Opportunistic Content-Aware Routing in Satellite-Terrestrial Integrated Networks

Jin Tang, Jian Li, *Senior Member, IEEE*, Lan Zhang, *Member, IEEE*, Xianhao Chen, *Member, IEEE*, Kaiping Xue, *Senior Member, IEEE*, Qibin Sun, *Fellow, IEEE*, and Jun Lu

***Abstract*—As a promising complement to terrestrial cellular networks, satellite networks have recently drawn increasing attention, offering seamless coverage cost-effectively. However, with the rapidly increasing users' demand for multimedia content, how to achieve efficient content transmission seamlessly becomes a critical but knotty problem. To provide an efficient solution from the routing perspective, in this paper, we propose an opportunistic content-aware routing scheme. Our scheme combines the features of in-network caching and content awareness of information-centric networking (ICN) architecture. The basic idea of the proposed scheme is to sense users' requests and find the optimal route solution with the largest potential gain. Moreover, considering the limitation of real-time signaling collection in satellite networks, we design a cached content prediction method. The method is capable of inferring the probability of content being cached based on historical popularity information, providing essential information for measuring potential gains. Extensive simulation results demonstrate that the proposed opportunistic content-aware routing scheme outperforms baseline approaches with significantly reduced delay and traffic consumption.**

***Index Terms*—Satellite-terrestrial integrated networks, content-aware, opportunistic routing, cache prediction.**

Manuscript received 15 August 2023; revised 10 January 2024; accepted 5 March 2024. Date of publication 19 March 2024; date of current version 3 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62201540, and in part by the Open Research Fund of State Key Laboratory of Integrated Services Networks under Grant ISN24-11. This paper was presented in part in the Proceedings of the IEEE Global Communications Conference (GLOBECOM) 2022 [DOI: 10.1109/GLOBECOM48099.2022.10001143]. Recommended for acceptance by C. M. Pinotti. *(Corresponding author: Jian Li.)*

Jin Tang is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, China, and also with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China.

Jian Li is with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China, and also with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China (e-mail: lijian9@ustc.edu.cn).

Lan Zhang is with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634 USA.

Xianhao Chen is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Pok Fu Lam, Hong Kong.

Kaiping Xue and Qibin Sun are with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China.

Jun Lu is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TMC.2024.3377729, provided by the authors.

Digital Object Identifier 10.1109/TMC.2024.3377729

I. Introduction

As a promising complement to the terrestrial cellular system, satellite networks have experienced rapid development to stride toward large-scale satellite-terrestrial integrated networks (STINs) in the 5G/6G era [2], [3], [4], [5]. Due to the characteristic of high altitude and broadcast nature, satellite networks can provide seamless coverage without the requirement of infrastructure [6], [7], which is attractive to terrestrial users located in rural areas, oceans and etc. Along with the rapid expansion of satellite constellation and tremendous advancements in communication technologies, it can be foreseen that terrestrial users can obtain ubiquitous services such as Internet access and content retrieval through STINs [8].

According to the latest report, by 2028, global mobile traffic is expected to increase about three times compared to today, and the number of IoT devices will reach more than 24 billion by 2030 [9], [10]. Consequently, the demand for multimedia content by users and devices expands tremendously in recent years [11]. Increasing content traffic generated by numerous applications challenges the capability of data delivery in satellite networks. The challenge mainly stems from the limitation of the capacity of the inter-satellite link (ISL) in the STIN: the limited link capacity is difficult to meet the ever-increasing content transmission requirements, that is, higher content transmission traffic burden and lower content transmission delay.

With the development of the storage capability of satellites [12], [13], [14], it becomes promising and feasible to develop on-board caching of low earth orbit (LEO) satellites in the STIN to improve the efficiency of content transmission. By actively caching frequently requested contents on edge satellites, popular contents can be delivered directly to users instead of being retrieved from remote servers through backhaul links. Thus precious on-board resources of LEO satellites, e.g., link capacity and communication energy, can be significantly saved [15], [16], [17], [18], [19], [20], [21]. On this basis, as the scale of the satellite network increases, in-network caching can bring new vitality to inter-satellite networking and routing. For example, the existing studies further adopt a new network architecture, combining the information-centric networking (ICN) [1], [22], [23], [24], [25], [26], [27] architecture with the STIN. The ICN architectures are characterized by two major features, i.e., routing-by-name and in-network caching, to offload the redundant traffic from users' requests for popular content and thus significantly improve the efficiency of content transmission in satellite networks.

Although the effectiveness of caching technology and ICN architecture has been verified in satellite networks, most of the existing research focuses on the optimization of caching and transmission strategies, and efficient routing design with content-aware functions is still lacking in satellite networks. In order to realize content-aware routing, it is necessary to collect global cache information, i.e., the cache status of cache-enabled nodes, in real-time to provide the necessary information for routing. Then satellites can more accurately route user requests to the nearest cache-enabled node to retrieve the desired content and deliver it to users. By doing this, the delay and traffic consumption of content transmission can be reduced. However, due to satellite networks' highly dynamic network topology and the limitation of communication overhead, global cache information is challenging to collect in real time through ISLs. Therefore, current content-aware routing designs in terrestrial networks cannot be directly applied to satellite networks [28], [29]. In this case, how to design an efficient content-aware routing scheme becomes a critical but unsolved problem.

In this paper, we propose an opportunistic content-aware routing and forwarding scheme. Considering the difficulty of the real-time collection of global cache information, we use medium earth orbit (MEO) and geostationary earth orbit (GEO) satellites to periodically collect and distribute signaling to reduce communication overhead. Signaling is the set of popularity tables for cache-enabled satellites, recording information for each request. In order to enable LEO satellites to know the cache status of cache-enabled satellites more accurately and provide helpful information for our content-aware routing, we propose a cache prediction method that can predict cache based on periodically distributed signaling. Furthermore, we propose a potential gains model to calculate the delay in retrieving contents from different nodes, including cache-enabled satellites and the ground station. Overall, the proposed scheme can route user requests to the optimal node in the network, that is, the node with the largest potential gains, thereby reducing content retrieval delay and traffic overhead. Finally, the scheme also gives the process of different satellites processing for requests and data packets, which provides a reference model for designing opportunistic content-aware routing schemes that utilize cache capabilities.

The contributions of this paper are summarized as follows:

- We design a cache prediction method to estimate the probability that contents are cached in the network through historical popularity information and satellites' cache strategy. The prediction method can obtain a more accurate cache status of cache-enabled nodes with less communication overhead, thus providing necessary information for subsequent decision-making.
- We design an opportunistic content-aware routing scheme in the satellite network, which can route user requests to the optimal node at a small cost. We also give the paradigm of opportunistic content-aware routing.
- We conduct extensive simulations compared with existing schemes to evaluate the proposed opportunistic content-aware routing performance. The results show a significant reduction of the proposed schemes in terms of content retrieval delay and traffic consumption compared with traditional routing and content-aware routing schemes.

The rest of the paper is organized as follows: The related works are introduced in Section II. The system model is proposed in Section III, and the cached content prediction method and theoretical analysis are discussed in Section IV. After that, the opportunistic content-aware routing design is proposed in Section V. Then, the performance evaluation and analysis are conducted in Section VI. Finally, the conclusions and future works are drawn in Section VII.

II. Related Works

The development of satellite storage capabilities has opened up new possibilities for improving content transmission efficiency in the STIN. By actively caching frequently requested content on satellites, popular content can be delivered directly to users, eliminating the need for retrieval from remote servers via backhaul links. Motivated by this advantage, Lai et al. [15] proposed a framework for building a global content distribution network using LEO satellites and cloud platforms, minimizing storage and traffic costs while maintaining low-latency content access. Han et al. [16] combined caching and multicast beamforming to propose a joint optimization problem. In summary, the long-term cache storage and the short-term content transmission problems were jointly considered to improve the cache utilization and spectrum efficiency in the STIN. Besides, the authors in [17] proposed a cooperative strategy to reduce satellite communication transmission delay, including cache placement algorithm for downlink and peer selection algorithm for uplink, thereby improving the performance in terms of delay, hit rate, channel rate, and handoff rate. An et al. [18] explored the incorporation of wireless content caching into hybrid satellite-to-ground relay networks, proposed two cache placement schemes, and demonstrated their significant performance improvements in terms of outage probability and spectral efficiency. Li et al. [19] proposed a cooperative transmission scheme by introducing a cache-enabled LEO satellite network as part of the radio access network (RAN). The cooperative transmission scheme achieved significant improvements in traffic offload and energy efficiency, which is suitable for coping with the mobile traffic explosion growing demand. In addition, Jiang et al. [20] divided the satellite network into blocks of different sizes and performed appropriate cache-enabled satellite deployment according to the ground user density. The proposed method can improve user experience and achieve efficient content distribution. Zhu et al. [21] proposed a framework for a three-tier cooperative caching that enables caching at satellites, base stations, and gateways. They formulated the cache placement problem as an optimization problem of minimizing the average retrieval latency over the network.

