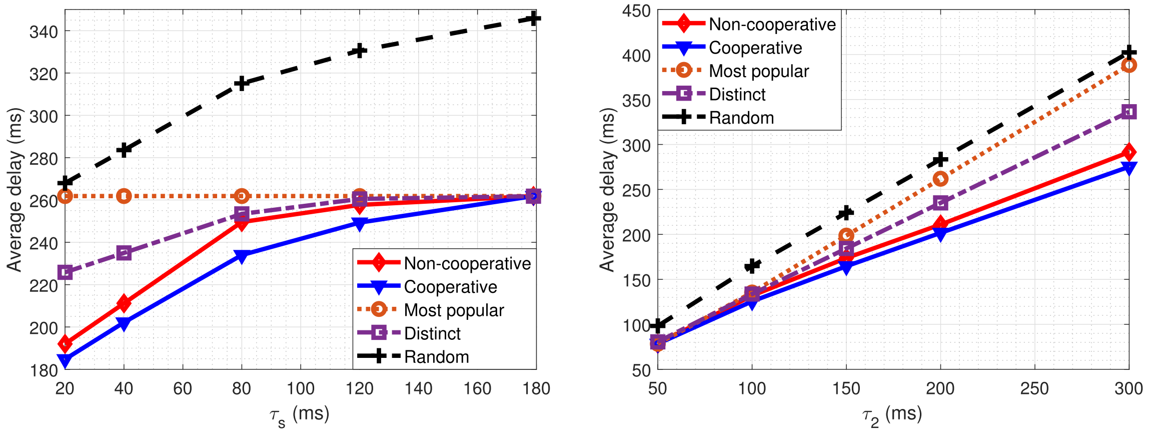
B. Performance Analysis

Fig. 3 shows the convergence process of the two proposed caching strategies. We can observe that both the two caching strategies converge fast for either large scale or small scale cases, within less than 10 iterations. In the non-cooperative caching strategy, each satellite updates the cache by the contents with the largest delay reduction gains in $F_{2,0+}^{part}$. In the cooperative caching strategy, each BS and satellite updates the cache by the contents with the largest delay reduction gains in $F_{1,0+}^{part}$ and $F_{2,0+}^{l,part}$. Both of the two caching strategies can be obtained within only several updating processes, demonstrating the feasibility of the proposed algorithms. Also, for the case of small scale, we find the optimal caching strategy via exhaustive search. It can be observed that the cooperative caching strategy performs very closely to the exhaustive algorithm, proving the optimality of the proposed algorithm.

Fig. 4 shows the performance comparison of the five caching strategies introduced above, in which different content numbers are considered. We can observe that the proposed cooperative caching strategy can achieve the lowest delay in all cases, while the non-cooperative caching strategy can achieve suboptimal performance. For $F = 1{,}000$, the proposed cooperative caching strategy can reduce the average content retrieving delay by 30% compared with the random caching strategy, by 25% compared with the most popular caching strategy, and by 15% compared with the distinct caching strategy. Also, we can observe that the proposed two caching strategies outperform other strategies more significantly when the content number is large. For small value of $F = 300$, since most of the contents can be cached, similar performance is achieved for the distinct caching strategy and the two proposed strategies. However, when F is large, cooperative caching is important for reducing the content retrieving delay. If a popular content is cached in the adjacent satellites of satellite l, satellite l or the BSs associated with it will not need to cache this content locally. Instead, other contents with larger delay reduction gains can be cached. In real cases, the size of the content set is generally much larger than the size of the cache. By fully utilizing the cooperative caching architecture in the integrated satellite-terrestrial network with the proposed algorithms, the content retrieving delay can be significantly reduced. In addition, compared with the cooperative caching strategy, there is about 5% performance loss for the non-cooperative caching strategy, since each satellite aims to minimize the average content retrieving delay of its local users, failing to optimize the global performance of the network. However, as discussed above, the computational complexity of the non-cooperative caching strategy is lower than the cooperative caching strategy. Suboptimal performance can be achieved by the non-cooperative caching strategy with lower complexity.

