# Edge Caching: On the Performance of Placement and Delivery Coding Schemes

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

***Abstract*—We study the performance of caching schemes based on two different coding techniques. A heterogeneous network is assumed, in which cache-aided relays are connected through a backhaul link to a master node while no connection exits between users and the master node. The first caching scheme considers to fill the cache with encoded content while the second scheme considers to encode the content during the delivery phase. We compare the schemes by characterizing the average transmission load when users ask for downloading content to the network. We provide an approximation of the expression of the average load which allows a fast evaluation of the network behavior for each scheme considered. We further assume a constraint on the capacity of the backhaul link and derive the expression of the outage probability, i.e. the probability that the system is not able to serve the entire set of users demands. Finally, we examine and discuss how the derived analysis of a two relays scenario can be scaled to study the performance of a network with a generic number of relays.**

## I. Introduction

The colossal growth of devices in the network, the ease of generating multimedia content, and the migration from traditional broadcasting services to streaming services are some of the causes that have led to an uncontrolled increase in traffic. To overcome the challenges of efficiently managing resources, reducing congestion in the network core, reducing latency and energy consumption, edge caching has been proposed as a key technology [1]. It has been proved that bringing the desired content to the edge of the network, i.e., memorizing copies of relevant information close to users, significantly improves the overall network performance [1].

The caching strategy is typically implemented in two-steps, 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*). The potential benefits of edge caching have been investigated in recent works in several ways. A significant body of work on caching has focused on the application of coding techniques to the cached content. The pioneering results were obtained in [2] by Caire et. al, where authors introduce the importance of coded content placement to reduce the download delay in mobile networks. Lately, the performance of the benchmark based caching scheme maximum distance separable (MDS) codes have been derived. MDS caching schemes have been studied in wireless networks to minimize the expected download time [3] or to reduce the amount of data to be sent [4]. Recently, practical caching schemes based on linear random codes in [5] and based on LT codes in [6] have also been proposed.

In contrast, another branch of research has focused on the limits of caching in a slightly different scenario. Specifically, Maddah-Ali et al. introduced in [7] the concept of *coded caching* where local caching directly on the user's device is considered. The main idea of the scheme is to deliver coded content and exploit the user's local cache to serve multiple users with a single transmission. Such technique has spurred an extensive body of research providing a solid understanding of the potential and limitations of caching, e.g. [8], [9]. Coded caching has been further studied for uncoded cache placement in [10]–[12] and for coded cache placement [13]–[16]. Most studies on coded caching focus on finding the maximum achivable rate that the network can support, without deriving essential performance metrics such as the average transmission rate over the backhaul link or the outage probability. Furthermore, coded caching has been studied especially in setups where few cache-aided users are connected to a common and unique server via a shared link. Instead, its potential and the trade-offs it may induce in other relevant scenarios remain unexplored to date. Another 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 [17]. 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 as it is assumed in [18]–[20]. A preliminary study considering the approach of coded caching in a two-tier caching heterogenous network for an arbitrary number of users was presented in [21]. In that 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. Important results were obtained showing the significant reduction in backhaul transmissions when the scheme of Maddah-Ali is in place. Enthusiastic about the obtained results, in the current manuscript we study in depth such coded caching set-up. We use the derivations made for the analysis of the average backhaul load to derive another essential metric for the system design, i.e. the outage probability.

This work was in part supported by the Federal Ministry of Education and Research of Germany in the programme of "Souveran. Digital. Vernetzt." Joint ¨ project 6G-RIC, project identification number: 16KISK022.

While an initial analysis of performance when coding is applied in a two tier network were first developed in our conference paper [21], this work makes new contributions beyond those of [21] in the following aspects

- an approximation to calculate the average transmission load for each coding scheme is derived. The new expression extremely reduces the computational demanding time to evaluate the average transmission load with respect to our solution given in [21] when the number of users in the system is large,
- evaluation of the outage probability for both schemes. The outage probability is an important metric to characterize the system performance, as it provides a tradeoff on how to choose the cache size when the capacity on the backhaul link is fixed,
- extension to a multi-relay scenario. In particular, we discuss and examine under which assumptions the model and analysis obtained for the two-relays scenario is scalable to a multi-relay network
- we provide a larger spectrum of results which are extensively discussed and provide useful hints at the time of design a cache-aided network.

As final remark, the present work employs results of balls and bins (BiB) as a tool to solve caching problems. The cast and use of BiB results in communication system problems is still unexplored and this fact can be proved by the little contribution that can be found in the literature.

*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 $\Pr\{X = x \,|\, Y = y\} = p_X(x|y)$. 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. This setup applies to different network configurations because all scenarios in which multiple users attempt to retrieve content from caches and cannot rely on a direct backhaul connection are feasible setups. For example, in beyond 5G-systems or in 6G networks, the role of the master node might be played by macro base stations (eNB) and cache-aided transmitter by small base stations (small/micro/pico BS). We will take as reference throughout our discussion the satellite topology illustrated in Fig. 1. Here, a satellite (S) has access to library $\mathcal{F} = \{f_1, \cdots, f_{\mathsf{N}}\}$ of equal-size files. On the ground, two cache-enabled relays ($\mathrm{R}_{\mathsf{B}}$ and $\mathrm{R}_{\mathsf{W}}$)[1] are connected via a backhaul link to S, and each one provides connectivity to some users. If it is speficied, we assume that the backhaul link has a limited capacity of $\mathsf{C}$ files, i.e. up to $\mathsf{C}$ files can be transmitted. As typical in current satellite-aided terrestrial networks, we assume that no direct link between users and the satellite 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 $h$ relays, where $h = \{1, 2\}$, whereas $\mathcal{U}_{\mathsf{B}}$ ($\mathcal{U}_{\mathsf{W}}$) denotes the subset of users connected only to relay $\mathrm{R}_{\mathsf{B}}$ ($\mathrm{R}_{\mathsf{W}}$). Note that the set of users connected to exactly one relay is $\mathcal{U}_1 = \mathcal{U}_{\mathsf{B}} \cup \mathcal{U}_{\mathsf{W}}$.

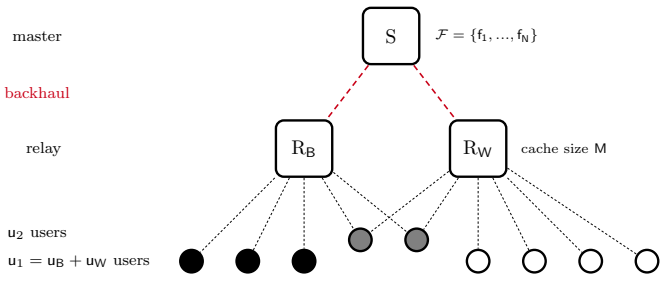


Fig. 1. System model: cache-aided relays are connected to the satellite through the backhaul link and users can be connected to one or more relays.

Each relay can store up to $M \leq \mathsf{N}$ files locally. With $\mathcal{Z}_{\mathsf{B}}$ ($\mathcal{Z}_{\mathsf{W}}$) we indicate the files present in the cache of $\mathrm{R}_{\mathsf{B}}$ ($\mathrm{R}_{\mathsf{W}}$). During the *placement phase*, which is carried out offline, each file $f_j \in \mathcal{F}$ is partitioned into $\mathsf{n_F}$ equally long fragments, i.e. $f_j$ is fragmented as $\{f_j^{(1)}, \cdots, f_j^{(\mathsf{n_F})}\}$. 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 $\max(0, \mathsf{n_F} - hF_j)$ missing fragments are forwarded by the relay after being retrieved by the satellite 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 $x \in \{1, 2, \mathsf{B}, \mathsf{W}\}$. 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 $\mathsf{u}$ the total number of terminals that concurrently request content from the library, each independently choosing a file to download. Specifically, we have that $\mathsf{u} = \mathsf{u}_1 + \mathsf{u}_2$ where $|\mathcal{U}_1| = \mathsf{u}_1$, are those connected to a single relay, while $|\mathcal{U}_2| = \mathsf{u}_2$ are those connected to both relays. We also have that $\mathsf{u}_1 = \mathsf{u}_{\mathsf{B}} + \mathsf{u}_{\mathsf{W}}$ where $\mathsf{u}_{\mathsf{B}}$ are the users requesting content only at $\mathrm{R}_{\mathsf{B}}$ and $\mathsf{u}_{\mathsf{W}}$ only at $\mathrm{R}_{\mathsf{W}}$. 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 [7], referred to as the *edge coded caching scheme* (ECC). To evaluate the performance of each scheme, we derive two metrics. First, we calculate the average backhaul transmission load $\mathsf{L}$, i.e. the average number of packets that the satellite should transmit in the backhaul link to satisfy user requests. We will focus

[1]The subscript $\mathsf{B}$ and $\mathsf{W}$ 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.

on the rv $L^{\mathrm{x}}$ which describes the number of transmissions required from the satellite for a given caching scheme x. We have that $\mathsf{L}^{\mathrm{x}} = \mathbb{E}[L^{\mathrm{x}}]$ where the operator $\mathbb{E}$ indicates the expected value. The metric $\mathsf{L}$ is used to compare the behavior against the benchmark given by MDS scheme. Second, we derive the probability of outage for each scheme defined as the probability that the amount of content that has to be sent over the backhaul link to serve user request is larger than its capacity. Let us discus both schemes in the following.

