Efficient Content Caching for Delivery Time Minimization in the LEO Satellite Networks

Sovit Bhandari, Thang X. Vu, Symeon Chatzinotas, and Björn Ottersten
Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg,
L-1855 Luxembourg. E-mail: {sovit.bhandari, thang.vu, symeon.chatzinotas, bjorn.ottersten}@uni.lu

***Abstract*—Next-generation multi-spot beam satellite systems open a new way to design low earth orbit (LEO) satellite communication systems with full flexibility in managing bandwidth, transmit power, and spot beam coverage, enabling the adoption of spatial multiplexing techniques to meet the unprecedented demand for future mobile traffic. However, conventional spatial multiplexing techniques perform poorly in satellite systems due to high correlation between the satellite channels, resulting in inefficient mitigation of inter-user interference. In this paper, we exploit the flexibility of multi-spot beam LEO satellites and consider the geographic distribution of users to improve the performance of LEO satellite-assisted edge caching systems. Our goal is to jointly optimize the beam coverage, bandwidth and transmit power and minimize the cache delivery time. In particular, the spot beam coverage is optimized by using the K-means algorithm applied to the realistic user demands, followed by a proposed successive convex approximation (SCA)-based iterative algorithm for optimizing the radio resources. Simulations shows that our optimal approach outperforms the conventional precoding-based approach and also shows a significant improvement in the minimization of the maximum content delivery time.**

***Index Terms*—caching, precoding, multicasting, optimization, LEO satellite, beamforming.**

I. Introduction

With the proliferation of connected devices and data-intensive applications, demand for mobile data traffic is expected to grow at an unprecedented rate. As per Ericsson, it is predicted that by the end of 2028, video traffic will account for 80% of the global mobile data traffic [1]. To solve the latency problem resulting from higher mobile data traffic, the concept of mobile edge computing (MEC) was introduced in early 2013. MEC is a decentralized version of cloud computing that has limited computational, signal processing, and storage capabilities. The access nodes with storing, computational, and signal processing capabilities in the MEC are called edge nodes (ENs) [2]. ENs can only store a limited amount of content, and this functionality of edge nodes is referred to as edge caching (EC) [3]. To make optimal use of the limited memory capacity of ENs for a better quality of service (QoS), popular content must be proactively stored in the cache memory of ENs such that the backhaul load can be minimized [4].

On the terrestrial network side, much research has been done on EC so that overall latency can be minimized [5]. However, for global connectivity, terrestrial networks are not the ideal solution because data cached in terrestrial networks must travel multiple hops unless the requesting user equipments (UEs) are adjacent to EN. To facilitate global connectivity with the goal of minimizing overall latency and thereby improving QoS, the concept of caching in satellite nodes (SNs) has become recognized as a promising technology [6], [7].

The round-trip delay for two-way communications is small for low earth orbit (LEO) satellites compared to other types of satellite constellations, making LEO constellations suitable for broadband services in under-serviced areas and may become a main feature of beyond fifth-generation (B5G) communications.

In [7], a two-stage cache-enabled satellite internet-of-things system is considered to analyze the successful delivery probability by storing different parts of the file content in the cache memory of the satellite and the base station (BS). In [8], a cooperative proactive caching scenario between terrestrial BS, satellite, and gateway (GW) is considered to minimize the content retrieval delay. In [9], an analytical expression for the outage probability is obtained for the hybrid satellite-based terrestrial relay network considering the non-caching and caching scenarios in the terrestrial relays. The works in [7]– [9] only considered the caching strategy in the integrated satellite-based terrestrial network with predefined beam coverage, transmit power and operating bandwidth. Although the impact of the bandwidth is numerically evaluated, these works do not exploit the flexible payload capability. It has been shown in our recent work [10], that joint bandwidth and transmit power design is beneficial in highly-loaded terrestrial

systems. Furthermore, the authors of [11] showed that satellite communications can be improved by using geographical user demands.

In this paper, we exploit the flexible payload capability and adaptive beam patterning to improve the performance of LEO-assisted caching systems. Our contributions are as follows:

- We formulate a joint design of spot beam coverage, operating bandwidth and precoding vectors to minimize the average delivery time in LEO-assisted caching networks. Although flexible bandwidth has been considered in satellite communications, to the best of our knowledge, this is the first work exploiting spatial multiplexing technique within each spot beam in LEO-enabled caching systems thanks to fully flexible payload capabilities.
- We propose to solve the joint optimization problem via two sub-problems, i.e., beam coverage design and radio resources allocations. To solve the first sub-problem, we propose a clustering technique based on K-Means++ [12] to optimize the spot beam coverage area. To tackle the non-convexity of the second sub-problem, we reformulate it using a difference-of-convex (DC) representation and propose an iterative algorithm using the successive convex approximation (SCA) method.
- Finally, the advantages of the proposed framework are shown by numerical results based on the realistic Movielens dataset [13]. The results show that our solutions achieve a performance gain of more than 28% compared to the zero-forcing (ZF)-based approach and reduce the average content delivery time by at least more than 71% when the backhaul rate is varied between (0.25–1) Gbps, assuming that the UEs requested contents are fully stored on LEO.

The remainder of this paper is organized as follows. Section II describes the system model and parameters. Section III presents the problem formulation and proposed solution. Section IV presents the numerical results. Finally, Section V provides conclusion.