The above studies concentrate on caching optimization, which is concerned on the cache performance on a single satellite. By caching popular contents on the cache-enabled satellites and responding to users when the requested content is cached,

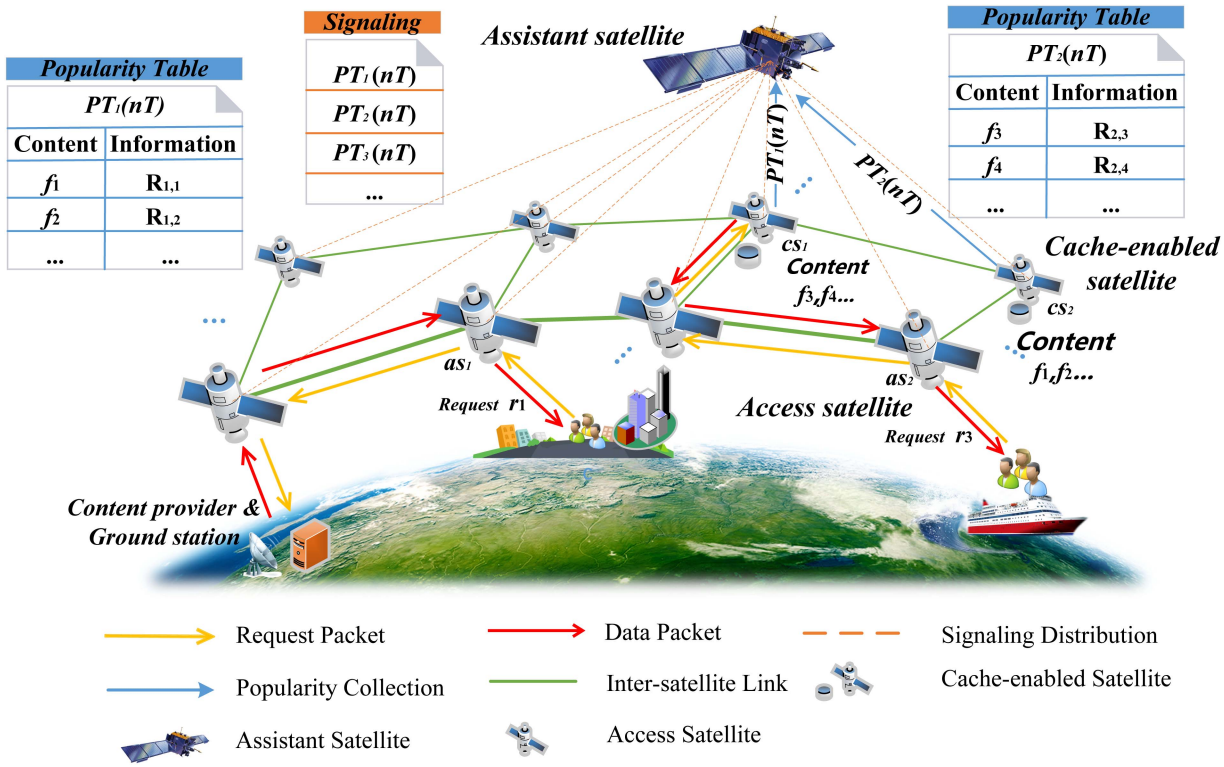


Fig. 1. Network architecture of our scheme.

it can improve the efficiency of content transmission. On this basis, as the scale of the satellite network increases, in-network caching can bring new vitality to inter-satellite networking and routing. For example, the existing studies further adopt a new network architecture, combining the ICN [22] architecture with the STIN. The ICN architectures are characterized by two major features, i.e., routing-by-name and in-network caching, to offload the redundant traffic from users' requests for popular content and thus significantly improve the efficiency of content transmission in satellite networks. For example, Galluccio et al. [23] and Tomaso et al. [24] first introduced ICN into satellite networks and designed the specific protocol architecture to achieve content-aware function and the extensive simulation results also showed effectiveness the ICN in satellite networks. Based on that, Li et al. [25] further proposed a novel architecture that combines software-defined networking (SDN) and ICN to provide flexible management and efficient content retrieval for the STIN. The MEO and GEO satellites were considered the controllers to globally control LEO satellites' caching strategy and content delivery process. After that, Yang et al. [26] focused on the specific content retrieval process in ICN-based satellite networks and devised a reliable and efficient content retrieval scheme through a coding-enabled multicast model to provide an interference-tolerant, low-latency, and efficient content delivery service. Tang et al. [1] proposed a content-aware routing scheme for satellite networks based on ICN architecture, using a cached content prediction model to route users' requests. Besides, Xu et al. [27] proposed a hybrid caching strategy for ICN-combined LEO satellite networks, considering node classification and popular content awareness, and dynamically divided satellite nodes into core nodes and edge nodes to improve content distribution efficiency and overcome satellite node mobility and dynamic topology brought challenges.

Most of the aforementioned studies have predominantly focused on the optimization of caching and transmission strategies, yet it still lacks the design of content routing in conjunction with satellites' cache. Therefore, in this paper, we present an opportunistic content-aware routing approach. By predicting satellites' cache, our routing scheme can route user requests to the optimal node, thereby reducing content retrieval delay and minimizing bandwidth consumption.

III. System Model

A. Network Model

As shown in Fig. 1, we consider a cache-enabled STIN consisting of terrestrial and satellite networks. The terrestrial network is composed of the content providers' server and its ground station, which is denoted as gs. The satellite network comprises LEO satellites and MEO/GEO satellites. LEO satellites are denoted as $\mathcal{L}$. The LEO satellites that provide Internet access services for ground users are called access satellites, denoted as $\mathcal{L}_{as} = \{as_1, as_2, \ldots as_N\}$. In addition, LEO satellites with cache capabilities are cache-enabled satellites, denoted as $\mathcal{L}_{cs} = \{cs_1, cs_2, \ldots, cs_{N_c}\}$. Note that all LEO satellites can provide Internet access for users, while constrained by high deployment costs, only part of LEO satellites are cached-enabled, i.e., $\mathcal{L}_{as} = \mathcal{L}, \mathcal{L}_{cs} \subseteq \mathcal{L}$. Furthermore, MEO/GEO satellites are

the assistant satellites in the network, responsible for periodically collecting global cache information and distributing it to all access satellites.

B. Popularity Table Model

The set of contents files requested by users is denoted as $\mathcal{M} = \{f_1, f_2, \ldots, f_M\}$. Without loss of generality, we assume that these contents have the same size, denoted as f bytes. The cache-enabled satellite has certain cache storage $c \cdot f$ bytes, i.e., one cache-enabled satellite can cache c pieces of contents at most. To record each cache-enabled satellite's cache status, we use the popularity table to describe the cache information of cache-enabled satellites. The structure of the cache-enabled satellite popularity table is shown in Fig. 1. Popularity table $PT_i(t)$ records the requested contents' information on the cache-enabled satellite cs_i until time t. For a specific content f_m, the requested information on cs_i is recorded as $R_{i,m}$, which includes the last accessed time $\tau_{i,m}$, the requested content local popularity $p_{i,m}$ and request average arrival rate $\lambda_{i,m}$. The last accessed time $\tau_{i,m}$ can be obtained from each request directly. Besides, the content's local popularity $p_{i,m} \in [0, 1]$, and can be calculated by

$$p_{i,m} = \frac{N_{i,m}}{N_i}, \tag{1}$$

where $N_{i,m}$ and N_i represent request times for content f_m and the total request times on cache-enabled satellite cs_i, respectively. We assume that the arrival process of requests on cache-enabled satellite cs_i obeys the Poisson process with arrival rate λ_i, and each content request arrival process obeys the Poisson process [30] with arrival rate $\lambda_{i,m}$, which satisfies

$$\lambda_{i,m} = \lambda_i \cdot p_{i,m}. \tag{2}$$

When cache-enabled satellites receive requests from other satellites or users, they should update the request information of the requested contents in their popularity table.

In addition, assistant satellites collect the popularity tables of all cache-enabled satellites in the entire network every T time and distribute them to all access satellites. The popularity tables of all cache-enabled satellites are recorded as the signaling, which is denoted as $PTs(nT) \triangleq \{PT_1(nT), PT_2(nT), \ldots, PT_{N_c}(nT)\}$.

C. Request Model

Users can directly send a request r_m for content f_m to the access satellite serving it. One request contains the content provider and its ground station's address as well as the requested content name so that access satellites can retrieve the correct content. We assume that each user can only connect to one access satellite. When the access satellite receives a request for content f_m from a ground user, it can forward it to the content provider's ground station, or forward it to a nearby cache-enabled satellite to retrieve the required content. As shown in Fig. 1, after the access satellite as_1 receives the user's request f_1, it forwards the request to the ground station to obtain the content from the content provider. On the contrary, access satellite as_2 forwards the request for f_3 to the nearby cache-enabled satellite cs_1 instead of forwarding the request to the ground station, thereby reducing content retrieval delay and saving traffic overhead.

Note that when the access satellite forwards the request to a cache-enabled satellite, the content is retrieved in an opportunistic manner. The opportunistic situation means the requested content may not be retrieved successfully due to the cache-enabled satellite does not cache the requested content. The reason for this phenomenon is that the access satellite can only obtain popularity tables of the cache-enabled satellite periodically and cannot know the real-time cache status of the cache-enabled satellite. Therefore, it is necessary and pivotal to correctly weigh the risks and benefits of going to different cache-enabled satellites to obtain the requested content. We describe in Section V how we comprehensively consider the risks and benefits of forwarding requests to cache-enabled satellites.

IV. Cache Prediction

In this section, we first introduce a local caching strategy for cache-enabled satellites in the STIN. Second, because the synchronization of cached information on the entire network is difficult to achieve in real time, especially for large-scale networks, we propose an online prediction method. The prediction method enables access satellites to deduce the probability of content being cached on cache-enabled satellites based on the historical popularity tables. Obtaining the probability that the content is cached by different cache-enabled satellites is a necessary decision basis for executing opportunistic content-aware routing.

A. Local Caching Strategy

The local caching strategy refers to the cache decision made locally by the cache-enabled nodes according to the arrival of requests. Common local caching strategies include least recently used (LRU), first in first out (FIFO), least frequently used (LFU), most popular caching (MPC), and random caching (RC) [31]. Compared with other caching strategies, the LRU caching strategy is simpler and easier to implement, and due to it being based on the least recently used criterion, the cached contents are often popular currently. Therefore, we assume the cache-enabled satellite replaces and deletes content according to the LRU cache policy. The main idea of the LRU caching strategy is to remove the content that has not been requested for the longest time when the cache is full and make space for caching the latest content. When the cache-enabled satellite is not full, it will cache any requested content.

Since the LRU strategy replaces the content that has not been requested for the longest time when the cache is full, we can conclude that in the cache of the cache-enabled satellite deploying LRU strategy, the cached c pieces of contents are the recently requested c contents. Let $rank_{i,m}(t) \in \{1, 2, \ldots, |\mathcal{M}|\}$ denote the rank of f_m's last accessed time in popularity table $PT_{i,m}(t)$ of cache-enabled satellite cs_i at time t. According to LRU caching strategy, if the content f_m is the latest requested content

at cache-enabled satellite cs_i, it means that the content f_m has the largest last accessed time $\tau_{i,m}$ and the rank $rank_{i,m}(t) = 1$. Further, we can conclude that if content f_m is cached on the cache-enabled satellite cs_i, its accessed time rank at time t should satisfy

$$rank_{i,m}(t) \leq c. \tag{3}$$

In other words, the top c pieces of contents with the largest last accessed time in the popularity table $PT_i(t)$ are the contents cached by cache-enabled satellite cs_i at time t. According to this conclusion and periodically distributed popularity tables of cache-enabled satellites, we can predict the cache of cache-enabled satellites.