*MDS Caching Scheme*

In the MDS caching scheme [4], the network caches and delivers packets that are encoded. In particular, $\mathsf{n_F}$ fragments of file $f_j$ are used to create $n > \mathsf{n_F}$ encoded packets using a $(n, \mathsf{n_F})$ 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 $\mathsf{n_F}$ different encoded packets [4].

If we assume a uniform distribution of the file requested then files are split into $\mathsf{n_F} = \mathsf{N}$ fragments. Each relay fills own cache with $F_j = \mathsf{M}$ 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 $n - 2\mathsf{M}$ 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 relay 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. Such event 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* $R_{\mathsf{B}}$ *and user* 2 *only to* $R_{\mathsf{W}}$*. Consider a memory size of* $\mathsf{M} = 1$ *and two equiprobable files split into* $\mathsf{n_F} = 2$ *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*

$$\mathcal{Z}_{\mathsf{B}} = \{e_1^{(1)}, e_2^{(1)}\} \text{ and } \mathcal{Z}_{\mathsf{W}} = \{e_1^{(2)}, e_2^{(2)}\}.$$

*To characterize the average backhaul load* $\mathsf{L}^{MDS}$*, we shall consider two cases. First, we suppose that users are requesting for different content, i.e. user 1 requests for* $f_1$ *to relay* $R_{\mathsf{B}}$ *and user 2 for* $f_2$ *to* $R_{\mathsf{W}}$*. Since each relay has one encoded packet of the requested file in cache, S should send one encoded packet to each relay. 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,* $\mathsf{L}$ *in MDS evaluates to*

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

*where we sum over the* $i$ *possibilities on how users can request for the library content. They ask for different files with* $p(l_1) = 1/2$ *while they ask for the same file with* $p(l_2) = 1/2$*.*

*Edge Coded Caching Scheme*

In the ECC caching 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 relay 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. The satellite is aware of which content has been stored in each relays. In the second phase, users make their requests to the corresponding relay. The delivery takes place in three stages. In the first stage, a user receives fragments directly from the cached content of the associated relay. In the second stage, the satellite is informed of users requests and provides missing content over the shared backhaul 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, the satellite 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). The satellite picks a fragment of a file requested at $\mathrm{R_B}$ and present in cache of $\mathrm{R_W}$ 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 relay by XORing the received packet with the corresponding fragment in cache obtains a fragment of the requested file. ECC scheme generates a gain over the MDS caching scheme whenever relays have disjoint requests. Let us clarify the last statement by considering the same setting discussed in Example 1.

**Example 2.** *Let us assume to have one user per relay and the following cache placement:*

$$\mathcal{Z}_{\mathsf{B}} = \{f_1^{(1)}, f_2^{(1)}\} \text{ and } \mathcal{Z}_{\mathsf{W}} = \{f_1^{(2)}, f_2^{(2)}\}.$$

*Let us first consider the case where users request different content. For example, user 1 requests for* $f_1$ *to* $R_{\mathsf{B}}$ *and user 2 for* $f_2$ *to* $R_{\mathsf{W}}$*. During the first stage, user* $i$ *receives* $f_i^{(i)}$ *from the cache of the related relay. At the second stage, S sends the*

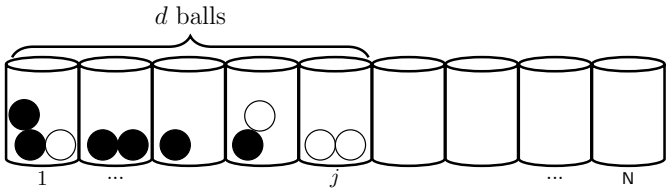


Fig. 2. Caching requests represented as the BiB problem. Bins represent files while balls represent users' requests. The occupancy problem lies on calculating the probability of having exactly $j$ bins not empty after throwing $d$ balls, i.e. the probability that $d$ users request exactly for $j$ different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

*following coded packet $p = f_2^{(1)} \oplus f_1^{(2)}$. Thus the number of packets transmitted over the backhaul, $l_1$, is*

$$l_1 = 1.$$

*$R_\mathsf{B}$ ($R_\mathsf{W}$) 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, $\mathsf{L}$ is then*

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

*Note that the study of a system in which both MDS and ECC schemes are implemented presents the same results as the ECC scheme. This can be shown by further considering to place encoded files in the cache in the current example.*

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 $\mathsf{L}$ 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 balls into bins (BiB) problem, see e.g. [22], consists in independently throwing $d$ balls into $\mathsf{N}$ 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 $\mathsf{N}$ non empty given that $d$ balls are thrown uniformly at random, which was derived in [23] and can be written as

$$\mathsf{P}_\mathsf{o}(j,d) = \binom{\mathsf{N}}{j} \Delta^d\, \mathbf{0}^j\, \mathsf{N}^{-d} \tag{1}$$

where $\Delta^m\, \mathbf{0}^n$ is known as *difference of zeros* [22], i.e.

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

In our setup, we further need to differentiate requests made to $\mathsf{R_B}$ from those made to $\mathsf{R_W}$ 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 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 requested at both relays. Having the $i$-th bin with only black (white) balls implies that the $i$-th file was requested only at relay $\mathsf{R_B}$ ($\mathsf{R_W}$).

Following this approach, a useful result is offered by the *multivariate occupancy problem* assuming that there are $\mathsf{N}$ bins and that $\mathsf{u_B}$ black balls have been thrown and have occupied $j$ different bins. The probability that, after throwing $\mathsf{u_W}$ white balls, there are exactly $k_\mathsf{B}$ bins containing only black balls and $k_\mathsf{W}$ bins containing only white balls is [22]

$$\mathsf{P_m}(k_\mathsf{B}, k_\mathsf{W}, \mathsf{u_W}|j) = \binom{j}{k_\mathsf{B}} \binom{\mathsf{N}-j}{k_\mathsf{W}} \Delta^{b_\mathsf{W}}\, \mathbf{0}^{\mathsf{u_W}}\, \mathsf{N}^{-\mathsf{u_W}} \tag{2}$$

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

Note that, in our setting, $\mathsf{P_m}(k_\mathsf{B}, k_\mathsf{W}, \mathsf{u_W}|j)$ provides the probability that exactly $k_\mathsf{B}$ files are requested only to relay $\mathsf{R_B}$ and $k_\mathsf{W}$ files are requested only to relay $\mathsf{R_W}$ when in total there are $\mathsf{u_1} = \mathsf{u_B} + \mathsf{u_W}$ users connected to exactly one relay.

