Caching in Heterogeneous Networks with Per-File Rate Constraints

Estefanía Recayte, *Student Member, IEEE* and Giuseppe Cocco, *Member, IEEE*

***Abstract*—We study the problem of caching optimization in heterogeneous networks with mutual interference and per-file rate constraints from an energy efficiency perspective. A setup is considered in which two cache-enabled transmitter nodes and a coordinator node serve two users. We analyse and compare two approaches: (i) a cooperative approach where each of the transmitters might serve either of the users and (ii) a non-cooperative approach in which each transmitter serves only the respective user. We formulate the cache allocation optimization problem so that the overall system power consumption is minimized while the use of the link from the master node to the end users is spared whenever possible. We also propose a low-complexity optimization algorithm and show that it outperforms the considered benchmark strategies.**

Our results indicate that significant gains both in terms of power saving and sparing of master node's resources can be obtained when full cooperation between the transmitters is in place. Interestingly, we show that in some cases storing the most popular files is not the best solution from a power efficiency perspective.

***Index Terms*—Caching Networks, Cooperative Communications, Energy Efficiency, Interference Channel, Rate Constraints.**

I. Introduction

Wireless data traffic has grown dramatically in recent years. It is foreseen that traffic demand in the fifth generation (5G) mobile communication networks will increase by 1000-fold by the year 2020 [1]. This huge demand will be mainly characterized by high-definition video contents and streaming applications which not only require a significant ammount of bandwidth but also have stringent quality of service requirements (QoS) [1]. In order to offload the network and to reduce the system level energy consumption, *cache-aided networks* have been studied extensively in the recent years [2]–[20]. One of the main advantages of caching is the possibility to spare backhaul resources by storing files during periods in which the network traffic is low and directly serve the users during high-load periods. Due to the limited storage capacity of the transmitters, the design of effective caching policies is required.

Cache-aided networks have been studied in literature with the aim of minimizing backhaul congestion [3] and energy consumption [4]–[6] or maximazing the throughput [7] [8]. The issue of interference in cache-aided networks has also been studied in many recent works [7]–[11]. However, most works consider an approach in which files are partitioned into sub-files and the same transmission rate over the channel is used for each of them.

In this paper, we focus on a full cooperative caching interference network. By jointly considering per file rate requirements and popularity ranking, we determine the content placement and the delivery strategies in order to optimize the energy efficiency of the system.

A. Related works

Caching optimization has been investigated in recent works for different kinds of heterogeneous networks. In [13] a two layer caching model that aims at minimizing the satellite bandwidth consumption by storing information both at ground stations and at the satellite is proposed. In [14] the authors consider cache-enable unmanned aerial vehicles (UAV) and study the optimal location, cache content and user-UAV association with the aim to maximize users' quality of experience while minimizing the transmit power. Most of the proposed solutions exploit the statistics of past users' file requests (often referred to as *popularity ranking*) for filling up local caches with the most popular contents [3] [15].

Caching from an energy efficiency perspective has also been extensively studied. In [6] the authors aim at reducing the total energy cost in a heterogeneous cellular network by applying a caching optimization algorithm for multicast transmission. Energy minimization in the backhaul link is studied in [4], where the authors propose a strategy for jointly optimizing content placement and delivery, which leads to minimize the energy consumption. Energy efficiency is also studied in [5], where a proactive downloading strategy to exploit the good channel conditions and minimize the total energy expenditure is proposed.

Due to the significant network densification expected in 5G and beyond, interference management represents a key challenge. In order to counteract interference, a combination of caching with interference alignment (IA) and zero-forcing (ZF) techniques has been recently proposed. In their seminal work [7], Maddah-Ali and Niesen leverage on IA and ZF to show that significant gains in terms of load balancing and degrees of freedom can be obtained when caches at both the transmitter and the receiver side are used. In [9] and [10] the authors extend the work in [7] considering a delivery phase in which different sub-files are transmitted over different channel

Estefanía Recayte is with the Institute of Communication and Navigation of the Deutsches Zentrum für Luft- und Raumfahrt (DLR), 82234 Weßling, Germany. Email: estefania.recayte@dlr.de. and with DEIS, University of Bologna, 40136 Bologna, Italy.

Giuseppe Cocco is with the LTS4 Signal Processing Laboratory and the Laboratory of Intelligent Systems (LIS) of the École Polytechnique Fédérale de Lausanne (EPFL). Email: giuseppe.cocco@epfl.ch.

Giuseppe Cocco is partly founded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Individual Fellowship grant agreement No. 751062.

blocks. In [11] authors propose a file placement strategy that allows to create cooperation opportunities between transmitters through interference cancellation and IA.

To the best of our knowledge, energy optimization in cooperative interference networks with caching capabilities only at the transmitter assuming a one-shot delivery phase and considering different per-file rate requirements and different popularity rankings has not yet been analysed in literature. Due to the one-shot delivery phase and to the different per-file rate requirements, the minimum sum-power needed to satisfy users requests depends on both the files' rates and the state of the links involved in the transmission. Since the requests are not known during the placement phase, a statistical optimization algorithm leveraging on the file popularity ranking has to be used to choose the cache allocation achieving the minimum power. Due to the discrete nature of the optimization along the cache allocation dimension and the non convexity of the objective function in the sum-power dimension, it is challenging to find a closed form solution or an algorithm with a complexity that scales well with the number of files.

B. *Contributions*

We optimize the placement and delivery phase for minimizing the power consumption of a cache-enabled network under QoS requirements and in the presence of interference. Our main contributions are the following:

- Unlike previous works, files are delivered over an interference channel and different PHY rate requirements per file have to be respected. While caching over interference channels has been formerly studied [20], in previous works the rate constraints are relative to the user rather than to the file, i.e., each file has the same rate requirement while the per-user rate over the channel can differ, in that it accounts for information relative to possibly more than one file. Moreover, in [20] only the broadcast (BC) strategy is considered and, unlike the present work, receivers are equipped with a cache, which allows them to efficiently perform interference cancellation.
- Unlike [6], [8], [11], we assume a one-shot delivery phase, in which transmission takes place in a single channel block and all selected files are transmitted at the same time.
- Depending on cache content and requests, transmitters can apply one of the following strategies: multiple input multiple output (MIMO) broadcast transmission, the Han-Kobayashi rate splitting approach [33], multiple input single output (MISO) transmission, broadcast and multicast (MC) transmission. Up to our knowledge no other paper so far jointly considered this set of techniques in the context of caching.
- Extending our preliminary work [18], we establish a correspondence between the cache allocation/file request and the strategy that is implemented using the set of cooperative techniques listed in the previous point and we derive the minimum power cost needed for satisfying the rate requirements.
- We propose an iterative algorithm for the Han-Kobayashi strategy that asymptotically converges to the minimum required sum-power and is significantly more efficient than an exhaustive search algorithm using the same granularity.
- We propose a low-complexity suboptimal algorithm to solve the cache allocation problem. The algorithm outperforms benchmark policies based on highest popularity and highest rate and has a complexity which grows linearly with the number of files.
- Our numerical results show that memorizing the most popular files is not always the most convenient strategy when transmitters have to respect per-file rate requirements. Furthermore, thanks to the considered cooperation strategies a significant saving of the coordinator node's resources can be achieved with respect to a non-cooperative system.