B. Cache Prediction

Considering that the real-time interaction of cached information on each cache-enabled satellite is a huge overhead, especially for large-scale satellite networks, a more practical approach is to introduce additional nodes for periodic collection and update. Therefore, we consider using satellites with wider coverage as assistant satellites and use interlayer links to periodically collect and distribute the popularity tables of cache-enabled satellites. Since this operation is not performed in real time, it is difficult for each access satellite to obtain the exact cache content of other cache-enabled satellites. For example, at time $t, nT \leq t < (n+1)T$, the popularity table of the cache-enabled satellite cs_i owned by access satellites is the popularity table $PT_i(nT)$ distributed by assistant satellites at time nT, not the real-time popularity table $PT_i(t)$ of the cache-enabled satellite. Consequently, the access satellite cannot know the real-time cache status of cache-enabled satellites. Thus, we design a cached content prediction method. In specific, for each access satellite as_n, the proposed method can predict the cached contents on other cache-enabled satellites, which refers to the probability that the content f_m has been cached on the cache-enabled satellite cs_i at time t.

On the cache-enabled satellite cs_i, the content set in its popularity table is recorded as $\mathcal{M}_i(t)$, which can be divided into two parts. One is the set $\mathcal{M}_i^{m-}(t)$, whose last accessed time rank is smaller than the content $f_{i,m}$, i.e., $rank_{i,m'}(t) < rank_{i,m}(t), f_{m'} \in \mathcal{M}_i^{m-}(t)$. Another part is set $\mathcal{M}_i^{m+}(t)$ whose last accessed time rank is not smaller than the content $f_{i,m}$, i.e., $rank_{i,m'}(t) \geq rank_{i,m}(t), f_{m'} \in \mathcal{M}_i^{m+}(t)$. The content set in cache-enabled satellite's popularity table satisfies $\mathcal{M}_i(t) = \mathcal{M}_i^{m-}(t) + \mathcal{M}_i^{m+}(t)$.

According to the definition before, the time for assistant satellites to periodically collect and distribute the signaling is kT, where $k = \{1, 2, \ldots, n, \cdots\}$ represents the number of periods. At the time $t, t = nT + \tau, (0 \leq \tau < T)$, after assistant satellites send the global cache information for the n-th time, let $P_{i,m}^{in}(t) \in [0,1]$ denote the probability of content f_m being cached on the cache-enabled satellite cs_i, and $P_{i,m}^{out}(t) \in [0,1]$ denote the probability of not being cached.

We consider the worst case for caching, i.e., the cache-enabled satellite cs_i does not receive any requests for content f_m after time nT and also the number of requests for different contents in the set $\mathcal{M}_i^{m-}(nT)$ are more than $c - |\mathcal{M}_i^{m+}(nT)|$ within (nT, t) time. In this situation, the last accessed time rank of content f_m in time t satisfies

$$\begin{aligned} rank_{i,m}(t) &> rank_{i,m}(nT) + c - \left|\mathcal{M}_i^{m+}(nT)\right| \\ &= \left|\mathcal{M}_i^{m+}(nT)\right| + c - \left|\mathcal{M}_i^{m+}(nT)\right| \\ &= c, \end{aligned} \tag{4}$$

which means the cached content f_m at time nT will be discarded and will not be cached by cache-enabled satellite cs_i at time t.

In the worst case, the probability that content f_m does not receive any request within (nT, t) is expressed as $P_{i,m}^0(t)$ where $P_{i,m}^0(t) \in [0,1]$. According to the previous definition, the arrival process of requests for content f_m on the cache-enabled satellite cs_i follows the Poisson process with parameters $\lambda_{i,m}$, hence

$$\begin{aligned} P_{i,m}^0(t) &= \Pr(N_{i,m}(nT, t) = 0) \\ &= \Pr(N_{i,m}(nT, nT + \tau) = 0) \\ &= e^{-\lambda_{i,m}\tau}. \end{aligned} \tag{5}$$

On the other hand, the probability that the number of requested content items in $\mathcal{M}_i^{m-}(nT)$ are greater than $c - |\mathcal{M}_i^{m+}(nT)|$, is expressed as $P_{i,m}^1(t)$. That is

$$P_{i,m}^1(t) = \Pr\left(K_{i,m}^1(t) \geq c - \left|\mathcal{M}_i^{m+}(nT)\right|\right), \tag{6}$$

where $K_{i,m}^1(t) = \sum_{m' \in \mathcal{M}_i^{m-}(nT)} \mathbb{I}_{(N_{i,m'}(nT,t) \geq 1)}$ and variable $\mathbb{I}_{(N_{i,m'}(nT,t) \geq 1)}$ satisfies

$$\mathbb{I}_{(N_{i,m'}(nT,t) \geq 1)} = \begin{cases} 1, & \text{when } N_{i,m'}(nT, t) \geq 1 \\ 0, & \text{when } N_{i,m'}(nT, t) = 0. \end{cases} \tag{7}$$

From the properties of the above functions, we can deduce that $\mathbb{I}_{(N_{i,m'}(nT,t) \geq 1)} + \mathbb{I}_{(N_{i,m'}(nT,t) = 0)} = 1$, then we can further represent $K_{i,m}^1(t)$,

$$\begin{aligned} K_{i,m}^1(t) &= \sum_{m' \in \mathcal{M}_i^{m-}(nT)} \mathbb{I}_{\left(N_{i,m'}(nT,t) \geq 1\right)} \\ &= \sum_{m' \in \mathcal{M}_i^{m-}(nT)} \left(1 - \mathbb{I}_{\left(N_{i,m'}(nT,t) = 0\right)}\right) \\ &= \left|\mathcal{M}_i^{m-}(nT)\right| - \sum_{\substack{m' \in \\ \mathcal{M}_i^{m-}(nT)}} \mathbb{I}_{\left(N_{i,m'}(nT,t) = 0\right)}. \end{aligned} \tag{8}$$

Bring the result of the (8) into the (6), we can obtain that

$$\begin{aligned} P_{i,m}^1(t) &= \Pr\left(K_{i,m}^1(t) \geq c - \left|\mathcal{M}_i^{m+}(nT)\right|\right) \\ &= \Pr\left(\sum_{m' \in \mathcal{M}_i^{m-}(nT)} \mathbb{I}_{\left(N_{i,m'}(nT,t) = 0\right)} \leq |\mathcal{M}_i(nT)| - c\right) \end{aligned} \tag{9}$$

Denote $X_{i,m}(nT, t) = \mathbb{I}_{(N_{i,m'}(nT,t) = 0)}$, which satisfies

$$X_{i,m'}(nT, t) = \begin{cases} 1, & \text{when } N_{i,m'}(nT, t) = 0 \\ 0, & \text{when } N_{i,m'}(nT, t) \geq 1. \end{cases} \tag{10}$$

Besides, the probability in (9) can be replaced by

$$P_{i,m}^{1}(t) = \Pr\left(K_{i,m}^{1}(t) \geq c - \left|\mathcal{M}_i^{m+}(nT)\right|\right)$$
$$= \Pr\left(\sum_{m' \in \mathcal{M}_i^{m-}(nT)} X_{i,m'}(nT,t) \leq |\mathcal{M}_i(nT)| - c\right). \quad (11)$$

For each item $X_{i,m'}(nT,t)$, if the content $f_{m'}$ is not requested once within time (nT,t), the value of $X_{i,m'}(nT,t)$ is 1, otherwise is 0. Then, according to the Poisson process property, we can deduce that

$$\Pr\left(N_{i,m'}(nT,t) = 0\right) = e^{-\lambda_{i,m'}\tau}, \quad (12)$$

and

$$\Pr\left(N_{i,m'}(nT,t) \geq 1\right) = 1 - e^{-\lambda_{i,m'}\tau}. \quad (13)$$

Therefore, variable $X_{i,m'}(nT,t)$ obeys the binomial distribution with mean value $\mu_{i,m'} = e^{-\lambda_{i,m'}\tau}$ and variance $\sigma_{i,m'}^2 = e^{-\lambda_{i,m'}\tau}(1 - e^{-\lambda_{i,m'}\tau})$, i.e., $X_{i,m'}(nT,t) \sim B(1, e^{-\lambda_{i,m'}\tau})$.

We then analyze that the sequence sum of random variables in (11), $\sum_{m' \in \mathcal{M}_i^{m-}(nT)} X_{i,m'}(nT,t)$, obeys Gaussian distribution through Theorem 1 and its proof, and then derive the probability $P_{i,m}^{1}(t)$.

Theorem 1: The sum of random variables

$$\Phi = \sum_{m' \in \mathcal{M}_i^{m-}(nT)} X_{i,m'}(nT,t), \quad (14)$$

follows a Gaussian distribution, and it's meanvalue is $\mu = \sum_{m' \in \mathcal{M}_i^{m-}(nT)} e^{-\lambda_{i,m'}\tau}$ and variance is $\sigma^2 = \sum_{m' \in \mathcal{M}_i^{m-}(nT)} e^{-\lambda_{i,m'}\tau}(1 - e^{-\lambda_{i,m'}\tau})$.

Proof. Please see Appendix A, available online, for details□

After proving that the sequence sum of random variables $\sum_{m' \in \mathcal{M}_i^{m-}(nT)} X_{i,m'}(nT,t)$ follows Gaussian distribution, we can further solve for the probability in (9) as

$$P_{i,m}^{1}(t) = \Pr\left(\Phi \leq |\mathcal{M}_i(nT)| - c\right)$$
$$= \Pr\left(\frac{\Phi - \mu}{\sqrt{\sigma}} \leq \frac{|\mathcal{M}_i(nT)| - c - \mu}{\sqrt{\sigma}}\right)$$
$$= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\left(|\mathcal{M}_i(nT)| - c - \mu\right)/\sqrt{\sigma}}{\sqrt{2}}\right)\right), \quad (15)$$

where $\mathrm{erf}(x)$ is the Gaussian error function, the parameter $\mu = \sum_{m' \in \mathcal{M}_i^{m-}(nT)} e^{-\lambda_{i,m'}\tau}$ and the parameter $\sigma^2 = \sum_{m' \in \mathcal{M}_i^{m-}(nT)} e^{-\lambda_{i,m'}\tau}(1 - e^{-\lambda_{i,m'}\tau})$.