Note also that, there is a relationship between the occupancy problem and the multivariate occupancy problem, in particular it is easily to verify that

$$\mathsf{P_m}(k_\mathsf{B}, k_\mathsf{W}, \mathsf{u_W}|j) = \binom{j}{k_\mathsf{B}} \binom{\mathsf{N}-j}{k_\mathsf{W}} \binom{\mathsf{N}}{j}^{-1} \mathsf{P_o}(b_\mathsf{W}, \mathsf{u_W}). \tag{3}$$

### A. Normal distribution approximation

Equations (1) and (2) require for each user $i$ evaluating the term $(-1)^i \binom{m}{i} (m-i)^m$ becoming computationally expensive when a large number of users (balls) is present in the system. To overcome this calculation effort, we provide an approximation function which allows us to fast estimate our original expression in one shot. The derivation cames from the balls and bins problem and here is appropriately adapted to our problem. In fact, in [22] it has been proved that the distribution of the number of non empty bins $j$ after the launches of $d$ balls can be well approximated to a normal distribution as follows

$$\mathsf{P_o}(j,d) \approx \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\Big\{ -\frac{(j-\mu_d)^2}{2\sigma_d^2} \Big\} = \mathsf{P_o^{\approx}}(j,d) \tag{4}$$

where mean and variance are defined as

$$\mu_d = \mathsf{N}\big(1 - e^{-\frac{d}{\mathsf{N}}}\big)$$
$$\sigma_d^2 = \mathsf{N} e^{-\frac{d}{\mathsf{N}}} (1 - e^{-\frac{d}{\mathsf{N}}}) - d e^{-\frac{2d}{\mathsf{N}}}.$$

**Example 3.** *Suppose to have a library of $\mathsf{N} = 100$ files in the system and $d = 50$ users. Assume that we want to calculate the probability that exactly $j = 39$ different files are requested, the result can be obtained by appyling equation* (1) *as follows*

$$\mathsf{P_o}(39, 50) = \binom{100}{39} \sum_{i=0}^{50} (-1)^i \binom{50}{i} (50-i)^{39} 100^{-50} = 0.165$$

TABLE I
LIST OF SOME OF THE RANDOM VARIABLES

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

*while the approximation is evaluated as*

$$\mathsf{P}_\mathsf{o}^{\approx}(39,50)=\frac{1}{\sqrt{2\pi\sigma_d^2}}\exp\big\{-\frac{(39-\mu_d)^2}{2\sigma_d^2}\big\}=0.168$$

*with mean value* $\mu_{50}=100(1-e^{-\frac{50}{100}})$ *and variance* $\sigma^2_{50}=100\,e^{-\frac{50}{100}}(1-e^{-\frac{50}{100}})-50e^{-2\frac{50}{100}}$. *As can be seen in this example, the approximation is not only close to the real value but also extremely reduces its computation. While equation* (1) *requires to calculate the sum of* $d+1$ *terms,* $\mathsf{P}_\mathsf{o}^{\approx}$ *only needs to evaluate a point given the mean and the variance.*

Similarly, in our problem by observing the relationship given in (3) , the multivariate occupancy problem can be approximated as follows

$$\begin{aligned}\mathsf{P}_\mathsf{m}(k_\mathsf{B},k_\mathsf{W},\mathsf{u}_\mathsf{W}|j)&\approx\frac{\binom{j}{k_\mathsf{B}}\binom{\mathsf{N}-j}{k_\mathsf{W}}}{\binom{\mathsf{N}}{j}}\frac{1}{\sqrt{2\pi\sigma^2_{\mathsf{u}_\mathsf{W}}}}\exp\Big\{-\frac{(b_\mathsf{W}-\mu_{\mathsf{u}_\mathsf{W}})^2}{2\sigma^2_{\mathsf{u}_\mathsf{W}}}\Big\}\\&=\mathsf{P}_\mathsf{m}^{\approx}(k_\mathsf{B},k_\mathsf{W},\mathsf{u}_\mathsf{W}|j)\end{aligned}\tag{5}$$

where mean and variance are defined as

$$\begin{aligned}\mu_{\mathsf{u}_\mathsf{W}}&=\mathsf{N}\big(1-e^{-\frac{\mathsf{u}_\mathsf{W}}{\mathsf{N}}}\big)\\\sigma^2_{\mathsf{u}_\mathsf{W}}&=\mathsf{N}e^{-\frac{\mathsf{u}_\mathsf{W}}{\mathsf{N}}}(1-e^{-\frac{\mathsf{u}_\mathsf{W}}{\mathsf{N}}})-\mathsf{u}_\mathsf{W}e^{-\frac{2\mathsf{u}_\mathsf{W}}{\mathsf{N}}}.\end{aligned}$$

## 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 $\mathsf{u}_1$ users connected to only one relay ($\mathcal{U}_1$) while $K_\mathsf{B}$ denotes the number of different files requested exclusively at $\mathsf{R}_\mathsf{B}$. 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 $\mathsf{R}_\mathsf{B}$ but not requested at $\mathsf{R}_\mathsf{W}$. Let us clarify all the mentioned quantities with an example.

**Example 4.** *Let us refer to Fig. 3 which illustrates 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* $R_\mathsf{B}$ *(black balls),* $\mathsf{u}_\mathsf{W}=4$ *only to* $R_\mathsf{W}$ *(white balls), while* $\mathsf{u}_2=7$ *are connected to both relays (grey).*

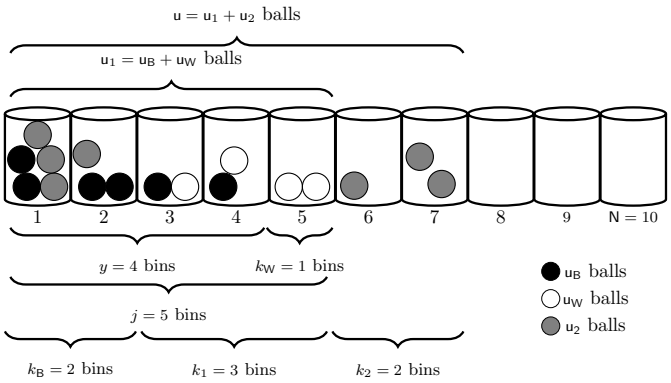


Fig. 3. Requests represented as the BiB problem. Black balls represents requests from users connected only to $\mathsf{R}_\mathsf{B}$, while balls request only to $\mathsf{R}_\mathsf{W}$ while gray balls represents request from users connected to both relays.

*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* $\mathsf{u}_1$ *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* $R_\mathsf{B}$ *are* $k_\mathsf{B}=2$*, i.e. the number of bins with black balls and without white balls, while those exclusively requested at* $R_\mathsf{W}$ *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. Approximation of the MDS Average Transmission Load

Let us recall that $J$ different files are requested by $\mathsf{u}_1$ 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 $\mathsf{M}$ are already provided to the user via the relay's cache.

The overall number of packets that S transmits to satisfy $\mathsf{u}_1$ requests is then expressed by the rv

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

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 $L_1^{\text{MDS}}$. 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 $(\mathsf{n_F}-2\mathsf{M})^+$ 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

$$L_2^{\text{MDS}} = (\mathsf{n_F}-2\mathsf{M})^+ K_2. \tag{4}$$

So that the average backhaul load $\mathsf{L}$ in the MDS is

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

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

$$\begin{aligned}\mathsf{L}_1^{\text{MDS}} &= \mathbb{E}\big[(\mathsf{n_F}-\mathsf{M})\,J\big] \\ &= (\mathsf{n_F}-\mathsf{M})\sum_{j=1}^{\beta_J} j\mathsf{P_o}(j,\mathsf{u}_1)\end{aligned}$$

where the quantity $\mathsf{P_o}(j,\mathsf{u}_1)$ was derived with BiB occupancy problem, see (1). Given (4), the approximation $\mathsf{L}_{\approx 1}^{\text{MDS}}$ of $\mathsf{L}_1^{\text{MDS}}$ can be written as

$$\begin{aligned}\mathsf{L}_{\approx 1}^{\text{MDS}} &= (\mathsf{N}-\mathsf{M})\sum_{j=1}^{\beta_J} j\mathsf{P}_{\mathsf{o}}^{\approx}(j,d) \\ &= (\mathsf{N}-\mathsf{M})\sum_{j=1}^{\beta_J} j\frac{1}{\sqrt{2\pi\sigma_d^2}}\exp\Big\{-\frac{(j-\mu_d)^2}{2\sigma_d^2}\Big\}\end{aligned} \tag{6}$$

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

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

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 $\mathsf{u}_2$ (grey) balls, conditioned on having already $j$ non empty bins occupied by $\mathsf{u}_1$ balls. As discussed, this results is offered by the multivariate occupancy problem, and we have

$$p_{K_2}(k_2|j) = \sum_{k_1=\alpha_1}^{j} \mathsf{P_m}(k_1,k_2,\mathsf{u}_2|j), \tag{8}$$

