Energy Efficiency and Traffic Offloading Optimization in Integrated Satellite/Terrestrial Radio Access Networks

Jian Li, Kaiping Xue, *Senior Member, IEEE*, David S. L. Wei, *Senior Member, IEEE*, Jianqing Liu, *Member, IEEE*, and Yongdong Zhang, *Senior Member, IEEE*

***Abstract*—In order to cope with the explosive growth of mobile traffic, many traffic offloading schemes such as heterogenous networks have been developed to enhance network capacity of the Radio Access Network (RAN). Among them, networking of Low-Earth Orbit (LEO) satellites promises to significantly improve the RAN performance due to its economical prospect and advantages in high bandwidth and low latency. In this paper, by introducing the cache-enabled LEO satellite network as a part of RAN, we propose an integrated satellite/terrestrial cooperative transmission scheme to enable an energy-efficient RAN by offloading traffic from base stations through satellite's broadcast transmission. Considering energy-constraints of satellites, we then formulate a nonlinear fractional programming problem aiming at optimizing transmission energy efficiency of the system. In order to effectively solve this problem, we transform it into an equivalent one, and then adopt iteration and sub-problem decomposition to obtain the optimal solution for each optimization variable, i.e., block placement, power allocation, and cache sharing variable. Numerical results show that compared with traditional terrestrial scheme, our cooperative transmission scheme achieves significant performance improvement in terms of traffic offloading and energy efficiency, especially in an environment of high request consistency degree.**

***Index Terms*—Radio access network, integrated satellite/terrestrial cooperative transmission, energy efficiency, traffic offloading.**

I. Introduction

WITH the rapid increase of wireless devices and mobile traffic, the conventional cellular network is facing severe challenges in terms of wireless/backhaul capacity, energy efficiency and so on [1]–[3]. To better handle these challenges, one straightforward solution is to deploy dense small cells to offload Base Station (BS)'s traffic by exploiting cooperation in heterogeneous networks such as different tiers of BSs and low-power Access Points (APs) [4]–[7]. However, the traditional offloading schemes in heterogenous networks can only provide limited improvement. On the one hand, the dense-deployed BSs and APs will bring in more notorious competition and interference given limited spectrum resources in the current Radio Access Network (RAN). On the other hand, due to the upper bound of energy efficiency and capacity improvement, there also exists densification limits for the conventional cellular network [4], [6]. In this case, some novel technologies, such as satellite communication, are also considered in heterogenous networks to bring in simple but efficient access approach and further enhance the performance of cellular network.

In the next generation cellular network, satellite communication has been considered as a promising complement approach, and satellite access technology should also be supported to assist terrestrial RAN [8], [9]. In the past, because of satellite's high cost, only a few researches considered to utilize satellite's advantages to assist terrestrial networks. Thanks to the rapid development of small satellite launch and maintenance technologies, Low-Earth Orbit (LEO) small satellites and networking, which can achieve low latency and high bandwidth [10], have become economic-friendly for industrial implementation. Many companies such as SpaceX and OneWeb, are planning to construct large-scale LEO satellite networks to provide worldwide Internet access service [11]–[14]. Thus, one can foresee that, considering its efficient broadcast transmission, satellite communication will become ubiquitous in the next few decades. In this case, more and more researches have been taking satellite into consideration in the design and implementation of a high performance and energy-efficient RAN.

While the benefit of satellite assistance has been proven in RAN [7], [8], [15], [16], introducing satellite also puts extra pressure on BSs and the performance of satellite access is limited by satellite backhaul link. To diminish the burden of BSs and break through the restriction of backhaul link, one effective solution is to take full advantage of each node's capability of storage, computation, and communication. For

Manuscript received January 22, 2019; revised June 12, 2019 and October 7, 2019; accepted December 23, 2019. Date of publication January 13, 2020; date of current version April 9, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0800301, in part by the National Natural Science Foundation of China under Grant 91538203 and Grant 61972371, and in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) under Grant 2016394. The associate editor coordinating the review of this article and approving it for publication was D. C. Popescu. *(Corresponding author: Kaiping Xue.)*

Jian Li, Kaiping Xue, and Yongdong Zhang are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: kpxue@ustc.edu.cn).

David S. L. Wei is with the Computer and Information Science Department, Fordham University, New York, NY 10458 USA.

Jianqing Liu is with the Department of Electrical and Computer Engineering, The University of Alabama in Huntsville, Huntsville, AL 35899 USA.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.2964236

example, in the conventional cellular network, the combination of caching, centralized/distributed management, and novel access technologies, e.g., millimeter-wave communication and Non-Orthogonal Multiple Access (NOMA), have become a common idea in existing works [4]–[6], [17]–[19]. With the development of satellite and Information and Communications Technology (ICT), onboard caching in satellite becomes a promising approach in practice. Many works consider how to leverage cache capacity to help optimize delivery, e.g., building Information Centric Networking (ICN)-based satellite networks [20]–[26]. For integrated satellite/terrestrial RAN, the combination of satellite and caching can both serve multiple small cells through broadcast transmission and provide efficient traffic offloading without introducing extra load on BSs nor getting pressure from the restriction of backhaul link. Moreover, with the help of some fine-grained caching schemes such as coded caching [5], [27], [28], satellite can achieve cooperative caching and transmission with terrestrial cellular network to further enhance the cache efficiency. Unfortunately, none of the existing works consider to adopt cooperative caching and transmission scheme to assist terrestrial RAN.

In order to enhance the capacity of conventional cellular network and overcome the limitations of traditional heterogeneous networks, in this paper, we consider a satellite network that consists of LEO satellites equipped with certain computing and storage resources to take full advantage of satellite's broadcast characteristic. By exploiting storage capabilities of satellite and dense-deployed APs, the proposed system can offload cellular traffic from BS in a cooperative manner with the help of coded caching technology. Considering the heavy load on backhaul and energy-saving concern, a joint optimization problem, which considers both traffic offloading and energy efficiency in integrated satellite/terrestrial RAN, is formulated. Furthermore, we also propose an effective algorithm that can significantly improve network performance in traffic offloading and energy efficiency compared with terrestrial scheme. For the convenience of the study of performance evaluation and to hightlight the advantage of employing the broadcast function of a satellite system, we define *request consistency degree* as the percentage of the cells that have the same request occurrences. Numerical results show that the proposed schemes can outperform terrestrial schemes in all environments of different request consistency degree.

The contributions of this paper can be summarized as follows:

- By introducing the cache-enabled LEO satellite network as a part of RAN, we propose an integrated satellite/terrestrial cooperative caching and transmission scheme, which leverages the broadcast characteristic in satellites to efficiently provide traffic offloading for terrestrial networks. Considering satellite's limited storage capability, a cache sharing scheme, which enables terrestrial AP to share its cached content to the satellite, is also proposed to better exploit satellite's broadcast transmission.
- Considering the energy-constraint of satellites, we formulate the offloading problem aiming at maximizing the energy efficiency of RAN in terms of block (i.e., a part of the original content) placement, power allocation, and cache sharing decisions. For this formulated nonlinear fractional programming problem, we exploit a parametric method to transform it into an equivalent one, which is further divided into three solvable sub-problems. Each sub-problem is associated with one variable to be optimized. Furthermore, we thoroughly analyze all these sub-problems, and propose corresponding optimization algorithms to solve them.
- We conduct evaluations to verify the superiority of our scheme in terms of energy efficiency and traffic offloading. Numerical results demonstrate that, compared with traditional terrestrial schemes, our proposed scheme can achieve several times of improvement in energy efficiency for the similar traffic offloading performance, and furthermore such advantages are more remarkable in the environment of high request consistency degree.

The rest of this paper is organized as follows. At first, the related works about terrestrial traffic offloading schemes and integrated satellite/terrestrial schemes are discussed in Section II. Then, the system model is given in Section III. To improve energy efficiency in our cooperative transmission scheme, a fractional optimization problem is formulated in Section IV, and the problem analysis and designed algorithms are given in Section V. At last, the performance evaluation is conducted in Section VI, and conclusions are drawn in Section VII.

II. Related Works

Considering the aggressive objectives in the next generation cellular network in terms of 1000-fold higher capacity, 100-fold higher energy efficiency, and 10-fold lower latency [6], [29], a lot of existing works try to enhance the system performance by exploiting novel access technologies (e.g., Multiple-Input Multiple-Output (MIMO) and NOMA). In terrestrial RAN, the densification of small cells and cooperation in heterogeneous networks have been considered as promising solutions. Xu *et al.* [30] proposed a cooperative NOMA transmission that combined Macro Base Station (MBS) and pico-BS, and formulated a resource allocation problem to improve the system capacity. Inspired by this, a cooperative access scheme that organizes multiple APs in ultra-dense network is investigated by Liu *et al.* [6]. However, simulation results show that the system capacity improvement will reach a bottleneck with the increasing density of APs. Since NOMA system exploits the power-domain for user multiplexing, the power allocation in NOMA system becomes an important issue, and most of the related works adopt centralized allocation scheme. To achieve distributed power control, Di *et al.* [31] proposed a novel NOMA-based scheme which combines centralized scheduling at the BS and distributed power control at the vehicles, and the results show that this scheme can significantly improve data rates compared with OMA scheme. Considering the phenomenon of multiple requests for popular contents in content distributions, in 2012, Golrezaei *et al.* [32] first proposed to deploy helpers with weak backhaul links but large storage capacity to assist the MBS offloading traffic of popular files that have been cached. In order to further improve cache efficiency and provide fine-granularity caching operations, Bioglio *et al.* [28] proposed a MDS-encoded caching scheme in heterogeneous network. Based on that, Xu *et al.* [5] further proposed a NOMA-based coded caching scheme among APs in small-cell networks, which means that each user can receive coded packets from multiple APs simultaneously. However, considering the spectrum resource competition and limited coverage of low power AP, the improvement of system capacity is indeed limited. At the same time, the explosive growth of mobile traffic has also triggered a dramatic increase in the energy consumption of RAN. Since energy consumption is a key concern for the operator as it is always the main part of the entire operational expenses [33], Prasad *et al.* [34] and Yunas *et al.* [35] investigated the influence of massive MIMO technique and the cooperation among dense deployed BSs in energy efficiency, respectively. Unfortunately, these existing works only consider terrestrial access approaches, none of them consider to leverage satellite's broadcast transmission to provide energy-efficient transmission for conventional cellular network.

Compared to the terrestrial dense-deployed APs, satellite has advantages in providing energy-efficient transmission through broadcast transmission. Some existing works have considered integrated satellite/terrestrial network and cache-enabled satellite to provide content access service. In the cellular network, Zhu *et al.* [7] first considered the the combination of satellite and terrestrial RAN, and proposed an integrated satellite/terrestrial architecture to cooperatively provide downlink transmission based on NOMA. Then, by distinguishing users' content requests and dividing users into different group, Zhu *et al.* [8] also proposed a downlink multicast transmission scheme to offload content traffic from BS. Considering the emerging C-RAN, a cloud based integrated satellite/terrestrial architecture was considered in [15]. By doing so, the operator can utilize centralized signal processing to achieve energy-efficient resource management at cloud and further achieve seamless coverage. Driven by the ongoing LEO constellation projects of SpaceX and OneWeb, Di *et al.* [36] proposed a terrestrial-satellite network architecture to achieve efficient data offloading. By utilizing ultra-dense LEO constellation, each terrestrial terminal can be served by multiple LEO satellites simultaneously, and further simulation results show that the integrated network significantly outperforms the non-integrated ones in terms of the sum data rate. Due to the long propagation delay (typical delay is 14ms [37]) of satellite communication, bent-pipe satellite suffers even longer transmission delay (over 50ms) since it needs to retrieve user's requested data from ground station. In order to alleviate the transmission delay and communication overhead between satellite and ground, researchers consider to combine satellite's wide-area coverage and in-network caching. Salvatore *et al.* [23] proposed a profile-aware satellite caching strategy. After that, Wu *et al.* [24] proposed a two-layer caching model for content distribution services and formulated a nonlinear integer programming problem. By considering both global and local content popularity, the caching strategy based on the genetics can significantly reduce both the downlink and uplink of bandwidth consumption. To further offload traffic from terrestrial network backhaul, Kalantari *et al.* [38] proposed an offline caching approach which considers both global content popularity with mono-beam satellite and local content popularity with multi-beam satellite in a hybrid satellite/terrestrial architecture. To alleviate the spectrum shortage and meet the requirements of improved spectral efficiency, An *et al.* [39] incorporated wireless content caching into satellite/terrestrial network. By storing frequently required contents and serving users' requests conveniently, the intrinsic issues of satellite/terrestrial network including bandwidth inefficiency and excessive propagation delay could be largely reduced. However, none of these existing works consider to adopt cache-enabled satellite as a part of RAN to cooperatively offload wireless traffic of APs from BSs.