Considering that we deduce the above process in the worst case, the probability that a certain cached content f_m in cache-enabled satellite cs_i at time nT will no longer be cached satisfies

$$P_{i,m}^{out}(t) \leq P_{i,m}^{0}(t) \cdot P_{i,m}^{1}(t)$$
$$= e^{-\lambda_{i,m}\tau} \cdot \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\left(|\mathcal{M}_i(nT)| - c - \mu\right)/\sqrt{\sigma}}{\sqrt{2}}\right)\right). \quad (16)$$

Therefore, the probability of cache-enabled satellite cs_i cache f_m satisfies

$$P_{i,m}^{in}(t) = 1 - P_{i,m}^{out}(t) \geq P_{i,m}^{0}(t) \cdot P_{i,m}^{1}(t)$$
$$= e^{-\lambda_{i,m}\tau} \cdot \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\left(|\mathcal{M}_i(nT)| - c - \mu\right)/\sqrt{\sigma}}{\sqrt{2}}\right)\right). \quad (17)$$

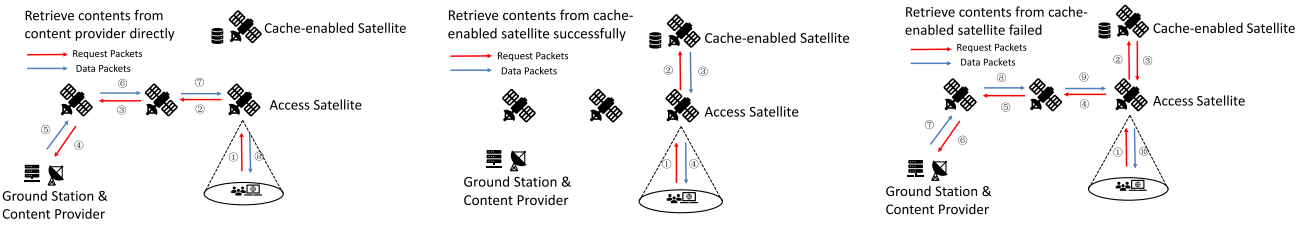
In order to conservatively estimate the probability, we might as well consider the probability of cache-enabled satellite cs_i still cache content f_m at time t is equal to $e^{-\lambda_{i,m}\tau} \cdot \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{(|\mathcal{M}_i(nT)| - c - \mu)/\sqrt{\sigma}}{\sqrt{2}}\right)\right)$ if $rank_{i,m} \leq c$. Besides, if a content is not cached on cs_i at time nT, we still think it will not be cached on cs_i at time t. Therefore, if $rank_{i,m}(nT) > c$, we have $P_{i,m}^{in}(t) = 0$.

In the analysis of this subsection, we propose a method that can predict the probability of content being cached on the cache-enabled satellite based on the regularly distributed global cache information $PT_s(nT)$. This approach avoids the huge overhead of real-time signaling interactions while ensuring efficient predictions of the probability of contents being cached. The probability that content is cached in different cache-enabled satellites significantly determines the risks and benefits of obtaining content from that cache-enabled satellite, which will have a non-negligible impact on our routing selection. Therefore, the accuracy of cache prediction will directly affect the overall performance of our proposed scheme: the more accurate the prediction, the more pronounced the performance improvement will be.

Nonetheless, we also recognize that there are factors that affect forecast accuracy. For example, a longer signaling distribution period, a smaller cache space, etc., will affect our prediction results and further affect the overall solution. We will elaborate on the impact of these factors in the simulation part in Section VI.

V. Opportunistic Content-Aware Routing

In the last section, we obtain the probability that cache-enabled satellites cache the content required by the user, which can also be considered as the probability of successfully retrieving the content from the cache-enabled satellite. In this section, based on the cache prediction method, we propose an opportunistic content-aware routing scheme. Opportunistic content-aware routing allows access satellites to retrieve the content from cache-enabled satellites with a certain probability, thereby reducing the content retrieval delay. First, we outline the overview of our routing scheme, which is divided into three routing ways. Then, we present an evaluation of the potential gains, which comprehensively consider both risks and benefits, when retrieving contents from different caching satellites with a certain probability. Based on this assessment of potential gains, we select the optimal node for each request. Finally, we analyze the workflow and the complexity of our routing algorithm.



(a) Directly retrieve the content from content provider.

(b) Successfully retrieve the content from the cache-enabled satellite.

(c) Failed retrieve the content from the cache-enabled satellite.

Fig. 2. Route the request and retrieve the content in different ways.

A. Overview of Routing

When an access satellite receives a user's request for a specific content, it performs opportunistic content-aware routing to select the most suitable destination node for this request. As depicted in Fig. 2(a), if the ground station is chosen as the destination node, the desired content can certainly be obtained. However, due to longer paths and limitations imposed by the terrestrial network, there is often a higher delay and greater traffic consumption. On the other hand, when attempting to retrieve the required content from a cache-enabled satellite, due to the inability to accurately know the real-time caching status of the satellite, there is only a certain probability of successfully retrieving the content. Nevertheless, compared to forwarding the request to the ground station, the path is typically shorter, and the delay is lower. For example, as shown in Fig. 2(b), once the content is successfully retrieved from the cache-enabled satellite, it significantly reduces the content retrieval delay and traffic consumption. However, if the request is mistakenly forwarded to a caching satellite that does not have the requested content cached, as shown in Fig. 2(c), rerouting becomes necessary.

Therefore, selecting the most suitable destination node, i.e., the optimal node, for each request becomes the core of our opportunistic content-aware routing algorithm. We need to assess the risks and rewards of retrieving contents from different cache-enabled satellites and then select the optimal node for each request.

B. Potential Gains Model

We utilize a potential gain model to assess the overall risk and reward of retrieving content from different cache-enabled satellites. The potential gains of the access satellite to retrieve the user-required content f_m from cache-enabled satellite cs_i at time t as $Gain_{i,m}(t)$ is denoted as

$$Gain_{i,m}(t) = P^{in}_{i,m}(t)S_{i,m}(t) - P^{out}_{i,m}(t)W_{i,m}(t). \quad (18)$$

Among them, $P^{in}_{i,m}(t)$ represents the probability that the content f_m is in the cache of cs_i at time t, which also represents the probability that the content f_m can be successfully retrieved on cs_i. While $P^{out}_{i,m}(t)$ means that the probability that content is not cached by cache-enabled satellite cs_i. In this situation, we cannot retrieve content f_m from cs_i. Besides, $S_{i,m}(t)$ represents the delay saved by successfully retrieving the content from the cache-enabled satellite cs_i without retrieving the content from the remote content provider. Similarly, suppose the content is not cached on the cache-enabled satellite cs_i and cannot be retrieved from it. In that case, the request must be forwarded back to the access satellite and then rerouted by the access satellite to the content provider's ground station. The wasted delay in this process is denoted as $W_{i,m}(t)$. Generally, the delay in the process of retrieving contents is mainly composed of the following three parts of delay, including propagation delay, transmission delay, and queuing delay. Note that our potential gain model can measure the gains from multiple dimensions. However, since we are primarily concerned with content retrieval delay, we only take the parameter of delay into consideration in our calculation. Specifically, we introduce how to calculate the saved delay $S_{i,m}(t)$ and wasted delay $W_{i,m}(t)$ in the following.

On one hand, in the STIN, the distance between satellites is usually thousands of kilometers, and the distance between multi-hop satellites is greater than this number, contributing to a nonnegligible propagation delay. Denote the propagation delay between the access satellite and the cache-enabled satellite cs_i as $\tau^{Prop}_i = \frac{d_i}{c_{radio}}$, and the propagation delay between the access satellite and the content provider's ground station as $\tau^{Prop}_g = \frac{d_g}{c_{radio}}$, where c_{radio} is the propagation rate of radio.

On the other hand, since the size of request packets and data packets are different, which contributes to different transmission delays in the forwarding request process and returning content process. On the other hand, there is a difference between the satellite's transmission rate and the ground station's transmission rate. When calculating the transmission delay, the factors mentioned above need to be considered. Denote the transmission rate of satellites as T^s and the transmission rate of ground stations as T^g. Denote the size of the request packet as s^r and the size of the data packet containing the requested content as s^d. Hence, on one satellite, the transmission delay of the request packet is $\tau^{Trans}_{r,s} = \frac{s^r}{T^s}$, the transmission delay of the data packet is $\tau^{Trans}_{d,s} = \frac{s^d}{T^s}$. Similarly, the transmission delay of the request packet in the ground station is $\tau^{Trans}_{r,g} = \frac{s^r}{T^g}$, and the transmission delay of the data packet is $\tau^{Trans}_{d,g} = \frac{s^d}{T^g}$.

Finally, we also consider the queuing delay on the node. Since the real-time queuing situation of each node cannot be obtained at a low cost, we use τ^{Queue}_s to represent the average queuing delay on a satellite and use τ^{Queue}_g to represent the average queuing delay on the ground station.

Note that the delay experienced by packets between ground stations and content provider servers is not considered in this paper since we mainly focus on the routing problem in STINs. Besides, the information related to delay calculation mentioned above, such as the path and number of hops, can be obtained based on the routing tables.

The distance from the access satellite to the cache-enabled satellite cs_i is denoted as d_i, and the number of hops experienced to cs_i is h_i. Similarly, the distance from the access satellite to the ground station is d_g, and the number of hops to gs is h_g, where the first $h_g - 1$ hops happen between satellites, the last hop happens between the satellite and the ground station. As shown in Fig. 2(b), when the access satellite successfully retrieves the content f_m from the cache-enabled satellite cs_i, instead of retrieving the content from gs, which is described in Fig. 2(a). The saved delay in this situation is:

$$S_{i,m}(t) = S_{i,m}^{Prop}(t) + S_{i,m}^{Trans}(t) + S_{i,m}^{Queue}(t). \tag{19}$$

Among them, the saved propagation delay $S_{i,m}^{Prop}(t)$ is represented as

$$S_{i,m}^{Prop}(t) = \tau_g^{Prop} - \tau_i^{Prop}. \tag{20}$$

The saved transmission delay can be represented as

$$\begin{aligned} S_{i,m}^{Trans}(t) = {} & (h_g - 1)\left(\tau_{r,s}^{Trans} + \tau_{d,s}^{Trans}\right) + \tau_{r,g}^{Trans} \\ & + \tau_{d,g}^{Trans} - h_i\left(\tau_{r,s}^{Trans} + \tau_{d,s}^{Trans}\right), \end{aligned} \tag{21}$$

which includes the saved transmission delay in both forwarding the request packet process and returning the data packet process. The saved queuing delay is represented as

$$S_{i,m}^{Queue}(t) = 2(h_g - 1)\tau_s^{Queue} + 2\tau_g^{Queue} - 2h_i\tau_s^{Queue}. \tag{22}$$