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 $\mathsf{R_B}$ represented by bins with only black balls). Accordingly,

$$\begin{aligned}\mathsf{L}_2^{\text{MDS}} &= \mathbb{E}_J\Big[(\mathsf{n_F}-2\mathsf{M})^+\sum_{k_2=0}^{\beta_2} k_2\, p_{K_2}(k_2|J)\Big] \\ &= (\mathsf{N}-2\mathsf{M})^+\sum_{j=0}^{\beta_J}\mathsf{P_o}(j,\mathsf{u}_1)\sum_{k_2=0}^{\beta_2}k_2\sum_{k_1=\alpha_1}^{j}\mathsf{P_m}(k_1,k_2,\mathsf{u}_2|j).\end{aligned} \tag{9}$$

The approximation $\mathsf{L}_{\approx 2}^{\text{MDS}}$ of (9) can be written, thanks to the derivations on (4) and (5), as

$$\begin{aligned}\mathsf{L}_{\approx 2}^{\text{MDS}} &= (\mathsf{N}-2\mathsf{M})^+\sum_{j=0}^{\beta_J}\mathsf{P}_{\mathsf{o}}^{\approx}(\mathsf{u}_1,j)\sum_{k_2=0}^{\beta_2}k_2\sum_{k_1=\alpha_1}^{j}\mathsf{P}_{\mathsf{m}}^{\approx}(k_1,k_2,\mathsf{u}_2|j) \\ &= \frac{(\mathsf{N}-2\mathsf{M})^+}{\sqrt{2\pi\sigma_{\mathsf{u}_1}^2}}\sum_{j=0}^{\beta_J}\exp\Big\{-\frac{(j-\mu_{\mathsf{u}_1})^2}{2\sigma_{\mathsf{u}_1}^2}\Big\}\sum_{k_2=0}^{\beta_2}k_2 \\ &\qquad \sum_{k_1=\alpha_1}^{j}\frac{\binom{j}{k_1}\binom{\mathsf{N}-j}{k_2}}{\binom{\mathsf{N}}{j}}\frac{1}{\sqrt{2\pi\sigma_{\mathsf{u}_2}^2}}\exp\Big\{-\frac{(b_{\mathsf{W}}-\mu_{\mathsf{u}_2})^2}{2\sigma_{\mathsf{u}_2}^2}\Big\}.\end{aligned} \tag{10}$$

Finally, by summing (6) and (10) we obtain the approximation of the load of the MDS scheme and normalizing by the number of files $\mathsf{N}$, we have that the approximation of the normalized average transmission load is

$$\bar{\mathsf{L}}_{\approx}^{\text{MDS}} = \frac{\mathsf{L}_{\approx 1}^{\text{MDS}} + \mathsf{L}_{\approx 2}^{\text{MDS}}}{\mathsf{N}}$$

and the complete expression is given in (11) at the top of the next page.

### *B. ECC Average Transmission Load*

Let us start by considering users in $\mathcal{U}_1$. Since $\mathsf{M}$ 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 $\mathsf{u_B}$ users connected only to $\mathsf{R_B}$ and by $K_{\mathsf{W}}$ the rv counting the number of files exclusively requested by the $\mathsf{u_W}$ connected only to $\mathsf{R_W}$ and not requested to $\mathsf{R_B}$, in total users have to receive $(\mathsf{n_F}-\mathsf{M})(Y+K_{\mathsf{W}})$ 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 $\omega_1$ XORed packets involving the two files. In particular,

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

where $\omega_1$ is the number of fragments per file combined in a coding opportunity and it depends on the cache size. For each transmission opportunity, when $\mathsf{M} \leq \mathsf{N}/2$; then in total $\mathsf{M}$ 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 $\omega_1$ 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

$$L_1^{\text{ECC}} = (\mathsf{n_F}-\mathsf{M})\,(Y+K_{\mathsf{W}}) - \omega_1\, Z$$

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

$$\bar{\mathsf{L}}_{\approx}^{\text{MDS}} = \frac{\mathsf{N}-\mathsf{M}}{\sqrt{2\pi\sigma_d^2}} \sum_{j=1}^{\beta_J} \exp\Big\{-\frac{(j-\mu_d)^2}{2\sigma_d^2}\Big\}\Big[j\Big(1-\frac{\mathsf{M}}{\mathsf{N}}\Big)+\Big(1-\frac{2\mathsf{M}}{\mathsf{N}}\Big)^{+}\sum_{k_1=\alpha_1}^{j}\frac{\binom{j}{k_1}}{\binom{\mathsf{N}}{j}}\sum_{k_2=0}^{\beta_2}\frac{\binom{\mathsf{N}-j}{k_2-1}}{\sqrt{2\pi\sigma_{\mathsf{u}_2}^2}}\exp\Big\{-\frac{(b_{\mathsf{W}}-\mu_{\mathsf{u}_2})^2}{2\sigma_{\mathsf{u}_2}^2}\Big\}\Big]. \quad (11)$$

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

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

The average bachkaul transmission load of the ECC is

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

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

$$\begin{aligned}\mathsf{L}_1^{\text{ECC}} &= \mathbb{E}_Y\Big[\mathbb{E}\big[(\mathsf{n_F}-\mathsf{M})\,(Y+K_{\mathsf{W}}) - \omega_1\, Z|Y\big]\Big] \\ &= \mathbb{E}_Y\Big[\mathbb{E}\big[(\mathsf{n_F}-\mathsf{M})\,(Y+K_{\mathsf{W}})|Y\big]\Big] - \mathbb{E}_Y\Big[\mathbb{E}\big[\omega_1\, Z|Y\big]\Big].\end{aligned} \quad (12)$$

Let us first focus on the conditional distribution of $K_{\mathsf{W}}$. Gi $Y = y$ different files requested from users in $\mathcal{U}_{\mathsf{B}}$, the probability $p_{K_{\mathsf{W}}}(k_{\mathsf{W}}|y)$ of having exactly $k_{\mathsf{W}}$ files requested only at $\mathsf{R}_{\mathsf{W}}$ can be derived from the BiB multivariate occupancy problem by considering all values that $k_{\mathsf{B}}$ can assume as

$$p_{K_{\mathsf{W}}}(k_{\mathsf{W}}|y) = \sum_{k_{\mathsf{B}}=\alpha_{\mathsf{B}}}^{y} \mathsf{P}_{\mathsf{m}}(k_{\mathsf{B}}, k_{\mathsf{W}}, \mathsf{u}_{\mathsf{W}}|y) \quad (13)$$

where $b_{\mathsf{W}} = y - k_{\mathsf{B}} + k_{\mathsf{W}}$.

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

$$p_Z(z|y) = \sum_{k_{\mathsf{W}}=z}^{\beta_{\mathsf{W}}-y+z} \mathsf{P}_{\mathsf{m}}(z, k_{\mathsf{W}}, \mathsf{u}_{\mathsf{W}}|y) + \sum_{k_{\mathsf{B}}=z+1}^{\min(y,\beta_{\mathsf{B}})} \mathsf{P}_{\mathsf{m}}(k_{\mathsf{B}}, z, \mathsf{u}_{\mathsf{W}}|y) \quad (14)$$

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

$$\mathsf{L}_1^{\text{ECC}} = \sum_{y=1}^{\beta_{\mathsf{B}}} \mathsf{P}_{\mathsf{o}}(y, \mathsf{u}_{\mathsf{B}})\Big[(\mathsf{N}-\mathsf{M})\Big(y + \sum_{k_{\mathsf{W}}=0}^{\beta_{\mathsf{W}}} k_{\mathsf{W}}\, p_{K_{\mathsf{W}}}(k_{\mathsf{W}}|y)\Big) - \omega_1 \sum_{z=0}^{y} z\, p_Z(z|y)\Big]. \quad (15)$$

The approximation $\mathsf{L}_{\approx 1}^{\text{ECC}}$ of (15) is

$$\sum_{y=1}^{\beta_{\mathsf{B}}} \mathsf{P}_{\mathsf{o}}^{\approx}(y, \mathsf{u}_{\mathsf{B}})\Big[\Big(y + \sum_{k_{\mathsf{W}}=0}^{\beta_{\mathsf{W}}} k_{\mathsf{W}} \sum_{k_{\mathsf{B}}=\alpha_{\mathsf{B}}}^{y} \mathsf{P}_{\mathsf{m}}^{\approx}(k_{\mathsf{B}}, k_{\mathsf{W}}, \mathsf{u}_{\mathsf{W}}|y)(\mathsf{N}-\mathsf{M})\Big) - \omega_1 \sum_{z=0}^{y} z \sum_{k_{\mathsf{W}}=z}^{\beta_{\mathsf{W}}-y+z} \mathsf{P}_{\mathsf{m}}^{\approx}(z, k_{\mathsf{W}}, \mathsf{u}_{\mathsf{W}}|y) + \sum_{k_{\mathsf{B}}=z+1}^{\min(y,\beta_{\mathsf{B}})} \mathsf{P}_{\mathsf{m}}^{\approx}(k_{\mathsf{B}}, z, \mathsf{u}_{\mathsf{W}}|y)\Big]. \quad (16)$$

