Coded Caching at the Edge of Satellite Networks

Estefanía Recayte

Institute of Communications and Navigation of DLR (German Aerospace Center),
Wessling, Germany. Email: {estefania.recayte}@dlr.de

Abstract—Caching multimedia contents at the network edge is a key solution to decongest the amount of traffic in the backhaul link. In this paper, we extend and analyze the coded caching technique [1] in an unexplored scenario, i.e. at the edge of two-tier heterogeneous networks with an arbitrary number of users. We characterize the performance of such scheme by deriving a closed-form expression of the average backhaul load and reveal a significant gain compared to other benchmark caching schemes proposed in the literature.

I. INTRODUCTION

Caching has emerged as one of the key technologies for next-generation wireless systems. Bringing the desired content to the edge of the network, i.e. memorizing copies of relevant information close to users, has been shown not only to alleviate the backhaul traffic, but also to significantly reduce latency and power consumption [2].

To achieve this goal, a two-step caching strategy is typically implemented, pre-fetching the content at the edge (e.g. at base stations, LEO satellites, relays or helpers) during network off-peak periods (*placement phase*), so as to serve the users without consuming backhaul capacity when the network is congested (*delivery phase*).

Several works have recently investigated the potential benefits of caching schemes at the edge. Interesting results were obtained in [3], where authors compared two different placement approaches, i.e. encoded and uncoded content placement to reduce the download delay. Caching scheme based on maximum distance separable (MDS) codes have been studied in wireless networks to minimize the expected download time [4] or to reduce the amount of data to be sent [5].

Furthermore, Maddah-Ali et al. introduced in their pioneering work [1] the concept of *coded caching*, considering local caching directly on the user's device. The idea is to deliver coded content and leverage the user's local content to serve multiple users with a single transmission. This technique has spurred an extensive body of research providing a solid understanding of the potential and limitations of caching, e.g. [6], [7]. However, coded caching has been studied especially in setups where few cache-aided users are connected to a common server via a shared link. Instead, its potential and the trade-offs it may induce in other relevant scenarios remain unexplored to date. An example of notable practical relevance is given by two-tier networks that foresee a satellite component, which will be an integrating part of 5G and 6G systems [8]. In these settings, commonly referred to as non-terrestrial networks, terminals may not be equipped with direct satellite connectivity, and the intermediate tier is responsible for forwarding content from one end to the other.

To shed light on these relevant design aspects, the present work considers the application of coded caching in a two-tier caching satellite network for an arbitrary number of users. Unlike previous work, caching is considered at the edge of the network (e.g. relays, helpers, LEO satellites) and multiple users are connected to one or more cache-aided relays. To analyze the system performance, we derive a closed-form expression of the average backhaul transmission load when coded caching is in place. We compare our results with the benchmark given by the MDS caching scheme proposed in literature and we show a reduction in backhaul transmissions. In particular, the presented scheme triggers a gain each time the mutual difference of sets representing the file requested at each relay is not empty. To quantify such gain, the distribution of how users request for content is casted onto a combinatorial balls into bins (BiB) setting. Due to complexity of the problem, closed form expressions are derived when the distribution of files requested is uniform, while we show via Monte-Carlo that the aforementioned gain is also present when files are not equiprobable. The significant enhancement obtained encourages to investigate further relevant aspect of the edge coded caching in satellite networks.

Notation: We use capital letters, e.g. X, for discrete random variables (rvs) and their lower case counterparts, e.g. x, for their realizations. The probability mass function (pmf) of the rv X is denoted as $p_X(x)$ and conditional pmfs as $\Pr\{X = x \mid Y = y\} = p_X(x|y)$. A set is denoted with calligraphic letters, e.g. $\mathcal{S}$. The cardinality of set $\mathcal{S}$ is indicated as $|\mathcal{S}|$.

II. SYSTEM MODEL

A two-tier heterogeneous network composed by a master node, two cache-enabled nodes and a number of end users are considered. While this setup applies to different network configurations, we will take as reference throughout our discussion the satellite topology illustrated in Fig. 1. Here, a satellite (S) stores a whole library $\mathcal{F} = \{f_1, \cdots, f_N\}$ of equal-size files. On the ground, two cache-enabled relays (R_B and R_W)[1] are connected via a backhaul link to S, and each one provides connectivity to some users. As typical in current satellite-aided terrestrial networks, we assume that no direct link between users and S is available[2]. Depending on their locations users (terminals) may be connected to one or both relays. $\mathcal{U}_h$ denotes the subset of users that are connected to h

---

[1]The subscript B and W have been chosen to facilitate the similarity between the caching scheme and BiB problem, as will become clear later.

[2]Note that the setup presented is not limited to this architecture. For instance, a possible scenario may consist of a GEO satellite which acts as master node connected via backhaul links to cache-enabled LEO satellites.
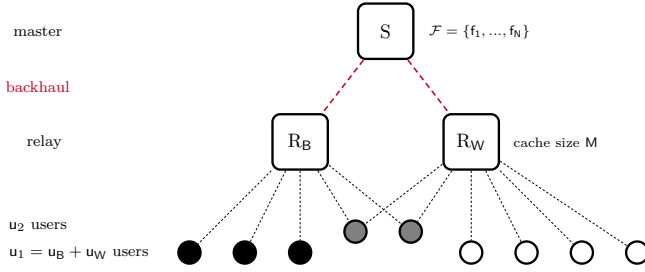
Fig. 1. System model: cache-aided relays are connected to the satellite through the backhaul link and users can be connected to one or more relays.

relays, where $h = \{1, 2\}$, whereas $\mathcal{U}_B$ ($\mathcal{U}_W$) denotes the subset of users connected only to relay R_B (R_W). Note that the set of users connected to exactly one relay is $\mathcal{U}_1 = \mathcal{U}_B \cup \mathcal{U}_W$.