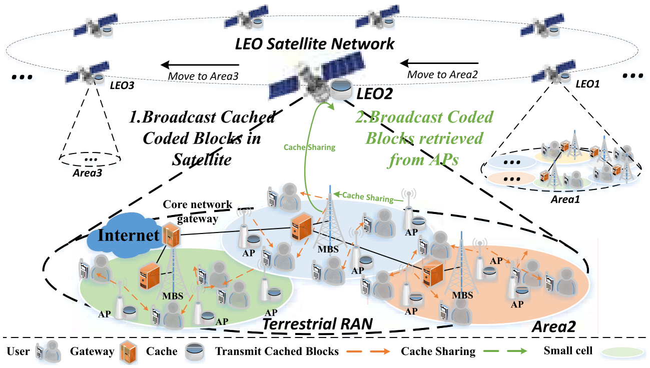


Fig. 1. The proposed system model of integrated satellite/terrestrial RAN.

III. System Model

A. Network Model

As illustrated in Fig. 1, we consider a downlink cooperative traffic offloading scheme in an integrated satellite/terrestrial cache-enabled RAN, where a number of APs and a LEO satellite network serve users in a cooperative manner. In each small cell, there are a MBS and multiple APs. APs and users are assumed to be spatially distributed on a plane according to two independent homogeneous poisson point processes with densities λ_a and λ_u, respectively. MBS is connected to the core network gateway via the reliable optical fiber backhaul link, while APs and each LEO satellite are with unreliable backhaul links and are connected to MBS through wireless links to update their cached contents. Note that we consider each satellite serves non-overlapping terrestrial area in this paper, that is to say, each user can only be served by one satellite anytime. Without loss of generality, we focus on the analysis of one satellite and multiple cells in its coverage area, and $\mathcal{J}$ and $\mathcal{U}$ are used to denote the cell set and the user set in the coverage, respectively.

Due to the rapid movement of LEO satellite, each LEO satellite can provide access service to terrestrial cells for a certain duration. The handover process and communication resource allocation shall be managed by Network Control Center (NCC) [40], [41]. According to satellite's predictable trajectory, each LEO satellite can provide access service for the same area periodically under the control of NCC. In Fig. 1,

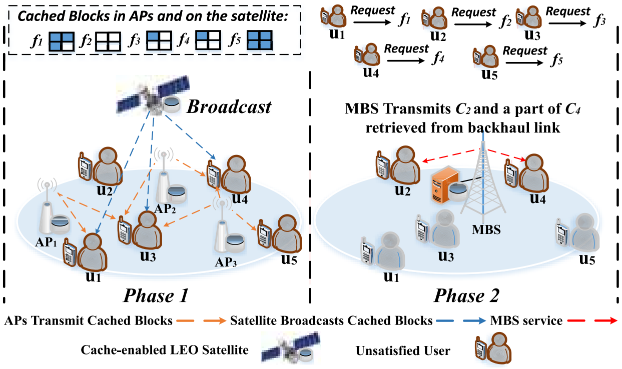

Cached Blocks in APs and on the satellite:
f1
f2
f3
f4
f5
Request
u1
u2
u3
u4
u5
Broadcast
MBS Transmits C2 and a part of C4 retrieved from backhaul link
AP1
AP2
AP3
MBS
Phase 1
Phase 2
APs Transmit Cached Blocks
Satellite Broadcasts Cached Blocks
MBS service
Cache-enabled LEO Satellite
Unsatisfied User

Fig. 2. The overall traffic offloading procedure for LEO satellite and APs when users request cached contents.

if we assume the start time of $LEO2$'s coverage for $Area2$ is t_1 and the coverage duration is T, then $LEO2$ will serve $Area2$ during $[t_1, t_1+T)$ and serve $Area3$ during $[t_1+T, t_1+2T)$. Actually, in dense deployed LEO networks such as Starlink project, thousands of satellites are deployed and the same location can be covered by multiple satellites simultaneously [13], [36]. Thus, seamless coverage of terrestrial RAN can be guaranteed. In this paper, we consider the single satellite scenario that each LEO satellite serves the terrestrial users during its service period.

In our integrated satellite/terrestrial RAN, each user is served by unicast transmissions of APs and/or broadcast transmissions of the satellite with their cached contents. Or the user could be served by MBS only when his/her demands can not be satisfied by APs or/and the satellite. In order to improve spectrum efficiency, NOMA-based transmission is adopted among APs. Thus, multiple APs will simultaneously serve the same user with their cached contents over the same spectrum resource. In order to avoid severe interference between satellite's broadcast and terrestrial transmission, we assume that LEO satellite operates over exclusive frequency bands such as S band (direct access service) and Ka band (indirect access service) [37], [42], [43]. Concerning the issue of what contribute to most traffic in content-based applications [1], we only concentrate on content distribution and retrieval applications in this paper.

To make our scheme easier to understand, we depict Fig. 2 to describe the basic procedure of our traffic offloading scheme. The basic idea behind our scheme is to utilize cache-enabled LEO satellite and terrestrial APs to offload traffic from base station and its backhaul link. In this case, after receiving user's request, LEO satellite and APs will first serve the user with their cached blocks (a part of user's requested content file) in a cooperative manner. If user's request can not be satisfied by the satellite and APs, e.g., content recovery failure or transmission failure, the user will be served by MBS for content retrieval through backhaul link. The specific cooperative transmission approach in Phase 1 will be described in more detail in Section III-D.

B. Coded Caching Model

In order to improve caching efficiency, a coded caching scheme similar to that of [5] is adopted in our work. We consider a content library consisting of N files, denoted by $\mathcal{F}$. All files are assumed to have the same size and each one is split into c equal-sized original blocks, where each one is with size L. Then c original blocks of each file are encoded into an arbitrarily large number of coded blocks to be cached by terrestrial APs and the satellite.

With coded caching, when a user receives an arbitrary number of blocks, n $(n \geq c)$, from APs and the satellite, he/she can recover the original file from the received blocks. By carefully selecting coding coefficients, each access node (i.e., AP or satellite) can provide linearly independent blocks to the user with high probability. In practice, it can be achieved through, for example, MDS (Maximum-Distance Separable) [28] and RLNC (Random Linear Network Coding) codes [5]. The local cache of each AP and satellite can store up to ML and $M^{\text{Sat}}L$ bits, respectively. We assume that all APs in the same cell j have the same cached blocks of i-th file in block placement phase, i.e., $m_{i,j}$. Thus, each user can access the same service from any AP in the same cell.

C. Channel Model

In the downlink channel of terrestrial APs, we consider both large-scale fading and small-scale fading. The large-scale fading is modeled by a standard distance-dependent power law path loss attenuation, while the Rayleigh fading is considered as small-scale fading. The transmission power of each AP for user u in cell j is the same, i.e., $P_{j,u}$. Thus, the received signal power at user u in cell j can be written as $|h|^2(1+r^{\alpha})^{-1}P_{j,u}$, where $h \sim \mathcal{CN}(0,1)$ represents fading coefficient, r denotes the distance between u and AP, and α is the path loss exponent.

Different from the terrestrial case, the satellite channel generally experiences less fluctuation considering the line-of-sight link between the satellite and terrestrial users. Thus, the channel model of the satellite can be considered as an Additive White Gaussian Noise (AWGN) channel [8]. According to the large-scale fading model in literatures [44], the received signal of a terrestrial user for i-th file can be computed by

$$y(x) = \frac{\lambda' G\sqrt{P_i^{\text{Sat}}}}{4\pi r^{\text{Sat}}}x_i + n_0, \tag{1}$$

where x_i denotes the symbol transmitted from the satellite for i-th file, λ' denotes the wave length, r^{Sat} denotes the distance between terrestrial users and the satellite, G is antenna gain and $n_0 \sim \mathcal{CN}(0, N_0)$ is the complex AWGN of power N_0. In order to successfully transmit cached blocks of i-th file from the satellite to terrestrial users, the minimum transmission rate $m_i^{\text{Sat}}L/t$ should be satisfied (i.e., $C \geq C_{min} = m_i^{\text{Sat}}L/t$), where m_i^{Sat} is cached blocks of i-th file on the satellite and t is the time slot in RAN. According to the Shannon equation, $C = W^{\text{Sat}}log(1+SNR)$, on-board transmission power of the satellite for i-th file can be computed by

$$P_i^{\text{Sat}} = N_0(\frac{4\pi r^{\text{Sat}}}{\lambda' G})^2 \cdot (2^{\frac{m_i^{\text{Sat}}L}{W^{\text{Sat}}t}} - 1), \tag{2}$$

where W^{Sat} is the subcarrier bandwidth in satellite system.

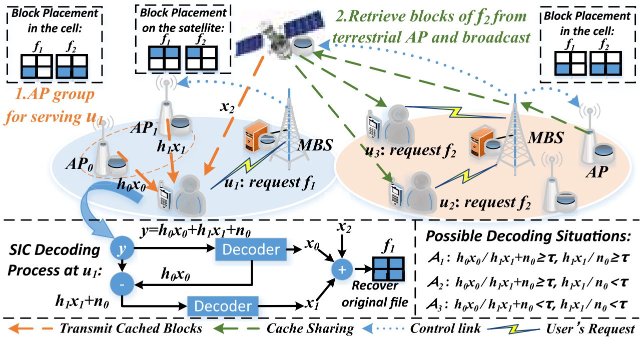

Block Placement in the cell:
f_1 f_2
Block Placement on the satellite:
f_1 f_2
2.Retrieve blocks of f_2 from terrestrial AP and broadcast
Block Placement in the cell:
f_1 f_2
1.AP group for serving u_1
AP_1
x_2
h_1x_1
MBS
AP_0
h_0x_0
u_1: request f_1
u_3: request f_2
MBS
AP
u_2: request f_2
SIC Decoding Process at u_1:
$y=h_0x_0+h_1x_1+n_0$
y
Decoder
x_0
x_2
f_1
+
-
h_0x_0
Recover original file
$h_1x_1+n_0$
Decoder
x_1
Possible Decoding Situations:
$\mathcal{A}_1$: $h_0x_0/h_1x_1+n_0 \geq \tau$, $h_1x_1/n_0 \geq \tau$
$\mathcal{A}_2$: $h_0x_0/h_1x_1+n_0 \geq \tau$, $h_1x_1/n_0 < \tau$
$\mathcal{A}_3$: $h_0x_0/h_1x_1+n_0 < \tau$, $h_1x_1/n_0 < \tau$
Transmit Cached Blocks
Cache Sharing
Control link
User's Request

Fig. 3. The proposed cooperative transmission and cache sharing scheme (split original blocks of each file $c = 4$ and cache size $M = M^{\mathrm{Sat}} = 3$).