Further, the saved delay in (19) can be further calculate as

$$\begin{aligned} S_{i,m}(t) = {} & \tau_g^{Prop} + (h_g - 1)(\tau_{r,s}^{Trans} + \tau_{d,s}^{Trans} + 2\tau_s^{Queue}) \\ & + \tau_{r,g}^{Trans} + \tau_{d,g}^{Trans} + 2\tau_g^{Queue} - \tau_i^{Prop} \\ & - h_i(\tau_{r,s}^{Trans} + \tau_{d,s}^{Trans} + \tau_s^{Queue}) - 2h_i\tau_s^{Queue}. \end{aligned} \tag{23}$$

Suppose that the access satellite forwards the request to the cache-enabled satellite cs_i, but at time t, the cache-enabled satellite cs_i does not cache content f_m, then the cache-enabled satellite will forward the request packet back to the access satellite. After that, the access satellite forwards the request to the content provider's ground station to retrieve the content. The above process is shown in Fig. 2(c), and the time wasted in this process is

$$W_{i,m}(t) = W_{i,m}^{Prop}(t) + W_{i,m}^{Trans}(t) + W_{i,m}^{Queue}(t). \tag{24}$$

Similar to computing the saved time, the wasted time in this process can be further expressed as

$$\begin{aligned} W_{i,m}(t) = {} & \tau_i^{Prop} + h_i(\tau_{r,s}^{Trans} + \tau_{d,s}^{Trans} + \tau_s^{Queue}) \\ & + 2h_i\tau_s^{Queue}. \end{aligned} \tag{25}$$

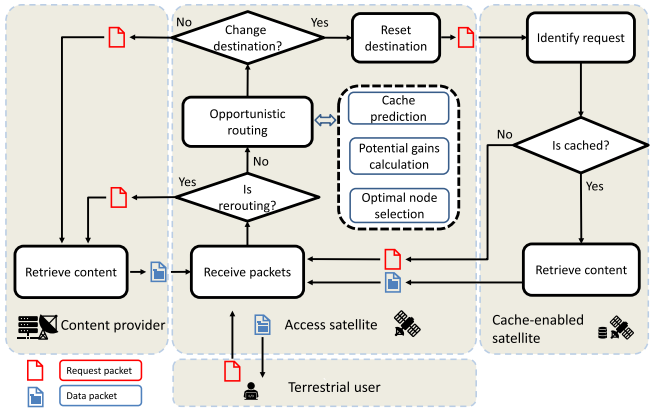


Fig. 3. Opportunistic content-aware routing workflow.

C. Optimal Node Selection

When the access satellite receives request r_m from users at time t, it calculates the potential gains of going to different cache-enabled satellites to retrieve content f_m according to the signaling distributed by assistant satellites at time nT. Denote cache-enabled satellite with largest potential gains as N_m^{cs}, which can be calculate as follows

$$N_m^{cs} = \arg\max\left(Gain_{i,m}(t)\right), \forall cs_i \in \mathcal{L}_{cs}, \tag{26}$$

and its potential gains is denoted as $Gain_m^{cs}$. Then we compare the potential gains $Gain_m^{cs}$ with 0. If it is bigger than 0, we select N_m^{cs} as the optimal node for request r_m because it can reduce content retrieval delay compared to gs. Otherwise, we still select gs as the optimal node for request r_m. Therefore, the optimal node for request r_m is

$$N_m^{optimal} = \begin{cases} N_m^{cs}, & \text{if } Gain_m^{cs} > 0 \\ gs, & \text{if } Gain_m^{cs} \le 0. \end{cases} \tag{27}$$

D. Workflow of Routing

The workflow for our proposed scheme can be described as shown in Fig. 3. When a user's request reaches the access satellite, the access satellite attempts to execute opportunistic routing decisions. This process mainly consists of the following steps: predicting the probability of different cache-enabled satellites caching the requested content, calculating the potential gain for different cache-enabled satellites, and selecting the optimal node with the largest potential gain. If the optimal node is still the ground station operated by the content provider, there is no need to modify the destination address of the request. If the optimal node is a cache-enabled satellite, the destination node of the request packet is modified to that cache-enabled satellite, and then the request is forwarded. It is important to note that opportunistic content-aware routing attempts are only made at the access satellite, and no further attempts are made after the first unsuccessful attempt.

When the request packet reaches the destination node, if the destination node is the content provider's ground station,

Algorithm 1: Opportunistic Content-Aware Routing.

Input: request r_m, ground station gs, current time $t = nT + \tau$, last popularity tables $PTs(nT)$;
Output: the path to the optimal node for retrieving content f_m: $P(N_m^{optimal})$;

1 ***Step1 Cache prediction:***
2 Initialize a set $\mathcal{L}_{cs,m} = \emptyset$;
3 **for** *cache-enabled satellite* $cs_i \in \mathcal{L}_{cs}$ **do**
4 Calculate the probability $P_{i,m}^{in}(t)$ according to (17);
5 **if** $P_{i,m}^{in}(t) > 0$ **then**
6 $\mathcal{L}_{cs,m} \leftarrow \mathcal{L}_{cs,m} + cs_i$
7 **end**
8 **end**
9 ***Step2 Potential gains calculation:***
10 Initialize $N_m^{optimal}$=gs, $Gain_m^{cs}$=0;
11 **for** *each cache-enabled satellite* $cs_i \in \mathcal{L}_{cs,m}$ **do**
12 Calculate the potential gains $Gain_{i,m}(t)$ according to (18);
13 **if** $Gain_{i,m}(t) > Gain_m^{cs}$ **then**
14 $N_m^{optimal}$=cs_i, $Gain_m^{cs}$=$Gain_{i,m}(t)$;
15 **end**
16 **end**
17 ***Step3 Optimal node selection:***
18 Select the optimal node $N_m^{optimal}$ for request r_m according to (26) (27);
19 ***Step4 Find path to the optimal node:***
20 Change the destination as $N_m^{optimal}$ for request r_m;
21 Find path $P(N_m^{optimal})$ according to routing tables;
22 **return** $P(N_m^{optimal})$

which can definitely be retrieved the content. It will fill the requested content into the data packet, then forward it to the access satellite. If the destination node is a cache-enabled satellite, the cache-enabled satellite will first identify the requested content and try to retrieve the content from its cache. Suppose that the content is cached on the cache-enabled satellite. The cache-enabled satellite will fill the data packet with the requested content and forward it back to the access satellite. If not cached, it will return the request packet to the access satellite. When the packet returns to the access satellite, it will update the probability according to the packet type to know the latest status of the cache-enabled satellite. If the packet is a request packet which means the cache-enabled satellite does not cache the requested content, the access satellite will update the probability that the cache-enabled satellite caches the content as 0 and then reroute the request to the ground station. However, if the packet is a data packet, the access satellite will update the probability as 1. After the access satellite receives the data packet, it will deliver the retrieved content to users.

The opportunistic content-aware routing is shown as Algorithm 1. Step 1, including lines 1 to 8, mainly involves predicting the probability of the requested content f_m being cached by different cache-enabled satellites at the current time. We predict the probabilities based on the global cache information $PTs(nT)$ from the previous period. Step 2, including lines 10 to 16, is to calculate the potential gains of retrieving content f_m from different cache-enabled satellites. The greater the potential gain in retrieving content f_m from a certain cache-enabled satellite, the greater the probability of successfully retrieving the desired content from that cache-enabled satellite with lower latency. Step 3, including lines 18 to 22, is involved in selecting the optimal node with the largest potential gains and modifying the destination address of the request packet to the selected optimal node. Step 4, lines 20 to 22, is to find the next hop that forwards request r_m to the optimal node according to the routing table.

E. Complexity Analysis

In the cache prediction method, i.e., step 1 of Algorithm 1, denote the cache-enabled satellites that cached content f_m at time nT as $\mathcal{L}_{cs,m}$. For each cache-enabled satellite cs_i in $\mathcal{L}_{cs,m}$, it needs $|\mathcal{M}_i^{m-}(nT)|$ additions operations to calculate the probability $P_{i,m}^{in}(t)$ that come from the process of calculating the parameters μ and σ^2 in (17). Therefore, the complexity of cache prediction method, is $\mathcal{O}(|\mathcal{L}_{cs}| \cdot |\mathcal{M}|)$. Step 2 of Algorithm 1 calculates the potential gains of going to different cache-enabled satellites belonging to $\mathcal{L}_{cs,m}$ to retrieve content f_m. Thus, the computational complexity of step 2 is $\mathcal{O}(|\mathcal{L}_{cs}|)$. For step 3, the complexity is $\mathcal{O}(1)$. For step 4, the complexity is also $\mathcal{O}(1)$ when there is a routing table. Therefore, the complexity of Algorithm 1 is $\mathcal{O}(|\mathcal{L}_{cs}| \cdot |\mathcal{M}|)$.

For each request, executing opportunistic content-aware routing requires knowing the probability of successfully retrieving the requested content on each cache-enabled satellite, i.e., requiring cache prediction. Based on the probability obtained by the cache prediction method, the potential gains of retrieving content from different cache-enabled satellites can be calculated, and further, the optimal node, that is, the cache-enabled satellite or the ground station with the largest potential gains, can be calculated. Considering that for each request, the cache prediction process and the optimal node selection process are required only once at the access satellite, and no additional processing is required during the forwarding process. Hence, the complexity $\mathcal{O}(|\mathcal{L}_{cs}| \cdot |\mathcal{M}|)$ is acceptable for a complete content retrieval process. To better investigate the overhead issues in practice, we provide the specific computational overhead of the proposed routing scheme in Section VI.

It should be noted that the proposed routing scheme is compatible with the traditional routing schemes without content-aware ability. Inherently, Algorithm 1 provides an extra content-aware routing computation process independent of the traditional routing process. From step 1 to step 3, for each content request, Algorithm 1 calculates its potential gains by exploring both risk and reward and then finding the optimal destination node to retrieve the content. In step 4, the routing table that provides the forwarding path can be updated by arbitrary classical routing algorithms, e.g., the Dijkstra algorithm in OSPF [32], [33] and the Bellman-Ford algorithm in RIP [34]. Thus, if steps 1-3 are not executed, the designed routing scheme can be easily degraded to the traditional routing process for non-content data transmission, which is friendly to practical implementation in STIN systems.