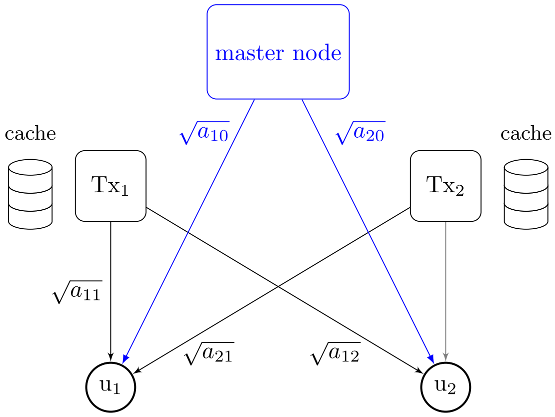


Figure 1: System model of the considered heterogeneous network with caches at the transmitters.

The remainder of the paper is organized as follows. In Section II we introduce the system model. In Section III the different transmission strategies are presented and the relative power optimization subproblems are addressed. The cache optimization is tackled in Section IV. V contains the numerical results while the conclusions are presented in Section VI.

II. System Model

A. *Preliminaries*

We consider a heterogeneous network composed by a transmitter master node which has access to a collection of files (eg. video files), two transmitter nodes with limited caching capabilities and two users u_n, $n = \{1, 2\}$. All nodes in the network are equipped with a single antenna. The analysis presented can be extended to a multi-antenna setup, in which case a lower power consumption can be achieved.

The study carried out in this paper can be extended to the case of a generic number of SBSs and users. In doing so, the Han-Kobayashi scheme can be replaced with a transmission approach similar to that considered in [27].

For ease of exposition in the following we assume a 5G network setup, with a macro base station (MBS) as master node and two small base stations (SBS) as transmitter nodes. However, the optimization problem and the transmission techniques presented in this work are not limited to a terrestrial architecture. In fact, in a heterogeneous satellite network a

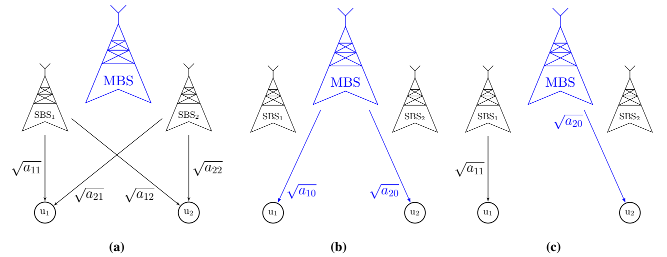


Figure 2: System model for the non cooperative approach: (a) interference as noise, (b) broadcast or multicast channel and (c) orthogonal channel

geostationary (GEO) satellite connected to a ground station could act as the master node with access to all the files while low Earth orbit (LEO) satellites connected to the respective cache-enabled ground stations act as transmitter nodes. Another possibility that is increasingly receiving attention both from the research community and the industry is to employ unmanned aerial vehicles (UAV) as mobile base stations. In such context, a high altitude UAV, a satellite or an MBS could act as master node while drones flying at lower altitude take the role of transmitters.

We assume that SBSs and MBS work in two separate frequency bands. The reason for choosing such setup is that it is likely to be implemented in the next generation of mobile networks (5G) [22]. In 5G, the SBSs will be assigned frequencies in the mmWave that, thanks to the large bandwidth available, allow to reach high data rates, while the MBS operate at lower frequency, providing continuous coverage to mobile users with lower data rates. An orthogonal bandwidth setup is also foreseen in next generation heterogeneous satellite networks [26] and could be as well employed in UAV-assisted networks.

We assume that user u_n is associated to SBS$_n$. Let us denote with $\sqrt{a_{n0}}$ the channel coefficient between u_n and the MBS while $\sqrt{a_{nm}}$ denotes the channel coefficient between u_n and SBS$_m$, $m = \{1,2\}$, as shown in Fig. 1. We also define $a_{+0} \triangleq \max\{a_{10}, a_{20}\}$ and $a_{-0} \triangleq \min\{a_{10}, a_{20}\}$ while R_+ (R_-) is the required rate of the file requested by the user with MBS channel coefficient a_{+0} (a_{-0}). Similarly P_+ and P_- are defined as the power used for files with rates R_+ and R_-, respectively. For SBSs we define $a_{+1} \triangleq \max\{a_{11}, a_{21}\}$, $a_{-1} \triangleq \min\{a_{11}, a_{21}\}$, $a_{+2} \triangleq \max\{a_{22}, a_{12}\}$, $a_{-2} \triangleq \min\{a_{22}, a_{12}\}$ while R_+, R_-, P_+ and P_- are defined for SBS$_1$ and SBS$_2$ similarly as for the MBS.

The MBS has access to a library of N files $\mathcal{F}$=$\{f_1, \dots, f_N\}$. We assume that all files have the same size. This assumption is justifiable in practice since files can be divided into blocks of the same size [19]. Each file f_i has a minimum required transmission rate R_i, i.e., the minimum rate measured in bits/s/Hz at which the file has to be transmitted to the user in order to satisfy the QoS constraints[1]. The probability that f_i is requested by a user is called *popularity ranking* and is indicated as q_i. We assume that SBS$_n$ is equipped with a local cache of size M the content of which is denoted by $\mathcal{M}_n$.

The binary variable x_{in} indicates whether file f_i is present in the memory of SBS$_n$ ($x_{in} = 1$) or not ($x_{in} = 0$). We further define $\bar{x}_{in} \triangleq 1 - x_{in}$. We denote with $\mathbf{x}_1$ and $\mathbf{x}_2$ the allocation vectors containing the values for x_{i1} and x_{i2}, respectively. Depending on whether the files requested by the users are present in the cache or not and on whether a cooperative scheme is considered or not, the system implements a different transmission approach. The possible approaches are explained in Section III. We indicate with $\mathbb{1}_\alpha$ the indicator function which takes value 1 if $\alpha > 0$ and 0 otherwise. We further define $\bar{\mathbb{1}}_\alpha \triangleq 1 - \mathbb{1}_\alpha$. We denote with $P_i^{(n)}$ the power used by the transmitter for sending f_i to user u_n at rate R_i. All transmission rates in the following are intended per complex dimension. We denote with $\mathbf{d} = [f_i \quad f_j]$ the users' demand vector. Without loss of generality we assume that f_i is requested by u_1 and f_j by u_2.