D. File Request and Delivery Model

For user u, there are two different phases during the transmission, i.e., file request and file delivery. User u first sends a request for i-th file to MBS during the request phase. Then, during the file delivery phase, MBS coordinates with the nearest APs and the satellite to transmit their cached blocks of i-th file to u. Note that APs and the satellite can serve the same user for the i-th file simultaneously over non-overlapping frequency band. If i-th file can be recovered from received blocks at u, the request is satisfied; Otherwise, the rest of the blocks will be served by MBS.

The file popularity is modeled as the Zipf distribution with shape parameter θ. Then the popularity of the i-th ranked file in cell j is

$$p_{i,j} = \frac{1/i^{\theta}}{\sum_{n=1}^{N} 1/n^{\theta}}. \tag{3}$$

We assume that in file request phase, each user in the same cell requests for i-th file with the same probability of popularity distribution, i.e., $p_{i,j} = p_{i,j,u}$, which denotes the request probability of each user for i-th file in cell j.

In file delivery phase, a user's request can be satisfied by three ways, i.e., only NOMA-based transmission from APs, only broadcast transmission from the satellite, or the combination of these two transmissions. Since we only consider AWGN channel with large-scale fading during broadcast transmission from the satellite, satellite's signal can be surely decoded when Eq. (2) is satisfied. Fig. 3 shows an example of cooperative transmission with both NOMA-based and broadcast transmission for u_1. Since MBS is in charge of the coordination of APs and the satellite, the number of control signal transmitted from MBS to APs and the satellite can be respectively calculated by $O_j = \sum_{i\in\mathcal{F}}\sum_{u\in\mathcal{U}_j} p_{i,j}\cdot sgn(m_{i,j})$ and $O^{\mathrm{Sat}} = \sum_{i\in\mathcal{F}} sgn(m_i^{\mathrm{Sat}})$, where function $sgn(\cdot)$ equals 1 when $m_{i,j} > 0$, otherwise, equals 0.

For APs, we consider a NOMA-based transmission scheme similar to [5], [6]. If the i-th file, requested by user u, is cached by APs in the cell, these APs can simultaneously transmit cached blocks of the requested file to u over the same frequency band. User u adopts successive interference cancellation receiver to decode the signals successively in the descending order of the average received signal strength. Assume that each AP adopts the same power $P_{j,u}$ to transmit signals to user u in cell j, so the Signal-to-Interference and Noise Ratio (SINR) at user u for decoding the signal from k-th AP can be computed by

$$\gamma_{u,k} = \frac{|h_k|^2(1+r_k{}^{\alpha})^{-1}P_{j,u}}{\sum\limits_{k'\in\mathcal{K}_u,k'>k}|h_{k'}|^2(1+r_{k'}{}^{\alpha})^{-1}P_{j,u}+N_0}, \tag{4}$$

where $\mathcal{K}_u$ represents the AP group for serving user u and N_0 is the AWGN power.

Considering the complexity of SIC receiver is related to the number of decoding layer in practice [45], we assume that at most 2 nearest APs can share the same frequency band to serve the same user, i.e., $|\mathcal{K}_u| \leq 2$.

In this case of 2 served APs for user u, there are three possible situations during the transmission: 1) Both transmissions are successful, i.e., $\mathcal{A}_1 : \gamma_{u,0}, \gamma_{u,1} \geq \tau$; 2) Only the transmission of the nearest AP is successful, i.e., $\mathcal{A}_2 : \gamma_{u,0} \geq \tau, \gamma_{u,1} < \tau$; 3) Both transmissions are failed, i.e., $\mathcal{A}_3 : \gamma_{u,0}, \gamma_{u,1} < \tau$. A decoding example is also shown in Fig. 3. Obviously, we have $Pr\{\mathcal{A}_1\} + Pr\{\mathcal{A}_2\} + Pr\{\mathcal{A}_3\} = 1$. Therefore, the average traffic of user u served by APs for i-th file in cell j, $q_{i,j,u}$, can be represented as

$$q_{i,j,u} = \begin{cases} m_{i,j}(2Pr\{\mathcal{A}_1\} + Pr\{\mathcal{A}_2\}), & m_{i,j} \in \mathbb{L}_1, \\ cPr\{\mathcal{A}_1\} + m_{i,j}Pr\{\mathcal{A}_2\}, & m_{i,j} \in \mathbb{L}_2, \end{cases} \tag{5}$$

where $m_{i,j}$ represents cached blocks of i-th file in every AP of cell j, $\mathbb{L}_1 = [0, c/2]$, $\mathbb{L}_2 = (c/2, c]$, and τ denotes the decoding threshold.

In order to take full advantage of satellite's broadcast capacity, we also consider a cache sharing scheme in the integrated satellite/terrestrial RAN considering the limited storage of LEO satellites. By sharing cached blocks of terrestrial AP to the satellite, some popular files can be cached by APs and the satellite can retrieve blocks from terrestrial AP on demand. An example is shown in Fig. 3. Once the cache sharing for i-th file is enabled, multiple users who request i-th file will only be served by the satellite.

IV. Problem Formulation

In this section, we first analyze the expected traffic offloading and energy consumption in our scheme, respectively. After that, energy efficiency optimization in our scheme, which jointly considers block placement, power allocation and cache sharing, is formulated as a nonlinear fractional programming problem.

A. Traffic Offloading

The total traffic served by APs can be computed by

$$B^{\mathrm{T}} = \sum_{i\in\mathcal{F}}\sum_{j\in\mathcal{J}}\sum_{u\in\mathcal{U}_j}(1-a_i^{\mathrm{toS}})\cdot p_{i,j}q_{i,j,u}, \tag{6}$$

where a_i^{toS} denotes cache sharing variable. When $a_i^{\mathrm{toS}} = 1$, AP will forward its cached i-th file to the satellite with the help of MBS, and users who request i-th file will be served by the satellite through broadcasting; otherwise, each user will be served by APs. In order to obtain the closed-form expression of $q_{i,j,u}$ in Eq. (6), we have the following proposition.

Proposition 1: For user u, there are three possible situations at u by decoding signals from two nearest APs with SIC receiver, and the probability of at least one successful decoding situations, i.e., $\mathcal{A}_1$ and $\mathcal{A}_2$, can be computed by

$$\begin{cases} Pr\{\gamma_{u,0} \geq \tau, \gamma_{u,1} \geq \tau\} = \varphi exp(-\lambda\omega), \\ Pr\{\gamma_{u,0} \geq \tau, \gamma_{u,1} < \tau\} = \varphi[exp(-\lambda d_0) - exp(-\lambda\omega)], \end{cases} \tag{7}$$

where $\varphi = \frac{d_1}{\tau d_0 + d_1}$, $\lambda = \frac{\tau N_0}{2\sigma^2 P_{j,u}}$, $\omega = d_0(\tau+1) + d_1$, $d_0 = (1 + r_0^\alpha)$, $d_1 = (1 + r_1^\alpha)$ and τ denotes decoding threshold.

Proof: According to the conditional probability formula and Eq. (4), when both transmissions are successful, we have

$$\begin{aligned} &Pr(\gamma_{u,0} \geq \tau, \gamma_{u,1} \geq \tau) \\ &= Pr(\gamma_{u,1} \geq \tau) Pr(\gamma_{u,0} \geq \tau | \gamma_{u,1} \geq \tau) \\ &= Pr(|h_1|^2 \geq \xi) \cdot Pr(|h_0|^2 \\ &\quad \geq \frac{\tau(|h_1|^2(1 + r_1^\alpha)^{-1} P_{ij} + N_0)}{(1 + r_0^\alpha)^{-1} P_{j,u}} | \gamma_{u,1} \geq \tau) \\ &\overset{(a)}{=} \int_\xi^\infty Pr(x) Pr(|h_0|^2 \geq \frac{\tau(x(1 + r_1^\alpha)^{-1} P_{j,u} + N_0)}{(1 + r_0^\alpha)^{-1} P_{j,u}} | x) dx \\ &\overset{(b)}{=} \frac{1}{2\sigma^2} \int_\xi^\infty exp(-\frac{x}{2\sigma^2}) exp(-\frac{\tau(x(1 + r_1^\alpha)^{-1} P_{j,u} + N_0)}{(1 + r_0^\alpha)^{-1} 2\sigma^2 P_{j,u}}) dx \\ &= \frac{1}{2\sigma^2} \int_\xi^\infty exp(-\frac{x[\tau \frac{d_0}{d_1} + 1] P_{j,u} + \tau N_0 d_0}{2\sigma^2} P_{j,u}) dx \\ &= \frac{P_{j,u}}{\tau P_{j,u}(\tau \frac{1 + r_0^\alpha}{1 + r_1^\alpha} + 1)} exp(-\frac{\tau N_0[\tau d_0 + d_1] + \tau N_0 d_0}{2\sigma^2 P_{j,u}}) \\ &= \varphi exp(-\frac{\tau N_0(\tau d_0 + d_1) + \tau N_0 d_0}{2\sigma^2 P_{j,u}}). \end{aligned} \tag{8}$$

Meanwhile, for the situation of successful transmission only from the nearest AP, we have

$$\begin{aligned} &Pr(\gamma_{u,0} \geq \tau, \gamma_{u,1} < \tau) \\ &\quad = Pr(\gamma_{u,1} < \tau) Pr(\gamma_{u,0} \geq \tau | \gamma_{u,1} < \tau) \\ &\quad \overset{(a)}{=} \int_0^\xi Pr(x) Pr(|h_0|^2 \geq \frac{\tau(x(1 + r_1^\alpha)^{-1} P_{ij} + N_0)}{(1 + r_0^\alpha)^{-1} P_{ij}} | x) dx \\ &\quad = \varphi[exp(\frac{-\tau N_0 d_0}{2\sigma^2 P_{j,u}}) - exp(-\frac{\tau N_0(\tau d_0 + d_1) + \tau N_0 d_0}{2\sigma^2 P_{j,u}})], \end{aligned} \tag{9}$$

where $\xi = \frac{\tau N_0}{(1 + r_1^\alpha)^{-1} P_{ij}}$. Here, we let $x = |h_1|^2$ in (a) and utilize the following conclusion in (b): For the square of Reyleigh fading variable y, the PDF can be computed by $f_Y(y) = \frac{1}{2\sigma^2} exp(-\frac{y}{2\sigma^2})$, and the CDF can be computed by $F_Y(y) = Pr(Y \leq y) = 1 - exp(\frac{y^2}{2\sigma^2})$. □

According to **Proposition 1**, by substituting Eq. (7) into Eq. (5), we have

$$q_{i,j,u} = \begin{cases} m_{i,j} \varphi(exp(-\lambda d_0) + exp(-\lambda\omega)), \ m_{i,j} \in \mathbb{L}_1, \\ \varphi(m_{i,j} exp(-\lambda d_0) + (c - m_{i,j}) \\ \qquad \cdot exp(-\lambda\omega)), \ m_{i,j} \in \mathbb{L}_2. \end{cases}$$

Now we can compute the total traffic served by APs.

For the satellite, broadcast transmission is used to serve users in its coverage region. Since the satellite can broadcast cached blocks and also cached blocks in terrestrial APs through cache sharing, the traffic offloaded from satellite through broadcast can be computed by

$$B^{\mathrm{Sat}} = \textstyle\sum_{i \in \mathcal{F}} (m_i^{\mathrm{Sat}} + m_i^* a_i^{\mathrm{toS}}) \rho_i^{\mathrm{Sat}} U_i, \tag{10}$$

where ρ_i^{Sat} denotes the probability of terrestrial users for successfully decoding the signal from the satellite, $U_i = \sum_{j \in \mathcal{J}} \sum_{u \in \mathcal{U}_j} p_{i,j}$ represents the average users that request i-th file in satellite coverage region and $m_i^* = \max_{j \in \mathcal{J}} \{m_{i,j}\}$ represents the maximum number of cached blocks of i-th file in APs. In order to avoid duplicate caching for the same file in APs and the satellite, we also have a constraint as

$$m_i^{\mathrm{Sat}} \leq 1 - \mathcal{K}_{i,u} m_i^*, \ \forall i \in \mathcal{F}, \ \forall u \in \mathcal{U}, \tag{11}$$

where $\mathcal{K}_{i,u}$ denotes the number of APs serving user u for i-th file and $\mathcal{K}_{i,u} = \begin{cases} 1, m_{i,j} = c \\ 2, 0 < m_{i,j} < c \end{cases}$ should be satisfied.