II. System Model

We consider the downlink of a cache-enabled LEO satellite network consisting of a LEO SN serving a set $\mathcal{U} = \{1, 2, ..., u, ..., U\}$ of U single-antenna UEs within the coverage, a GW, and a central cloud (CC), as shown in Fig. 1. The LEO is equipped with a flexible payload that can generate M spot beams of arbitrary shapes [14] to serve UEs at position $\boldsymbol{y}_u$ (2D coordinate vector) within its footprint.

For simplicity, the shape of spot beam m is considered circular and the spot beam gain G_m can be written as $4D_m^2/r_m^2$ [15], where r_m is the spot beam's radius and

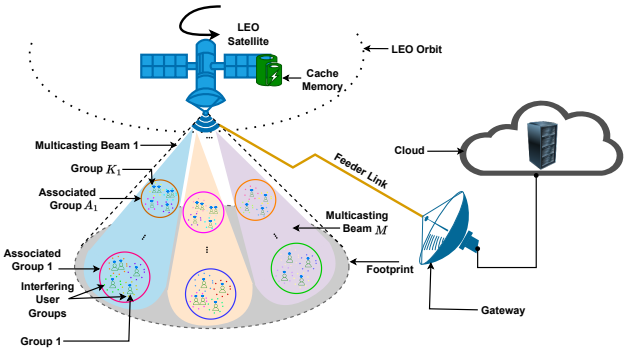


Fig. 1. A downlink cache enabled multicasting LEO satellite system.

D_m is the slant distance between LEO SN and the m-th spot beam. Due to the geographical distributions of the UEs, we assume non-overlapped spot beams and inter-beam interference is negligible. As a result, full-frequency reuse is applied. Thanks to the advanced payload technology, each LEO can deliver up to N spatial multiplexing data streams in each spot beam [16].

A. Caching Model

By equipping with the advanced flexible regenerative payload, the LEO SN is able to process data and has a limited cache memory of C bits. The U UEs are interested in the content library of $\mathcal{F} = \{1, 2, ..., f, ..., F\}$ at the CC consisting of F file. Due to the non-geostationary nature of LEO, the limited time for cache placement is considered τ from the total connection time T in one orbital revolution, so the total delivery time is $T - \tau$, which is considered as the accumulation of multiple timeslots (short-term delivery period) t. We consider offline caching in which the demand vectors $\boldsymbol{L}$ is obtained in advance, e.g., via historical average or prediction model[1]. Based on $\boldsymbol{L}$, the cache placement is performed.

B. User Grouping

To exploit the flexible multi-beam capability, the users are served in groups depending on their geographical locations and requested contents. The set of UEs in each spot beam m is denoted as $\mathcal{U}_m \subseteq \mathcal{U}$ which are divided into K_m. UEs within the same spot beam requesting the same file are assigned to the same group. If $K_m \leq N$, the all user groups can be served simultaneously using the whole bandwidth B. Otherwise, K_m groups are divided into $A_m = \lceil \frac{K_m}{N} \rceil$ associate user groups, which are served via orthogonal frequency bandwidths. Let $\mathcal{A}_m = \{a_1^m, a_2^m, ..., a_{A_m}^m\}$ denote the set of A_m associate user groups in beam m, and $\mathcal{K}_{a,m}$ denote the set of

[1]The joint design for on-line caching setting is left for future work.

users belong to the associate user group a of spot beam m. To illustrate, suppose 8 users $\{u_1, u_2, ..., u_8\}$ within spot beam m requests the corresponding files $\{f_1, f_2, f_3, f_3, f_4, f_5, f_5, f_6\}$, which includes $K_m = 6$ distinct files. If $N = 4$, then there are two associate user groups, which are: $\mathcal{K}_{1,m} = \{u_1, u_2, u_3, u_4, u_5\}$ and $\mathcal{K}_{2,m} = \{u_6, u_7, u_8\}$.

C. Transmission Model

The signal received by the UE u under group k ($k \in \mathcal{K}_{a,m}$) of associated group a of the multicast spot beam m can be written as follows:

$$y_{u,k,a,m} = \boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k,a,m} s_{k,a,m} + \sum_{k' \in \mathcal{K}_{a,m} \setminus \{k\}} \boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k',a,m} s_{k',a,m} + n_u, \quad (1)$$

where $\boldsymbol{h}_{u,k,a,m} \in \mathbb{C}^{N \times 1}$ is the downlink channel coefficient from spot beam m to UE u of multicast group k of associated group a, $\boldsymbol{w}_{k,a,m} \in \mathbb{C}^{N \times 1}$ is the precoding vector from spot beam m to multicast group k of associated group a, $s_{k,a,m} \in \mathbb{C}$ is the data symbol requested by UEs of group k of associated group a via multicast spot beam m with $\mathbb{E}[|s_{k,a,m}|^2] = 1$, and $n_u \sim \mathcal{CN}(0, \sigma_u^2)$ is the additive white gaussian noise (AWGN).

The downlink channel coefficient between the LEO SN and UE u, after the Doppler compensation, can be written as, $\boldsymbol{h}_{u,k,a,m} = g_u \cdot \boldsymbol{v}_u(\varphi_u)$, where g_u is the channel gain and $\boldsymbol{v}_u(\varphi_u) \in \mathbb{C}^{N \times 1}$ is the downlink array response vector for UE u, wherein φ_u is the angle of departure (AoD) [17].