Since $\mathsf{L}_{\approx 2}^{\text{ECC}} = \mathsf{L}_{\mathsf{A}2}^{\text{MDS}}$, by adding $\mathsf{L}_{\approx 2}^{\text{ECC}}$ to (16) the final result in (17) is obtained which represents the approximation of the average transmission load of the ECC scheme normalized to the number of files

$$\bar{\mathsf{L}}_{\approx}^{\text{ECC}} = \frac{\mathsf{L}_{\approx 1}^{\text{ECC}} + \mathsf{L}_{\approx 2}^{\text{ECC}}}{\mathsf{N}}.$$

## V. Outage Probability

The outage probability is a fundamental metric at the time of designing a cache-aided system. The system is said to be in *outage* if the network cannot serve the requests of $\mathsf{u}$ users, such event occurs whenever the amount of content that has to be sent through the backhaul link exceeds its capacity $\mathsf{C}$. The expression of the outage probability depends on the caching scheme considered. In the following, we derive the probability of outage for the MDS caching scheme and for the ECC caching scheme. For an easier derivation, in the following calculations it is assumed that the total number of users in the system coincides with the number of users connected only to one relay, i.e. $\mathsf{u} = \mathsf{u}_1$.

### A. MDS Outage Probability

In the MDS caching scheme the system needs to send in total $Y + K_{\mathsf{W}}$ different files given $\mathsf{u}_1$ requests. Each cache stores the portion $\frac{\mathsf{M}}{\mathsf{N}}$ per each file therefore the satellite should send the remaining portion $1 - \frac{\mathsf{M}}{\mathsf{N}}$ through the backhaul link for each content requested. Thus, the outage probability can be written as

$$\begin{aligned}\mathsf{P}_{\text{out}}^{\text{MDS}} &= \Pr\Big\{(Y+K_{\mathsf{W}})\big(1-\frac{\mathsf{M}}{\mathsf{N}}\big) > \mathsf{C}\Big\} \\ &= 1 - \Pr\Big\{(Y+K_{\mathsf{W}}) \le \frac{\mathsf{C}}{1-\frac{\mathsf{M}}{\mathsf{N}}}\Big\}.\end{aligned}$$

If we condition to the quantity $J = Y + K_{\mathsf{W}}$ which represents the rv of total number of files requested by $\mathsf{u}_1$ users then we can write

$$\begin{aligned}\mathsf{P}_{\text{out}}^{\text{MDS}} &= 1 - \Pr\Big\{j \le \frac{C}{1-\frac{\mathsf{M}}{\mathsf{N}}}\Big| J = j\Big\}\Pr\big\{J = j\big\} \\ &= 1 - \sum_{j=1}^{\min(\eta, \beta_{\mathsf{B}}+\beta_{\mathsf{W}})} \mathsf{P}_{\mathsf{o}}(j, \mathsf{u}_1)\end{aligned}$$

where $\eta = \frac{\mathsf{C}}{1-\mathsf{M}/\mathsf{N}}$.

### B. ECC Outage Probability

In the ECC caching scheme, the coding opportunities during the delivery phase should be considered for the outage calculations. The random variable $Z$ accounts for the number of combined transmissions and the quantity $Z\frac{\omega_1}{\mathsf{N}}$ is the total

gain obtained by the scheme. The outage probability in the ECC scheme can be written as

$$\begin{aligned}\mathsf{P}_{\mathsf{out}}^{\mathsf{ECC}} &= \Pr\Big\{J\big(1-\frac{\mathsf{M}}{\mathsf{N}}\big)-Z\frac{\omega_1}{\mathsf{N}} > \mathsf{C}\Big\}\\ &= 1-\Pr\Big\{J\big(1-\frac{\mathsf{M}}{\mathsf{N}}\big)-Z\frac{\omega_1}{\mathsf{N}} \le \mathsf{C}\Big\}.\end{aligned}$$

If we condition to $J$ we obtain

$$\mathsf{P}_{\mathsf{out}}^{\mathsf{ECC}}=1-\sum_{j=1}^{\min(\mathsf{u},\mathsf{N})}\Pr\Big\{j\big(1-\frac{\mathsf{M}}{\mathsf{N}}\big)-Z\frac{\omega_1}{\mathsf{N}}\le\mathsf{C}\Big|J=j\Big\}\Pr\big\{J=j\big\}$$

$$=1-\Bigg[\sum_{j=\eta+1}^{\min(\mathsf{u},\mathsf{N})}\Pr\Big\{j\big(1-\frac{\mathsf{M}}{\mathsf{N}}\big)-Z\frac{\omega_1}{\mathsf{N}}\le\mathsf{C}\Big|J=j\Big\}\Pr\big\{J=j\big\}+\sum_{j=1}^{\eta}\Pr\Big\{J=j\Big\}\Bigg]$$

where the sum was divided into two terms given that when the number of total request $J$ is less than $\eta=\mathsf{C}/(1-\mathsf{M}/\mathsf{N})$ then the network always suceeds regardless of whether there is a gain or not. Now let us define $\gamma=\frac{j\left(1-\frac{\mathsf{M}}{\mathsf{N}}\right)-\mathsf{C}}{\omega_1/\mathsf{N}}$ then

$$\mathsf{P}_{\mathsf{out}}^{\mathsf{ECC}}=1-\Bigg[\sum_{j=1}^{\eta}\mathsf{P}_{\mathsf{o}}(j,\mathsf{u})+\sum_{j=\eta+1}^{\min(\mathsf{u},\mathsf{N})}\Pr\Big\{Z\ge\gamma\Big|J=j\Big\}\mathsf{P}_{\mathsf{o}}(j,\mathsf{u})\Bigg]$$

$$=1-\Bigg[\sum_{j=1}^{\eta}\mathsf{P}_{\mathsf{o}}(j,\mathsf{u})+\sum_{j=\eta+1}^{\min(\mathsf{u},\mathsf{N})}\sum_{z=\gamma}^{\min(k_W,y)}\Pr\Big\{Z=z\Big|J=j\Big\}\mathsf{P}_{\mathsf{o}}(j,\mathsf{u})\Bigg]$$

where $\Pr\Big\{Z=z\Big|J=j\Big\}$ indicates the probability of having $Z$ delivery opportunities given that in total $J$ files have been requested. Note that $J=Y+K_{\mathsf{W}}$, that is the sum between number of total files requested from users connected to relay $R_{\mathsf{B}}$ and the number of exclusive files requested by users connected to $R_{\mathsf{W}}$. For deriving the probability mass function of the rv $Z$ all possible combinations of $Y$ and $K_{\mathsf{W}}$ should be considered and due to its complexity, the term $\Pr\Big\{Z=z\Big|J=j\Big\}$ is derived by Monte-Carlo.