Each relay can store up to $M \leq N$ files locally. With $\mathcal{Z}_B$ ($\mathcal{Z}_W$) we indicate the files present in the cache of R_B (R_W). During the *placement phase*, which is carried out offline, each file $f_j \in \mathcal{F}$ is partitioned into n_F equally long fragments, i.e. f_j is fragmented as $\{f_j^{(1)}, \cdots, f_j^{(n_F)}\}$. Each cache stores F_j fragments related to file f_j according to one of the caching schemes that will be introduced later in this section. During the *delivery phase*, the network serves the user's requests. Each user picks a file according to the file distribution considered (e.g. uniform or Zipf distribution). A user connected to h relays which request for f_j receives $h F_j$ fragments directly from the cached content and the $\max(0, n_F - hF_j)$ missing fragments are forwarded by the R after being retrieved by S via the backhaul link. The transmission technique in the backhaul link depends on the caching scheme considered. With $\mathcal{D}_x \subseteq \mathcal{F}$ we denote the subset of files requested by the set of users $\mathcal{U}_x$, where $x \in \{1, 2, B, W\}$. For example, the cardinality of $\mathcal{D}_1$ is the number of different files requested by users connected to a single relay (users in $\mathcal{U}_1$).

In such configuration, we indicate with u the total number of terminals that concurrently request content from the library, each independently choosing a file to download. Specifically, we have that $u = u_1 + u_2$ where $|\mathcal{U}_1| = u_1$, are those connected to a single relay, while $|\mathcal{U}_2| = u_2$ are those connected to both relays. We also have that $u_1 = u_B + u_W$ where u_B are the users requesting content only at R_B and u_W only at R_W. A relay directly delivers content present in its cache and retrieves content that is not available locally via the backhaul link. For simplicity we assume that all transmissions are error-free.

Following this notation, we analyze a coded caching scheme at the edge based on the strategy proposed in [1], referred to as the *edge coded caching scheme* (ECC). To evaluate the scheme, we derive the average backhaul transmission load L, i.e. the average number of packets that S should transmit in the backhaul link to satisfy user requests. We will focus on the rv L^x which describes the number of transmissions required from S for a given caching scheme x. We have that $L^x = \mathbb{E}[L^x]$ where the operator $\mathbb{E}$ indicates the expected value. The metric L is used to compare the behavior against the benchmark given by MDS scheme. Let us discus both schemes in the following.

MDS Caching Scheme

In the MDS scheme [5], the network works by caching and delivering packets that are encoded. In particular, n_F fragments of file f_j are used to create $n > n_F$ encoded packets using a (n, n_F) MDS code. The set of encoded packets related to f_j can be written as $e_j = \{e_j^{(1)}, \cdots, e_j^{(n)}\}$ where $e_j^{(i)}$ and $f_j^{(i)}$ are equally long for every i and j. With the MDS coding technique, a user can reconstruct successfully the requested file by receiving any subset of n_F encoded packets [5].

If we assume a uniform distribution of the file requested then it is also assumed that files are split into $n_F = N$ fragments. Each relay fills own cache with $F_j = M$ encoded packets per file such that $\mathcal{Z}_1 \cap \mathcal{Z}_2 = \emptyset$, i.e. relays store a different subset of encoded packets for the same file. The satellite keeps $n - 2M$ encoded packets for every file. The delivery phase is split into the following stages. First, users receive content from the relays' cache, subsequently the missing encoded packets are sent by S through the backhaul link to the R which forwards them to the users. The benefit of this strategy is based on being able to serve both relays in parallel with a single transmission via the backhaul link. This occurs whenever there are requests for the same content at both relays. To clarify, consider the following numerical example.

Example 1. *Let us assume to have two users: user 1 is connected only to R_B and user 2 only to R_W. Consider a memory size of $M = 1$ and two equiprobable files split into $n_F = 2$ fragments*

$$f_1 = \{f_1^{(1)}, f_1^{(2)}\} \text{ and } f_2 = \{f_2^{(1)}, f_2^{(2)}\}.$$

Let us consider a $(3, 2)$ MDS code such that we can write the encoded packets as

$$e_1 = \{e_1^{(1)}, e_1^{(2)}, e_1^{(3)}\} \text{ and } e_2 = \{e_2^{(1)}, e_2^{(2)}, e_2^{(3)}\}.$$

We further set $\mathcal{Z}_B = \{e_1^{(1)}, e_2^{(1)}\}$ and $\mathcal{Z}_W = \{e_1^{(2)}, e_2^{(2)}\}$.

To characterize the average backhaul load L^{MDS}, we shall consider two cases. First, we suppose that users are requesting for different content, i.e. user 1 (user 2) requests for f_1 (f_2) to relay R_B (R_W). Since each R has one encoded packet of the requested file in cache, S should send one encoded packet to each R. Hence, the number of required backhaul transmissions, i.e. the realization of l of the rv L^{MDS} takes value

$$l_1 = 2.$$

Each user is able to reconstruct the file by receiving one encoded packet directly from the cache and the other forwarded by the relay. If, instead, both users request for the same content, S can only transmit the encoded packet $e_i^{(3)}$ to both and they will successfully decode the requested content. In this case, the number of packets to transmit is

$$l_2 = 1.$$

Combining the two cases, L in MDS evaluates to

$$L^{MDS} = \sum_i p_L(l_i) \, l_i = \frac{1}{2} \, l_1 + \frac{1}{2} \, l_2 = 1 + \frac{1}{2} = \frac{3}{2}$$

where we sum over the i possibilities on how users can request for the library content. They ask for different files with

probability $p(l_1) = 1/2$ *while they ask for the same file with* probability $p(l_2) = 1/2$.