B. Energy Consumption

For energy efficiency concerns in integrated satellite/terrestrial RAN, we compute the energy consumed by both APs and the satellite. Since APs can provide multiple access to each user and cache sharing to the satellite, the energy consumption of terrestrial transmission can be computed by

$$\begin{aligned} P^{\mathrm{T}} = &\sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{J}} \sum_{u \in \mathcal{U}_j} \mathcal{K}_{i,u} P_{j,u} p_{i,j} (1 - a_i^{\mathrm{toS}}) \\ &+ \sum_{i \in \mathcal{F}} P_i^{\mathrm{toS}}(m_i^*) a_i^{\mathrm{toS}}, \end{aligned} \tag{12}$$

where $P_i^{\mathrm{toS}}(m_i^*)$ denotes the power consumption for forwarding m_i^* cached blocks of i-th file from AP to the satellite through MBS.

For the satellite, energy consumption is an important concern since it has the direct impact on the system design and satellite lifespan. In order to offload traffic from cellular network, it is inevitable that the satellite's energy consumption will increase for more content data transmission over high speed channel. The energy consumption of satellite is composed of two parts, i.e., broadcast cached blocks in the satellite and blocks retrieved from APs. Thus, the energy consumption of satellite can be computed by

$$P^{\mathrm{Sat}} = \sum_{i \in \mathcal{F}} Pr(U_i \geq 1)[P_i^{\mathrm{Sat}}(m_i^{\mathrm{Sat}}) + P_i^{\mathrm{Sat}}(m_i^*) a_i^{\mathrm{toS}}], \tag{13}$$

where $Pr(U_i \geq 1) = (1 - \Pi_{j \in \mathcal{J}} (1 - p_{ij})^{|\mathcal{U}_j|})$ denotes the probability that at least one user requests i-th file, and P_i^{Sat} is the power consumption for broadcast m_i^{Sat} blocks of i-th file and it can be computed by Eq. (2).

C. Objective

After we obtain the formulation of traffic offloading and energy consumption from APs and satellite, the total traffic served by APs and satellite can be computed by

$$\mathcal{B}_{\mathrm{total}} = B^{\mathrm{T}} + B^{\mathrm{Sat}}, \tag{14}$$

and the total energy consumption can be computed by

$$\mathcal{P}_{\mathrm{total}} = P^{\mathrm{T}} + P^{\mathrm{Sat}}. \tag{15}$$

In order to maximize the energy efficiency, our objective can be formulated as

$$\textbf{Problem 1}: \max_{\mathbf{a},\mathbf{m},\mathbf{P}} \quad \frac{\mathcal{B}_{\text{total}}(\mathbf{a},\mathbf{m},\mathbf{P})}{\mathcal{P}_{\text{total}}(\mathbf{a},\mathbf{m},\mathbf{P})}, \tag{16}$$

$$s.t. \quad \sum\nolimits_{i\in\mathcal{F}} p_{i,j} = 1, \ \forall j \in \mathcal{J}, \tag{C1}$$

$$\sum\nolimits_{i\in\mathcal{F}} m_{i,j} \le M, \ \forall j \in \mathcal{J}, \tag{C2}$$

$$\sum\nolimits_{i\in\mathcal{F}} m_i^{\text{Sat}} \le M^{\text{Sat}}, \tag{C3}$$

$$m_i^{Sat} \le 1 - \mathcal{K}_{i,u} m_i^*, \ \forall i \in \mathcal{F}, \ u \in \mathcal{U}, \tag{C4}$$

$$\sum\nolimits_{i\in\mathcal{F}} (P_i^{\text{Sat}}(m_i^{\text{Sat}}) + a_i^{\text{toS}} P_i^{\text{Sat}}(m_i^*)) \le P_{\text{total}}^{\text{Sat}}, \tag{C5}$$

$$P_{j,u} \le P_{\max}, P_i^{\text{Sat}} \le P_{\max}^{\text{Sat}}, \tag{C6}$$

$$a_i^{\text{toS}} \in \{0,1\}, m_i^{\text{Sat}}, m_{i,j} \in \mathcal{N}^+, \tag{C7}$$

where $\mathbf{a} = \{a_i^{\text{toS}}\}, \mathbf{m} = \{m_{i,j}, m_i^{\text{Sat}}\}$ and $\mathbf{P} = \{P_{j,u}\}$ $(i \in \mathcal{F}, \ j \in \mathcal{J}, \ u \in \mathcal{U})$ represent cache sharing vector, block placement vector and power allocation vector, respectively. $P_{\text{total}}^{\text{Sat}}$ denotes the power constraint of each LEO satellite, and $P_{\max}$ and $P_{\max}^{\text{Sat}}$ represent the maximum transmission power for AP and the satellite, respectively.

V. Problem Analysis and Algorithm Design

A. Problem Analysis

Problem 1 is a nonlinear fractional programming problem consisting of integer variables and also continuous variables, which is difficult to derive a global optimal solution. Considering that a number of cells and APs are involved in an integrated satellite/terrestrial RAN, in order to solve the problem in polynomial time, we introduce a parametric method to transform the original nonlinear fractional objective function [46], [47] in **Problem 1** and respectively consider three solvable subproblems. At first, a problem equivalence theorem is given as follows. For notation convenience, $\mathcal{G}$ is defined as the set of feasible solutions of **Problem 1** and $g = (\mathbf{a}, \mathbf{m}, \mathbf{P}) \in \mathcal{G}$.

Theorem 1 (Problem Equivalence): η^* is achieved if and only if

$$\max_{g^*} \ \mathcal{B}_{total}(g) - \eta^* \mathcal{P}_{total}(g) = \mathcal{B}_{\text{total}}(g^*) - \eta^* \mathcal{P}_{\text{total}}(g^*) = 0, \tag{17}$$

*where g^* is the optimal solution of **Problem 1** and $\eta^* = \mathcal{B}_{\text{total}}(g^*)/\mathcal{P}_{\text{total}}(g^*)$.*

Proof: By referring to [47], [48], we give the following proof process.

At first, we prove the sufficient condition of **Theorem 1**. Since $\eta^* = \mathcal{B}_{\text{total}}(g^*)/\mathcal{P}_{\text{total}}(g^*)$ represents the maximal energy efficiency performance, it is obvious that η^* holds, i.e.,

$$\eta^* = \frac{\mathcal{B}_{total}(g^*)}{\mathcal{P}_{total}(g^*)} > \frac{\mathcal{B}_{total}(g)}{\mathcal{P}_{total}(g)}, \tag{18}$$

where $g \in \mathcal{G}$ is the feasible solution for solving **Problem 1**, and $\mathcal{P}_{total}(g') > 0$ (total power consumption can not be negative). Then, according to (18), we can derive the following formulas

$$\begin{cases} \mathcal{B}_{\text{total}}(g) - \eta \mathcal{P}_{\text{total}}(g) \le 0, \\ \mathcal{B}_{\text{total}}(g^*) - \eta^* \mathcal{P}_{\text{total}}(g^*) = 0. \end{cases} \tag{19}$$

Consequently, we can conclude that $\max_{g^*} \ \mathcal{B}_{total}(g) - \eta^* \mathcal{P}_{total}(g) = 0$ can be achieved by the optimal solution g^*. The sufficient condition is proved.

Secondly, we prove the necessary condition of **Theorem 1**. Let $\widetilde{g}$ be the optimal solution of the transformed objective function, we can have $\mathcal{B}_{\text{total}}(\widetilde{g}) - \eta^* \mathcal{P}_{\text{total}}(\widetilde{g}) = 0$. For any feasible solution g, they can be expressed as

$$\mathcal{B}_{\text{total}}(g) - \eta^* \mathcal{P}_{\text{total}}(g) \le \mathcal{B}_{\text{total}}(\widetilde{g}) - \eta^* \mathcal{P}_{\text{total}}(\widetilde{g}) = 0. \tag{20}$$

The above inequality can be derived as

$$\frac{\mathcal{B}_{total}(g)}{\mathcal{P}_{total}(g)} \le \eta^* \text{ and } \frac{\mathcal{B}_{total}(\widetilde{g})}{\mathcal{P}_{total}(\widetilde{g})} = \eta^*. \tag{21}$$

Therefore, the optimal solution $\widetilde{g}$ for the transformed objective function is also the optimal solution for the original objective function. The necessary condition of **Theorem 1** is proved. □

Based on the optimal condition stated in **Theorem 1**, the original problem can be transformed to **Problem 2** if we can find the optimal value η^*.

$$\textbf{Problem 2}: \max_{\mathbf{a},\mathbf{m},\mathbf{P}} \quad \mathcal{B}_{\text{total}}(g) - \eta^* \mathcal{P}_{\text{total}}(g),$$
$$s.t. \text{ Conditions } C1, C2, C3, C4, C5, C6, C7 \text{ are satisfied.} \tag{22}$$

Then an iteration algorithm is proposed to update η in **Algorithm 1** and ε is a constant which can make the algorithm converge to η^* with a superlinear convergence rate [47]. Inspired by Block Coordinate Descent (BCD) [49], [50], we divide problem 2 into three sub-problems according to the relationship of each variable. To optimize a multi-variable objective in BCD method, the coordinates of variables are first partitioned into blocks.[1] Through another iterative process (i.e., line 4-8 in **Algorithm 1**), we optimize the objective in terms of one of the coordinate blocks while the other coordinates are fixed at each iteration. Moreover, three algorithms are proposed to solve three sub-problems of block placement, power allocation and cache sharing, respectively. Then we can apply them to **Algorithm 1** to obtain a global sub-optimal solution in an iterative manner.

B. Block Placement Algorithm

In order to obtain the optimal block placement vector $\mathbf{m}^*$ in the integrated satellite/terrestrial RAN, we consider a two-step solving process as described in **Algorithm 2**. The basic idea is that, we first find the optimal block placement vector for terrestrial APs in each cell. Then, we try to find the optimal block placement vector on the satellite. Since the terrestrial block placement $m_{i,j}$ and satellite block placement m_i^{Sat} are coupled, we also consider the influence between $m_{i,j}$ and m_i^{Sat}. In the following, we will specifically analyze the property of block placement sub-problem and explain why our algorithm can obtain the optimal placement vector $\mathbf{m}^*$.

[1] Note that the block here is different from content block we mentioned in Section III-B (Coded Caching Model).