For communication with handhelds with wider mainlobes, we assume a Rician fading channel with a large-scale fading distribution in which $g_u = \sqrt{\beta_u}\hat{g_u}$, where $\beta_u = G_m G_u M \lambda^2 / (4\pi D_u)^2$ is a large-scale fading due to large propagation distance, wherein G_m, G_u are transmit and receive antenna gains, respectively, and D_u is the slant distance between LEO SN and UE. $\hat{g}_u = \alpha_u(h_{LoS}\sqrt{\kappa_u/(\kappa_u+1)} + h_{NLoS}\sqrt{1/(\kappa_u+1)})$ denotes the small-scale fading channel model, wherein κ_u represents the Rician factor, $\alpha_u = \mathbb{E}\{|\hat{g}_u|^2\}$ is the average channel power, h_{LoS} is the deterministic line-of-sight (LoS) part, and h_{NLoS} represents the non-LoS (NLoS) component. The real and imaginary parts of $\hat{g}_u$ are independently and identically distributed as $\mathcal{N}(\sqrt{\kappa_u \alpha_u / 2(\kappa_u+1)},\ \alpha_u/2(\kappa_u+1))$. The signal-to-interference-plus-noise ratio (SINR) of UE u of the associated group a at splot beam m is calculated as follows:

$$\gamma_{u,k,a,m} = \frac{|\boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k,a,m}|^2}{\sum_{k' \in \mathcal{K}_{a,m} \setminus \{k\}} |\boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k',a,m}|^2 + N_0 b_{a,m}}, \quad (2)$$

where N_0 is the noise spectral density and $b_{a,m}$ is the bandwidth of associated group a within spot beam m.

The transmission rate of a group k within the associated group a, determined by smallest SINR, in spot beam m calculated as follows:

$$R_{k,a,m} = b_{a,m} \log_2(1 + \min_u\{\gamma_{u,k,a,m}\}). \quad (3)$$

III. Problem Formulation and Proposed Solution

A. Problem Formulation

We aim at jointly optimizing the spot beam coverage r_m, allocated bandwidth, and precoding vectors to minimizing the worst-case average delivery latency. The joint optimization is formulated as follows:

$$\mathcal{P}: \min_{\{\boldsymbol{w},\boldsymbol{b},\boldsymbol{r},\boldsymbol{K}\}} \quad t(\boldsymbol{w},\boldsymbol{b},\boldsymbol{r},\boldsymbol{K}) \quad (4a)$$

$$\text{s.t.} \quad R_{k,a,m} \geq \eta_{req}, \quad \forall k,a,m, \quad (4b)$$

$$\sum_{k \in \mathcal{K}_{a,m}} \sum_{a \in \mathcal{A}_m} \|\boldsymbol{w}_{k,a,m}\|^2 \leq P_{tot}(K_m/K), \forall m, \quad (4c)$$

$$\sum_{a \in \mathcal{A}_m} b_{a,m} \leq B, \quad \forall m, \quad (4d)$$

$$\bigcup_{m \in \mathcal{M}} \pi r_m^2 \geq A_{tot}, \quad (4e)$$

where $\boldsymbol{w} \coloneqq \{\boldsymbol{w}_{k,a,m}\}_{\forall k,a,m}$, $\boldsymbol{b} \coloneqq \{b_{a,m}\}_{\forall a,m}$, $\boldsymbol{r} \coloneqq \{r_m\}_{m=1}^M$, and $\boldsymbol{K} \coloneqq \{K_m\}_{m=1}^M$ are the short-hand notations for the precoding vectors, bandwidth allocation, spot beam radius, and group allocation, respectively; η_{req} is the minimum QoS requirement; $\mathcal{A}_m$ is the set of associated groups in spot beam m; P_{tot} is the total transmit power of LEO SN; $K = \sum_{m=1}^M K_m$; and A_{tot} is the total service area of LEO SN. The objective function of (4), i.e., $t(\boldsymbol{w},\boldsymbol{b},\boldsymbol{r},\boldsymbol{K})$, is the end-to-end latency, assuming the FastForward full-duplex capability [18], is computed as follows:

$$t(\boldsymbol{w},\ \boldsymbol{b},\ \boldsymbol{r},\ \boldsymbol{K}) = \max_{\{k,a,m\}} \Big(\max\Big(\Big(\frac{q_k}{R_{k,a,m}} + \frac{D_k}{c} \Big), \Big(\frac{q_k}{R_{BH}} + \frac{D_k}{c} \Big)\Big(1 - \mu_k\Big)\Big)\Big) \quad (5)$$

where q_k is the file size, $D_k = \max_u(D_u)$ is the slant distance between LEO SN and k-th group; c is the speed of light; $\mu_k \in [0,1]$ denotes the fraction of the k-th file on LEO SN; R_{BH} is the backhaul rate between CC to SN; $q_k/R_{k,a,m}$ and D_k/c are the transmission and the propagation delays, respectively, incurred while sending files from SN to UEs of group k; and q_k/R_{BH} is the transmission delay incurred while sending files from SN to GW.