Edge Coded Caching Scheme

In the ECC scheme, caches are filled with non-encoded fragments, while the encoding takes place in the delivery phase. S creates coded delivery opportunities so that with a unique transmission both relays are able to recover the desired information also when different content is requested.

In the placement phase, each R fills its cache with $F_j \, \mathsf{n_F}$ exclusive fragments of file f_j so that relays have different fragments of the same file. S is aware of which content has been stored in each R. In the second phase, users make their requests to the corresponding R. The delivery can be split into three stages. In the first stage, a user receives fragments directly from the cached content of the associated R. In the second stage, S is informed of users requests and provides missing content over the shared bachkaul link by creating coded multicast opportunities transmissions when is possible. In this stage the relays decode the transmission and forward the desired packets to users. In the third and last stage, S sends the remaining content in a non-encoded transmission, and relays forward this to users.

A coded multicast opportunity allows both relays to retrieve file fragments with a single transmission. In particular, S creates a coded packet by XORing two fragments (i.e., a bitwise operation). S picks a fragment of a file requested at R_B and present in cache of R_W and vice-versa and combines them for delivering. In this way, each R receives a coded packet which is composed by a fragment present in own memory and a desired fragment. Each R by XORing the received packet with the corresponding fragment in cache obtains a fragment of the requested file. ECC generates a gain over the MDS caching scheme whenever relays have disjoint requests. Let us clarify the last statement by considering the setting discussed in Example 1.

Example 2. *Let us assume to have one user per relay and the following cache placement:* $\mathcal{Z}_B = \{f_1^{(1)}, f_2^{(1)}\}$ *and* $\mathcal{Z}_W = \{f_1^{(2)}, f_2^{(2)}\}$.

Consider first the case where users request for different content. For example, user 1 (user 2) requests for f_1 (f_2) to R_B (R_W). During the first stage, user i receives $f_i^{(i)}$ from the cache of the related R. At the second stage, S sends the following coded packet $p = f_2^{(1)} \oplus f_1^{(2)}$. Thus the number of packets transmitted over the backhaul, l_1, is

$$l_1 = 1.$$

R_B (R_W) reconstructs the missing fragment by computing $p \oplus f_2^{(1)}$ ($p \oplus f_1^{(2)}$). Similarly, when users request for the same file, both are satisfied with a single coded transmission, i.e.

$$l_2 = 1.$$

In the ECC scheme, L *is then*

$$\mathsf{L}^{ECC} = \sum_i p_L(l_i) \, l_i = \frac{1}{2} \, l_1 + \frac{1}{2} \, l_2 = 1$$
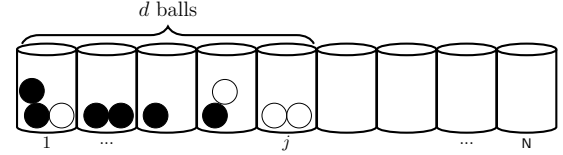

Fig. 2. Caching requests represented as the BiB problem. Bins represent files while balls represent users' requests. The occupancy problem lies on calculating the probability of having exactly j bins not empty after throwing d balls, i.e. the probability that d users request exactly for j different files.
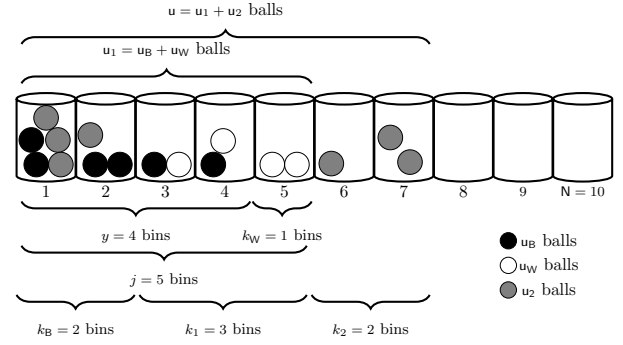

Fig. 3. Requests represented as the BiB problem. Black balls represents requests from users connected only to R_B, while balls request only to R_W while gray balls represents request from users connected to both relays.

With the presented examples, we observe that there exists a gain in the ECC scheme whenever there are requests for different files at relays. To understand the potential of this gain, we derive L in both schemes in a more general setting. To this aim, we start by recalling some useful results of the BiB problem, which will be later applied to our derivations.