The value $c_T(i,j)$ indicates the minimum power consumed for serving the users when the system implements the transmission approach T. Such value is the sum of two transmission powers $c_T(i,j) = P_i^{(1)} + P_j^{(2)}$ if two transmitters are active while in case one transmitter takes care of both requests, the power cost is denoted as $c_T(i,j) = P_{i,j}^{(1,2)}$. Finally, we indicate with $Q_c(\mathbf{x}_1, \mathbf{x}_2)$ and $Q_{nc}(\mathbf{x}_1, \mathbf{x}_2)$ the expected power cost for serving a request in the cooperative and in the non-cooperative approach, respectively.

B. System Architecture

We assume that the SBSs are connected to the MBS through a backhaul link while users have a direct radio frequency link to both the SBSs and the MBS.

In the placement phase the SBSs fill their own caches without deterministic knowledge of the user demands. Our aim is to understand which is the best strategy to choose the files for each transmitter's cache so that the average power consumption is minimized during the delivery phase.

In the delivery phase each user requests a file. We assume that the SBSs and the MBS can listen to the requests of both users and the communication is error-free[2]. We assume that SBSs and MBS estimate the channel during the request and the channel does not vary during the file transmission while changes between two consecutive transmissions. Channel state information at the transmitter (CSIT) is assumed. The master node has knowledge of all the channel coefficients, the cache content and the user requests. According to this knowledge the MBS decides how the users have to be served and, eventually, coordinates the transmitters. The file request and the file transmission take place in consecutive time slots and each user has to wait until the end of the transmission slot before placing a new request[3].

[1] Although we consider a constraint in terms of transmission rate over the channel, this can be easily translated into a constraint in terms of delay and quality, which is particularly suited to model video and image transmission.

[2] In practice this can be achieved by noting that the size of the request message is small and assuming a low channel code rate.

[3] Since files have the same size but different rate constraints, the transmission of the file with highest rate terminates before the other file's. This implies an interference-free transmission during the part of the slot for the latter. In this paper we look for an upper bound on the minimum required power and assume that the interference is constant through the transmission

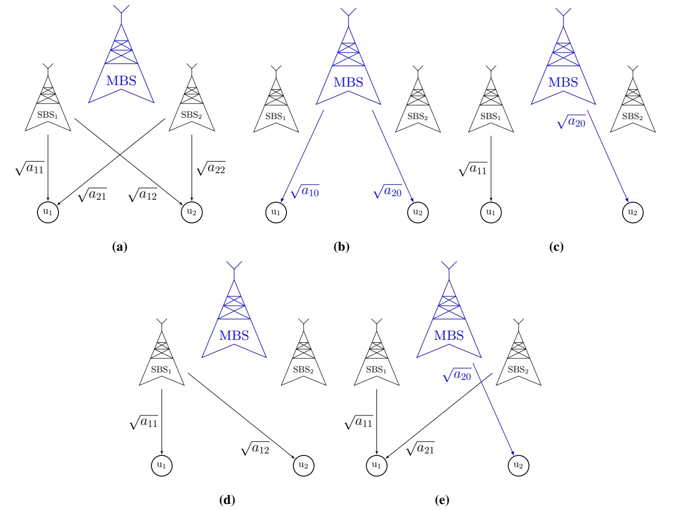


Figure 3: System model for the cooperative approach: (a) interference channel without common information or interference channel with common information or MIMO broadcast channel, (b) broadcast or multicast channel from MBS, (c) orthogonal channel, (d) multicast or broadcast channel from SBS and (e) MISO channel.

Two different approaches are considered: a non-cooperative approach (NCA) and a cooperative approach (CA). In the NCA user u_n is served by SBS$_n$ if the requested file is present in the cache of SBS$_n$ and channel conditions (interference plus noise and fading levels) allow reliable communication at the required rate. If this is not possible, the user is served directly by the MBS. In the CA user u_n can be served by either the SBSs or by the MBS if none of the SBSs has the requested file. The CA has two goals. On the one hand it aims at reducing the overall system's energy consumption and on the other hand it tries to spare the MBS resources as much as possible. As a matter of facts in future heterogeneous terrestrial networks a single MBS will serve a large number of SBSs [22]. Thus, in order to avoid the MBS being a bottleneck, priority is always given to SBS transmissions whenever possible.

III. Channel Models and Power Minimization Subproblems

In this section we introduce the different transmission approaches with the relative channel models. For each of them we minimize the power $c_T(i,j)$ required to transmit f_i to u_1 and f_j to u_2 at rate R_i and R_j, respectively. In practical systems, a maximum per-node power constraint is present due to the physical limitations of the transmitters. Unless otherwise stated, in the following we assume that such maximum power is larger than that required for each considered setup. This allows us to get an insight on the power efficiency and MBS resource sparing, which is a step towards the full understanding of interference-limited caching systems. Note that using such assumption is not new in the literature related to caching. In [20], for instance, bounds on the minimum average required power in a caching systems are studied with no explicit limits in terms of maximum per-node power.

The MBS decides which of the strategies described in the following has to be applied. The chosen strategy depends on which file has been requested and on the cache content of the SBSs. From now on, the subscripts in $P_i^{(1)}$ and $P_j^{(2)}$ are removed if there is no ambiguity.

A. Gaussian Interference Channel with Common Information (GI$_c$)

This channel is considered in the CA when the demand vector of the users is $\mathbf{d} = [f_i \quad f_i]$ and $x_{i1} = 1$, $x_{i2} = 1$, (Fig. 3-a). We consider the interference channel model in standard form, which was introduced in [23]. Starting from the physical model, the standard form is obtained as follows. Let us consider the physical channel model

$$\begin{aligned} y_1 &= \sqrt{a_{11}} \cdot x_1 + \sqrt{a_{12}} \cdot x_2 + z_1, \\ y_2 &= \sqrt{a_{21}} \cdot x_1 + \sqrt{a_{22}} \cdot x_2 + z_2, \end{aligned} \tag{1}$$

where z_1 and z_2 are independent Gaussian variables with zero mean and unit variance. From the point of view of the achievable rates such model is equivalent to [23]

$$\begin{aligned} y_1 &= x_1 + \sqrt{c_{12}} \cdot x_2 + z_1, \\ y_2 &= \sqrt{c_{21}} \cdot x_1 + x_2 + z_2, \end{aligned} \tag{2}$$

where $c_{12} = a_{12}/a_{22}$, $c_{21} = a_{21}/a_{11}$. Let us denote with $P^{(1)}$ ($P^{(2)}$) the physical power of x_1 (x_2). Given the power $\tilde{P}^{(n)}$ in the standard model, the physical power is: $P^{(1)} = \tilde{P}^{(1)}/a_{11}$ and $P^{(2)} = \tilde{P}^{(2)}/a_{22}$. In the GIc the same file f_i has to be transmitted to both users at rate $R_i > 0$. Using the results in [24] and noting that in our case only a common message is to be transmitted, it can be easily shown that the power minimization problem can be written as

$$\begin{aligned} \underset{P_i, P_j}{\text{minimize}} \quad & \frac{P_i^{(1)}}{a_{11}} + \frac{P_j^{(2)}}{a_{22}}, \\ \text{subject to} \quad & \frac{1}{2}\log_2\left(1 + \left(\sqrt{P_i^{(1)}} + \sqrt{c_{12} \cdot P_j^{(2)}}\right)^2\right) \geq R_i, \\ & \frac{1}{2}\log_2\left(1 + \left(\sqrt{P_j^{(2)}} + \sqrt{c_{21} \cdot P_i^{(1)}}\right)^2\right) \geq R_i, \\ & P_i^{(1)}, P_j^{(2)} \geq 0. \end{aligned}$$

The minimum power cost $c_{GIc}(i,j) = P_i^{(1)} + P_j^{(2)}$ is derived in Appendix I.