In problem $\mathcal{P}$, the constraint (4b) satisfies the minimum QoS requirement. The constraint (4c) limits the power allocated to each LEO SN spot beam. The constraint (4d) states that the sum of bandwidths allocated to all associated groups of the spot beam m of LEO SN must not exceed the total allocated bandwidth B. The constraint (4e) ensures that the non-overlapping union of the coverage areas of a total number of spot beams covers at least the total service area of LEO SN.

Difficulty to solve problem $\mathcal{P}$: The problem $\mathcal{P}$ cannot be solved directly due to the non-convexity of the objective function as well as the constraint (4b), also the constraints (4c) and (4d) boundary value is unknown, and the constraint (4e) is a disjoint set partition problem which is non-deterministic polynomial time (NP) hard problem.

B. Proposed Solution

To transform the problem $\mathcal{P}$ into a more tractable form, the constraint (4e) is separated from the rest of the problem as the spot beam coverage area is fixed prior to the initiation of the data transmission process. Therefore, the problem $\mathcal{P}$ is decomposed into two sub-problems as:

1) Min. of the Spot Beam Coverage Area: The spot beam coverage area minimization sub-problem is formulated based on (4e) as follows:

$$\mathcal{P}_1 : \min_{\{r_m\}} \sum\nolimits_{m\in\mathcal{M}} \pi r_m^2 \tag{6a}$$

$$\text{s.t.} \quad \bigcup\nolimits_{m\in\mathcal{M}} \pi r_m^2 \geq A_{tot}, \quad \forall m. \tag{6b}$$

The problem $\mathcal{P}_1$ is NP-hard and can only be solved using clustering techniques, hence the K-Means++ [12] clustering technique is used to solve it. The clustering is done based on the position of U UEs that demand the service in the timeslot t, so the spot beam center is likely to point in the direction where the number of UEs is dominant. Since there are M spot beams serving U UEs in timeslot t, the U UEs are categorized into M clusters such that m-th spot beam serves m-th cluster. The problem $\mathcal{P}_1$ can be reformulated in terms of clustering as follows:

$$\mathcal{P}_1^{'} : \min_{\{\mathcal{U}_m, \boldsymbol{c}_{\mathcal{U}_m}, r_m\}} \sum\nolimits_{m\in\mathcal{M}} \pi r_m^2 \tag{7a}$$

$$\text{s.t.} \quad \bigcup\nolimits_{m\in\mathcal{M}} \mathcal{U}_m = \mathcal{U}, \tag{7b}$$

$$(\boldsymbol{y}_u - \boldsymbol{c}_{\mathcal{U}_m})(\boldsymbol{y}_u - \boldsymbol{c}_{\mathcal{U}_m})' \leq r_m^2, \quad \forall u, m, \tag{7c}$$

where $\mathcal{U}_m$ is the list of UEs in m-th cluster and $\boldsymbol{c}_{\mathcal{U}_m}$ is the 2D centroid of m-th cluster. The constraint (7b) ensures that all unique UEs lie within the total service area of LEO SN. The constraint (7c) guarantees that UEs are clustered based on the Euclidean distance between $\boldsymbol{y}_u$ and $\boldsymbol{c}_{\mathcal{U}_m}$, which is bounded by the radius of coverage of the cluster, i.e., r_m.

The procedure to obtain $\mathcal{U}_m$, $\boldsymbol{c}_{\mathcal{U}_m}$, and r_m is shown in Algorithm 1. To find the boundaries of the clusters, Voronoi tessellation technique is used, where the boundaries of the Voronoi polygons are computed using $\boldsymbol{c}_{\mathcal{U}_m}$. However, for mathematical tractability, the coverage area of the spot beam is considered circular, without taking the curvature of the earth into account. Using the outputs of Algorithm 1, user grouping is done as shown in Section II(B) to get $\mathcal{A}_m$, $\mathcal{K}_{a,m}$, and K_m, which are used in solving second sub-problem.

Algorithm 1 Iterative Algorithm to Solve (7a)

Input: U, A_{tot}, $\boldsymbol{y}_u$, M
Output: $\mathcal{U}_m$, $\boldsymbol{c}_{\mathcal{U}_m}$, r_m
Init : $\{\boldsymbol{c}_{\mathcal{U}_m}\}_{m=1}^{M}$, $i = 1$, I, $dis = [\,]$, ϵ, $err = 1$
Based on $\boldsymbol{y}_u$, apply the K-Means++ cluster Alg.
while $err > \epsilon$ and $i < I$ **do**
  Calculate $\mathcal{U}_m$ using K-Means++ Alg.
  Compute $\boldsymbol{c}_{\mathcal{U}_m}^{(i)} = \text{mean}\{\boldsymbol{y}_u \mid u \in \mathcal{U}_m\}$
  Compute $err = |\boldsymbol{c}_{\mathcal{U}_m}^{(i)} - \boldsymbol{c}_{\mathcal{U}_m}^{(i-1)}|$
  Update $\boldsymbol{c}_{\mathcal{U}_m}^{(i-1)} \leftarrow \boldsymbol{c}_{\mathcal{U}_m}^{(i)}$; $i \leftarrow i + 1$
end while
Calculate radius r_m of the cluster $\mathcal{U}_m$
for $m = 1$ to M **do**
  for $j = 1$ to length$\{\mathcal{U}_m\}$ **do**
    Compute $dis^{(j)} = \sqrt{(\boldsymbol{y}_u - \boldsymbol{c}_{\mathcal{U}_m})(\boldsymbol{y}_u - \boldsymbol{c}_{\mathcal{U}_m})'}$
  end for
  Compute $r_m = \lceil(\max\{dis\})\rceil$
end for