Algorithm 1 Iteration Algorithm for **Problem 2**

```
Input: The maximum number of iteration $T_1$, $T_2$ and
        convergence factor $\delta$, $\delta'$;
Output: Optimal solution $g^*$ for each sub-problem;
Initialize: Take $\eta_0$ and set $t = t' = 0$;
while $\mathcal{B}_{\text{total}} - \eta_t \mathcal{P}_{\text{total}} > \delta$ and $t < T_1$ do
    Set $t' = 0$;
    while $\mathcal{E}_{t'} - \mathcal{E}_{t'-1} < \delta'$ and $t' < T_2$ do
        Apply sub-algorithm to obtain solution of
        $\mathbf{a}_{t'+1}, \mathbf{m}_{t'+1}$ and $\mathbf{P}_{t'+1}$ for each sub-problem
        while fixing $(\mathbf{m}_{t'}, \mathbf{P}_{t'})$, $(\mathbf{a}_{t'+1}, \mathbf{P}_{t'})$ and
        $(\mathbf{a}_{t'+1}, \mathbf{m}_{t'+1})$ respectively;
        $\mathrm{g}_{t'+1} = (\mathbf{a}_{t'+1}, \mathbf{m}_{t'+1}, \mathbf{P}_{t'+1})$;
        $\mathcal{E}_{t'+1} = \mathcal{B}_{\text{total}}(\mathrm{g}_{t'+1}) - \eta_t \mathcal{P}_{\text{total}}(\mathrm{g}_{t'+1})$, $t' = t' + 1$;
    end
    Let $g_t = (\mathbf{a}_{t'}, \mathbf{m}_{t'}, \mathbf{P}_{t'})$, take $\eta_{t+1} = \frac{\mathcal{B}_{\text{total}}(g_t)}{\mathcal{P}_{\text{total}}(g_t)} + \varepsilon$,
    $t = t + 1$;
end
```

1) Terrestrial Block Placement: At first, we only consider the optimal block placement $m_{i,j}$ for APs, and the objective function related to $m_{i,j}$ in Eq. (22) can be written as

$$Q_{i,j}^1(m_{i,j}, \eta_t) = \sum_{u \in \mathcal{U}_j} (1 - a_i^{\text{toS}})(q_{i,j,u} - \eta_t \mathcal{K}_{i,u} P_{j,u}) p_{i,j} - \eta_t P_i^{\text{toS}}(m_i^*) a_i^{\text{toS}}. \quad (23)$$

Terrestrial block placement can be classified as a multiple-choice knapsack problem. By utilizing the monotonicity of $Q_{i,j}^1$, a specific process of finding the optimal terrestrial block placement is given in line 2-18 of **Algorithm 2**. For each loop in line 3-17, the algorithm first calculates the gain of caching one more block for each file in line 5, and finds the most valuable caching block with maximum gain by implementing line 7-16. Note that only one block will be selected to be added to cache for each loop in line 3-17, and $\mathcal{I}$ represents the set of files that could not be cached anymore. The analysis of the property of $Q_{i,j}^1$ is described as follows.

Proposition 2: $q_{i,j,u}$ is a monotonic increasing and piece-wise function of variable $m_{i,j}$ with fixed SINR threshold τ.

Proof: At first, we relax $m_{i,j}$ as continuous variable and let $f(m_{i,j})$ represent $q_{i,j,u}$. When $m_{i,j} \in \mathbb{L}_1 = [0, c/2]$, the derivative $\frac{df(m_{i,j})}{dm_{i,j}} = \frac{d_1}{\tau d_0 + d_1}[exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}} d_0) + exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}}(d_0(\tau + 1) + d_1))] > 0$; When $m_{i,j} \in \mathbb{L}_1 = (c/2, c]$, $\frac{df(m_{i,j})}{dm_{i,j}} = \frac{d_1}{\tau d_0 + d_1}[exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}} d_0) - exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}}(d_0(\tau + 1) + d_1))]$, due to $d_0 < d_0(\tau + 1) + d_1$, we have $-\frac{\tau N_0}{2\sigma^2 P_{j,u}} d_0 > -\frac{\tau N_0}{2\sigma^2 P_{j,u}}(d_0(\tau + 1) + d_1)$. Thus, $exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}} d_0) > exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}}(d_0(\tau + 1) + d_1))$, and the derivative $\frac{df(m_{i,j})}{dm_{i,j}} > 0$ can be obtained. Meanwhile, we assume δ is an arbitrary small value, then we have $f(c/2 + \delta) - f(c/2) = \delta exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}} d_0) - \delta exp(-\frac{\tau N_0}{2\sigma^2 P_{j,u}}(d_0(\tau + 1) + d_1)) > 0$. In this case, **Proposition 2** is proved. □

Corollary 1: $Q_{i,j}^1$ is a monotonic increasing and piece-wise function of variable $m_{i,j}$ with fixed SINR threshold τ.

Proposition 3: The optimal placement vector for maximizing $Q_{i,j}^1$ satisfies that, $m_{1,j} > m_{2,j}$ for any given file popularity $p_{1,j} > p_{2,j}$ when $a_1^{\text{toS}} = a_2^{\text{toS}} = 0$.

Proof: We prove this proposition by contradiction. Let $a < b$ and $a, b \in \mathcal{N}^+$. Assume that there are two files $f_1, f_2 \in \mathcal{F}$ and $p_{1,j} > p_{2,j}$ in cell j. We further suppose that $\mathbf{m}^* = \{m_{1,j} = a, m_{2,j} = b\}$ is an optimal caching strategy. Accordingly, we have another caching strategy $\mathbf{m}' = \{m_{1,j} = b, m_{2,j} = a\}$, which is obviously not the optimal caching strategy. Based on the above supposition, we have

$$Q_j^1(\mathbf{m}^*) = \sum\nolimits_{u \in \mathcal{U}_j} [p_{1,j} q_{i,j,u}\ (a) + p_{2,j} q_{i,j,u}(b) - \eta \mathcal{K}'_{a,b} P_{j,u}].$$

Then we exchange the placement for file f_1, f_2, and we have

$$Q_j^1(\mathbf{m}') = \sum\nolimits_{u \in \mathcal{U}_j} [p_{1,j} q_{i,j,u}\ (b) + p_{2,j} q_{i,j,u}(a) - \eta \mathcal{K}'_{b,a} P_{j,u}],$$

where $\mathcal{K}'_{a,b} = \mathcal{K}_{1,u}(a) + \mathcal{K}_{2,u}(b)$ and $\mathcal{K}'_{b,a} = \mathcal{K}_{1,u}(b) + \mathcal{K}_{2,u}(a)$. Then, we have $\mathcal{K}'_{a,b} = \mathcal{K}'_{b,a}$. At last, we make the comparison between $Q_j^1(\mathbf{m}^*)$ and $Q_j^1(\mathbf{m}')$:

$$\begin{aligned} & Q_j^1(\mathbf{m}^*) - Q_j^1(\mathbf{m}') \\ & = p_{1,j} q_{i,j,u}(a) + p_{2,j} q_{i,j,u}(b) - [p_{1,j} q_{i,j,u}(b) + p_{2,j} q_{i,j,u}(a)], \\ & = p_{1,j}[q_{i,j,u}(a) - q_{i,j,u}(b)] - p_{2,j}[q_{i,j,u}(a) - q_{i,j,u}(b)], \\ & = \sum\nolimits_{u \in \mathcal{U}_j} (p_{1,j} - p_{2,j})[q_{i,j,u}(a) - q_{i,j,u}(b)]. \end{aligned} \quad (24)$$

From the assumption and **Proposition 2**, we know that $p_{1,j} > p_{2,j}$ and $q_{i,j,u}(a) < q_{i,j,u}(b)$. Then we can further get $Q_j^1(\mathbf{m}^*) < Q_j^1(\mathbf{m}')$, so $\mathbf{m}'$ represents a better block placement strategy, which leads to a contradiction to our previous assumption that the optimal block placement $\mathbf{m}^*$ should have $m_{1,j} < m_{2,j}$ when $p_{1,j} > p_{2,j}$. Thus, the optimal placement should satisfy $m_{1,j} > m_{2,j}$ for any given file popularity $p_{1,j} > p_{2,j}$ when $a_1^{\text{toS}} = a_2^{\text{toS}} = 0$. **Proposition 3** is thus proved. □

2) Satellite Block Placement: Considering the broadcast transmission of satellite, satellite block placement will influence the terrestrial block placement and constraint (C4) should be satisfied. Similar to Eq. (23), the objective related to the satellite can be written as

$$\begin{aligned} & Q_i^2(m_i^{\text{Sat}}, \eta_t) \\ & \quad = (m_i^{\text{Sat}} + m_i^* a_i^{\text{toS}}) U_i - \eta_t (P_i^{\text{Sat}}(m_i^{\text{Sat}}) + P_i^{\text{Sat}}(m_i^*) a_i^{\text{toS}}). \end{aligned} \quad (25)$$

Proposition 4: Q_i^2 is a concave function of variable m_i^{Sat}.

Proof: P_i^{Sat}, related to variable m_i^{Sat} in Q_i^2, can be obtained by Eq. (2). By computing second order derivative of P_i^{Sat}, we can obtain that $d^2 P_i^{\text{Sat}} / d(m_i^{\text{Sat}})^2 > 0$ and $-\eta P_i^{\text{Sat}}$ must be a concave function. Considering that the $m_i^{\text{Sat}} U_i$ is a linear function, Q_i^2 is a concave function can be proven according to convexity preserving property [51]. □

Proposition 5: The optimal placement vector for maximizing $Q_{i,j}^2$ satisfies that, $m_1^{\text{Sat}} > m_2^{\text{Sat}}$ for any given file popularity $U_1 > U_2$ when $a_1^{\text{toS}} = a_2^{\text{toS}} = 0$.

Proof: Similar to the proof of **Proposition 3**. □

Algorithm 2 Global Block Placement Algorithm

```
   Input: Cache Size $M$ and $M^{\text{Sat}}$, number of original
          blocks $c$ and $\eta_t$ at $t$-th iteration;
   Output: The optimal block placement vector $\mathbf{m}^*$;
 1 Initialize:
   Set $\mathcal{I} = \oslash$, $m_{i,j} = 0$ and $m_i^{\text{Sat}} = 0$, $\forall i \in \mathcal{F}, j \in \mathcal{J}$;
   Sort the file popularity of each cell;
 2 for $j \in \mathcal{J}$ do
 3     while $\sum_{i\in\mathcal{F}} m_{i,j} \le M$ and $|\mathcal{I}| < |\mathcal{F}|$ do
 4         for $i \in \mathcal{F}$ do
 5             Compute
               $\delta_i = Q^1_{i,j}(m_{i,j}+1, \eta_t) - Q^1_{i,j}(m_{i,j}, \eta_t)$;
 6         end
 7         while true do
 8             Let $i' = \mathrm{argmax}_{\mathrm{i}\in\{\mathcal{F}-\mathcal{I}\}}\delta_{\mathrm{i}}$;
 9             if $m_{i',j} < c$ then
10                 Set $m_{i',j} = m_{i',j} + 1$;
11                 break;
12             end
13             else
14                 Set $\mathcal{I} = \mathcal{I}\bigcup\{i'\}$;
15             end
16         end
17     end
18 end
19 Sort the file set $\mathcal{F}'$ according to the value of $U_i$;
20 for $i \in \mathcal{F}'$ do
21     while $\sum_{i\in\mathcal{F}} m_i^{\text{Sat}} \le M^{\text{Sat}}$ do
22         Compute the derivative of $Q_i^2(m_i^{\text{Sat}}, \eta_t)$ at
           $m_i^{\text{Sat}} = 0$;
23         if $f'(m_i^{\text{Sat}}) \le 0$ then
24             Continue loop in line 20;
25         end
26         if $m_i^{\text{Sat}} + 1 > c - \mathcal{K}_{i,u} m_i^*$ then
27             Compute the differences $\Delta_1 =$
               $\sum_{j\in\mathcal{J}^*}[Q^1_{i,j}(m_i^*, \eta_t) - Q^1_{i,j}(m_i^* - \mathcal{K}_{i,u}, \eta_t)]$,
               and $\Delta_2 = Q_i^2(m_i^{\text{Sat}}+1, \eta_t) - Q_i^2(m_i^{\text{Sat}}, \eta_t)$;
28             Find the optimal placement $\Delta m_{i',j}$ for APs to
               replace $\mathcal{K}_{i,u}$ cached blocks of $i$-th file in $j \in \mathcal{J}^*$
               cell with $\mathcal{K}_{i,u}$ blocks of $i'$-th file in $j \in \mathcal{J}^*$ cell;
29             Compute the increasing gain
               $\Delta_1 = \Delta_1 + \sum_{j\in\mathcal{J}^*}[Q^1_{i',j}(m_{i',j} + \Delta m_{i',j}, \eta_t) -$
               $Q^1_{i',j}(m_{i',j}, \eta_t)]$;
30             if $\Delta_2 > \Delta_1$ then
31                 if $m_i^{\text{Sat}} + 1$ satisfies constraint (C4)
                   and (C5) then
32                     Set $m_i^{\text{Sat}} = m_i^{\text{Sat}} + 1$;
33                     Set $m_{i',j} = m_{i',j} + \Delta m_{i',j}$ and
                       $m_{i,j} = m_{i,j} - \mathcal{K}_{i,u}$ for $j \in \mathcal{J}^*$;
34                 end
35             end
36         end
37         else
38             if $m_i^{\text{Sat}} + 1$ satisfies constraint (C4) and (C5)
               then
39                 Set $m_i^{\text{Sat}} = m_i^{\text{Sat}} + 1$;
40             end
41         end
42     end
43 end
```