B. Gaussian Interference as Noise (GIN)

This channel is considered in the NC approach when $x_{i1} = 1$, $x_{j2} = 1$ and $\mathbb{1}_\alpha = 1$ (Fig. 2-a). where $\mathbb{1}_\alpha$ is the indicator function and $\alpha = a_{22} + \frac{a_{21} \cdot a_{12}}{a_{11}} \cdot (2^{2Rj} - 1) \cdot (1 - 2^{2R_i})$.

The channel model is given by (1). In this transmission scheme, each SBS considers the interference generated by the other SBS as noise. The minimization problem is presented in [18] and not reported here for a matter of space. The minimum cost is achieved when

$$P_i = \frac{1}{a_{11}} \cdot (2^{2R_i} - 1) \cdot (a_{12} \cdot P_j + 1)$$

and

$$P_j = \frac{(2^{2R_j} - 1) \cdot \left[\frac{a_{21}}{a_{11}} \cdot (2^{2R_i} - 1) + 1\right]}{a_{22} + \frac{a_{21} \cdot a_{12}}{a_{11}} \cdot (2^{2R_j} - 1) \cdot (1 - 2^{2R_i})}. \tag{3}$$

The value of P_j in (3) is positive if and only if $\alpha > 0$. In such case the SBSs are able to successfully deliver their files to the respective users. On the contrary, if $\alpha \leq 0$ no $P_i^{(1)}$ and $P_j^{(2)}$ exist such that the minimum required rate can be achieved. In such case, the SBSs are not able to serve their users and the MBS will serve both users either applying a broadcast (if files are different) or a multicast transmission (if files are the same). The cost of the GIN channel is

$$c_{GIN}(i,j) = \frac{(2^{2R_j}-1)\left[\frac{a_{21}}{a_{11}}(2^{2R_i}-1)+1\right]}{a_{22}+\frac{a_{21}a_{12}}{a_{11}}(2^{2R_j}-1)(1-2^{2R_i})} + \frac{(2^{2R_i}-1)}{a_{11}}\left(\frac{a_{12}(2^{2R_j}-1)\left[\frac{a_{21}}{a_{11}}(2^{2R_i}-1)+1\right]}{a_{22}+\frac{a_{21}a_{12}}{a_{11}}(2^{2R_j}-1)(1-2^{2R_i})}+1\right).$$

C. Multicast Channel

This channel is considered when $\mathbf{d} = [f_i \quad f_i]$ in the following cases: (i) in both CA and NCA approaches when requests are satisfied by the MBS ($x_{i1} = 0$, $x_{i2} = 0$), in which case, since both users require the same file, the MBS sends the file at the level of power needed to satisfy the u_n experiencing the worst channel condition, (ii) in the CA when both requests are satisfied by one SBS ($x_{i1} = 1$, $x_{i2} = 0$) or ($x_{i1} = 0$, $x_{i2} = 1$) and (iii) in the NCA when $\mathbb{1}_\alpha = 0$. The CA is represented in Fig. 3-b,c while the NCA in Fig. 2-a.

For the case of multicast or broadcast transmission the channel model is

$$\begin{aligned} y_1 &= \sqrt{a_{10}} \cdot x_1 + z_1, \\ y_2 &= \sqrt{a_{20}} \cdot x_2 + z_2. \end{aligned} \tag{4}$$

The minimum power required for u_1 is $P_i^{(1)} = \frac{2^{2R_i}-1}{a_{10}}$ while for u_2 is $P_i^{(2)} = \frac{2^{2R_i}-1}{a_{20}}$. Thus, we have $P_i^{(1,2)} = \max\left\{\frac{2^{2R_i}-1}{a_{10}}, \frac{2^{2R_i}-1}{a_{20}}\right\} = \frac{2^{2R_i}-1}{a_{-0}}$. The cost of this channel is

$$c_{\text{MC}}^{\text{MBS}}(i,j) = \frac{2^{2R_i}-1}{a_{-0}}.$$

Similarly, for a multicast transmission from SBS$_n$ we have that $c_{\text{MC}}^{\text{SBS}}(i,j) = \frac{2^{2R_i}-1}{a_{-n}}$.

D. Broadcast Channel

In the broadcast channel, we have one transmitter and two receivers which request different files [28]. This type of channel is implemented by the MBS or one of the SBSs when $\mathbf{d} = [f_i \quad f_j]$, $i \neq j$, in the following cases: (i) in the CA when files are not present in SBSs caches ($x_{i1} = 0$, $x_{i2} = 0$, $x_{j1} = 0$, $x_{j2} = 0$), in this case the MBS serves both users, (ii) in the CA when one SBS has both files while the other has none of them, that is when ($x_{i1} = 1$, $x_{i2} = 1$, $x_{j1} = 0$, $x_{j2} = 0$) or ($x_{i1} = 0$, $x_{i2} = 0$, $x_{j1} = 1$, $x_{j2} = 1$), in this case a SBS serves both users and (iii) in NCA when the file requested by u_n is not present in SBS_n for $n = 1, 2$ thus the MBS serves both users ($x_{i1} = 0$, $x_{j2} = 0$). Finally, (iv) the broadcast channel is implemented in the NCA when $\mathbb{1}_\alpha = 0$. The CA is represented in Fig. 3-b,c while the NCA in Fig. 2-a.

The power minimization problem to be solved, considering an MBS transmission, can be found in [18] and thus is not reported here since it is formally the same as in case of SBS transmission. Let P_-, P_+ be the powers related to the worst and best channel, respectively, as defined in Section II-A. The overall cost of this channel is

$$c_{\text{BC}}^{\text{MBS}}(i,j) = \frac{2^{2R_+}-1}{a_{+0}} + \frac{(2^{2R_-}-1)(1+a_{-0}P_+)}{a_{-0}}.$$

Similarly, for the broadcast transmission from SBS$_n$ we have:

$$c_{\text{BC}}^{\text{SBS}}(i,j) = \frac{2^{2R_+}-1}{a_{+n}} + \frac{(2^{2R_-}-1)(1+a_{-n}P_+)}{a_{-n}}.$$

E. MIMO Broadcast

This channel is considered in the CA when $\mathbf{d} = [f_i \quad f_j]$, for $i \neq j$, and each SBS has both requested files ($x_{i1} = 1, x_{i2} = 1, x_{j1} = 1, x_{j2} = 1$), as depicted in Fig.3-a. In this case, the MBS coordinates the SBSs transmission and provides the information needed (e.g. the channel matrices) such that each SBS calculates the portion of the MIMO codeword to be transmitted. This channel is not considered in [18] and is analyzed in the following.