## VI. Multi-Relay Extended Scenario

In this Section, we present and briefly discuss how the analysis for the model illustrated in Section II can be extended to a scenario in which the network consists of a generic number of relays. To this end, we shall introduce further assumptions for the new system model. Firstly, we assume that the scaled network can be represented as a chain of an arbitrary number $\mathsf{T}$ of relays $(\mathsf{R}^1,...,\mathsf{R}^{\mathsf{T}})$ all with cache size $\mathsf{M}$, as illustrated in Fig. (4). Each relay should be classified either as black $\mathsf{R}_{\mathsf{B}}$ or as white $\mathsf{R}_{\mathsf{W}}$ in such way that two neighbour relays cannot have the same colour. For each scheme, it is assumed that during the placement phase all black relays fill their cache with the same content and that all white relays do so as well. The number of users connected to the i-th relay can be written as $\mathsf{u}_{\mathsf{B}}^{\mathsf{i}}$ if the relay is $\mathsf{R}_{\mathsf{B}}^{\mathsf{i}}$ while $\mathsf{u}_{\mathsf{W}}^{\mathsf{i}}$ if it is $\mathsf{R}_{\mathsf{W}}^{\mathsf{i}}$. The number of users connected to both relays $\mathsf{R}^{\mathsf{i}}$ and $\mathsf{R}^{\mathsf{i}+1}$ is denoted by $\mathsf{u}_2^{\mathsf{i}}$ for $\mathsf{i}=\{1,...,\mathsf{T}-1\}$. In this multi-relay scenario, $\mathsf{u}_{\mathsf{B}}$ indicates total number of users connected only to black relays and $\mathsf{u}_{\mathsf{W}}$ only to white relays such that

$$\mathsf{u}_{\mathsf{B}}=\sum_{\mathsf{i}=1}^{\mathsf{T}}\mathsf{u}_{\mathsf{B}}^{\mathsf{i}}\quad\text{and}\quad\mathsf{u}_{\mathsf{W}}=\sum_{\mathsf{i}=1}^{\mathsf{T}}\mathsf{u}_{\mathsf{W}}^{\mathsf{i}}. \tag{18}$$

Note that $\mathsf{u}_{\mathsf{B}}^{\mathsf{i}}=0$ $(\mathsf{u}_{\mathsf{W}}^{\mathsf{i}}=0)$ if the i-th relay is white (black). Instead, the parameter $\mathsf{u}_1$ indicates the total number of users in the network connected to only one relay while $\mathsf{u}_2$ the total number of users connected to both relays and we have that

$$\mathsf{u}_1=\mathsf{u}_{\mathsf{B}}+\mathsf{u}_{\mathsf{W}}=\sum_{\mathsf{i}=1}^{\mathsf{T}}\mathsf{u}_{\mathsf{B}}^{\mathsf{i}}+\sum_{\mathsf{i}=1}^{\mathsf{T}}\mathsf{u}_{\mathsf{W}}^{\mathsf{i}}\quad\text{and}\quad\mathsf{u}_2=\sum_{\mathsf{i}=1}^{\mathsf{T}-1}\mathsf{u}_2^{\mathsf{i}}. \tag{19}$$

As for the two-relay scenario, in the extended scenario, during the delivery phase when users request files, they are first served by the corresponding content present in cache which they are connected, then the master node aggregates all requests and transmits the missing content to the relays according to the MDS or the ECC approach. The master does not need to know from which relay a certain content was requested, but only needs to know for the ECC case, as before, whether it was requested from a black or a white relay. Each relay will then forward the content received from the master to its users. The performance of the multi-relay scenario above presented with $\mathsf{T}$ relays and where the value of the parameters $\mathsf{u}_{\mathsf{W}},\mathsf{u}_{\mathsf{B}},\mathsf{u}_1$ and $\mathsf{u}_2$ takes into account all the users present, calculated as in (18) and (19) is equivalent to the performance of a network with only two relays, as presented in Section II. In fact, it is easy to show that for each scheme, the number of transmissions through the backhaul link does not depend on network topology but on the number of users to connected each coloured relay. At this point, the average transmission rate, the outage probability and its approximations can be calculated as presented in Section IV and in Section V, respectively.

In the MDS caching scheme, the master node transmits $\mathsf{n}_{\mathsf{F}}-\mathsf{M}$ encoded packets over the backhaul link for each content that have been requested only by users connected to one relay, while it transmits $\mathsf{n}_{\mathsf{F}}-2\mathsf{M}$ encoded packets for

$$\begin{aligned}\bar{\mathsf{L}}_{\approx}^{\mathsf{ECC}} = &\sum_{y=1}^{\beta_{\mathsf{B}}}\frac{1}{\sqrt{2\pi\sigma_{\mathsf{u}_{\mathsf{B}}}^2}}\exp\Big\{-\frac{(y-\mu_{\mathsf{u}_{\mathsf{B}}})^2}{2\sigma_{\mathsf{u}_{\mathsf{B}}}^2}\Big\}\Big[(\mathsf{N}-\mathsf{M})\Big(y+\frac{1}{\sqrt{2\pi\sigma_{\mathsf{u}_{\mathsf{W}}}^2}}\sum_{k_{\mathsf{W}}=0}^{\beta_{\mathsf{W}}}\sum_{k_{\mathsf{B}}=\alpha_{\mathsf{B}}}^{y}\frac{\binom{y}{k_{\mathsf{B}}}\binom{\mathsf{N}-y}{k_{\mathsf{W}}-1}}{\binom{\mathsf{N}}{y}}\exp\Big\{-\frac{(b_{\mathsf{W}}-\mu_{\mathsf{u}_{\mathsf{W}}})^2}{2\sigma_{\mathsf{u}_{\mathsf{W}}}^2}\Big\}\Big)\\ &-\omega_1\binom{\mathsf{N}}{y}^{-1}\sum_{z=0}^{y}\sum_{k_{\mathsf{W}}=z}^{\beta_{\mathsf{W}}-y+z}\binom{y}{z-1}\binom{\mathsf{N}-y}{k_{\mathsf{W}}}\exp\Big\{-\frac{(b_{\mathsf{B}}-\mu_{\mathsf{u}_{\mathsf{W}}})^2}{2\sigma_{\mathsf{u}_{\mathsf{W}}}^2}\Big\}+\sum_{k_{\mathsf{B}}=z+1}^{\min(y,\beta_{\mathsf{B}})}\binom{y}{k_{\mathsf{B}}}\binom{\mathsf{N}-y}{z-1}\exp\Big\{-\frac{(b_{\mathsf{W}}-\mu_{\mathsf{u}_{\mathsf{W}}})^2}{2\sigma_{\mathsf{u}_{\mathsf{W}}}^2}\Big\}\Big] \quad (17)\\ &+(\mathsf{N}-2\mathsf{M})^+\frac{1}{\sqrt{2\pi\sigma_{\mathsf{u}_1}^2}}\sum_{j=0}^{\beta_{\mathsf{J}}}\exp\Big\{-\frac{(j-\mu_{\mathsf{u}_1})^2}{2\sigma_{\mathsf{u}_1}^2}\Big\}\sum_{k_2=0}^{\beta_2}\sum_{k_1=\alpha_1}^{j}\binom{j}{k_1}\binom{\mathsf{N}-j}{k_2-1}\binom{\mathsf{N}}{j}^{-1}\frac{1}{\sqrt{2\pi\sigma_{\mathsf{u}_2}^2}}\exp\Big\{-\frac{(b_{\mathsf{W}}-\mu_{\mathsf{u}_2})^2}{2\sigma_{\mathsf{u}_2}^2}\Big\}.\end{aligned}$$

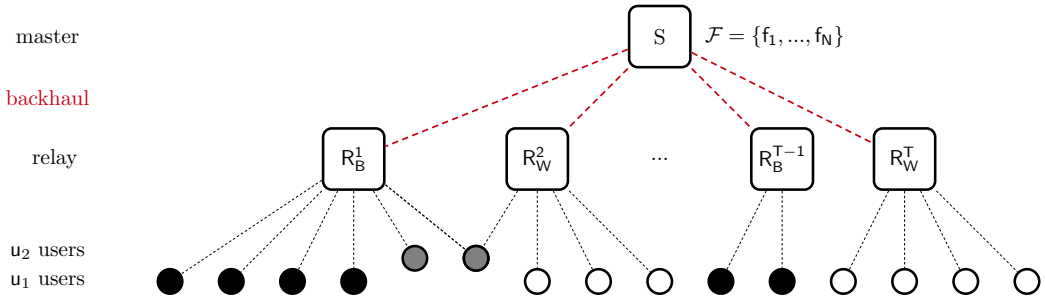


Fig. 4. Extended system model for a generic number of $T$ relays in which $u_1$ users are connected to only one relay and $u_2$ users connected to two relays. In the illustrated scenario, if we consider $T = 4$ then we have that $u_B^1 = 4, u_W^1 = 3, u_B^2 = 2, u_W^2 = 4$ and in total the number of users connected to a black relay $u_B = 6$ and to a white relay is $u_W = 7$ such that $u_1 = 13$ while $u_2 = 2$.

each content that has been requested only by users connected to both relays. Since there are no transmission opportunities, the number of backhaul transmissions depends only on the number of different file requested by the total number of users connected to one relay $u_1$ and those to two relays $u_2$, regardless of how they are distributed between the black and white relays. In the ECC scheme, the master node gains a transmission whenever there is a content pair of which one file has been requested exclusively at relay black and the other exclusively at the white relay. In order to evaluate the gain achieved in ECC, it is necessary to know $u_B$ and $u_W$ while knowledge of how these users are connected to is not relevant. We explain with the following simple example that there is no analysis distinction between the two relay network and its extended version.