2) Min. of Short-term Content Delivery Period: The short-term content delivery minimization sub-problem is formulated based on (4a)–(4d) as follows:

$$\mathcal{P}_2 : \min_{\{\boldsymbol{w},\boldsymbol{b}\}} \max_{k,a,m}(\max(\Delta_1, \Delta_2)) \tag{8a}$$

$$\text{s.t. } \mathrm{b}_{a,m}\log_2(1 + \min_u\{\gamma_{u,k,a,m}\}) \geq \eta_{req}, \forall \quad k, a, m, \tag{8b}$$

$$(4c), (4d),$$

where the objective function (8a) is obtained by replacing the values of (5) and (3) in (4a), wherein $\Delta_1 = q_k/(b_{a,m}\log_2(1 + \min_u\{\gamma_{u,a,k,m}\}))$ and $\Delta_2 = (q_k/R_{BH} + D_k/c)(1 - \mu_k)$, and the constraint (8b) is obtained by replacing the value of (3) in (4b).

The problem $\mathcal{P}_2$ is non-convex due to the objective function and the constraint (8b). Thus, to solve problem $\mathcal{P}_2$, following the relaxation method, the minimal operators $\min_u\{\gamma_{u,k,a,m}\}$ in the objective function is replaced by positive auxiliary variable $\gamma_{k,a,m}$. Also, $b_{a,m}\log_2(1 + \min_u\{\gamma_{u,k,a,m}\})$ in the objective function of problem $\mathcal{P}_2$ is replaced by the positive auxiliary variable $z_{k,a,m}$ to transform the problem $\mathcal{P}_2$ into a more tractable form. As a result, the problem $\mathcal{P}_2$ can be equivalently reformulated as:

$$\mathcal{P}_2^{'} : \min_{\{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{z}\}} \max_{k,a,m}(\max(\Pi_1, \Delta_2)) \tag{9a}$$

$$\text{s.t.} \quad b_{a,m}\log_2(1 + \gamma_{k,a,m}) \geq z_{k,a,m}, \forall k, a, m, \tag{9b}$$

$$(|\boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k,a,m}|^2)/(\textstyle\sum_{k'\in\mathcal{K}_{a,m}\setminus k}|\boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k',a,m}|^2 + N_0 b_{a,m}) \geq \gamma_{k,a,m}, \quad \forall u, k, a, m, \tag{9c}$$

$$b_{a,m}\log_2(1 + \gamma_{k,a,m}) \geq \eta_{req}, \quad \forall k, a, m, \tag{9d}$$

$$(4c), (4d),$$

where $\boldsymbol{\gamma} \coloneqq \{\gamma_{k,a,m}\}_{\forall k,a,m}$ and $\boldsymbol{z} \coloneqq \{z_{k,a,m}\}_{\forall k,a,m}$ are the short-hand notations for auxiliary variables of

SINR and channel capacity, respectively, and $\Pi_1 = (q_k/z_{k,a,m} + D_k/c)$.

The main challenge in solving problem $\mathcal{P}_2^{'}$ lies in the first three constraints, i.e., (9b), (9c), and (9d). We can handle the constraint (9b) by considering the slack variable $x_{k,a,m}$, which can be reformulated as:

$$\log_2(1+\gamma_{k,a,m}) \geq x_{k,a,m}, \tag{10}$$

$$b_{a,m}x_{k,a,m} \geq z_{k,a,m}. \tag{11}$$

Constraint (10) is convex, to deal with constraint (11), we use an equivalent representation as:

$$(11) \Leftrightarrow (b_{a,m}+x_{k,a,m})^2 \geq 2z_{k,a,m} + b_{a,m}^2 + x_{k,a,m}^2, \tag{12}$$

which has a DC form as both sides are convex functions. The DC programming in constraint (12) can be easily tackled using the iterative-based SCA method by taking the first-order approximation of the LHS of the constraint (12). Let $\hat{b}_{a,m}$ and $\hat{x}_{k,a,m}$ be the feasible values of the constraint (12) in the current iteration. In the next iteration, the constraint (12) can be approximated as a convex constraint as:

$$2(b_{a,m}+x_{k,a,m})(\hat{b}_{a,m}+\hat{x}_{k,a,m}) - (\hat{b}_{a,m}+\hat{x}_{k,a,m})^2 \geq 2z_{k,a,m} + b_{a,m}^2 + x_{k,a,m}^2. \tag{13}$$

To tackle the non-convexity of constraint (9c), we represent it in an equivalent form as:

$$(|\boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k,a,m}|^2)/\gamma_{k,a,m} \geq \sum_{k' \in \mathcal{K}_{a,m}\setminus\{k\}} |\boldsymbol{h}_{u,k,a,m}^H \boldsymbol{w}_{k',a,m}|^2 + N_0 b_{a,m}. \tag{14}$$