We denote with $\mathbf{H}_1 = [\sqrt{a_{11}} \quad \sqrt{a_{21}}]^T$ and $\mathbf{H}_2 = [\sqrt{a_{12}} \quad \sqrt{a_{22}}]^T$ the channel matrices between each SBSs and the two users. In the MIMO BC channel the two single-antenna SBSs collaborate for transmitting both files acting as a single multi-antenna terminal. The dirty paper coding (DPC) approach is used. DPC consists in a layered precoding of the data so that the effect of the self-induced interference can be cancelled out [29]. User messages are sequentially encoded. Let assume that the message for u_1 is encoded first. The decoding is such that a message experiences interference only from messages encoded after him. In this case, the message for u_1 is interfered by the message for u_2 and u_2 experiences no interference.

Let us define π_m with $m = \{1, 2\}$ as a permutation describing the possible encoding orders such that $\pi_1 = \{2, 1\}$ and $\pi_2 = \{1, 2\}$ and let $\pi_m(l)$ indicate the l-th element of the encoding order m, e.g. $\pi_1(1) = 2$ and $\pi_1(2) = 1$. The minimization problem to be solved is

$$\begin{aligned} \underset{S_1,S_2,\pi_m}{\text{minimize}} \quad & \text{tr}\left(\sum_{j=1}^{2}\mathbf{S}_j\right) \\ \text{subject to} \quad & \frac{1}{2}\log_2\left|\mathbf{I}+\widetilde{\mathbf{H}}_{\pi_m(2)}^T\mathbf{S}_{\pi_m(2)}\widetilde{\mathbf{H}}_{\pi_m(2)}\right| = R_{\pi_m(2)}, \\ & \frac{1}{2}\log_2\frac{\left|\mathbf{I}+\sum_{j=1}^{n}\widetilde{\mathbf{H}}_{\pi_m(1)}\mathbf{S}_j\widetilde{H}_{\pi_m(1)}\right|}{\left|\mathbf{I}+\widetilde{\mathbf{H}}_{\pi_m(1)}^T\mathbf{S}_{\pi_m(2)}\widetilde{\mathbf{H}}_{\pi_m(1)}\right|} = R_{\pi_m(1)}, \\ & m = 1, 2 \end{aligned} \tag{5}$$

where $\widetilde{\mathbf{H}}_n$ is a square channel matrix obtained by opportunely modifying $\mathbf{H}_n$. Such modification is done as suggested in [30], by introducing a virtual receive antenna at each user with virtual channel gain equal to ϵ, with ϵ much smaller than any of the channel coefficients. Thus, $\widetilde{\mathbf{H}}_1 = \left(\begin{smallmatrix}\sqrt{a_{11}} & \epsilon \\ \sqrt{a_{21}} & \epsilon\end{smallmatrix}\right)$ and $\widetilde{\mathbf{H}}_2 = \left(\begin{smallmatrix}\sqrt{a_{12}} & \epsilon \\ \sqrt{a_{22}} & \epsilon\end{smallmatrix}\right)$. $\mathbf{S}_1$ and $\mathbf{S}_2$ are the positive semi-definite input covariance matrices for u_1 and u_2 respectively while $\mathbf{I}$ is the noise covariance matrix. To solve (5) we transform the BC optimization problem into a dual MAC optimization problem

[31]. Then, the equivalent convex optimization problem can be written as

$$
\begin{aligned}
\underset{\mathbf{S}_1^M, \mathbf{S}_2^M, \pi_m}{\text{minimize}} \quad & \text{tr}\left(\sum_{j=1}^{2} \mathbf{S}_j^M\right) \qquad (6)\\
\text{subject to} \quad & \frac{1}{2}\log_2 |\mathbf{I} + \sum_{j=1}^{2} \widetilde{\mathbf{H}}_j \mathbf{S}_j^M \widetilde{\mathbf{H}}_j^T| = \bar{R}^M_{\pi_m(2)},\\
& \frac{1}{2}\log_2 |\mathbf{I} + \widetilde{\mathbf{H}}_{\pi_m(1)} \mathbf{S}^M_{\pi_m(1)} \widetilde{\mathbf{H}}^T_{\pi_m(1)}| = \bar{R}^M_{\pi_m(1)},\\
& m = 1, 2,
\end{aligned}
$$

where $\mathbf{S}_j^M$ indicates the dual MAC covariance matrix. The following holds: $\text{tr}\left(\sum_{j=1}^{2} \mathbf{S}_j\right) = \text{tr}\left(\sum_{j=1}^{2} \mathbf{S}_j^M\right)$ and $\bar{R}_i^M = \sum_{j=i}^{2} R_i$. It is not straightforward to find the solution to problem (6). A suboptimal solution $\left(\mathbf{S}_1^M, \mathbf{S}_2^M\right)$ can be obtained using the iterative algorithm suggested in [30]. The minimum power is readily obtained as $P_i + P_j = \text{tr}\left(\mathbf{S}_1 + \mathbf{S}_2\right) = \text{tr}\left(\mathbf{S}_1^M + \mathbf{S}_2^M\right)$ and the overall cost of the channel is $c_{\text{MIMO}}(i, j) = P_i + P_j$.

F. Orthogonal Channel

This channel is considered when u_1's (u_2's) request is satisfied by the SBS and u_2's (u_1's) request is satisfied by the MBS using orthogonal channels. In the CA this occurs when $(x_{i1} = 1, \; x_{i2} = 0, x_{j1} = 0, x_{j2} = 0)$, $(x_{i1} = 0, x_{i2} = 0, x_{j1} = 0, x_{j2} = 1)$ or $(x_{i1} = 0, x_{i2} = 1, x_{j1} = 0, \; x_{j2} = 0)$, $(x_{i1} = 0, x_{i2} = 0, x_{j1} = 1, x_{j2} = 0)$. In the NCA such channel is considered when $(x_{i1} = 1, x_{j2} = 0)$ or $(x_{i1} = 0, x_{j2} = 1)$ (Fig. 2-c and Fig. 3-c). Let us assume, without loss of generality, that u_1 is served by SBS$_1$ and u_2 by the MBS. The channel model is the following

$$
\begin{aligned}
y_1 &= \sqrt{a_{11}} \cdot x_1 + z_1,\\
y_2 &= \sqrt{a_{20}} \cdot x_2 + z_2.
\end{aligned}
$$

The cost of this channel is

$$
c_{\perp}(i, j) = \frac{2^{2R_i} - 1}{a_{11}} + \frac{2^{2R_j} - 1}{a_{20}}.
$$