Fig. 5 shows the performance comparison of the five caching strategies for different Zipf exponents. We can observe that the proposed caching strategies can achieve superior performance for any value of δ. When δ is small, the popularity of different contents distributes more evenly. In this case, since most of the contents will be requested frequently, increasing the diversity of cached contents can help to reduce the average content retrieving delay by means of cooperative content retrieving. On the contrary, long content retrieving delay will be brought if applying the benchmark most popular caching strategy in this case. For $\delta = 0.2$, the proposed cooperative caching strategy outperforms the most popular caching strategy by 25%, outperforms the random caching strategy by 15%, and outperforms the distinct caching strategy by 13%. However, When δ is large, the popularity of different contents distributes more unevenly, in which case the request will be concentrated on a few popular contents. The delay reduction gains of these contents will be larger than all other contents, whether these contents are cached in adjacent satellites or not. Then, simply caching these popular contents can achieve the optimal performance. When $\delta = 1.6$, all the caching strategies, except the random caching strategy, can achieve similar performance.

Fig. 6 shows the performance comparison for different settings of link delays. We can also observe that the proposed cooperative caching strategy can achieve the lowest delay for different settings, while the proposed non-cooperative caching strategy can achieve suboptimal performance. Fig. 6 (a) shows the performance comparison for different inter-satellite delay τ_s. When τ_s is small, retrieving contents from adjacent satellites is of lower delay, and more adjacent satellites are accessible with larger $N_{h,max}$. We can observe that the proposed caching strategies can achieve more superior performance for small value of τ_s. For $\tau_s = 20$ ms, the proposed cooperative caching strategy can reduce the average content retrieving delay by 30% compared with the random caching strategy, by 28% compared with the most popular caching strategy, and by 18% compared with the distinct caching strategy. When $\tau_s = \tau_3 = 180$ms, no delay reduction gain can be obtained from cooperative caching. Then, caching the most popular contents can achieve the optimal performance. Also, when τ_s decreases from 40 ms to 20 ms, the average content retrieving delay is reduced by 10 %. For satellite constellations of the same orbit altitude, increasing the number of satellites can reduce the delay of ISLs, and then reduce the content retrieving delay. Fig. 6 (b) shows the performance comparison for different satellite-ground delay. We can observe that the



(a) Different inter-satellite delay.

(b) Different satellite-ground delay.

Fig. 6. Performance comparison for different settings of link delays.

proposed caching strategies outperform other caching strategies more significantly for large value of τ_2. For $\tau_2 = 300$ ms, the proposed cooperative caching strategy can reduce the average content retrieving delay by 32% compared with the random caching strategy, by 29% compared with the most popular caching strategy, and by 18% compared with the distinct caching strategy. Also, when τ_2 decreases from 200 ms to 100 ms, the average content retrieving delay is reduced by 35%. Generally, lower satellite-ground delay can be achieved by satellites of lower orbit altitudes. For LEO satellites on the orbit of 1,000 km, the propagation delay brought to τ_2 is 7 ms, while the propagation delay brought to τ_2 is 133 ms for MEO satellites on the orbit of 20,000 km. Thus most satellite constellation projects established now are composed of LEO satellites, such as Starlink, OneWeb, and Telesat [5].

VII. Conclusion

In this paper, we investigated the cooperative multilayer edge caching in the integrated satellite-terrestrial network. Based on the proposed three-layer cooperative caching model, we analyzed the content retrieving process and derived the cache hit probability for different caching locations. Considering each satellite acts selfishly to minimize the average content retrieving delay of its local users, we first proposed the non-cooperative multilayer caching strategy. We proved that the non-cooperative caching strategy of the BS is to cache the most popular contents in all BSs, while the non-cooperative caching strategy of the satellite was obtained by introducing the concept of delay reduction gains. Considering each satellite acts cooperatively to minimize the average content retrieving delay of the network, we then proposed the cooperative multilayer caching strategy to fully utilize the advantage of cooperation. An iterative algorithm was proposed to calculate the optimal caching strategy of the entire network. Numerical results showed that the proposed three-layer cooperative caching model can significantly reduce the content retrieving delay in the integrated satellite-terrestrial network. Based on the performance comparison, it was demonstrated that the proposed cooperative caching strategy can achieve the lowest delay in all cases, while the proposed non-cooperative caching strategy can achieve suboptimal performance with lower complexity. Besides, since cooperative caching among satellites is enabled by the ISLs, the design of the satellite network topology is also an important issue for delay optimization, which can be further explored in future research.

References

[1] L. Kuang, X. Chen, C. Jiang, H. Zhang, and S. Wu, "Radio resource management in future terrestrial-satellite communication networks," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 81–87, Oct. 2017.