III. BALLS INTO BINS PROBLEM APPLIED TO CACHING

To instantiate such calculations, it is convenient to map our setting onto a balls into bins setup. The general BiB problem, see e.g. [9], consists in independently throwing d balls into N bins. As illustrated in Fig. 2, this can be cast to our caching problem by having each bin associated to a file of the library, and by having balls which represent user requests. Following this parallel, the possibility for more balls to land into the same bin corresponds to have multiple users asking for a common library element. A first useful result is given by the probability of having exactly j bins out of N non empty given that d balls are thrown uniformly at random, which was derived in [10] and can be written as

$$p_J(j|d) = \frac{\binom{\mathsf{N}}{j} S(d,j) \, j!}{\mathsf{N}^d} \tag{1}$$

where $S(d, j)$ is the Stirling number of second kind, i.e.,

$$S(d,j) = (j!)^{-1} \sum_{i=0}^{d} (-1)^i \binom{d}{i} (j-i)^d.$$

In our setup, we further need to differentiate requests made to R_B from those made to R_W and, similarly, requests made by users connected to one relay (set $\mathcal{U}_1$), from those made by users connected to two relays (set $\mathcal{U}_2$). To this aim we

TABLE I
LIST OF SOME OF THE RANDOM VARIABLES

rv	Definition	Alphabet
J	$|\mathcal{D}_1|$	$\{1, ..., \beta_J = \min(\mathsf{u}_1, \mathsf{N})\}$
Y	$|\mathcal{D}_\mathsf{B}|$	$\{1, ..., \beta_Y = \min(\mathsf{u}_\mathsf{B}, \mathsf{N})\}$
K_B	$|\mathcal{D}_\mathsf{B}\backslash\mathcal{D}_\mathsf{W}|$	$\{\alpha_\mathsf{B} = \max(0, y - \mathsf{u}_\mathsf{W} + k_\mathsf{W}),$ $..., \beta_\mathsf{B} = \min(\mathsf{u}_\mathsf{B}, \mathsf{N})\}$
K_W	$|\mathcal{D}_\mathsf{W}\backslash\mathcal{D}_\mathsf{B}|$	$\{0, ..., \beta_\mathsf{W} = \min(\mathsf{u}_\mathsf{W}, \mathsf{N} - \beta_\mathsf{B})\}$
K_1	$|\mathcal{D}_1\backslash\mathcal{D}_2|$	$\{\alpha_1 = \max(0, j - \mathsf{u}_2),$ $..., \beta_1 = \min(\mathsf{u}_1, \mathsf{N})\}$
K_2	$|\mathcal{D}_2\backslash\mathcal{D}_1|$	$\{0, ..., \beta_2 = \min(\mathsf{u}_2, \mathsf{N} - \beta_1)\}$
Z	$\min\{K_\mathsf{B}, K_\mathsf{W}\}$	$\{0, ..., y\}$

distinguish requests at different relays, as illustrated in Fig. 3, by considering balls of two different colors, e.g. black and white balls. A bin containing black and white balls indicates that the same file is required at both relays. Having the i-th bin with only black (white) balls implies that the i-th file was requested only at relay R_B (R_W).

Following this approach, a useful result is offered by the *multivariate occupancy problem* assuming that there are N bins and that u_B black balls have been thrown and have occupied j different bins. The probability that, after throwing u_W white balls, there are exactly k_B bins containing only black balls and k_W bins containing only white balls is [9]

$$\mathsf{P}_\mathsf{oc}(j, k_\mathsf{B}, k_\mathsf{W}, \mathsf{u}_\mathsf{W}) = \binom{j}{k_\mathsf{B}}\binom{\mathsf{N}-j}{k_\mathsf{W}}\frac{\Delta^{b_\mathsf{W}}\mathbf{0}^{\mathsf{u}_\mathsf{W}}}{\mathsf{N}^{\mathsf{u}_\mathsf{W}}}$$

where b_W is the number of bins containing the u_W white balls, i.e. $b_\mathsf{W} = j - k_\mathsf{B} + k_\mathsf{W}$ and the quantity

$$\Delta^m\mathbf{0}^n := \sum_{i=0}^{m}(-1)^i\binom{m}{i}(m-i)^n$$

is known as *difference of zeros* [9]. Note that, in our setting, $\mathsf{P}_\mathsf{oc}(j, k_\mathsf{B}, k_\mathsf{W}, \mathsf{u}_\mathsf{W})$ provides the probability that exactly k_B files are requested only to relay R_B and k_W files are requested only to relay R_W when in total there are $\mathsf{u}_1 = \mathsf{u}_\mathsf{B} + \mathsf{u}_\mathsf{W}$ users connected to exactly one relay.

IV. AVERAGE BACKHAUL TRANSMISSION LOAD

Leaning on the parallel with the BiB problem, we now derive the mean number of packets/fragments that S needs to send via the backhaul link to satisfy u requests.

For convenience, we list in Table I the rvs needed for our derivations together with their definition and alphabet. The first column indicates the notation of the rv, the second its definition and the last its alphabet. For instance, the rv J denotes the number of different files requested by u_1 users connected to only one relay ($\mathcal{U}_1$) while K_B denotes the number of different files requested exclusively at R_B. Instead, the notation $\mathcal{D}_\mathsf{B}\backslash\mathcal{D}_\mathsf{W}$ indicates the set difference and that is the set of file requested at R_B but not requested at R_W. Let us clarify all the mentioned quantities with an example.

Example 3. *Let us refer to Fig. 3 which íllustrates a library of $\mathsf{N} = 10$ files (bins) and $\mathsf{u} = 17$ users (balls). There are $\mathsf{u}_\mathsf{B} = 6$ users connected only to R_B (black balls), $\mathsf{u}_\mathsf{W} = 4$ only*

to R_W *(white balls), while $\mathsf{u}_2 = 7$ are connected to both relays (grey).*

We have that $\mathsf{u}_\mathsf{B} = 6$ users requested in total $y = 4$ different files (represented by the four bins with black balls). Users in $\mathcal{U}_2$ asked in total for three files and are represented by bins 3, 4 and 5. The files requested by the u_1 users connected to only one relay are in total $j = 5$, i.e. the number of bins with black or with balls. Out of those, the files requested exclusively at R_B are $k_\mathsf{B} = 2$, i.e. the number of bins with black balls and without white balls, while those exclusively requested at R_W are $k_\mathsf{W} = 1$, i.e. the number of bins with white balls and without black balls. Since the minimum number of mono-colour bins is one, then $z = 1$, i.e. $z = \min\{k_\mathsf{B}, k_\mathsf{W}\}$.