G. MISO Orthogonal

The multiple input single output orthogonal channel is considered in the CA when both SBSs have the file requested by one user but do not have the file requested by the other user $(x_{i1} = 1, x_{i2} = 1, x_{j1} = 0, x_{j2} = 0$ and $x_{i1} = 0, x_{i2} = 0, x_{j1} = 1, x_{j2} = 1)$ (Fig.3-e). Without loss of generality, let us assume that u_1 is served by the two SBSs and u_2 is served by the MBS. We recall that all nodes have a single antenna but, thanks to the coordination of the MBS, the two transmitters act as a multi-antenna terminal. The channel model is the following

$$
\begin{aligned}
y_1 &= \sqrt{a_{11}} \cdot x_1 + \sqrt{a_{12}} \cdot x_1 + z_1,\\
y_2 &= \sqrt{a_{20}} \cdot x_2 + z_2.
\end{aligned}
$$

The problem to be solved is [32]

$$
\begin{aligned}
\underset{P_i, P_j}{\text{minimize}} \quad & P_i + P_j,\\
\text{subject to} \quad & \frac{1}{2}\log_2\left(1 + (a_{11} + a_{12}) \cdot P_i\right) \geq R_i,\\
& \frac{1}{2}\log_2\left(1 + a_{20} \cdot P_j\right) \geq R_j,\\
& P_i, P_j \geq 0,
\end{aligned}
$$

where the solution is $P_i = \frac{2^{2R_i} - 1}{a_{11} + a_{12}}$, $P_j = \frac{2^{2R_j} - 1}{a_{20}}$ and the cost is

$$
c_{\text{MISO}}(i, j) = \frac{2^{2R_i} - 1}{a_{11} + a_{12}} + \frac{2^{2R_j} - 1}{a_{20}}.
$$

H. Gaussian Interference Channel without Common Information (GI$_{wc}$)

In this channel the SBSs implement the Han-Kobayashi rate-splitting strategy [33]. Such cooperative scheme consists in splitting each message rate into two parts. One part of the message is *private* and is decoded only by the intended user while the other part of the message is *common* and is decoded by both receivers. The sum of the rates of such parts is equal to the rate of the original message. Decoding part of the interfering signal allows to expand the capacity region with respect to the GIN case. This scheme is considered when $\mathbf{d} = [f_i \quad f_j]$, with $i \neq j$, and requests are satisfied by the two SBSs ($x_{i1} = 1$, $x_{j2} = 1$, $x_{i2} = 0$, $x_{j1} = 0$) and ($x_{j1} = 1$, $x_{i2} = 1$, $x_{i1} = 0$, $x_{j2} = 0$). In the latter case, each SBS does not have the file requested from the corresponding user but has the file requested by the other user. Thus SBS$_1$ serves u_2 and SBS$_2$ serves u_1. (Fig. 3-a).

Solving analytically the power minimization subproblem for the case of GIwc is a challenging task due to the non convexity of the problem and to the intricacy of the conditions (see expressions (14)-(18) and Appendix II). In order to calculate the cost $c_{GIwc} = P_i + P_j$, we developed the algorithm calculate_Min_Ptot_HK, shown in Algorithm 1 in pseudocode.

Let us indicate with ΔP^{tot} the granularity at which the sum power is evaluated, with ΔP the granularity with which the sum power is divided between the SBSs P_i and P_j and with $\Delta\lambda$ the granularity in which power is split between the private and the common message.[4] At initialization the algorithm performs the search over a grid of points[5] spaced apart by the largest granularity specified in the input in order to speed-up convergence. The granularity is progressively decreased to the minimum value specified in the input to enhance the accuracy of the solution. Note that the minimum power c_{GIwc} is lower bounded by $c_{\perp}$ and upper bounded by $c_{GIN}(i, j)$.

The algorithm is much more efficient than exhaustive search over a grid with same minimum granularity. Specifically, if $\Delta P^{\text{tot}}_{\text{min}} \ll \Delta P^{\text{tot}}_{\text{max}}$, the gain in terms of number of operations is

[4] The three parameters are bounded as follows: $\Delta P^{\text{tot}} > 0$, $0 < \Delta P < 1$ and $0 < \Delta\lambda < 1$. The algorithm takes as input the minimum and maximum granularities of ΔP^{tot}, ΔP and $\Delta\lambda$ as well as the required file rates R_i and R_j and returns the minimum value of c_{GIwc} that satisfies the rate constraints for the two cached files.

[5] As an example, the first point of the grid for a given P^{tot} is (P_1, P_2, λ_1, λ_2) = ($\Delta P \cdot P^{\text{tot}}$, $(1 - \Delta P)P^{\text{tot}}$, $\Delta\lambda$, $\Delta\lambda$), the second point is (P_1, P_2, λ_1, λ_2) = ($2\Delta P \cdot P^{\text{tot}}$, $(1 - 2\Delta P)P^{\text{tot}}$, $\Delta\lambda$, $\Delta\lambda$) and so on.

Algorithm 1 c_{GIwc} = calculate_Min_Ptot_HK($\Delta P^{\text{tot}}_{\text{min}}$, $\Delta P^{\text{tot}}_{\text{max}}$, ΔP_{min}, ΔP_{max}, $\Delta\lambda_{\text{min}}$, $\Delta\lambda_{\text{max}}$, R_i, R_j)

```
$c_{GIwc} = -1$; $P_{\text{tot}} = \Delta P^{\text{tot}}_{\text{max}}$; $\Delta P_{\text{tot}} = \Delta P^{\text{tot}}_{\text{max}}$;
while $\Delta P^{\text{tot}} > \Delta P^{\text{tot}}_{\text{min}}$ do    ▷ Go on until the minimum granularity for $\Delta P^{\text{tot}}$ is reached
    flag = 0
    $\Delta P = \Delta P_{\text{max}}$; $\Delta\lambda = \Delta\lambda_{\text{max}}$    ▷ Reset $\Delta P$ and $\Delta\lambda$
    while flag == 0 do
        create $G(\Delta P, \Delta\lambda)$    ▷ Create search grid with current granularity excluding points previously checked
        for $g \in G(\Delta P, \Delta\lambda)$ do    ▷ Each point $g$ is defined by coordinates ($P_i$, $P_j$, $\lambda_1$, $\lambda_2$)
            if $g$ allows to achieve the pair $(R_i, R_j)$ then
                flag = 1
                $c_{GIwc} = P^{\text{tot}}$
                $\Delta P^{\text{tot}} = \Delta P^{\text{tot}}/2$
                $P^{\text{tot}} = P^{\text{tot}} - \Delta P^{\text{tot}}$    ▷ Check if $(R_i, R_j)$ are achievable with a smaller $P^{\text{tot}}$
                break for
            end if
        end for
        if (flag == 0) then
            if ($\Delta P > \Delta P_{\text{min}}$) then
                $\Delta P = \Delta P/2$
            else if ($\Delta\lambda > \Delta\lambda_{\text{min}}$) then
                $\Delta\lambda = \Delta\lambda/2$
            else if $c_{GIwc} \neq -1$ then    ▷ Check if so far ($R_i$, $R_j$) has been achieved
                $\Delta P^{\text{tot}} = \Delta P^{\text{tot}}/2$
                $P^{\text{tot}} = P^{\text{tot}} + \Delta P^{\text{tot}}$
            else
                $P^{\text{tot}} = P^{\text{tot}} + \Delta P^{\text{tot}}$    ▷ If ($R_i$, $R_j$) is not achievable with $P^{\text{tot}}$, the same holds for a smaller $P^{\text{tot}}$
                break inner while
            end if
        end if
    end while
end while
return $c_{GIwc}$
```