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Constellation Type	WalkerStar
Height of Orbits	1050km
Orbit Angle	90°
Number of LEO	100
Number of Planes	10
Transmission Rate of Satellites	50Mbps
Transmission Rate of Groud Station	200Mbps
Size of Request	5KB
Size of Content	0.1MB
Size of Popularity Table	0.1MB
Request Rate	$0.2s^{-1}$
Caching Strategy	LRU
Number of Request	100000

VI. PERFORMANCE EVALUATION

We adopt an open-source large-scale satellite network simulator (LSNS) [35] as our simulation environment. We take a real-world dataset based on MovieLens [36], which has 100,000 requests for 1682 movies. In the experiments, we randomly select access satellites for each request, i.e., the request's locations are randomly distributed. At the same time, in order to ensure the validity of the dataset, we generate requests according to the actual request order of the dataset and let the terrestrial users send the requests to the access satellites. All experiments were performed on a PC with four 3.6 GHz CPUs, 16 GB RAM, and Windows 10 OS. The simulation parameters are shown as Table I.

For performance comparison, we adopt the following five schemes:

- *SPFR*: Shortest path first routing, which is searched by the Dijkstra algorithm without content-aware capability.
- *CR*: Content-aware routing, which is without cache prediction and considers cache-enabled satellites' caches unchanged.
- *OCR*: Opportunistic content-aware routing, which is based on the cache prediction method proposed in this paper.
- *Optimal*: Real-time content-aware routing, which can obtain cache-enabled satellites' cache status in real time. This scheme is the optimal case for content-aware routing.
- *NCPCA*: Node classification and popular content awareness scheme, which is derived from [27]. This scheme selects the core nodes as cache-enabled satellites and takes a probabilistic caching scheme on cache-enabled satellites.

For strategy SPFR, the access satellite receives the user's request without any processing and directly forwards the user's request to the original destination node, i.e., the content provider's ground station. Strategy CR believes that the cache of the cache-enabled satellite does not change, so at time t, it believes that the cache status of the cache-enabled satellite is still consistent with time nT. Hence, it will forward the request r_m to the closest cache-enabled satellite, which cached content f_m at time nT.

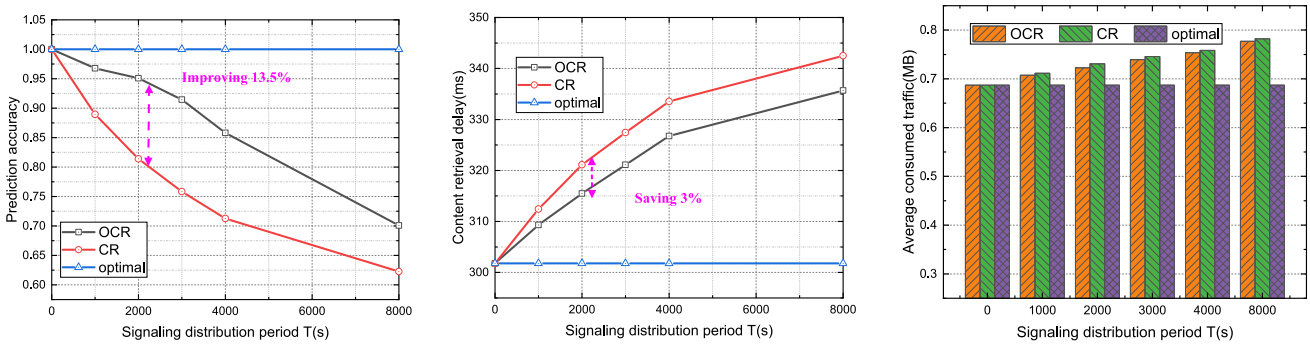
Strategy OCR is our proposed solution, which considers the change of cache-enabled satellites' cache status. After receiving the request r_m from users, it will predict the probability of content being cached by our cache prediction method and select the optimal node according to our proposed potential gains model. The optimal strategy can obtain the cache status of cache-enabled satellites in real-time and forward the request r_m to the closest cache-enabled satellite, which caches the content f_m at time t. Note that this strategy is ideal due to the huge overhead of obtaining the real-time cache status of all cache-enabled satellites in a highly dynamic satellite network.

A. Prediction Accuracy Analysis

Fig. 4 shows the effect of the signaling distribution period T on the prediction accuracy. Define $\beta = \frac{N_{succ}}{N_{succ}+N_{fail}}$ as the prediction accuracy [1], where N_{succ} represents the number of requests that the requested content was successfully retrieved in the cache-enabled satellite and N_{fail} is the number of requests that the requested content was failed retrieved from the cache-enabled satellite. It can be observed that as the collecting and distributing time of signaling increases, the accuracy of opportunistic routing requests decrease. This phenomenon is because the period of the signaling collected and distributed by assistant satellites is longer, making it impossible for access satellites to obtain the more current cache status of cache-enabled satellites. Consequently, access satellites mistakenly think that some contents that have been replaced are still in the cache of cache-enabled satellites and forward the requests to these cache-enabled satellites. This failed way increases the proportion of prediction errors. Compared with the CR scheme without prediction, our proposed scheme can significantly improve the performance of opportunistic routing's accuracy and avoid the situation that the required content cannot be retrieved on the optimal node. In addition, when the period of signaling collection and distribution is $T = 1000\ s$ or $T = 2000\ s$, the prediction accuracy of our proposed scheme OCR is above 95%, and the performance in terms of content retrieval delay and traffic consumption is very close the optimal strategy.

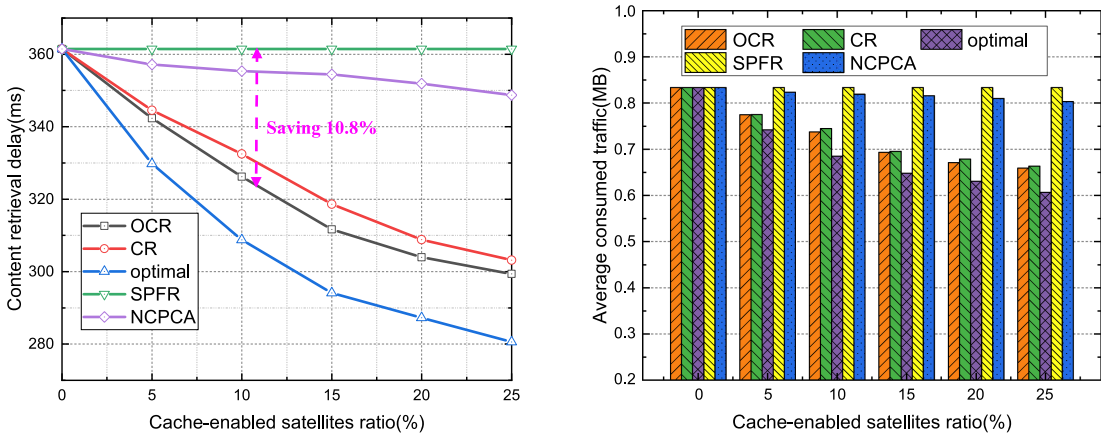
B. Performance Under Different Cache Settings

In this subsection, we evaluate the performance of different cache settings, including the cache size and the number of cache-enabled satellites. Fig. 5 shows the effect of the number of nodes with caching enabled on the scenario. The situation is similar to terrestrial content distribution networks, which allow access satellites to retrieve contents from closer satellites rather than the remote content provider. In addition, compared with SPFR, CR, and NCPCA, OCR is closer to the theoretical optimal scheme. Fig. 6 shows the performance of different cache sizes. It can be observed that content retrieval delay and traffic consumption decrease as the cache space increases. With the increase of the cache size, cache-enabled satellites can cache more popular contents. Hence, access satellites are more likely to successfully retrieve content from cache-enabled satellites after receiving users' requests. It should be noted that when the cache space is 10, the content retrieval delays of OCR and CR are higher than NCPCA. What's more, the content retrieval delay of CR is even higher than SPFR without cache capability. The reason for this phenomenon is that due to the tiny cache space,



(a) Prediction accuracy verse different T. (b) Content retrieval delay verse different T. (c) Consumed traffic verse different T.

Fig. 4. Performance under different signaling distribution interval T.



(a) Content retrieval delay verse different cache-enabled satellites number. (b) Consumed traffic verse different cache-enabled satellites number.

Fig. 5. Performance under different cache-enabled satellite scale.

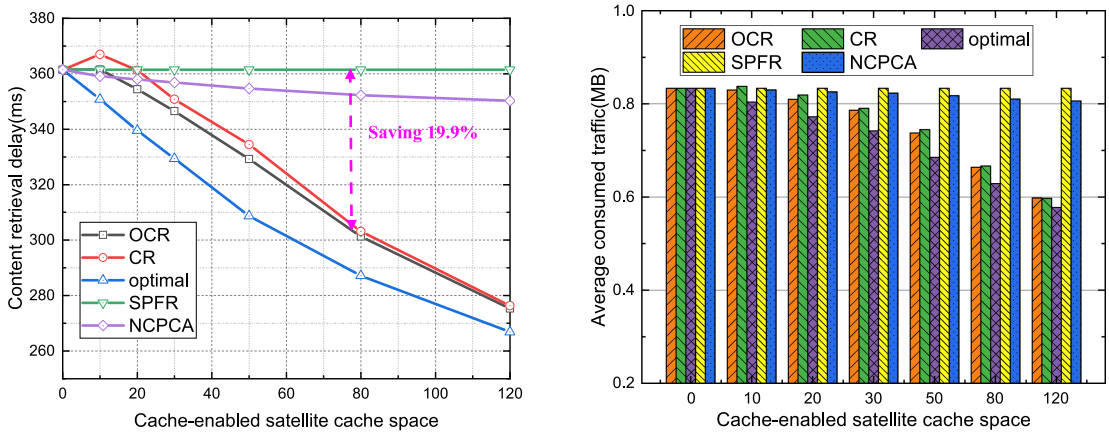
the cached contents set of the cache-enabled satellite will always be in a highly variable state under the LRU caching strategy. Therefore, there is a big difference between the cache set at nT time and the cache set at $t = nT + \tau$ time. Scheme CR considers that the cache of the default node does not change, which leads to incorrectly forwarding the request to a cache node that does not cache the content and further contributes to an increase in content retrieval delay. Fortunately, with the proposed cache prediction method, i.e., in scheme OCR, the content retrieval delay and traffic consumption can be reduced even though the cache space is tiny and the gap with NCPCA and the optimal solution is also decreasing. However, when the cache space is large, since the cached contents set of cache-enabled satellites will not change widely, the OCR is significantly better than NCPCA and also better than the solution CR without cache prediction.