Since the constraint (14) is also in DC form, we use the SCA method to solve it iteratively. Taking $\hat{w}_{k,a,m}$ and $\hat{\gamma}_{k,a,m}$ as the feasible value, (14) can be approximated as:

$$\frac{2\boldsymbol{w}_{k,a,m}^H \boldsymbol{H}_{u,k,a,m}\hat{\boldsymbol{w}}_{k,a,m}}{\hat{\gamma}_{k,a,m}} - \gamma_{k,a,m}\frac{\hat{\boldsymbol{w}}_{k,a,m}\boldsymbol{H}_{u,k,a,m}\hat{\boldsymbol{w}}_{k,a,m}}{\hat{\gamma}_{k,a,m}^2} \geq \sum_{k' \in \mathcal{K}_{a,m}\setminus\{k\}} \boldsymbol{w}_{k',a,m}^H \boldsymbol{H}_{u,k,a,m}\boldsymbol{w}_{k',a,m} + N_0 b_{a,m}, \tag{15}$$

where $\boldsymbol{H}_{u,k,a,m} \coloneqq \boldsymbol{h}_{u,k,a,m}\boldsymbol{h}_{u,k,a,m}^H$.

Using (13) and (15), the problem $\mathcal{P}_2^{'}$ can be approximated by a convex optimization problem $\mathcal{P}_2^{''}$ as:

$$\mathcal{P}_2^{''(\hat{\boldsymbol{w}},\hat{\boldsymbol{b}},\hat{\boldsymbol{x}},\hat{\boldsymbol{\gamma}})}: \min_{\{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{x},\boldsymbol{z}\}} \max_{k,a,m} (\max(\Pi_1, \Delta_2)) \tag{16a}$$

$$\text{s.t. (4c), (4d), (10), (13), (15),}$$

$$\log_2(1+\gamma_{k,a,m}) \geq \eta_{req}/b_{a,m} \; \forall k,a,m, \tag{16b}$$

where $\boldsymbol{x} \coloneqq \{x_{k,a,m}\}_{\forall k,a,m}$ is the short-hand notation for slack variables and (16b) is obtained from (9d) by dividing its two sides by a positive number $b_{a,m}$.

The problem $\mathcal{P}_2^{''}$ is a convex problem and it can be solved directly using interior point method. Since the solutions of problem $\mathcal{P}_2^{''}$ should satisfy all the constraints of problem $\mathcal{P}_2$, the solution provided by problem $\mathcal{P}_2^{''}$ is sub-optimal for problem $\mathcal{P}_2$ and also depends largely on the initialization of the parameters $\hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{x}}$, and $\hat{\boldsymbol{\gamma}}$. Therefore, to solve (8), Algorithm 2 is proposed.

Algorithm 2 Iterative Algorithm to Solve (8a)

Input: $\mathcal{A}_m$, $\mathcal{K}_{a,m}$, K_m, $\boldsymbol{h}_{u,k,a,m}$, μ_k, D_k, c, R_{BH}

Output: $\boldsymbol{w}_{k,a,m}^*$, $b_{a,m}^*$, $x_{k,a,m}^*$, $\gamma_{k,a,m}^*$, $z_{k,a,m}^*$

Init: $\hat{\boldsymbol{w}}_{k,a,m}$, $\hat{b}_{a,m}$, $\hat{x}_{k,a,m}$, $\hat{\gamma}_{k,a,m}$, $\hat{z}_{k,a,m}$, $i = 1$, I, ϵ, $err = 1$

1: **while** $err > \epsilon$ and $i < I$ **do**
2: Solve (16a) to get $\boldsymbol{w}_{k,a,m}^*$, $b_{a,m}^*$, $x_{k,a,m}^*$, $\gamma_{k,a,m}^*$, $z_{k,a,m}^*$
3: Compute $t^{(i)}$
4: Compute $err = |t^{(i)} - t^{(i-1)}|$
5: Update $\hat{\boldsymbol{w}}_{k,a,m} \leftarrow \boldsymbol{w}_{k,a,m}^*$; $\hat{b}_{a,m} \leftarrow b_{a,m}^*$; $\hat{x}_{k,a,m} \leftarrow x_{k,a,m}^*$; $\hat{\gamma}_{k,a,m} \leftarrow \gamma_{k,a,m}^*$; $t^{(i-1)} \leftarrow t^{(i)}$; $i \leftarrow i+1$
6: **end while**

IV. Performance Evaluation on Realistic System Parameters

A. LEO SN Footprint

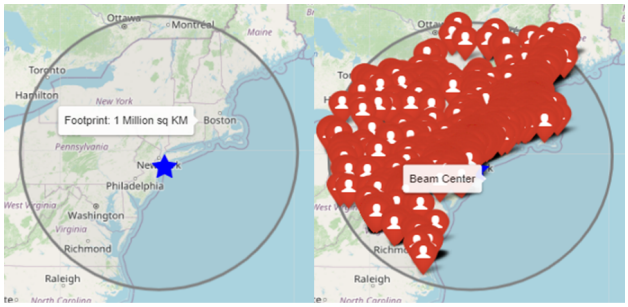


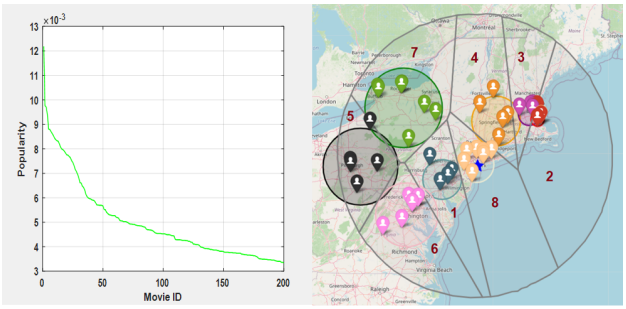
Fig. 2. LEO SN beam coverage and location-based UE traffic. (a) Footprint. (b) Traffic.