[2] X. Zhu, C. Jiang, L. Kuang, Z. Zhao, and S. Guo, "Two-layer game based resource allocation in cloud based integrated terrestrial-satellite networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 509–522, Jun. 2020.

[3] D. Bogdan-Martin, *Measuring Digital Development: Facts and Figures*. Geneva, Switzerland: International Telecommunication Union, 2019.

[4] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband LEO satellite communications: Architectures and key technologies," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 55–61, Apr. 2019.

[5] I. D. Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta Astronautica*, vol. 159, pp. 123–135, Jun. 2019.

[6] M. Latva-aho and K. Leppänen, "Key drivers and research challenges for 6G ubiquitous wireless intelligence," Oulu, Finland, Sep. 2019.

[7] L. Kuang, C. Jiang, Y. Qian, and J. Lu, *Terrestrial-Satellite Communication Networks: Transceivers Design and Resource Allocation*. Cham, Switzerland: Springer, 2017.

[8] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[9] A. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A survey of mobile cloud computing application models," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 393–413, 1st Quart., 2014.

[10] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[11] Z. Zhang, W. Zhang, and F.-H. Tseng, "Satellite mobile edge computing: Improving QoS of high-speed satellite-terrestrial networks using edge computing techniques," *IEEE Netw.*, vol. 33, no. 1, pp. 70–76, Jan. 2019.

[12] J. Montalban *et al.*, "Multimedia multicast services in 5G networks: Subgrouping and non-orthogonal multiple access techniques," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 91–95, Mar. 2018.

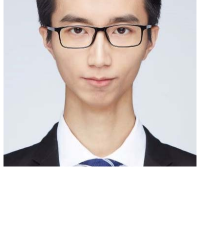
[13] M. Amjad, M. H. Rehmani, and S. Mao, "Wireless multimedia cognitive radio networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1056–1103, 2nd Quart., 2018.

[14] X. Zhang and Q. Zhu, "Hierarchical caching for statistical QoS guaranteed multimedia transmissions over 5G edge computing mobile wireless networks," *IEEE Wireless Commun. Mag.*, vol. 25, no. 3, pp. 12–20, Jun. 2018.

[15] J. Yao, T. Han, and N. Ansari, "On mobile edge caching," *IEEE Commun. Surveys Tuts.*, vol. 21, pp. 2525–2553, 3rd Quart., 2019.

[16] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[17] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 80–87, Jun. 2018.
[18] A. Kalantari, M. Fittipaldi, S. Chatzinotas, T. X. Vu, and B. Ottersten, "Cache-assisted hybrid satellite-terrestrial backhauling for 5G cellular networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.
[19] C. Jiang and Z. Li, "Decreasing big data application latency in satellite link by caching and peer selection," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2555–2565, Oct. 2020.
[20] K. An, Y. Li, X. Yan, and T. Liang, "On the performance of cache-enabled hybrid satellite-terrestrial relay networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1506–1509, Oct. 2019.
[21] E. Wang, X. Lin, and S. Zhang, "Content placement based on utility function for satellite networks," *IEEE Access*, vol. 7, pp. 163150–163159, 2019.
[22] C. Qiu, H. Yao, F. R. Yu, F. Xu, and C. Zhao, "Deep Q-learning aided networking, caching, and computing resources allocation in software-defined satellite-terrestrial networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5871–5883, Jun. 2019.
[23] S. Liu, X. Hu, Y. Wang, G. Cui, and W. Wang, "Distributed caching based on matching game in leo satellite constellation networks," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 300–303, Feb. 2018.
[24] E. Wang, H. Li, and S. Zhang, "Load balancing based on cache resource allocation in satellite networks," *IEEE Access*, vol. 7, pp. 56864–56879, 2019.
[25] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, "UAV-relaying-assisted secure transmission with caching," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3140–3153, May 2019.
[26] H. Wu, J. Li, H. Lu, and P. Hong, "A two-layer caching model for content delivery services in satellite-terrestrial networks," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.
[27] X. Artiga *et al.*, "Shared access satellite-terrestrial reconfigurable backhaul network enabled by smart antennas at mmWave band," *IEEE Netw.*, vol. 32, no. 5, pp. 46–53, Sep./Oct. 2018.
[28] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. S. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.
[29] S. Zhang, N. Zhang, P. Yang, and X. Shen, "Cost-effective cache deployment in mobile heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11264–11276, Dec. 2017.
[30] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *J. Internet Services Appl.*, vol. 5, no. 1, p. 8, Dec. 2014.
[31] T. X. Tran and D. Pompili, "Adaptive bitrate video caching and processing in mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 18, no. 9, pp. 1965–1978, Sep. 2019.
[32] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
[33] S. Zhang, W. Sun, and J. Liu, "Spatially cooperative caching and optimization for heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11260–11270, Nov. 2019.
[34] Q. Li, W. Shi, X. Ge, and Z. Niu, "Cooperative edge caching in software-defined hyper-cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2596–2605, Nov. 2017.
[35] R. Liu *et al.*, "Capacity of two-layered satellite networks," *Wireless Netw.*, vol. 23, no. 8, pp. 2651–2669, Nov. 2017.