Considering the property of $Q_i^2(m_i^{\text{Sat}}, \eta_t)$ and the coupling of terrestrial and satellite block placement, **Algorithm 2** cooperatively considers terrestrial and satellite block placement to obtain the optimal solution after obtaining terrestrial block placement in line 2-18. Let $\mathcal{J}^*$ represent terrestrial cells which cache m_i^* blocks of i-th file. According to the average request number of i-th file, i.e., U_i, **Algorithm 2** adds blocks one by one and computes the corresponding caching gain based on the ordered sequence of $\mathcal{F}'$ in line 19. In line 22-25, the algorithm first computes the derivative of $Q_i^2(m_i^{\text{Sat}}, \eta_t)$. If the derivative is not positive, it means that there is no gain to cache i-th file on the satellite. Thus, the algorithm would terminate the rest of the procedure and continue another loop in line 20-43. In line 26, $c - \mathcal{K}_{i,u} m_i^*$ represents the number of unique blocks that can be directly cached on the satellite. If $m_i^{\text{Sat}}+1 \le c-\mathcal{K}_{i,u}m_i^*$, i-th file could be directly cached on the satellite once, and constraints (C4) and (C5) are satisfied. Otherwise, the algorithm compares the gain of caching i-th file on the satellite and the gain of caching i-th file in terrestrial APs. According to the comparison result of the gain of caching i-th file in terrestrial APs and on the satellite (i.e., Δ_1 and Δ_2), when $\Delta_2 > \Delta_1$ and constraints (C4) and (C5) are satisfied, i-th file will be cached on the satellite (i.e., $m_i^{\text{Sat}} = m_i^{\text{Sat}}+1$) as shown in line 30-34. Accordingly, when $\Delta_2 \le \Delta_1$ or constraints (C4) and (C5) are not satisfied, i-th file would not be cached on the satellite and the algorithm continues another loop in line 20-43. In this way, the algorithm can find the optimal block placement vector $\mathbf{m}^*$ in a greedy manner.

3) Computational Complexity: The complexity of the optimal exhaustive search for $\mathbf{m}^*$ is $\mathcal{O}(2^{(M\cdot|\mathcal{J}|+M^{\text{Sat}})\cdot N})$ and the complexity of our block placement algorithm is $\mathcal{O}((M + M^{\text{Sat}}) \cdot (NlogN) \cdot |\mathcal{J}|)$.

C. Power Allocation Algorithm

Since satellite channel model is considered as AWGN channel with large-scale fading, the transmission power of the satellite can be computed by Eq. (2). In this case, we only consider the power allocation of APs. Here, we have the following theorem about power allocation sub-problem.

Theorem 2: Objective function (22) is a concave function when power allocation variable satisfying $P_{j,u} \ge \tau N_0 \omega_u/2, \forall j, u$; Objective function (22) is a convex function when power allocation variable satisfying $P_{j,u} \le \tau N_0 d_0/2, \forall j, u$, where $\omega_u = d_{u,0}(\tau+1) + d_{u,1}$.

Proof: Let $f(P_{j,u})$ represents the objective in Eq. (22), then by computing the second order derivative of $f(P_{j,u})$, we have

$$\frac{d^2 f(P_{j,u})}{dP_{j,u}^2} = a_1 exp(-\lambda_1/P_{j,u})(\tau N_0 d_0 - 2P_{j,u}) \\ +a_2 exp(-\lambda_2/P_{j,u})(\tau N_0 \omega_u - 2P_{j,u}),$$

where $a_1 = \frac{\tau N_0 d_0 (2\sigma^2)^2 m_{i,j}\varphi_u}{(2\sigma^2 P_{j,u})^4}$, $a_2 = \frac{\tau N_0 \omega_{j,u} m_{i,j}\varphi_u}{(2\sigma^2 P_{j,u})^4}$ or $\frac{\tau N_0 \omega_{j,u}(N-m_{i,j})\varphi_u}{(2\sigma^2 P_{j,u})^4}$, $\lambda_1 = \tau N_0 d_{u,0}/2\sigma^2$, $\lambda_2 = \tau N_0 \omega_u/2\sigma^2 > \lambda_1$ and $\omega_u = d_{u,0}(\tau+1) + d_{u,1}$. Since $\tau N_0 \omega_u > \tau N_0 d_{u,0}, \forall u \in \mathcal{U}$ is satisfied, when $P_{j,u} \ge \tau N_0 \omega_u/2$, we have

$\frac{d^2 f(P_{j,u})}{dP_{j,u}^2} \geq 0$; When $P_{j,u} \leq \tau N_0 d_0/2$, we have $\frac{d^2 f(P_{j,u})}{dP_{j,u}^2} \leq 0$. Thus, **Theorem 2** is proven. □

As described in Theorem 2, power allocation sub-problem satisfies concavity or convexity when $P_{j,u} \in \mathcal{P} = [P_{max}, \tau N_0 \omega_u/2] \cup [\tau N_0 d_{u,0}/2, 0]$. In order to efficiently solve power allocation sub-problem, we assume that the optimal power allocation could be found in partial solution space $\mathcal{P}$ with a great probability. If we assume the gap between $\tau N_0 \omega_u/2$ and $\tau N_0 d_{u,0}/2$ as β, $\beta = \tau N_0 \omega_u/2 - \tau N_0 d_{u,0}/2 = \tau N_0(\tau d_{u,0} + d_{u,1})/2$. Typically, $N_0 \approx 3.981 * 10^{-14}$ (power spectral density of noise $= -174dBm/Hz \approx 3.981 * 10^{-21}W/Hz$, and subcarrier bandwidth = $10MHz$) and $\tau = 0.414$ (file size = $50kb$, time slot = $10ms$ and subcarrier bandwidth = $10MHz$). Considering $d_{u,0}, d_{u,1} \leq 1$ in small cell ($1km * 1km$ cell size is considered in our evaluation), the gap β will be smaller than $1 * 10^{-13}$. In this case, our assumption is reasonable. Hence, the objective function becomes a combination of a concave and convex function. By setting variable value range, interior-point method can be applied to the problem and obtain the optimal power allocation efficiently.

According to our numerical results, when request consistency degree becomes low, the performance of the proposed algorithm in terms of traffic offloading will decrease and become worse than the baseline.[2] We further consider that the performance of traffic offloading is also important for practical application. In this case, we set the minimum threshold of transmission power to improve the performance of energy efficiency without any sacrifice of traffic offloading compared with the baseline. By doing this, regardless of the degree of request consistency, our numerical results show that the proposed algorithm with threshold can outperform the baseline both in terms of energy efficiency and traffic offloading.

D. Cache Sharing Algorithm

In order to determine cache sharing decision, the related objective function can be written as

$$Q_i^3(a_i, \eta_t) = a_i^{\text{toS}}[m_i^* U_i - \eta_t(P_i^{\text{Sat}} + P_i^{\text{toS}}(m_i^*))] + \sum_{j\in\mathcal{J}} \sum_{u\in\mathcal{U}_j} (1 - a_i^{\text{toS}})[p_{i,j} q_{i,j,u} - \mathcal{K}_{i,u} P_{j,u} p_{i,j}]. \quad (26)$$

Considering the constraint in (C7), the cache sharing sub-problem can be treated as 0-1 knapsack problem with power constraint. Although there are some general algorithms to obtain the optimal solution [52], these algorithms can only provide non-polynomial complexity such as $\mathcal{O}(2^N \cdot |\mathcal{J}|)$ for branch and search algorithm. In order to accelerate **Algorithm 1** in each iteration, we propose **Algorithm 3** to efficiently obtain a sub-optimal solution through a greedy approach. The algorithm first finds maximum cached number of blocks for each file, i.e., $m_i^* = \max_{j\in\mathcal{J}}\{m_{i,j}\}$ in line 2. To obtain the gain by enabling cache sharing for i-th file in line 3-8, the original gain by caching i-th file in terrestrial APs can be deducted. After that, according to the sorting result in line 10, the algorithm further tries to find the best choice for cache sharing with maximum gain. If the benefit of enabling cache sharing of i-th file is positive (i.e., $\mathcal{E}_i > 0$) and $a_i^{\text{toS}} + 1$ also satisfies constraint (C5), i-th file will be selected for cache sharing. Otherwise, the algorithm continues the loop in line 11-15 to find other files for possible cache sharing. The complexity of cache sharing algorithm is $\mathcal{O}((N \cdot (logN + |\mathcal{J}|))$. Meanwhile, we also provide performance comparison between the proposed scheme and the scheme without cache sharing in the evaluation. Results show that our proposed **Algorithm 3** can also achieve significant improvement compared with the scheme without cache sharing.

Algorithm 3 Cache Sharing Algorithm

```
Input: $\eta_t$ at $t$-th iteration and block placement vector $\mathbf{m}^*$;
Output: Cache sharing vector $\mathbf{a}^*$;
1  for $i \in \mathcal{F}$ do
2      $m_i^* = \max_{j\in\mathcal{J}}\{m_{i,j}\}$;
3      if $m_i^* == 0$ then
4          $\mathcal{E}_i = 0$;
5      else
6          Compute the gain $\mathcal{E}_i = Q_i^3(a_i^{\text{toS}} + 1, \eta_t) - \sum_{j\in\mathcal{J}} Q_{i,j}^1(m_{i,j}, \eta_t)$;
7      end
8  end
9  end
10 Sort the file set $\mathcal{F}'$ according to the value of $\mathcal{E}_i$;
11 for $i \in \mathcal{F}'$ do
12     if $\mathcal{E}_i > 0$ and $a_i^{\text{toS}} + 1$ satisfies power constraint (C5) then
13         $a_i^{\text{toS}} = a_i^{\text{toS}} + 1$;
14     end
15 end
```

VI. Numerical Evaluation

In this section, we evaluate the performance of the proposed algorithm compared with traditional terrestrial scheme. We consider a LEO satellite covering an urban area with a number of cells, the square area of each cell is $1km \times 1km$. The distribution of users and APs in each cell follows two independent homogeneous Poisson point processes. Table I summarizes the key evaluation parameters, and the comparison schemes are given as follows.

- ***Random Caching***: Instead of performing **Algorithm 2**, APs and the satellite randomly choose files to cache.
- ***Pure Terrestrial***: Using our proposed algorithm in the environment with only APs.
- ***Non-Cooperative Scheme***: Only local popularity is considered in each cell and the satellite without considering constraint (C5), and cache sharing between APs and the satellite is also disabled.

Note that we first conduct the experiments that the users' requests in each cell j satisfy $p_{i,j} = p_i$ for the results in Fig. 5 and Fig. 6. After that, we investigate the system

[2] For the purpose of conducting performance comparison, the performance of traditional terrestrial schemes (i.e., Pure Terrestrial in the evaluation) is considered as the baseline.