Users connected to both relays requested in total for 4 further files, represented by bins 1, 2, 6 and 7. Then the files requested only at one R are bins 3, 4 and 6 so in total $k_1 = 3$, while files requested only by users connected to both relays are bins 6 and 7 such that $k_2 = 2$.

In the next derivations we assume that files are equiprobable, each file is split into $\mathsf{n}_\mathsf{F} = \mathsf{N}$ fragments and the number of files stored at each relay is $F_j = \mathsf{M}$ for all j.

A. MDS Average Transmision Load

Let us recall that J different files are requested by u_1 users in $\mathcal{U}_1$. By the working principle of the MDS scheme, for each file requested, S has to send in the backhaul $\mathsf{n}_\mathsf{F} - \mathsf{M}$ packets, whereas the remaining M are already provided to the user via the relay's cache. The overall number of packets that S transmits to satisfy u_1 requests is then expressed by the rv

$$L_1^{\mathrm{MDS}} = (\mathsf{n}_\mathsf{F} - \mathsf{M})\,J. \tag{3}$$

To complete the analysis, we derive the number of packets needed to satisfy users connected to both relays (users in $\mathcal{U}_2$). Note that in this calculation it is needed to take into account only the K_2 *aggregated* requests, i.e. the new files requested by $\mathcal{U}_2$ users but not requested by $\mathcal{U}_1$. In fact, whenever a file requested by users in $\mathcal{U}_2$ coincides with a user request from $\mathcal{U}_1$, both requests are satisfied with the same backhaul transmission already accounted for by L_1^{MDS}. Observing that each user in $\mathcal{U}_2$ receives in total $2\mathsf{M}$ different fragments of the respective file from relays, the number of packets that S has to send for each aggregated file is $(\mathsf{n}_\mathsf{F} - 2\mathsf{M})^+$ where $(x)^+ := \max(0, x)$. Note that whenever $\mathsf{M} \geq \mathsf{N}/2$, no transmission in needed.

Combining these remarks, the transmissions that S has to perform to satisfy the aggregated requests can be expressed as

$$L_2^{\mathrm{MDS}} = (\mathsf{n}_\mathsf{F} - 2\mathsf{M})^+\,K_2. \tag{4}$$

So that the average backhaul load L in the MDS is

$$\mathsf{L}^{\mathrm{MDS}} = \mathbb{E}\big[L_1^{\mathrm{MDS}}\big] + \mathbb{E}\big[L_2^{\mathrm{MDS}}\big]. \tag{5}$$

Let us now calculate the two addends of equation (5).

The average transmission load for users in $\mathcal{U}_1$ can be computed by simply averaging over J to obtain

$$\overline{\mathsf{L}}^{\mathrm{MDS}} = \sum_{j=1}^{\beta_J} \frac{\binom{\mathsf{N}}{j} S(\mathsf{u}_1,j)j!}{\mathsf{N}^{\mathsf{u}_1}} \Big[j\big(1-\frac{\mathsf{M}}{\mathsf{N}}\big) + \big(1-\frac{2\mathsf{M}}{\mathsf{N}}\big)^+ \sum_{k_1=\alpha_1}^{j} \binom{j}{k_1} \sum_{k_2=0}^{\beta_2} \binom{\mathsf{N}-j}{k_2-1} \sum_{i=0}^{\mathsf{N}-b_2} (-1)^i \binom{\mathsf{N}-b_2}{i} \Big(\frac{\mathsf{N}-b_2-i}{\mathsf{N}}\Big)^{\mathsf{u}_2} \Big]. \quad (10)$$

$$\begin{aligned}
\mathsf{L}_1^{\mathrm{MDS}} &= \mathbb{E}\big[(\mathsf{n}_F - \mathsf{M})\, J\big] \\
&= (\mathsf{n}_F - \mathsf{M}) \sum_{j=1}^{\beta_J} p_J(j|\mathsf{u}_1) \cdot j \\
&= (\mathsf{N} - \mathsf{M}) \sum_{j=1}^{\beta_J} \frac{\binom{\mathsf{N}}{j} S(\mathsf{u}_1,j)\, j!}{\mathsf{N}^{\mathsf{u}_1}} \cdot j
\end{aligned} \quad (6)$$

where the quantity $p_J(j|\mathsf{u}_1)$ was derived with BiB occupancy problem, see (1).

The average transmission load given by the aggregated files requested by $\mathcal{U}_2$ can be computed by conditioning to J, i.e.

$$\mathsf{L}_2^{\mathrm{MDS}} = \mathbb{E}_J\Big[\mathbb{E}\big[(\mathsf{n}_F - 2\mathsf{M})^+ K_2|J\big]\Big]. \quad (7)$$