The LEO SN is assumed to be in orbit just above New York (NY) at time t. The LEO SN is at an altitude (H_s) of 550 km just above Earth's surface. The central angle ($\beta_\circ$) is assumed to be 5.2°. Based on H_s and $\beta_\circ$, A_{tot} is about 1.05 million (M) km^2 with NY as the beam center and a coverage radius (R_{SN}) of ≈ 578 km [19]. The footprint of the LEO SN is shown in Fig. 2(a). Simulations for Algorithm 1 and 2 are performed in Matlab 2022a using a built-in K-Means++ clustering algorithm and external CVX (optimization toolbox) solver, respectively, on a laptop with Intel i9-11950H processor and 32GB of RAM.

B. Content Popularity Based on the Movielens Dataset

The content popularity is generated from the location-based Movielens dataset, in which 1M movie ratings

are provided. The dataset contains UE IDs, UE locations (ZIP code), movie IDs, movie genres, and rating time, from which we can calculate the distribution of requests in any given time period. We use the ZIP code information to accurately determine the geographic distribution of requests by mapping the ZIP codes with the corresponding latitude and longitude. Since UEs in the 1M dataset are distributed across the entire U.S., while our LEO SN at time t only covers the upper part of the U.S. East Coast. The distribution of the UEs within the LEO SN beam coverage region is shown in Fig. 2(b). After calculating the content popularity $\boldsymbol{L}$ within the preferred region, only top 200 movie ID's are only taken into account. The most popular movie is indexed as 1, while the least popular is indexed as 200. The popularity of the top 200 movies within LEO SN beam coverage region is shown in Fig. 3(a). For each t duration, both the location and the content requested by UEs are randomly changed based on the historical probability distribution.



(a) (b)

Fig. 3. Content popularity and LEO SN spot beam at time t. (a) Content popularity. (b) Spot beam.

C. Performance Evaluations

TABLE I
SIMULATION PARAMETERS

Parameters	Value
Rician factor κ_u	10 dB
AoD φ_u	U [-0.5, 0.5]
User's elevation angle ϵ_u	U [40°, 90°]
Carrier frequency f_c	2 GHz
Bandwidth B	100 MHz
Number of spot beams M	8
Parallel data streams N	4
Number of UEs U	45
Transmit power P_{tot}	30 dBW – 40 dBW
QoS η_{req}	$0.25 \times b_{a,m}$
Backhaul rate R_{BH}	0.25 Gbps – 1 Gbps
Size of movies files q_k	U [0.5 GB, 1 GB]
Cache storage capacity C	150 GB

In this part, we conduct the numerical results considering a scenario where LEO SN has a total of 8 spot beams and the total number of UEs requesting data is 45. The 45 UEs are randomly distributed in the coverage area as shown in Fig. 3(b). It is assumed that each spot beam of LEO SN is capable of transmitting 4 parallel data streams. The coverage area and the number of UEs within each spot beam is calculated using Algorithm 1. The spot beam with the lowest number of UEs is marked as 1, while the spot beam with the highest number of UEs is marked as 8. The number 8 is shown in the spot beam where the beam center is located. The parameters used in the simulations are summarized in Table I. The simulation results are averaged over the 500 random channel realization. In each realization, the position of UEs and their requested contents are varied randomly. For the group-based precoding design, we compare the proposed optimal precoding approach with the ZF-based method.

Fig. 4 shows the mean/min data rate and outage probability as a function of P_{tot} for both the optimal and ZF-based precoding design approaches. From the figure, it can be seen that the data rate achieved by the optimal precoding approach is significantly better than the ZF-based approach because optimal precoding-based approach eliminates the inter-user interference more efficiently than the ZF-based method. Moreover, there is no outage scenario in the optimal precoding-based design, while there are massive outages in the ZF-based approach. When P_{tot} is increased, the outage probability decreases, and therefore the data rate improves comparatively more in the ZF-based approach than in the optimal precoding approach.

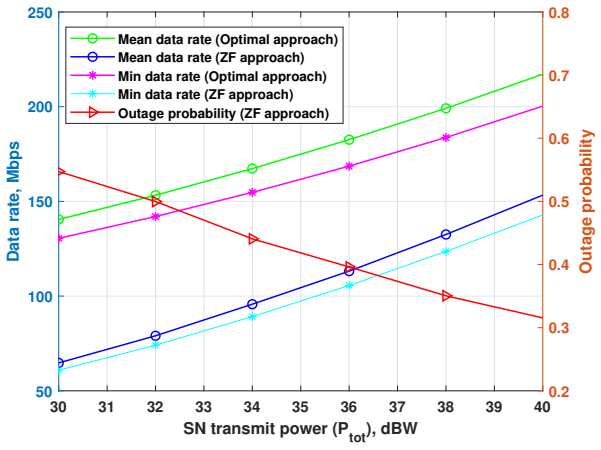


Fig. 4. The data rate and the outage probability versus the total transmit power of the LEO SN.