on the order of $(\Delta P^{\text{tot}}_{\text{max}}/\Delta P^{\text{tot}}_{\text{min}} - 1) \times 1/\Delta\lambda_{\text{min}} \times 1/\Delta P_{\text{min}}$. The algorithm is based on the fact that, if a given sum power P^{tot} does not allow to satisfy the rate constraints for any rate- and power-split, then no $\tilde{P}^{\text{tot}} < P^{\text{tot}}$ can satisfy such constraints. This can be derived from the conditions in Appendix II. The algorithm converges to the absolute minimum asymptotically as $\Delta P^{\text{tot}}_{\text{min}} \to 0$, $\Delta\lambda_{\text{min}} \to 0$, $\Delta P_{\text{min}} \to 0$. Note that from conditions in Appendix II follows that a solution to the minimization problem always exist. Therefore, there is no need to check for the existence of a solution as done for the GIN channel in the NCA, where the indicator function $\mathbb{1}_\alpha > 0$ was required. The proof is not reported here for a matter of space.

IV. Optimization Problem

A. Cooperative Approach

In this section we present the caching optimization problem for the CA. In the CA a user can be served by the MBS or any of the SBSs. The expected power cost for serving a request $Q_c(\mathbf{x}_1, \mathbf{x}_2)$ can be written as in (7). In the following, given a variable $h = \{1, 2\}$ we indicate with $\bar{h}$ its complementary value, i.e. if $h = 1$ then $\bar{h} = 2$ and vice-versa. Denoting the capacity of a point to point Gaussian channel with signal-to-noise ratio x as $C(x) = \frac{1}{2}\log_2(1 + x)$, the optimization problem for the CA can be formulated as follows

$$\underset{P^{(1)}, P^{(2)}, \mathbf{x}_1, \mathbf{x}_2}{\text{minimize}} \quad Q_{\text{c}}(\mathbf{x}_1, \mathbf{x}_2), \tag{8}$$

$$\text{subject to: } \sum_{f_i \in \mathcal{F}} x_{in} \leq M_n, \qquad n = \{1, 2\}, \tag{9}$$

$$\bar{x}_{i1} \cdot \bar{x}_{i2} \cdot \bar{x}_{j1} \cdot \bar{x}_{j2} \cdot R_- \leq C(a_{-0}, P_-) \tag{10}$$

$$\bar{x}_{i1} \cdot \bar{x}_{i2} \cdot \bar{x}_{j1} \cdot \bar{x}_{j2} \cdot R_+ \leq C\left(a_{+0} \cdot P_+\right), \tag{11}$$

$$x_{lw} \cdot \bar{x}_{l\bar{w}} \cdot \bar{x}_{\bar{l}w} \cdot \bar{x}_{\bar{l}\bar{w}} \cdot R_{\bar{l}} \leq C\left(a_{\bar{w}0} P^{(\bar{w})}_{\bar{l}}\right), \tag{12}$$

$$w = \{1, 2\}, l = \{i, j\},$$

$$x_{lw} \cdot \bar{x}_{l\bar{w}} \cdot \bar{x}_{\bar{l}w} \cdot \bar{x}_{\bar{l}\bar{w}} \cdot R_l \leq C\left(a_{wk} P^{(w)}_l\right), \tag{13}$$

$$\text{if } l = i \text{ then } k = 1 \text{ else } k = 2,$$

$$x_{ir} \cdot x_{j\bar{r}} \cdot \bar{x}_{i\bar{r}} \cdot \bar{x}_{jr} R_i \leq \rho_1(c_{r\bar{r}}, c_{\bar{r}r}, \tilde{P}^{(1)}_i, \tilde{P}^{(2)}_j), \tag{14}$$

$$r = \{1, 2\},$$

$$x_{ir} \cdot x_{j\bar{r}} \cdot \bar{x}_{i\bar{r}} \cdot \bar{x}_{jr} R_j \leq \rho_2(c_{r\bar{r}}, c_{\bar{r}r}, \tilde{P}^{(1)}_i, \tilde{P}^{(2)}_j) \tag{15}$$

$$x_{ir} \cdot x_{j\bar{r}} \cdot \bar{x}_{i\bar{r}} \cdot \bar{x}_{jr} (R_i + R_j) \leq \rho_{12}(c_{r\bar{r}}, c_{\bar{r}r}, \tilde{P}^{(1)}_i, \tilde{P}^{(2)}_j) \tag{16}$$

$$x_{ir} \cdot x_{j\bar{r}} \cdot \bar{x}_{i\bar{r}} \cdot \bar{x}_{jr} (2R_i + R_j) \leq \rho_{10}(c_{r\bar{r}}, c_{\bar{r}r}, \tilde{P}^{(1)}_i, \tilde{P}^{(2)}_j) \tag{17}$$

$$x_{ir} \cdot x_{j\bar{r}} \cdot \bar{x}_{i\bar{r}} \cdot \bar{x}_{jr} (R_i + 2R_j) \leq \rho_{20}(c_{r\bar{r}}, c_{\bar{r}r}, \tilde{P}^{(1)}_i, \tilde{P}^{(2)}_j) \tag{18}$$

$$x_{i1} \cdot x_{i2} \cdot x_{j1} \cdot x_{j2} \cdot R_i \leq C(\pi_m, \mathbf{H}_1, \mathbf{H}_2, \mathbf{S}_1, \mathbf{S}_2), \tag{19}$$

$$x_{i1} \cdot x_{i2} \cdot x_{j1} \cdot x_{j2} \cdot R_j \leq C(\pi_m, \mathbf{H}_1, \mathbf{H}_2, \mathbf{S}_1, \mathbf{S}_2) \tag{20}$$

$$x_{iz} \cdot x_{jz} \cdot \bar{x}_{i\bar{z}} \cdot \bar{x}_{j\bar{z}} \cdot R_- \leq C\left(a_{-z} P_- / (1 + a_{-z} P_+)\right), \tag{21}$$

$$z = \{1, 2\},$$

$$x_{iz} \cdot x_{jz} \cdot \bar{x}_{i\bar{z}} \cdot \bar{x}_{j\bar{z}} \cdot R_+ \leq C\left(a_{+z} P_+\right), \tag{22}$$

$$x_{i1} \cdot x_{i2} \cdot \bar{x}_{j1} \cdot \bar{x}_{j2} \cdot R_i \leq C\left((a_{11} + a_{12}) \cdot P^{(1)}_i\right), \tag{23}$$