TABLE I

THE SIMULATION PARAMETERS

Parameter	Value
The APs density λ_a	20APs/km^2
The users density λ_u	10-80users/km^2
Path loss exponent	2
Files in the content library	100
Original blocks	8
Satellite attitude	780km
Time slot	10ms
Carrier center frequency/Subcarrier bandwidth of satellite system	3GHz/10MHz
Carrier center frequency/Subcarrier bandwidth of terrestrial APs	2GHz/1MHz
The maximum transmission power $P_{\max}^{\text{Sat}}$ of satellite	46dBm
The maximum transmission power $P_{\max}$ of AP	24dBm
Power spectral density of noise	-174dBm/Hz

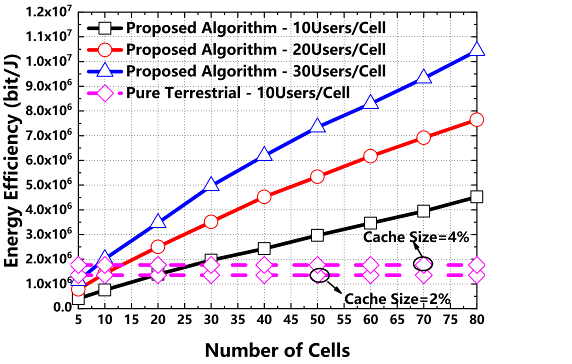


Fig. 4. Energy efficiency comparison between the proposed cooperative scheme and the pure terrestrial scheme (request consistency degree=0%).

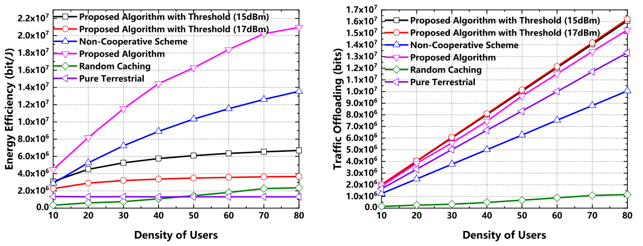
performance versus request inconsistency in Fig. 7, Fig. 8 and Fig. 9. The cache size in the evaluation represents storage capacity of the percentage of content library. Without special notification in the figures, our parameters are set as: shape parameter in Zipf distribution θ=1, cache size=2%, number of cells=50, and user density=40 $users/cell$. Different from terrestrial communication, the power spectral density of noise for satellite communication is set as -171dBm/Hz [53], [54].

A. Energy Efficiency vs Satellite Coverage

We make a preliminary comparison between our cooperative scheme and the pure terrestrial scheme. As shown in Fig. 4, the proposed scheme can achieve 3.3-7.7 times improvement with different user density in 80 km^2 coverage, and acquire approximately linear increase of energy efficiency along with the increase of satellite's coverage compared with the pure terrestrial scheme. However, the proposed scheme is worse than terrestrial scheme when users in the satellite coverage are less than 200. The reason is that the transmission power of satellite is greater than AP, then the energy efficiency of the satellite transmission is greater than the pure terrestrial scheme only if there are certain number of users served by satellite's broadcast.

B. Energy Efficiency & Traffic Offloading vs User Density

We make a performance comparison among different schemes in terms of energy efficiency as well as traffic offloading as shown in Fig. 5(a) and Fig. 5(b). Compared to the pure terrestrial scheme, our cooperative scheme can achieve 15.9 times improvement of energy efficiency and 14.6% improvement of traffic offloading with 80 users in each cell. For non-cooperative scheme, the performance is much worse than our cooperative scheme, which proves the effectiveness of our global block placement algorithm.



(a) User density vs energy efficiency (b) User density vs traffic offloading

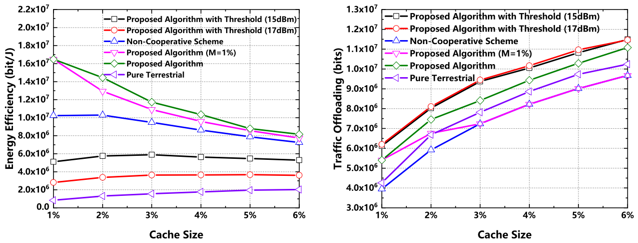
Fig. 5. Performance comparison vs user density (cache size=2%, number of cells=50, request consistency degree=100%).

In Fig. 5(b), we also find out that the performance of non-cooperative scheme in terms of traffic offloading is worse than the pure terrestrial scheme, which can be explained as follows. Due to the low energy efficiency of AP's unicast transmission, our proposed algorithm tends to satisfy users's requests through the satellite and avoid terrestrial transmission. Meanwhile, non-cooperative scheme disables cache sharing, which makes the satellite only broadcast the blocks cached by its own to users. Thus, extra satellite can barely bring any gains in terms of overall traffic offloading to terrestrial users.

Since the performance of traffic offloading is also important in practical application, we also set a minimum transmission power threshold of AP in our power allocation algorithm to achieve further traffic offloading improvement and make a comparison. In Fig. 5(a) and Fig. 5(b), we set it as 15dBm and 17dBm, respectively. Compared with the pure terrestrial scheme, results show that they can respectively achieve 20.5% and 21.4% improvements of traffic offloading and 2.8 and 5.1 times increasing of energy efficiency with 80 users in each cell.

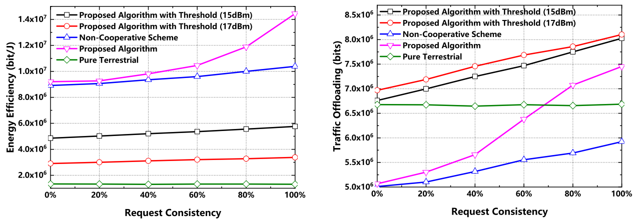
C. Energy Efficiency & Traffic Offloading vs Cache Size

To investigate the influence of cache size at APs and the satellite, we make a performance comparison as shown in Fig. 6, and "Proposed Algorithm ($M = 1\%$)" represents the cache size of APs in our proposed algorithm is fixed and only the cache size of satellite is changing. As shown in Fig. 6(a), energy efficiency of the pure terrestrial scheme slowly increases with the increase of AP's cache size, However, energy efficiency of our cooperative scheme will decline with the increase of the satellite's cache size. The reason is that the request possibility of files is exponentially decreasing according to Zipf distribution, and considering satellite serves terrestrial users through broadcast transmission, satellite's unit cache gain is also decreasing with the increasing size of cache size. In this case, from the perspective of traffic offloading and energy efficiency, simply increasing satellite's cache size is not



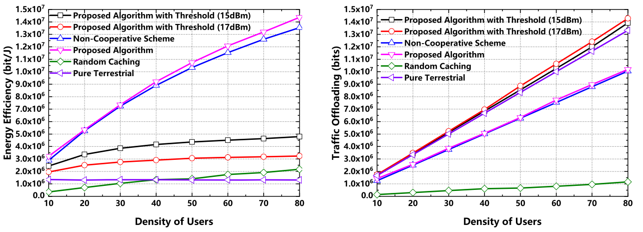
(a) Cache size vs energy efficiency (b) Cache size vs traffic offloading

Fig. 6. Performance comparison vs cache size (number of cells=50, user density=40 $users/cell$, request consistency degree=100%).



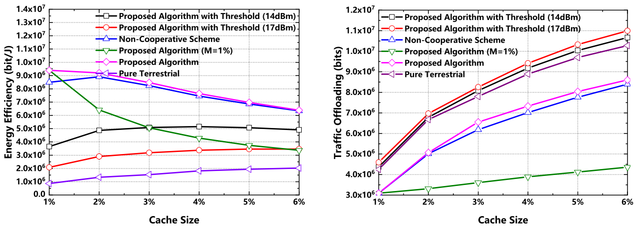
(a) Energy efficiency vs request consistency degree (b) Traffic offloading vs request consistency degree

Fig. 7. Performance comparison vs request consistency degree (cache size=2%, number of cells=50 and user density=40 $users/cell$).



(a) Energy efficiency vs user density (b) Traffic offloading vs user density

Fig. 8. Performance comparison vs user density (cache size=2%, number of cells=50, request consistency degree=0%).



(a) Energy efficiency vs cache size (b) Traffic offloading vs cache size

Fig. 9. Performance comparison vs cache size (number of cells=50, user density=40 $users/cell$, request consistency degree=0%).

an efficient way to improve the system performance according to the results in Fig. 6(a) and Fig. 6(b).

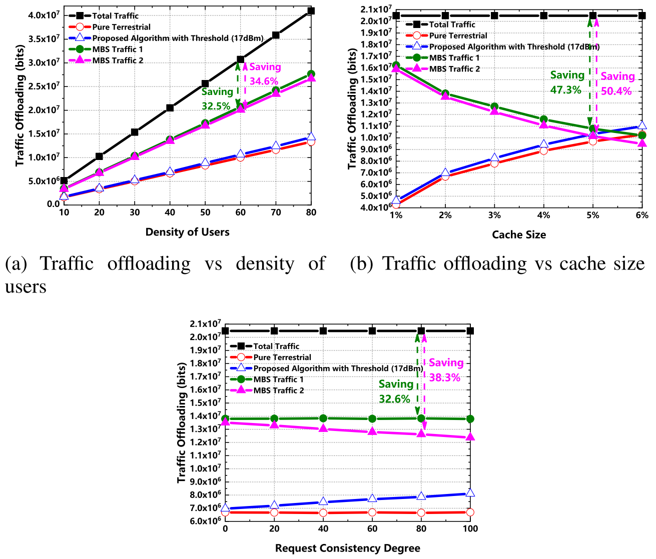
D. Energy Efficiency & Traffic Offloading vs Request Consistency Degree

During our evaluation process, we also found out an interesting phenomenon that the performance of our cooperative transmission scheme can be significantly influenced by request consistency degree. Here, we give an explanation. For all users in satellite coverage, if users' request preference in each cell is the same, then we consider these users' requests have highest request consistency degree; otherwise, these users' requests have lowest request consistency degree. The degree of request consistency in Fig. 7 represents the percentage of cells follows the same request preference. To be specific, we have $p_{i,j} = p_{i,j'}, \forall j, j' \in \mathcal{J}$ when request consistency degree $=$ 100% and $p_{i,j} \neq p_{i,j'}, \forall j, j' \in \mathcal{J}$ when request consistency degree $=$ 0%.

According to our explanation, it is clear that higher request consistency degree can make popular files in all cells concentrate on less files and satellite's broadcast transmission can be benefited. Results in Fig. 7 also verify this fact. In Fig. 7, all schemes consisting of the satellite will be influenced by the request consistency degree in various degrees. For our proposed cooperative scheme, which takes full advantage of satellite's broadcast transmission, it is significantly influenced by request consistency degree and has 22.2% and 19.1% performance difference in terms of energy efficiency and traffic offloading between the highest and lowest request consistency degree, respectively. When request consistency degree is lower than 70%, it can observed that the performance of our proposed algorithm in terms of traffic offloading becomes worse than the pure terrestrial scheme in Fig. 7(b). In this case, by setting minimum transmission power threshold of APs, regardless of the degree of request consistency, our proposed algorithm with threshold can guarantee the improvement on both energy efficiency and traffic offloading compared with the pure terrestrial scheme.

In order to show performance comparisons in request inconsistency, we also make performance comparison among different schemes in the lowest request consistency degree environment as shown in Fig. 8 and Fig. 9. In this case, without minimum transmission power threshold of AP, although our proposed algorithm can provide maximum 10.9 times improvement of energy efficiency compared with the pure terrestrial scheme, it also has maximum 13.3% decreasing of traffic offloading. Thus, in a low request consistency degree environment, if we want to make a tradeoff between energy efficiency and traffic offloading, we have to enforce the proposed algorithm to use low energy-efficient terrestrial transmissions. By setting minimum transmission power threshold as 15dBm, our proposed algorithm can provide 3.6 times improvement in energy efficiency and also 4.4% increase of traffic offloading with 80 users in each cell compared with the pure terrestrial scheme in Fig. 8.