**Example 5.** *Let us assume to have* $T = 4$*, and four users connected to only one relay such that* $u_B = u_W = 2$*, and non users are connected to both relays. Let assume that each relay has cache size* $M = 2$ *and the library size is* $N = 4$*. As the scheme foresees, a possible cache placement might be*

$$\mathcal{Z}_B = \{f_1^{(1)}, f_2^{(1)}, f_3^{(1)}, f_4^{(1)}\} \text{ and } \mathcal{Z}_W = \{f_1^{(2)}, f_2^{(2)}, f_3^{(2)}, f_4^{(2)}\}$$

Let us now assume that all files are requested such that users connected to $u_B$ ask for $f_1, f_2$ while users connected to $u_W$ ask for $f_3, f_4$. The master in order to satisfy the users request can send either the following two transmissions

$$f_1^{(2)} \oplus f_3^{(1)} \text{ and } f_2^{(2)} \oplus f_4^{(1)} \tag{20}$$

or the following two

$$f_1^{(2)} \oplus f_4^{(1)} \text{ and } f_3^{(1)} \oplus f_2^{(2)}. \tag{21}$$

Since content in cache of relays of equal color is the same then transmission (20) or (21) will satisfy user request in the case that we have one user per relay or both $u_B$ users connected to the same $R_B$ or for any kind of combination of how $u_B = u_W = 2$ users can be connected.

As a final remark, a cache-aided network composed with $T$ relays can be easily studied by simplifying it under the new assumptions to a two-relay network, as presented in the current section.

## VII. Results

We start by evaluating the approximation average backhaul load and by comparing the MDS scheme and the ECC scheme.

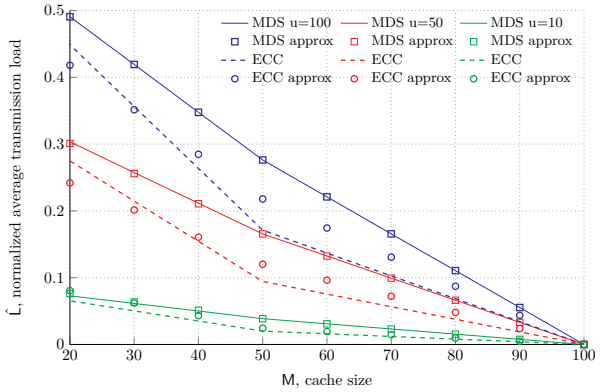


Fig. 5. The normalized average transmission load $\bar{L}$ as a function of cache size $M$ for $u = 10, 50, 100$ for $N = 100$. 40% of the users are connected to $R_B$, 40% to $R_W$ and 20% to both relays. Given $u$, the marked and dashed curves indicate the results for the ECC and the MDS scheme respectively.

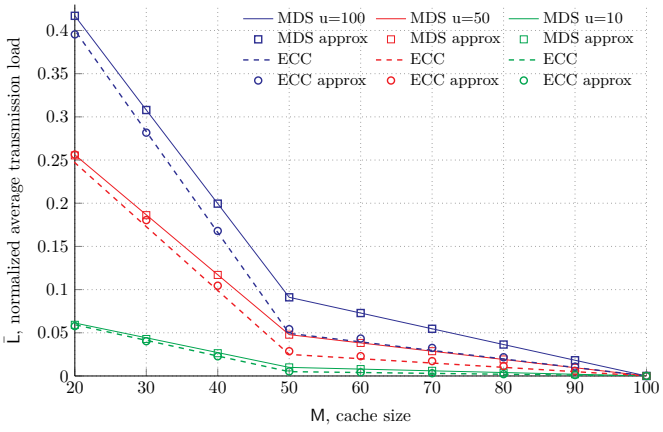


Fig. 6. The normalizzed average transmission load $\bar{L}$ as a function of cache size $M$ for $u = 10, 50, 100$ for $N = 100$. 10% of the users are connected to $R_B$, 10% to $R_W$ and 80% to both relays. Given $u$, the marked and dashed curves indicate the results for the ECC and the MDS scheme respectively.

In our fist scenario, we assume that the library size $N = 100$ 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 $R_B$ and half only to $R_W$.

In Fig. 5, the normalized average backhaul load as a function of the cache size $M$ for different number of users $u$ present in the system is plotted. Solid curves represent the

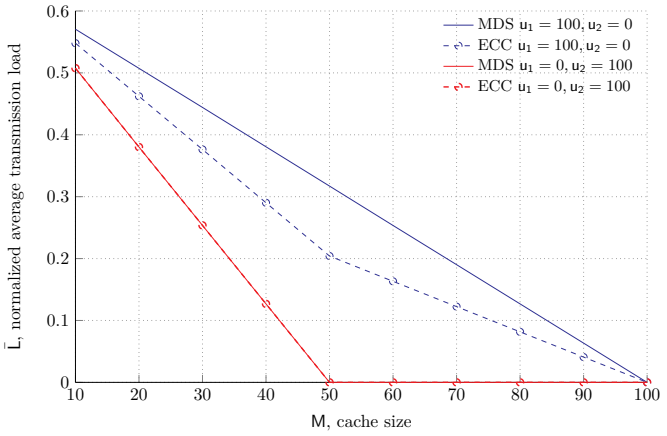


Fig. 7. The normalizzed average transmission load $\bar{L}$ as a function of cache size $M$ for the ECC and MDS schemes. Blue curves represents a scenario where each relays serves 50% of the usres and no users are connected to both relays while red curves when all users are connected to both relays.

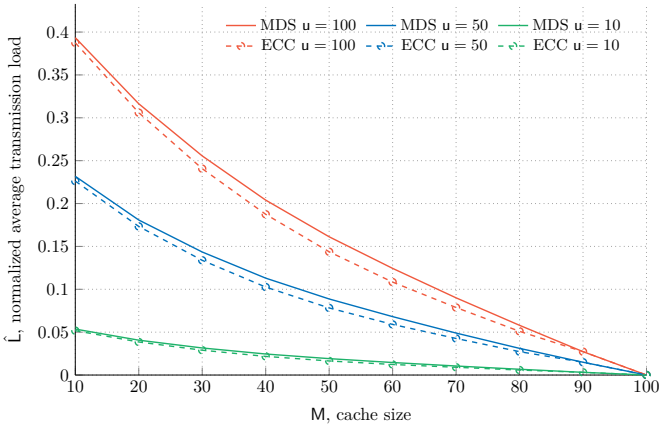


Fig. 8. The normalizzed average backhaul load $\bar{L}$ versus $M$ for $u = 10, 50, 100$ for $N = 100$ when file requests follows a Zipf distribution with $\alpha = 0.8$. 40% of users are connected to $R_1$, 40% to $R_2$ and 20% to both Rs. The dot marked and dashed curves indicate the results obtained for the ECC and MDS scheme respectively.

performance of the MDS scheme, dashed curves represents the performance of the ECC scheme while markers represent the results obtained by the presented approximation. Blue curves indicate that the number of users are $u = 100$, red curves indicate $u = 50$ while green curves indicate $u = 10$. The ECC scheme outperforms the benchmark MDS caching scheme for every number of users $u$ considered in the network. In fact, the ECC scheme requires the use of fewer backhaul resources for serving users compared to the MDS scheme. The approximation of the MDS scheme (square markers) coincides with the exact analysis (solid curves). Good results are instead obtained for the ECC scheme which the approximation is not exact but gives a good hint on the performance of the scheme. We believe that the average load of the MDS has a better approximation due to the fact that the expression has less approximated terms. As expected, by fixing $u$ requests, $\bar{L}$ 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 $M = N/2$, 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.

In Fig. 6 the normalized average backhaul load as a function of the cache size $M$ when 80% of the users are connected to both relays, 10% of the users are connected only to $R_W$ and 10% to $R_B$ is plot. Also in this case, the ECC scheme presents a better performance with respect to the MDS scheme. We can observe that the gain given by the ECC scheme is smaller with respect to the previous scenario. As explained, this is due to the fact that the gain depends on the number of exclusive files requested by users connected to only one relay. In this scenario, the our approximation is tighter as fewer terms have to be calculated for the gain due to the small number of users connected to one relay.