Let us first focus on the inner expectation, and derive the conditional pmf $p_{K_2}(k_2|j)$, i.e. the probability that users connected to both relays request for exactly k_2 new files given that j different files have been requested by users connected to one relay. To help the reader, we refer to Fig. 3 the sought probability can be computed in the BiB setup as the probability of having $j+k_2$ non empty bins after throwing u_2 (grey) balls, conditioned on having already j non empty bins occupied by u_1 balls. As discussed, this results is offered by the multivariate occupancy problem, and we have

$$p_{K_2}(k_2|j) = \sum_{k_1=\alpha_1}^{j} \mathsf{P}_{\mathrm{oc}}(j,k_1,k_2,\mathsf{u}_2), \quad (8)$$

where the correspondent number of file requested at both relays is $b_2 = j - k_1 + k_2$. In (8) we are summing up all the possible values that k_1 can assume (i.e. files exclusively requested at R_B represented by bins with only black balls). Accordingly,

$$\begin{aligned}
\mathsf{L}_2^{\mathrm{MDS}} &= \mathbb{E}_J\Big[(\mathsf{n}_F - 2\mathsf{M})^+ \sum_{k_2=0}^{\beta_2} k_2\, p_{K_2}(k_2|J)\Big] \\
&= (\mathsf{N} - 2\mathsf{M})^+ \sum_{j=0}^{\beta_J} p_J(j|\mathsf{u}_1) \sum_{k_2=0}^{\beta_2} k_2 \sum_{k_1=\alpha_1}^{j} \mathsf{P}_{\mathrm{oc}}(j,k_1,k_2,\mathsf{u}_2).
\end{aligned} \quad (9)$$

By inserting (6) and (9) onto (5) we obtain $\mathsf{L}^{\mathrm{MDS}}$ and normalizing by the number of fragments N, we have that the normalized average transmission load, $\overline{\mathsf{L}}^{\mathrm{MDS}} = \mathsf{L}^{\mathrm{MDS}}/\mathsf{N}$, in the MDS scheme is given in (10) at the top of the page.

B. ECC Average Transmission Load

Let us start by considering users in $\mathcal{U}_1$. Since M fragments of the requested files are obtained from the relay's cache then each user needs $\mathsf{n}_F - \mathsf{M}$ additional fragments. Let us calculate the number of packets that S should transmit to satisfy these requests by considering the coded caching opportunities. Denoting by Y the rv counting the number of different files requested by the u_B users connected only to R_B and by K_W

the rv counting the number of files exclusively requested by the u_W connected only to R_W and not requested to R_B, in total users have to receive $(\mathsf{n}_F - \mathsf{M})(Y + K_W)$ fragments in order that all their requests are satisfied. However, note that transmissions given by the coded opportunities should not be counted. As for Example 2, a coded transmission opportunity take places each time that a file is requested at one relay and not in the other and vice-versa. S by XORing the corresponding content present at each cache can make a transmission useful to both relays. Each coded transmission opportunity allows the S to generate ω_1 XORed packets involving the two files. In particular,

$$\omega_1 = \min(\mathsf{M}, \mathsf{N} - \mathsf{M}),$$

where ω_1 is the number of fragments per file combined in a coding opportunity and it depends on the cache size. For each transmission opportunity, when $\mathsf{M} \le \mathsf{N}/2$; then in total M fragments per file are XORed, whereas if $\mathsf{M} > \mathsf{N}/2$ then requests are satisfied by combining $\mathsf{N} - \mathsf{M}$ fragments per file.

In summary, each coded transmission opportunity allows S to combine ω_1 packets where a packet is formed by two encoded fragments. Accordingly, the overall number of transmission needed in the backhaul to serve users in $\mathcal{U}_1$ is

$$L_1^{\mathrm{ECC}} = (\mathsf{n}_F - \mathsf{M})(Y + K_W) - \omega_1 Z$$

where Z is the rv denoting the number of coded opportunities.

Let us now consider the users connected to both relays, i.e. the set $\mathcal{U}_2$. In this case, we simply observe that no gain opportunity emerges from the aggregated requests by such users. In fact, users already receive content from both caches. Therefore, the value of the backhaul transmissions is the same as computed for the MDS scheme and we get

$$L_2^{\mathrm{ECC}} = L_2^{\mathrm{MDS}}.$$

The average bachkaul transmission load of the ECC is

$$\mathsf{L}^{\mathrm{ECC}} = \mathbb{E}\big[L_1^{\mathrm{ECC}}\big] + \mathbb{E}\big[L_2^{\mathrm{ECC}}\big]. \quad (11)$$

where we need to derive only $\mathbb{E}\big[L_1^{\mathrm{ECC}}\big]$. Conditioning on Y, we have

$$\begin{aligned}
\mathsf{L}_1^{\mathrm{ECC}} &= \mathbb{E}_Y\Big[\mathbb{E}\big[(\mathsf{n}_F - \mathsf{M})(Y + K_W) - \omega_1 Z|Y\big]\Big] \\
&= \mathbb{E}_Y\Big[\mathbb{E}\big[(\mathsf{n}_F - \mathsf{M})(Y + K_W)|Y\big]\Big] - \mathbb{E}_Y\Big[\mathbb{E}\big[\omega_1 Z|Y\big]\Big].
\end{aligned} \quad (12)$$