C. Performance Under Different Dataset

To verify our proposed scheme's generality, we adopt different synthetic datasets, which obey the Zipf [37] distribution with different parameters. The larger the parameter of the Zipf distribution, the higher the proportion of requests for certain contents in the request data set, i.e., the more concentrated the popular contents. The first four columns in Fig. 7 represent the requested datasets of different exponent parameters in Zipf distributions, and the last column represents the real-world dataset. We can see that OCR has less content retrieval delay and traffic consumption compared to SPFR, CR, and NCPCA, regardless of whatever dataset is conducted. Meanwhile, it can be observed that OCR has different gain effects on datasets with different popularity distributions. With the increase of the Zipf distribution parameter, the performance improvement for using OCR is more prominent. This phenomenon is because users' requests converge on several popular contents with the increasing distribution parameter α. Thus cache-enabled satellites prefer to cache them for lower content retrieval delay and traffic consumption. Since a small number of content accounts for the vast majority of requests, cache-enabled satellites also cache these contents most of the time, and access satellites can successfully retrieve these popular contents from cache-enabled satellites.

D. Performance Under Large-Scale Network

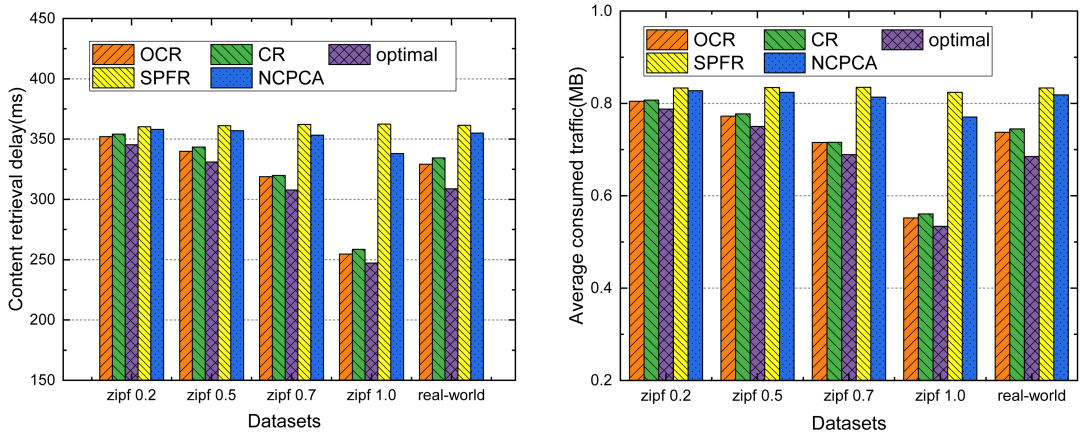
In order to verify the performance of our scheme in a large-scale LEO satellite network, we adopted the OneWeb constellation design [38], which has 648 LEO satellites, a total of 18 orbital planes, and each orbital plane has 36 satellites and



(a) Content retrieval delay verse different cache space.

(b) Consumed traffic verse different cache space.

Fig. 6. Performance under different cache space c.



(a) Content retrieval delay verse different datasets.

(b) Consumed traffic verse different datasets.

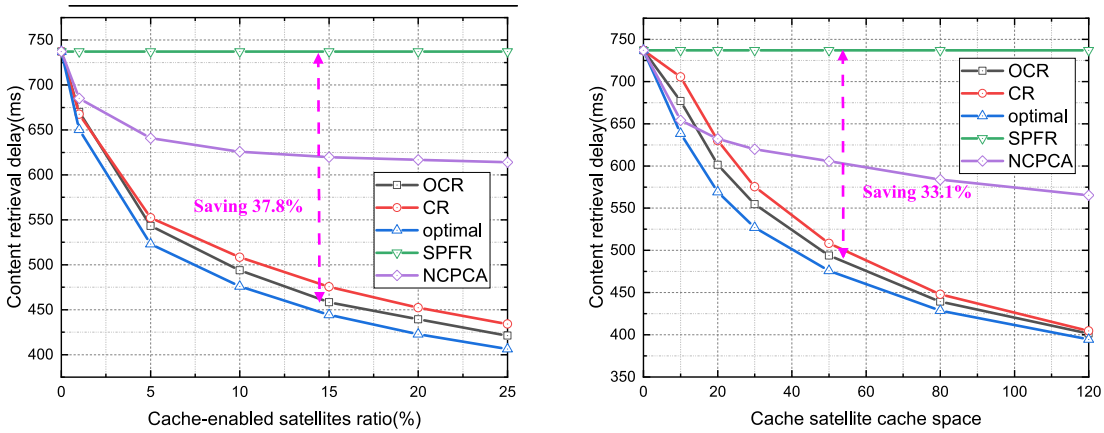
Fig. 7. Performance under different datasets.

orbit height is 1200 km. Fig. 8 shows the performance under different cache settings in the large-scale network. It can be found that the scheme with caching enabled prominently reduces the content retrieval delay compared with the scheme without caching enabled. Even when the number of cache nodes is small, or the cache space is small, the performance of content-aware routing is significant compared with scheme SPFR. This phenomenon shows that in large-scale networks, enabling caching can bring considerable benefits. Besides, the performance will be further improved and closer to the theoretically optimal scheme if combined with our cache prediction method and potential gains model. As shown in Fig. 8(b), compared with NCPCA, our proposed solutions, OCR and CR, perform better when the number of cache-enabled satellites and the cache space are larger. However, when the cache space is limited, similar to Fig. 6(a), the performance of CR and OCR is worse than the NCPCA scheme. This is attributed to the fact that the cache of a cache-capable satellite changes dynamically when the cache space is small. Consequently, it frequently occurs that the required content cannot be retrieved from the cache-enabled satellite by our schemes, leading to the necessity of rerouting and resulting in suboptimal performance. In large-scale networks, i.e., Fig. 8(b), this phenomenon is more apparent compared to Fig. 6(a) due to the larger number of cache-enabled satellites.

E. Additional Computational and Transmission Overhead

Our proposed scheme incurs additional computation and transmission overhead compared to the basic routing scheme SPFR. On the one hand, the additional computational overhead mainly comes from the process of cache prediction in Algorithm 1. On the other hand, the additional transmission overhead mainly comes from the interactions of global cache information. Table II shows the additional computational and transmission overhead required by OCR, where cache space c is 50 and the number of LEO satellites is 100.

For each request, the access satellite selects the optimal node according to Algorithm 1 and forwards it according to the routing table. In the case where the network has $|\mathcal{L}_{cs}|$ cache-enabled



(a) Content retrieval delay verse different cache-enabled satellites number.

(b) Content retrieval delay verse different cache space.

Fig. 8. Performance under different cache settings in the large-scale LEO satellites network.

TABLE II
ADDITIONAL COMPUTATIONAL AND TRANSMISSION OVERHEAD

Cache-enabled Satellites	$\lvert\mathcal{L}_{cs}\rvert = 5$	$\lvert\mathcal{L}_{cs}\rvert = 10$	$\lvert\mathcal{L}_{cs}\rvert = 15$	$\lvert\mathcal{L}_{cs}\rvert = 20$	$\lvert\mathcal{L}_{cs}\rvert = 25$
Computational Overhead by SPFR (CPU Cycles)	9.82×10^5	9.83×10^5	9.81×10^5	1.04×10^6	1.07×10^6
Computational Overhead by OCR (CPU Cycles)	1.16×10^6	1.29×10^6	1.42×10^6	1.61×10^6	1.76×10^6
Additional Computational Overhead (CPU Cycles)	1.8×10^5	3.1×10^5	4.4×10^5	5.8×10^5	6.9×10^5
Consumed Traffic by SPFR (GB)	84.332	84.332	84.332	84.332	84.332
Consumed Traffic by OCR (GB)	77.852	73.062	69.732	67.292	65.132
Saved Traffic (GB)	6.42	11.27	14.60	17.04	19.20
Additional Transmission Overhead by OCR (GB)	0.25	0.50	0.75	1.00	1.25

satellites, a total content set the size of $|\mathcal{M}|$, and a routing table, the complexity of executing Algorithm 1 is $\mathcal{O}(|\mathcal{L}_{cs}| \cdot |\mathcal{M}|)$ as we analyze in Section V-E. The calculation process of Algorithm 1 is the additional computational overhead of our scheme compared to the basic routing scheme SPFR. In our simulation environment, we evaluate the computational overhead for each request generated by scheme OCR and scheme SPFR. The results are shown in the 2-4 lines of Table II. We can find that the more cache-enabled satellites, the greater the computational overhead OCR needs. This is because our scheme requires cache prediction for each cache-enabled satellite, which increases the computational overhead. Considering that Algorithm 1 only needs to be executed once at the access satellite and does not need to be processed during the forwarding process. Hence, the additional computational overhead is acceptable for us. Besides, we believe that the additional computational overhead can be better solved with the development of computer hardware technology.

Compared with the basic routing scheme SPFR, our OCR scheme requires additional transmission overhead to interact with the global cache information, providing necessary conditions for opportunistic content-aware routing. The additional transmission overhead mainly comes from the collection and broadcast of global cache information by assistant satellites. When the number of cache-enabled satellites in the network is $|\mathcal{L}_{cs}|$, the size of popularity table $PT_i(nT)$ of cache-enabled satellite is S MB, we can deduce that the transmission overhead for the interaction of global cache information in one period is $2 \cdot |\mathcal{L}_{cs}| \cdot S$ MB. The 7-8 lines of Table II compare the saved traffic and the additional transmission overhead caused by the interaction of global cache information. The comparison shows that the additional transmission overhead caused by global cache information interaction is far less than the traffic saved by our proposed scheme OCR. This is because we use broadcasting to deliver global cache information and collect and distribute signaling at a larger interval simultaneously. Thus, the transmission overhead caused by global cache information exchange can be significantly reduced.

VII. CONCLUSION

In this paper, we proposed an opportunistic content-aware routing scheme to provide efficient content delivery service in the STIN. The basic idea is to leverage the cached contents on cache-enabled satellites and find the optimal route solution for each request with the largest potential gains, i.e., how much delay can be reduced. In addition, considering the overhead of real-time information collection in dynamic satellite networks, each satellite can only be informed of the cache status of other

satellites periodically. Thus, we also designed a cache prediction method based on historical popularity information. That is, a satellite can predict the probability of a certain content being cached in other cache-enabled satellites according to the cache information periodically collected and distributed by assistant satellites. Extensive simulations showed that the opportunistic content-aware routing scheme based on the cached contents prediction method outperforms the traditional shortest path first scheme and content-aware routing scheme regarding content retrieval delay and traffic consumption.

In our future work, we will further improve the performance of the solution and gradually reduce the additional overhead it brings while ensuring low content retrieval delay and traffic consumption.