Fig. 5 shows the delay as a function of normalized cache capacity for the caching models such as most popular caching (MPC), uniform caching (UC), and random caching (RC). From figure, it can be seen that t is lower in the MPC approach compared to the UC and

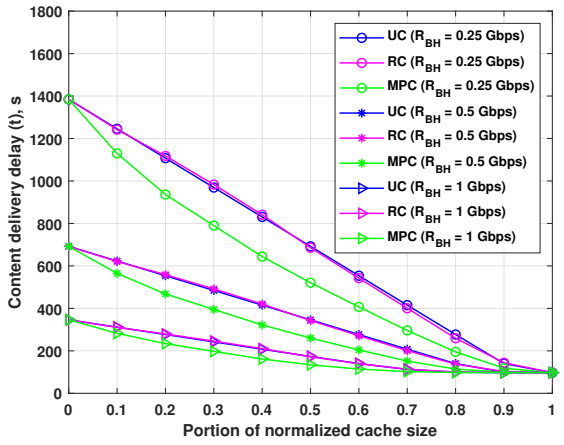


Fig. 5. Content delivery delay versus the portion of normalized cache size for different generic caching scenarios.

RC methods when the fraction of normalized cache size is smaller, and RC almost approaches the UC method due to the averaging of a large number of channel realizations. When the fraction of the normalized cache size of LEO SN is increased, t in the MPC approach tries to converge to the methods UC and RC. It can also be seen from the figure that R_{BH} also significantly affects t for the different values of the cache portion size.

V. Conclusion

In this paper, we have developed a joint adaptive beamforming, bandwidth allocation, and precoding optimization algorithm that fully exploits the LEO SN beam amplification, spatial multiplexing gain, and frequency allocation considering realistic system parameters. Using numerical results, we have shown that the proposed optimal precoding design for the multicasting beams significantly outperforms the ZF-based precoding design for the multicasting beams in terms of both rate improvement and reduction in the overall average maximum content delivery time from LEO SN to the requesting UE. Moreover, we have shown that the MPC-based caching method outperforms caching strategies based on RC and UC and significantly improves the average latency for content delivery compared to a scenario without caching.

Acknowledgment

This work is supported by the Luxembourg National Research Fund (FNR) via project INSTRUCT, ref. IPBG19/14016225/INSTRUCT, and project RUTINE, ref. C22/IS/17220888/RUTINE.

References

[1] "Traffic by application – mobility report," Ericsson. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/traffic-by-application.

[2] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[3] J. Yao, T. Han, and N. Ansari, "On Mobile Edge Caching," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 3, pp. 2525–2553, 2019.

[4] S. Bhandari et al., "Deep Learning-Based Content Caching in the Fog Access Points," *Electronics*, vol. 10, no. 4, p. 512, 2021.

[5] T. X. Vu et al., "Latency Minimization for Content Delivery Networks with Wireless Edge Caching," *2018 IEEE International Conference on Communications (ICC)*, USA, 2018, pp. 1–6, 2018.

[6] T. X. Vu et al., "Modeling and Implementation of 5G Edge Caching over Satellite," *International Journal of Satellite Communications and Networking*, vol. 38, no. 5, pp. 395–406, 2020.

[7] Q. T. Ngo et al., "Two-Tier Cache-Aided Full-Duplex Hybrid Satellite–Terrestrial Communication Networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 3, pp. 1753–1765, 2022.

[8] X. Zhu et al., "Cooperative Multilayer Edge Caching in Integrated Satellite-Terrestrial Networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 2924–2937, 2022.

[9] K. An et al., "On the Performance of Cache-Enabled Hybrid Satellite-Terrestrial Relay Networks," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1506–1509, 2019.

[10] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Dynamic Bandwidth Allocation and Precoding Design for Highly-Loaded Multiuser MISO in Beyond 5G Networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1794–1805, 2022.

[11] P. J. Honnaiah et al., "Demand-Based Adaptive Multi-Beam Pattern and Footprint Planning for High Throughput GEO Satellite Systems," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1526–1540, 2021.

[12] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[13] "Movielens 1M Dataset," GroupLens. [Online]. Available: https://grouplens.org/datasets/movielens/1m/.

[14] "SES-14 Satellite," SES. [Online]. Available: https://www.ses.com/our-coverage/explore/satellite/369.

[15] Avionics Department of NAWCWD, "Electronic Warfare and Radar Systems Engineering Handbook," Naval Air Warfare Center Weapons Division, USA, Fourth Edition, 2013.

[16] "5G NR: Multiplexing and channel coding," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 138.211, version 16.2.0.

[17] K. -X. Li et al., "Downlink Transmit Design for Massive MIMO LEO Satellite Communications," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 1014–1028.

[18] B. Dinesh and K. Sachin, "FastForward: Fast and constructive full duplex relays," *Proc. ACM Conf. SIGCOMM*, USA, 2014, pp. 199–210.

[19] S. Cakaj, "The parameters comparison of the 'starlink' leo satellites constellation for different orbital shells," *Frontiers in Communications and Networks*, vol. 2, 2021.