In Fig. 7 the average transmission load $\bar{L}$ as a function of cache size $M$ for two boarderline scenarios. We assume to have a library of $N = 100$ files and $u = 100$ users. Blue curves indicate that there are no users connected to both relays and each relay serves 50% of the users. Red curves indicate that all users are connected to both relays. As aspected, when all users are connected to both relays, there is no gain for the ECC scheme and the average load matches that of the MDS scheme. Instead, when non of the users are connected to both relays and each relays serves 50% of them then the ECC reaches the maximum gain compared to the MDS scheme. The ECC scheme reaches a gain of more than 10% when size cache is half of the library size.

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. 8. 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 [4]. 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.

In Fig. 9 the outage probability $P_{out}$ as a function of cache size $M$ for the ECC and MDS schemes for different capacity $C$ constraints is plot. It is assumed a library size of $N = 100$ and 50 users connected at each relay. Solid curves represent the results obtained for the MDS caching scheme while dashed curves for the ECC scheme. As normal, the probabilty of outage decreases by increasing the memory $M$ or the capacity in the backhaul link $C$. Also in this scenario can be observed a better performance of the ECC with respect to the MDS scheme. The gain allows to operate to a much lower outage probability for a given cache size $M$ and capacity constraint $C$. For example, assume that we have a memory cache of $M = 20$ and the capacity in the backahaul link is $C = 40$, then the value

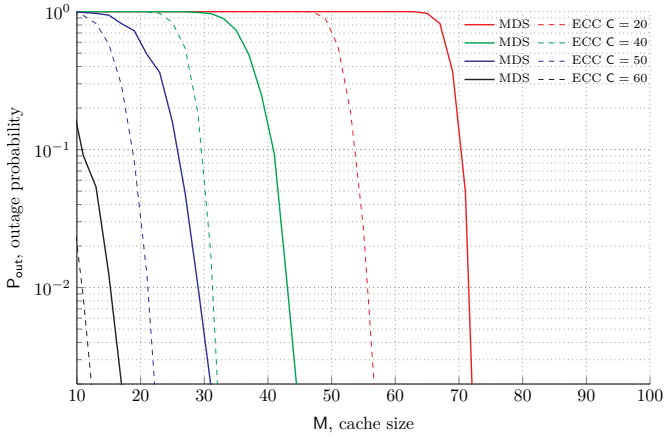


Fig. 9. The outage probability $\mathsf{P_{out}}$ as a function of cache size $\mathsf{M}$ for the ECC (dashed curves) and MDS (solid curves) schemes for different capacity constraints $\mathsf{C}$. In this scenario is considered a library size $\mathsf{N} = 100$ and $\mathsf{u} = 50$ users connected per relay.

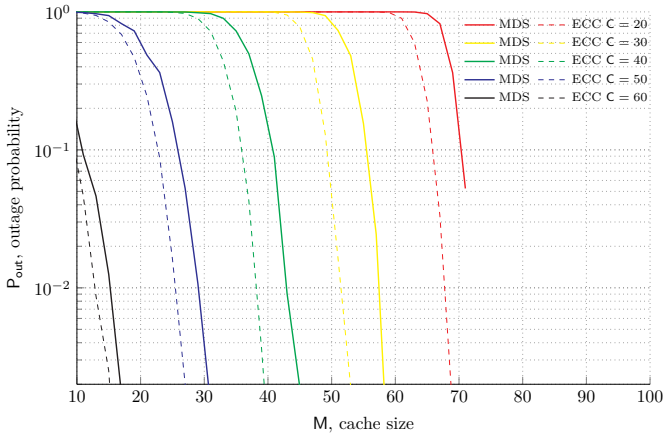


Fig. 10. The outage probability $\mathsf{P_{out}}$ as a function of cache size $\mathsf{M}$ for the ECC (dashed curves) and MDS (solid curves) schemes for different capacity constraints $\mathsf{C}$. In this scenario is considered a library size $\mathsf{N} = 100$ and $\mathsf{u}_B = 80$ users connected to one relay and $\mathsf{u}_W = 20$ to the other.

of the outage probability for the MDS is $\mathsf{P_{out}} = 5 \cdot 10^{-1}$ while the ECC reduce the outage of one order of magnitude, i.e. $\mathsf{P_{out}} = 2 \cdot 10^{-2}$.

The perfomance of our last scenario is represented in Fig. 10 where it is assumed a library size is $\mathsf{N} = 100$ and $\mathsf{u_B} = 80$ users connected to one relay and $\mathsf{u_W} = 20$ users to the other. The outage probability $\mathsf{P_{out}}$ as a function of cache size $\mathsf{M}$ for the ECC and MDS schemes for different capacity $\mathsf{C}$ constraints is plot. The trend of the outage illustrated is similar to our previous scenario and even if the number of coding opportunity in the delivery is reduced, the ECC is still showing better perfomance with respect to the MDS caching scheme. We clearly see that for a given probability of outage and a given capacity constrait the memory size required in the ECC scheme is smaller than in the MDS scheme.

## VIII. Conclusions

Considering a two-tier network with caching at the edge, an analysis between coding schemes for the placement and for the delivery was presented. In particular, the benchmark of the MDS coding caching scheme was studied for the placement phase while the Maddah-Ali caching scheme for the delivery phase. We identify and quantified the nature of the gain obtained when coding takes places in the delivery phase by casting out problem with known results obtained in the balls and bins setting. We provided an approximation of the average backhaul transmission load which allows a fast evaluation of the performance of each caching scheme. The results shows the perfect match of the approximation and the results for the MDS scheme while a good performance was obtained for the ECC scheme. The results further illustrate that the bachkaul transmissions can be reduced by applying the edge coded caching scheme whenever there are users connected to a single relay in the system. We also elavuated the outage probability of both schemes as a function of the memory size in two different scenarios. The good performance obtained validates the coded caching scheme in satellite networks and suggest its investigation in more sophisticated scenarios. Finally, we derived how a network with a generic number of relays can be simplified to a network with two-relays and studied with the derivations obtained in this work.

## References


[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] E. Recayte, F. Lázaro, and G. Liva, "Caching in heterogeneous satellite networks with fountain codes," *International Journal of Satellite Communications and Networking*, vol. n/a, no. n/a, pp. 1–10, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sat.1323

[6] E. Recayte, F. Lázaro, and G. Liva, "Caching at the edge with lt codes," in *2018 IEEE 10th International Symposium on Turbo Codes Iterative Information Processing (ISTC)*, 2018, pp. 1–5.

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

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

[9] 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.

[10] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *2016 IEEE Information Theory Workshop (ITW)*, 2016, pp. 161–165.

[11] M. M. Amiri, Q. Yang, and D. Gündüz, "Coded caching for a large number of users," in *2016 IEEE Information Theory Workshop (ITW)*, 2016, pp. 171–175.

[12] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1281–1296, 2018.

[13] C. Tian and J. Chen, "Caching and delivery via interference elimination," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1548–1560, 2018.

[14] S. Sahraei and M. Gastpar, "K users caching two files: An improved achievable rate," in *2016 Annual Conference on Information Science and Systems (CISS)*, 2016, pp. 620–624.

[15] M. Mohammadi Amiri and D. Gündüz, "Fundamental limits of coded caching: Improved delivery rate-cache capacity tradeoff," *IEEE Transactions on Communications*, vol. 65, no. 2, pp. 806–815, 2017.

[16] J. Gómez-Vilardebó, "Fundamental limits of caching: Improved bounds with coded prefetching," *arXiv preprint arXiv:1612.09071*, 2016.

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

[18] A. Kalantari, M. Fittipaldi, S. Chatzinotas, T. X. Vu, and B. Ottersten, "Cache-assisted hybrid satellite-terrestrial backhauling for 5g cellular networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[19] H. Wu, J. Li, H. Lu, and P. Hong, "A two-layer caching model for content delivery services in satellite-terrestrial networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[20] K. An, Y. Li, X. Yan, and T. Liang, "On the performance of cache-enabled hybrid satellite-terrestrial relay networks," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1506–1509, 2019.

[21] E. Recayte, "Coded caching at the edge of satellite networks," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 1124–1130.

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

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