Let us first focus on the conditional distribution of K_W. Given $Y = y$ different files requested from users in $\mathcal{U}_B$, the probability $p_{K_W}(k_W|y)$ of having exactly k_W files requested only at R_W can be derived from the BiB multivariate occupancy problem by considering all values that k_B can assume as

$$p_{K_W}(k_W|y) = \sum_{k_B=\alpha_B}^{y} \mathsf{P}_{\mathrm{oc}}(y,k_B,k_W,\mathsf{u}_W) \quad (13)$$

where $b_W = y - k_B + k_W$.
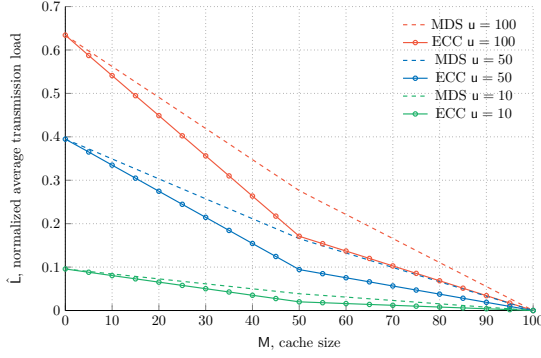
Fig. 4. $\hat{\mathsf{L}}$ versus M for $\mathsf{u} = 10, 50, 100$ for $\mathsf{N} = 100$. 40% of the users are connected to $\mathsf{R_B}$, 40% to $\mathsf{R_W}$ and 20% to both relays. Given u, the marked and dashed curves indicate the results for the ECC and the MDS scheme respectively.

Similarly, the probability $p_Z(z|y)$ of having $Z = z$ coded transmission opportunities conditioned on $Y = y$ files, can be computed considering two disjoint events. The first is that $\mathsf{u_B}$ users ask exclusively for exactly z files at $\mathsf{R_B}$ and $\mathsf{u_W}$ users have ask at least z exclusively files at $\mathsf{R_W}$. The second is the probability that $\mathsf{u_B}$ users ask exclusively for more than z files at $\mathsf{R_B}$ and $\mathsf{u_W}$ users ask exactly z exclusively files at $\mathsf{R_W}$. Thus, we can write

$$p_Z(z|y) = \sum_{k_\mathsf{W}=z}^{\beta_\mathsf{W}-y+z} \mathsf{P_{oc}}(y,z,k_\mathsf{W},\mathsf{u_1}) + \sum_{k_\mathsf{B}=z+1}^{\min(y,\beta_\mathsf{B})} \mathsf{P_{oc}}(y,k_\mathsf{B},z,\mathsf{u_1}) \tag{14}$$

where $b_\mathsf{W} = y - z + k_\mathsf{W}$ and $b_\mathsf{B} = y - k_\mathsf{B} + z$. If we now plug (13) and (14) into (11) and we remove the condition on Y, we obtain

$$\mathsf{L_1^{ECC}} = \sum_{y=1}^{\beta_\mathsf{B}} p_Y(y|\mathsf{u_B}) \Big[(\mathsf{N} - \mathsf{M})\Big(y + \sum_{k_\mathsf{W}=0}^{\beta_\mathsf{W}} k_\mathsf{W}\, p_{K_\mathsf{W}}(k_\mathsf{W}|y) \Big)$$
$$- \min(\mathsf{M}, \mathsf{N} - \mathsf{M}) \sum_{z=0}^{y} z\, p_Z(z|y) \Big]. \tag{15}$$

The final result in (16) is obtained by adding the expression of $\mathsf{L_2^{ECC}}$ to (15) which is the average transmission load of the ECC scheme normalized to the number of fragment $n_\mathsf{F} = \mathsf{N}$, $\bar{\mathsf{L}}^{ECC} = \mathsf{L}^{ECC}/\mathsf{N}$.

V. RESULTS

We evaluate the average backhaul load for the benchmark MDS scheme and compare it with the one of the proposed ECC scheme. In both cases quantities are normalized to the library size N, i.e. $\hat{\mathsf{L}}^{MDS} = \mathsf{L}^{MDS}/\mathsf{N}$ and $\hat{\mathsf{L}}^{ECC} = \mathsf{L}^{ECC}/\mathsf{N}$. We assume the library size $\mathsf{N} = 100$ and 20% of the users to be connected to both relays while 80% to a single relay. For simplicity, we consider that of half users in $\mathcal{U}_1$ are connected only to $\mathsf{R_B}$ and half only to $\mathsf{R_W}$.

In Fig. 4, the normalized average backhaul load as a function of the cache size M for different number of users u is plotted. As shown, the ECC scheme outperforms the benchmark MDS caching scheme for every number of users u considered. As expected, by fixing u requests, $\bar{\mathsf{L}}$ decreases by increasing M, since more content directly from the cache can be provided. Given M, the gain between ECC and the MDS scheme is higher when the number of total users u is greater because more transmission opportunities take place. The maximum gain is obtained when $\mathsf{M} = \mathsf{N}/2$, in fact, this cache operating point encodes half of file content (the maximum portion of a file that can be combined) in a transmission opportunity.

Motivated by the good performance obtained, we also show Monte-Carlo results when file request distribution is not equiprobable. The normalized average backhaul load in this case is reported in Fig. 5. It is assumed that users request for content according the Zipf distribution with $\alpha = 0.80$ and each relay optimizes own cache content according the algorithm given in [5]. In this set up, we can appreciate the efficiency of the caching placement due to the not uniform demands. In fact, given the number of users u, a cache size M and a scheme then L is lower than in our previous scenario. A gain on the ECC with respect to MDS is still present. Due to the lower number of coding opportunities and due to the placement considered such gain is smaller with respect to our previous results.