Different from the highest request consistency degree environment in Fig. 6, in the lowest request consistency degree environment as shown in 9(a) and 9(b), our proposed algorithm has 27.4%-16.3% decreasing of traffic offloading compared with the pure terrestrial scheme. It can also provide 9.9-2.1 times improvement of energy efficiency, and the improvement is decreasing with the increasing of cache size. By setting minimum transmission power threshold of AP as 15dBm, our proposed algorithm can provide less improvement of energy efficiency but more increasing of traffic offloading. The specific number becomes 3.8%-1.7% increasing of traffic



(a) Traffic offloading vs density of users

(b) Traffic offloading vs cache size

(c) Traffic offloading vs request consistency degree

Fig. 10. Traffic served by MBS, APs, and the satellite (number of cells=50, user density=40 $users/cell$, request consistency degree=0%).

offloading and 3.2-1.4 times improvement of energy efficiency with different cache size. Furthermore, comparing the performance of 15dBm with 17dBm, it can be also seen that the higher transmission power threshold of terrestrial AP will lead to less energy efficiency and more traffic offloading for the proposed cooperative transmission scheme.

The reason of such phenomenon in the lowest request consistency degree environment can be explained as follows. Since the difference of the most popular files in each terrestrial cell become larger in lower request consistency degree environment, the total offloading traffic through broadcast transmission is declining and the advantage of satellite's broadcast transmission also becomes weaker. In this situation, NOMA-based transmission of terrestrial APs becomes a more effective approach to offload traffic in each cell. Although it will degrade the performance of energy efficiency, it can satisfy more users' requests in each terrestrial cell and significantly improve performance of traffic offloading. Meanwhile, thanks to the assistance of satellite's broadcast, the performance of our cooperative transmission scheme in terms of energy efficiency is always better than pure terrestrial scheme with or without minimum terrestrial transmission power threshold.

E. Performance of Traffic Offloading in MBS

Since the objective of the proposed scheme is to offload traffic from MBS, we also evaluate the traffic served by MBS in different scenarios. Note that "MBS Traffic 1" and "MBS Traffic 2" represent the traffic served by MBS for pure terrestrial scheme and the proposed scheme respectively. The performance comparison is depicted in Fig. 10. In Fig. 10(a), with the increasing of user density in each cell, the percentage of offloading traffic from MBS remains stable, changes from 32.6% to 32.5% for pure terrestrial scheme and 34% to 34.9% for the proposed scheme. In Fig. 10(b), both pure terrestrial scheme and the proposed scheme will offload more traffic from MBS with stronger cache ability, the offloaded traffic changes from 20.7% to 50.1% and 22.4% to 53.6%, respectively. In Fig. 10(c), the performance of pure terrestrial scheme is not influenced by request consistency degree in which offloaded traffic remains 32.6%. For the proposed scheme, the offloaded traffic changes from 34% to 39.5%.

In conclusion, the numerical results show that our proposed algorithm can efficiently solve the problem in Eq. (16) to maximize the energy efficiency in integrated satellite/terrestrail RAN, and the proposed algorithm can also provide significant improvement in both traffic offloading and energy efficiency compared with pure terrestrial scheme. However, in order to ensure better performance in terms of traffic offloading in the environment with low request consistency degree, setting minimum transmission power threshold of terrestrial APs in the algorithm is also required.

VII. Conclusion

In this paper, we introduced a LEO satellite network to cooperatively serve users with terrestrial APs in cache-enabled RAN. We first formulated a nonlinear fractional programming problem to maximize the energy efficiency in the integrated satellite/terrestrial RAN. We further adopted a parametric method to obtain an equivalent problem and divided it into three sub-problems, and we also designed efficient algorithms to obtain the optimal solution for each sub-problem. The numerical results show that our scheme can provide significant improvement of energy efficiency with similar traffic offloading performance compared with traditional terrestrial scheme in a cooperative manner. For the next generation cellular system, an integrated satellite/terrestrial RAN can also be an effective solution to alleviate the competition of terrestrial MBS and APs and provide high energy efficiency service for terrestrial users.

Acknowledgment

The authors sincerely thank the editor, Dr. Dimitrie C. Popescu, and the anonymous referees for their invaluable suggestions that have led to the present improved version from the original manuscript.

References

[1] Cisco. (2018). *Cisco Visual Networking Index: Forecast and Methodology 2017–2022*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[2] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.

[4] X. Ge, S. Tu, G. Mao, and C. X. Wang, "5G ultra-dense cellular networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[5] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, May 2017.

[6] Y. Liu, X. Li, F. R. Yu, H. Ji, H. Zhang, and V. C. M. Leung, "Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2295–2311, Oct. 2017.

[7] X. Zhu, C. Jiang, L. Kuang, N. Ge, and J. Lu, "Non-orthogonal multiple access based integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2253–2267, Oct. 2017.

[8] X. Zhu, C. Jiang, L. Yin, L. Kuang, N. Ge, and J. Lu, "Cooperative multigroup multicast transmission in integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 981–992, May 2018.

[9] *Study on Architecture for Next Generation System*, document TR 23.799, 3GPP, 2016.

[10] 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.

[11] J. N. Pelton and B. Jacqué, "Distributed Internet-optimized services via satellite constellations," in *Handbook of Satellite Applications*. New York, NY, USA: Springer, 2016, pp. 1–21.

[12] T. Lee, "Building technical communities for the benefit of humanity [2015 president's column]," *IEEE Microw. Mag.*, vol. 17, no. 1, pp. 8–11, Dec. 2016.

[13] *Starlink FCC Application*. Accessed: Sep. 25, 2019. [Online]. Available: https://fcc.report/IBFS/SAT-MOD-20181108-00083

[14] J. Li, H. Lu, K. Xue, and Y. Zhang, "Temporal netgrid model-based dynamic routing in large-scale small satellite networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6009–6021, Jun. 2019.

[15] X. Zhu, C. Jiang, L. Kuang, N. Ge, and J. Lu, "Energy efficient resource allocation in cloud based integrated terrestrial-satellite networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[16] K. An, T. Liang, G. Zheng, X. Yan, Y. Li, and S. Chatzinotas, "Performance limits of cognitive-uplink FSS and terrestrial FS for Ka-band," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2604–2611, Oct. 2019.

[17] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[18] X. Yan, H. Xiao, C.-X. Wang, and K. An, "Outage performance of NOMA-based hybrid satellite-terrestrial relay networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 538–541, Aug. 2018.

[19] X. Yan *et al.*, "The application of power-domain non-orthogonal multiple access in satellite communication networks," *IEEE Access*, vol. 7, pp. 63531–63539, 2019.

[20] 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.*, to be published, doi: 10.1109/MNET.2019.1900138.

[21] 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.

[22] 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.

[23] S. D'Oro, L. Galluccio, G. Morabito, and S. Palazzo, "SatCache: A profile-aware caching strategy for information-centric satellite networks," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 4, pp. 436–444, Apr. 2014.

[24] 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 Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[25] T. De Cola *et al.*, "Network and protocol architectures for future satellite systems," *Found. Trends Netw.*, vol. 12, nos. 1–2, pp. 1–161, 2017.

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

[27] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[28] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[29] L. Su, C. Yang, and I. Chih-Lin, "Energy and spectral efficient frequency reuse of ultra dense networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5384–5398, Apr. 2016.

[30] Y. Xu, H. Sun, R. Q. Hu, and Y. Qian, "Cooperative non-orthogonal multiple access in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[31] B. Di, L. Song, Y. Li, and G. Y. Li, "Non-orthogonal multiple access for high-reliable and low-latency V2X communications in 5G systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2383–2397, Oct. 2017.

[32] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[33] E. Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy consumption analysis and minimization in multi-layer heterogeneous wireless systems," *IEEE Trans. Mobile Comput.*, vol. 14, no. 12, pp. 2474–2487, Dec. 2015.

[34] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 86–94, Jun. 2017.

[35] S. Yunas, M. Valkama, and J. Niemelä, "Spectral and energy efficiency of ultra-dense networks under different deployment strategies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 90–100, Jan. 2015.

[36] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.

[37] *Study on New Radio (NR) to Support Non Terrestrial Networks*, document TR 38.811, 3GPP, 2018.

[38] 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. (GLOBECOM)*, 2017, pp. 1–6.

[39] 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.

[40] L. Bertaux *et al.*, "Software defined networking and virtualization for broadband satellite networks," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 54–60, Mar. 2015.

[41] J. Bao, B. Zhao, W. Yu, Z. Feng, C. Wu, and Z. Gong, "OpenSAN: A software-defined satellite network architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 347–348, 2014.

[42] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document TR 38.913, 3GPP, 2018.

[43] O. Y. Kolawole, S. Vuppala, M. Sellathurai, and T. Ratnarajah, "On the performance of cognitive satellite-terrestrial networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 668–683, Dec. 2017.

[44] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Multicast multigroup precoding and user scheduling for frame-based satellite communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4695–4707, Sep. 2015.

[45] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[46] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[47] I. M. Stancu-Minasian, *Fractional Programming* (Mathematics and Its Applications), vol. 409. Dordrecht, The Netherlands: Springer, 1997, ch. 4.

[48] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[49] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[50] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

[51] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[52] E. Horowitz and S. Sahni, "Computing partitions with applications to the knapsack problem," *J. ACM*, vol. 21, no. 2, pp. 277–292, Apr. 1974.

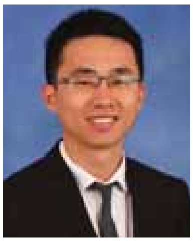
[53] J. Do, D. M. Akos, and P. K. Enge, "L and S bands spectrum survey in the San Francisco bay area," in *Proc. IEEE Position Location Navigat. Symp. (PLANS)*, Apr. 2004, pp. 566–572.

[54] M. J. Miller, B. Vucetic, and L. Berry, *Satellite Communications: Mobile and Fixed Services*. Berlin, Germany: Springer, 1993.

Jian Li received the bachelor's degree from the Department of Electronics and Information Engineering, Anhui University, Hefei, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research interests include satellite networks, cooperative transmission in integrated satellite/terrestrial networks, and cache-enabled wireless heterogeneous networks.

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 Ph.D. degree from the Department of Electronic Engineering and Information Science (EEIS), USTC, in 2007. From May 2012 to May 2013, he was a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, University of Florida. He is currently an Associate Professor with the School of Cyber Security, Department of EEIS, USTC. His research interests include the next-generation Internet, distributed networks, and network security. He is an IET fellow. He serves on the Editorial Board of several journals, including the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (TWC), the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT (TNSM), *Ad Hoc Networks*, IEEE ACCESS, and *China Communications*. He has also served as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) and a Lead Guest Editor for *IEEE Communications Magazine*.

David S. L. Wei (Senior Member, IEEE) received the Ph.D. degree in computer and information science from the University of Pennsylvania in 1991. From May 1993 to August 1997, he was on the Faculty of Computer Science and Engineering, The University of Aizu, Japan as an Associate Professor and then a Professor. He is currently a Professor with the Computer and Information Science Department, Fordham University. He has authored and coauthored more than 100 technical articles in various archival journals and conference proceedings. His research interests include cloud computing, big data, the IoT, and cognitive radio networks. He was a guest editor or a lead guest editor for several special issues in the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, and the IEEE TRANSACTIONS ON BIG DATA. He also served as an Associate Editor for the IEEE TRANSACTIONS ON CLOUD COMPUTING from 2014 to 2018, and an Associate Editor of the *Journal of Circuits, Systems and Computers* from 2013 to 2018.



Jianqing Liu (Member, IEEE) received the B.Eng. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2013, and the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA, in 2018. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The University of Alabama in Huntsville, Huntsville. His research interests include wireless networking and network security in cyber-physical systems.

Yongdong Zhang (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor with the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China (USTC). He has authored more than 100 refereed journals and conference papers. His current research interests are in the fields of multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. He is a member of the Editorial Board of the IEEE TRANSACTIONS ON MULTIMEDIA and *Multimedia Systems* journal. He was a recipient of the Best Paper Awards in PCM 2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate at ICME 2011.