REFERENCES

[1] J. Tang, J. Li, L. Zhang, K. Xue, Q. Sun, and J. Lu, "Content-aware routing based on cached content prediction in satellite networks," in *Proc. IEEE Glob. Commun. Conf.*, 2022, pp. 6541–6546.

[2] 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.

[3] H. Guo, J. Li, J. Liu, N. Tian, and N. Kato, "A survey on space-air-ground-sea integrated network security in 6G," *IEEE Commun. Surv. Tut.*, vol. 24, no. 1, pp. 53–87, First Quarter 2022.

[4] 3GPP, "NR NTN (non-terrestrial networks) enhancements," Sophia Antipolis, France, Tech. Rep. TR 38.801, 2022.

[5] F. Wang, D. Jiang, Z. Wang, J. Chen, and T. Q. Quek, "Seamless handover in LEO based non-terrestrial networks: Service continuity and optimization," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1008–1023, Feb. 2023.

[6] X. Zhu, C. Jiang, L. Kuang, N. Ge, S. Guo, and J. Lu, "Cooperative transmission in integrated terrestrial-satellite networks," *IEEE Netw.*, vol. 33, no. 3, pp. 204–210, May/Jun. 2019.

[7] Y. Zhu, W. Bai, M. Sheng, J. Li, D. Zhou, and Z. Han, "Traffic allocation for heterogeneous links in satellite data relay networks," *IEEE Trans. Veh. Technol*, vol. 70, no. 8, pp. 8065–8079, Aug. 2021.

[8] Z. Jia, M. Sheng, J. Li, D. Zhou, and Z. Han, "Joint HAP access and LEO satellite backhaul in 6G: Matching game-based approaches," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1147–1159, Apr. 2021.

[9] The mobile economy 2023, 2023. [Online]. Available: https://www.gsma.com/mobileeconomy/wp-content/uploads/2023/03/270223-The-Mobile-Economy-2023.pdf/

[10] T. Li, J. Yuan, and M. Torlak, "Network throughput optimization for random access narrowband cognitive radio internet of things (NB-CR-IoT)," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1436–1448, Jun. 2018.

[11] M. Zhang, Y. Xiong, S. X. Ng, and M. El-Hajjar, "Content-aware transmission in UAV-assisted multicast communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7144–7157, Nov. 2023.

[12] Spacebelt aims to store data in satellites, 2020. [Online]. Available: https://blocksandfiles.com/2020/04/21/spacebelt-store-data-in-satellites-analysis/

[13] H. Huang, S. Guo, and K. Wang, "Envisioned wireless big data storage for low-earth-orbit satellite-based cloud," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 26–31, Feb. 2018.

[14] X. Jia, T. Lv, F. He, and H. Huang, "Collaborative data downloading by using inter-satellite links in LEO satellite networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1523–1532, Mar. 2017.

[15] Z. Lai, H. Li, Q. Zhang, Q. Wu, and J. Wu, "STARFRONT: Cooperatively constructing pervasive and low-latency CDNs upon emerging LEO satellites and clouds," *IEEE/ACM Trans. Netw.*, vol. 31, no. 6, pp. 2559–2574, Dec. 2023.

[16] D. Han, W. Liao, H. Peng, H. Wu, W. Wu, and X. Shen, "Joint cache placement and cooperative multicast beamforming in integrated satellite-terrestrial networks," *IEEE Trans. Veh. Technol*, vol. 71, no. 3, pp. 3131–3143, Mar. 2022.

[17] 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, Fourth Quarter 2020.

[18] 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.

[19] J. Li, K. Xue, D. S. Wei, J. Liu, and Y. Zhang, "Energy efficiency and traffic offloading optimization in integrated satellite/terrestrial radio access networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2367–2381, Apr. 2020.

[20] D. Jiang et al., "QoE-aware efficient content distribution scheme for satellite-terrestrial networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 443–458, Jan. 2023.

[21] X. Zhu, C. Jiang, L. Kuang, and Z. Zhao, "Cooperative multilayer edge caching in integrated satellite-terrestrial networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 2924–2937, May 2022.

[22] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 26–36, Jul. 2012.

[23] L. Galluccio, G. Morabito, and S. Palazzo, "Caching in information-centric satellite networks," in *Proc. IEEE Int. Conf. Commun.*, 2012, pp. 3306–3310.

[24] T. De Cola and A. Blanco, "ICN-based protocol architectures for next-generation backhauling over satellite," in *Proc. IEEE Int. Conf. Commun.*, 2017, pp. 1–6.

[25] J. Li, K. Xue, J. Liu, Y. Zhang, and Y. Fang, "An ICN/SDN-based network architecture and efficient content retrieval for future satellite-terrestrial integrated networks," *IEEE Netw.*, vol. 34, no. 1, pp. 188–195, Jan./Feb. 2020.

[26] Y. Yang, T. Song, W. Yuan, and J. An, "Towards reliable and efficient data retrieving in ICN-based satellite networks," *J. Netw. Comput. Appl.*, vol. 179, 2021, Art. no. 102982.

[27] R. Xu et al., "A hybrid caching strategy for information-centric satellite networks based on node classification and popular content awareness," *Comput. Commun.*, vol. 197, pp. 186–198, 2023.

[28] Y. Nishiyama, M. Ishino, Y. Koizumi, T. Hasegawa, K. Sugiyama, and A. Tagami, "Proposal on routing-based mobility architecture for ICN-based cellular networks," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2016, pp. 467–472.

[29] L. Zhang et al., "Named data networking," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 66–73, 2014.

[30] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," in *Proc. IEEE Conf. Comput. Commun.*, 2014, pp. 2040–2048.

[31] S. Podlipnig and L. Böszörmenyi, "A survey of web cache replacement strategies," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 374–398, 2003.

[32] J. Moy, "RFC 2328, OSPF version 2," Apr. 1998. [Online]. Available: http://www.rfc-editor.org/rfc/rfc2328.txt

[33] A. Roy and M. Chandra, "Extensions to OSPF to support mobile ad hoc networking," Internet Engineering Task Force, Request For Comments (Experimental) RFC 5820, Mar. 2010.

[34] G. Malkin, "RFC 2453, RIP version 2," Nov. 1998. [Online]. Available: http://www.rfc-editor.org/rfc/rfc2453.txt

[35] Large-scale satellite networks simulator, 2019. [Online]. Available: https://github.com/infonetlijian/ONE-Extended-Simulator

[36] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interactive Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2015.

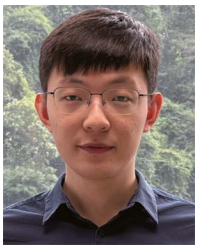
[37] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Phys.*, vol. 46, no. 5, pp. 323–351, 2005.

[38] OneWeb satellite constellation, 2021. [Online]. Available: https://en.wikipedia.org/wiki/OneWeb#OneWeb_satellite_constellation

Jin Tang received the bachelor's degree in communication engineering from Beijing Jiao Tong University, in July 2021. He is currently working toward the graduate degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). His research interests include satellite-terrestrial integrated network, content delivery, and edge caching.

Jian Li (Senior Member, IEEE) received the bachelor's degree from the Department of Electronics and Information Engineering, Anhui University, in 2015, and the doctor's degree from the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China (USTC), in 2020. From 2019 to 2020, he was a visiting scholar with the Department of Electronic and Computer Engineering, University of Florida. From 2020 to 2022, he was a post-doctoral researcher with the School of Cyber Science and Technology, USTC. He is currently an associate researcher with the School of Cyber Science and Technology, USTC. He serves as an editor of the *China Communications*. His research interests include quantum networks wireless networks, and next-generation Internet.

Lan Zhang (Member, IEEE) received the BE and MS degrees from the University of Electronic Science and Technology of China, in 2013 and 2016, respectively, and the PhD degree from the University of Florida, in 2020. She has been a tenure-track assistant professor with the Department of Electrical and Computer Engineering, Clemson University since 2024. Before that, she was an assistant professor with the Department of Electrical and Computer Engineering, Michigan Technological University from 2020 to 2023. Her research interests include wireless communications, distributed machine learning, and cybersecurity for various Internet-of-Things applications.



Xianhao Chen (Member, IEEE) received the BEng degree from Southwest Jiaotong University, in 2017, and the PhD degree from the University of Florida, in 2022. He is currently an assistant professor with the Department of Electrical and Electronic Engineering, University of Hong Kong. His research interests include wireless networking and machine learning.

Kaiping Xue (Senior Member, IEEE) received the bachelor's degree from the Department of Information Security, University of Science and Technology of China (USTC), in 2003, and the doctor's degree from the Department of Electronic Engineering and Information Science (EEIS), USTC, in 2007. From 2012 to 2013, he was a postdoctoral researcher with the Department of Electrical and Computer Engineering, University of Florida. Currently, he is a professor with the School of Cyber Science and Technology, USTC. He is also a director of Network and Information Center, USTC. His research interests include next-generation Internet architecture design, transmission optimization, and network security. His work won best paper awards in IEEE MSN 2017 and IEEE HotICN 2019, the Best Paper Honorable Mention in ACM CCS 2022, the Best Paper Runner-Up Award in IEEE MASS 2018, and the best track paper in MSN 2020. He serves on the Editorial Board of several journals, including the *IEEE Transactions on Dependable and Secure Computing* (TDSC), *IEEE Transactions on Wireless Communications* (TWC), and *IEEE Transactions on Network and Service Management* (TNSM). He has also served as a (lead) guest editor for many reputed journals/magazines, including the *IEEE Journal on Selected Areas in Communications* (JSAC), *IEEE Communications Magazine*, and *IEEE Network*. He is an IET fellow.

Qibin Sun (Fellow, IEEE) received the PhD degree from the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China (USTC), in 1997. From 1996 to 2007, he was with the Institute for Infocomm Research, Singapore, where he was responsible for industrial, as well as academic research projects in the area of media security, image and video analysis, etc. He was the head of Delegates of Singapore in ISO/IEC SC29 WG1(JPEG). He worked with Columbia University during 2000–2001 as a research scientist. Currently, he is a professor with the School of Cyber Security, USTC. His research interests include multimedia security, network intelligence and security and so on. He led the effort to successfully bring the robust image authentication technology into ISO JPEG2000 standard Part 8 (Security). He has published more than 120 papers in international journals and conferences.

Jun Lu received the bachelor's degree from Southeast University, in 1985, and the master's degree from the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China (USTC), in 1988. Currently, he is a professor with the Department of EEIS, USTC. His research interests include theoretical research and system development in the field of integrated electronic information systems. He is an academician of the Chinese Academy of Engineering (CAE).