VI. CONCLUSIONS

We applied the Maddah-Ali caching scheme at the edge of a two-tier heterogeneous satellite network with multiple users. A closed-form expression of the average backhaul transmission load for the ECC scheme was derived. The performance of the scheme was compared with those of the benchmark given by the MDS one. We quantified the nature of the transmission gain obtained by casting out problem with known results obtained in the BiB setting. Results shows a gain in terms of backhaul transmissions for any number of users considered in the system.

$$\bar{\mathsf{L}}^{ECC} = \sum_{y=1}^{\beta_\mathsf{B}} \frac{\binom{\mathsf{N}}{y} S(\mathsf{u_B}, y)\, y!}{\mathsf{N}^{\mathsf{u_B}}} \Bigg\{ \sum_{k_\mathsf{W}=0}^{\beta_\mathsf{W}} \binom{\mathsf{N}-y}{k_\mathsf{W}-1}\Big[(y+k_\mathsf{W})\Big(1-\frac{\mathsf{M}}{\mathsf{N}}\Big)\Big] \Bigg[\sum_{k_\mathsf{B}=\alpha_\mathsf{B}}^{y} \binom{y}{k_\mathsf{B}} \sum_{i=0}^{\mathsf{N}-b_2}(-1)^i \binom{\mathsf{N}-b_2}{i}\Big(\frac{\mathsf{N}-b_2-i}{\mathsf{N}}\Big)^{\mathsf{u}} \Bigg] - \frac{\omega_1}{\mathsf{N}}$$

$$\sum_{z=1}^{y} z\Bigg[\binom{y}{z}\sum_{k_\mathsf{W}=z}^{\beta_\mathsf{W}-y-z}\binom{\mathsf{N}-y}{k_\mathsf{W}}\sum_{i=0}^{\mathsf{N}-b_\mathsf{W}}(-1)^i\binom{\mathsf{N}-b_\mathsf{W}}{i}\Big(\frac{\mathsf{N}-b_\mathsf{W}-i}{\mathsf{N}}\Big)^{\mathsf{u_W}} + \binom{\mathsf{N}-y}{z}\sum_{k_\mathsf{B}=z+1}^{\min(y,\beta_\mathsf{B})}\binom{y}{k_\mathsf{B}}\sum_{i=0}^{\mathsf{N}-b_\mathsf{B}}\binom{\mathsf{N}-b_\mathsf{B}}{i}(-1)^i \tag{16}$$

$$\Big(\frac{\mathsf{N}-b_\mathsf{B}-i}{\mathsf{N}}\Big)^{\mathsf{u_W}}\Bigg]\Bigg\} + \Big(1-\frac{2\mathsf{M}}{\mathsf{N}}\Big)^+ \sum_{j=1}^{\beta_J} \frac{\binom{\mathsf{N}}{j} S(\mathsf{u_1}, j)\, j!}{\mathsf{N}^{\mathsf{u_1}}}\Bigg[\sum_{k_1=\alpha_1}^{j}\binom{j}{k_1}\sum_{k_2=0}^{\beta_2}\binom{\mathsf{N}-j}{k_2-1}\sum_{i=0}^{\mathsf{N}-b_2}(-1)^i\binom{\mathsf{N}-b_2}{i}\Big(\frac{\mathsf{N}-b_2-i}{\mathsf{N}}\Big)^{\mathsf{u_2}}\Bigg].$$
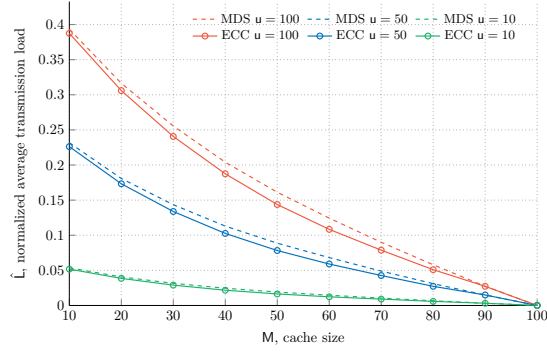
Fig. 5. $\hat{L}$ versus M for $u = 10, 50, 100$ for $N = 100$ when file requests follows a Zipf distribution with $\alpha = 0.8$. 40% of users are connected to R_1, 40% to R_2 and 20% to both Rs. The dot marked and dashed curves indicate the results obtained for the ECC and MDS scheme respectively.

The relevant reduction of load backhaul transmission obtained validates the coded caching scheme in satellite networks and suggest its investigation in more sophisticated scenarios.

VII. Acknoledgment

This work was supported by the Federal Ministry of Education and Research (BMBF, Germany) as part of the 6G Research and Innovation Cluster 6G-RIC under Grant 16KISK020K.

References

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, 2014.

[2] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, 2014.

[3] 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, 2013.

[4] A. Piemontese and A. G. i. Amat, "MDS-coded distributed storage for low delay wireless content delivery," in *Int. Symp. on TC and Iter. Inf. Proc.(ISTC)*, Brest, France, Sep. 2016.

[5] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *Proc. IEEE Globecomm*, San Diego, U.S.A., Dec. 2015.

[6] C. Tian, "On the fundamental limits of coded caching and exact-repair regenerating codes," in *2015 Inter. Symp. on Net. Cod. (NetCod)*, 2015.

[7] K. Vijith, B. K. Rai, and T. Jacob, "Fundamental limits of coded caching: The memory rate pair (k - 1 - 1/k, 1/(k-1))," in *2019 IEEE Inter. Sym. on Inf. Theory (ISIT)*, 2019.

[8] M. Giordani and M. Zorzi, "Non-terrestrial networks in the 6G era: Challenges and opportunities," *IEEE Network*, vol. 35, no. 2, 2021.

[9] S. K. L Johnson, *Urn Models and Their Application*. New York: John Wiley & Sons, 1977, chapter 6.

[10] E. Recayte and A. Munari, "Caching at the edge: Outage probability," in *2021 IEEE Wireless Commun. and Networking Conf. (WCNC)*, 2021.