Xiangming Zhu received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, China, in 2014 and 2019, respectively. He currently holds a post-doctoral position with Zhejiang Lab and Beijing National Research Center for Information Science and Technology, Tsinghua University. His major research interests include satellite networking, integrated terrestrial-satellite communications, and edge computing. He received the Best Paper Award IWCMC in 2017.

Chunxiao Jiang (Senior Member, IEEE) received the B.S. degree (Hons.) in information engineering from Beihang University, Beijing, in 2008, and the Ph.D. degree (Hons.) in electronic engineering from Tsinghua University, Beijing, in 2013. He is currently an Associate Professor with the School of Information Science and Technology, Tsinghua University. His research interests include application of game theory, optimization, and statistical theories to communication, networking, and resource allocation problems, in particular space networks and heterogeneous networks. He was a recipient of the Best Paper Award from IEEE GLOBECOM in 2013, the Best Student Paper Award from IEEE GlobalSIP in 2015, the IEEE Communications Society Young Author Best Paper Award in 2017, the Best Paper Award IWCMC in 2017, the IEEE ComSoc TC Best Journal Paper Award of the IEEE ComSoc TC on Green Communications and Computing 2018, the IEEE ComSoc TC Best Journal Paper Award of the IEEE ComSoc TC on Communications Systems Integration and Modeling 2018, and the Best Paper Award from ICC 2019. He received the Chinese National Second Prize in the Technical Inventions Award in 2018 and the Natural Science Foundation of China Excellent Young Scientists Fund Award in 2019. He has also served as a member for the Technical Program Committee as well as the Symposium Chair for a number of international conferences, including the IEEE CNS 2020 Publication Chair, the IEEE WCSP 2019 Symposium Chair, the IEEE ICC 2018 Symposium Co-Chair, the IWCMC 2020/19/18 Symposium Chair, the WiMob 2018 Publicity Chair, the ICCC 2018 Workshop Co-Chair, and the ICC 2017 Workshop Co-Chair. He has served as an Editor for IEEE INTERNET OF THINGS JOURNAL, *IEEE Network*, and IEEE COMMUNICATIONS LETTERS, and a Guest Editor for *IEEE Communications Magazine*, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING.

Linling Kuang (Member, IEEE) received the B.S. and M.S. degrees from the National University of Defense Technology, Changsha, China, in 1995 and 1998, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2004. Since 2007, she has been with Tsinghua Space Center, Tsinghua University. Her research interests include wireless broadband communications, signal processing, and satellite communication. She is a member of the IEEE Communications Society.

Zhifeng Zhao (Member, IEEE) received the B.E. degree in computer science, the M.E. degree in communication and information system, and the Ph.D. degree in communication and information system from the PLA University of Science and Technology, Nanjing, China, in 1996, 1999, and 2002, respectively. From 2002 to 2004, he acted as a Post-Doctoral Researcher with Zhejiang University, Hangzhou, China, where his researches were focused on multimedia next generation networks (NGNs) and soft-switch technology for energy efficiency. From 2005 to 2006, he acted as a Senior Researcher with the PLA University of Science and Technology, where he performed research and development on advanced energy-efficient wireless router, *ad hoc* network simulator, and cognitive mesh networking test-bed. From 2006 to 2019, he was an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University. He is currently with Zhejiang Lab, Hangzhou. His research area includes software defined networks (SDN), wireless network in 6G, computing networks, and collective intelligence. He is the Symposium Co-Chair of ChinaCom 2009 and 2010. He is the Technical Program Committee (TPC) Co-Chair of the 10th IEEE International Symposium on Communication and Information Technology (ISCIT 2010).