$$x_{i1} \cdot x_{i2} \cdot \bar{x}_{j1} \cdot \bar{x}_{j2} \cdot R_j \leq C(a_{20} \cdot P^{(2)}_j), \tag{24}$$

$$x_{j1} \cdot x_{j2} \cdot \bar{x}_{i1} \cdot \bar{x}_{i2} \cdot R_i \leq C(a_{10} \cdot P^{(1)}_i), \tag{25}$$

$$x_{j1} \cdot x_{j2} \cdot \bar{x}_{i1} \cdot \bar{x}_{i2} \cdot R_j \leq C\left((a_{22} + a_{21}) \cdot P^{(2)}_j\right), \tag{26}$$

$$x_{ip} \cdot \bar{x}_{i\bar{p}} \cdot R_i \leq C(a_{pp} \cdot P^{(p)}_i), \qquad p = \{1, 2\}, \tag{27}$$

$$x_{ip} \cdot \bar{x}_{i\bar{p}} \cdot R_i \leq C(a_{\bar{p}p} \cdot P^{(\bar{p})}_j), \tag{28}$$

$$\bar{x}_{i1} \cdot \bar{x}_{i2} \cdot \bar{x}_{j1} \cdot \bar{x}_{j2} \cdot R_i \leq C\left(a_{10} \cdot P^{(1)}_i\right), \qquad i = j, \tag{29}$$

$$\bar{x}_{i1} \cdot \bar{x}_{i2} \cdot \bar{x}_{j1} \cdot \bar{x}_{j2} \cdot R_j \leq C\left(a_{20} \cdot P^{(2)}_j\right), \qquad i = j, \tag{30}$$

$$x_{i1} \cdot x_{i2} \cdot R_i \leq C\Bigg(\bigg(\sqrt{\tilde{P}^{(1)}_i} + \sqrt{c_{21} \cdot \tilde{P}^{(2)}_j}\bigg)^2\Bigg), i = j, \tag{31}$$

$$x_{i1} \cdot x_{i2} \cdot R_j \leq C\Bigg(\bigg(\sqrt{\tilde{P}^{(2)}_j} + \sqrt{c_{12} \cdot \tilde{P}^{(1)}_i}\bigg)^2\Bigg), i = j, \tag{32}$$

$$P^{(1)}_i \geq 0, P^{(2)}_j \geq 0, \tag{33}$$

$$P_- \geq 0, P_+ \geq 0, \tag{34}$$

$$x_{in} \in \{0, 1\}, x_{jn} \in \{0, 1\}, \tag{35}$$

$$f_i, f_j \in \mathcal{F}. \tag{36}$$

$$Q_{\mathrm{c}}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{f_i\in\mathcal{F}}\sum_{\substack{f_j\in\mathcal{F}\\ f_i\neq f_j}} q_iq_j\Big\{c_{\mathrm{GI_c}}(i,j)\Big[\sum_{t=i,j}\sum_{n=1,2}x_{tn}x_{\bar{t}n}x_{t\bar{n}}\bar{x}_{\bar{t}\bar{n}}\Big]+c^{\mathrm{SBS}}_{\mathrm{BC}}(i,j)\Big[\sum_{n=1,2}x_{in}x_{jn}\bar{x}_{i\bar{n}}\bar{x}_{j\bar{n}}\Big]+c^{\mathrm{MBS}}_{\mathrm{BC}}(i,j)\Big[\bar{x}_{i1}\bar{x}_{i2}\bar{x}_{j1}\bar{x}_{j2}\Big]+$$
$$c_{\perp}(i,j)\Big[\sum_{t=i,j}\sum_{n=1,2}x_{tn}\bar{x}_{\bar{t}n}\bar{x}_{t\bar{n}}\bar{x}_{\bar{t}\bar{n}}\Big]+c_{\mathrm{GI_{wc}}}(i,j)\Big[\sum_{n=1,2}x_{in}x_{j\bar{n}}\bar{x}_{i\bar{n}}\bar{x}_{jn}\Big]+c_{\ \mathrm{MISO}}(i,j)\Big[\sum_{t=i,j}x_{t1}x_{t2}\bar{x}_{\bar{t}1}\bar{x}_{\bar{t}2}\Big]+$$
$$c_{\ \mathrm{MIMO}}(i,j)\Big[x_{i1}x_{i2}x_{j1}x_{j2}\Big]\Big\}+\sum_{\substack{f_j\in\mathcal{F}\\ f_i=f_j}}q_i^2\Big\{c^{\mathrm{MBS}}_{\mathrm{MC}}(i,j)[\bar{x}_{i1}\bar{x}_{i2}]+c^{\mathrm{SBS}}_{\mathrm{MC}}[x_{i1}\bar{x}_{i2}+x_{i2}\bar{x}_{i1}]+c_{\mathrm{GI_c}}(i,j)[x_{i1}x_{i2}]\Big\}. \qquad (7)$$

Inequality (9) ensures that the total amount of data stored in a cache does not exceed its size, (10)-(11) denote the capacity region of the Gaussian broadcast channel of the MBS, while (12)-(13) represent the conditions for the case of orthogonal channels, (14)-(18) are the Han-Kobayashi conditions and define the best known achievable rate of the GI$_{wc}$ (the explicit expressions for such conditions are given in the Appendix II), (19)-(20) define the capacity region for the MIMO broadcast channel, (21)-(22) denote the capacity region for the broadcast channel of SBS$_n$, (23)-(24) and (25)-(26) define the capacity region of the orthogonal channel when the SBSs implement a MISO channel to serve u_1 (u_2) while u_2 (u_1) is served by the MBS, (27)-(28) denote the capacity region of the multicast channel of SBS$_n$, (31)-(32) define an achievable rate region of the GIc when both SBSs have to transmit the same file, (29)-(30) define the capacity region of the multicast channel of the MBS, (33) is imposed to guarantee the non negativity of the transmit power, while (35) represents the caching choice and accounts for the discrete nature of the optimization variable. Note that the power variables in constraints (14)-(18) and (31)-(32) refer to the normalized channel model ($\tilde{P}_i^{(1)}$, $\tilde{P}_j^{(2)}$). Since we are interested in evaluating the physical power consumption, the power values in the calculation of the cost have to be scaled as $P_i^{(1)}=\tilde{P}_i^{(1)}/a_{11}$ and $P_j^{(2)}=\tilde{P}_j^{(2)}/a_{22}$. As a side remark, note that there is a significant difference between the MISO and the GI$_c$ channel, despite the fact that in both cases the same file is transmitted by both SBSs. In fact, in the GI$_c$ channel SBSs transmit to both users and powers are adjusted such that both users achieve the QoS required while in the MISO channel the SBSs transmit only to one user.

B. Non-Cooperative Approach

In the NCA each user is served either by the SBS to which it is associated or by the MBS. This approach considers for all transmissions the interference of the other SBS as noise. The MBS transmits to the user when the file requested by u_n is not present in the cache of SBS$_n$ or when the channel conditions do not allow to have a reliable transmission. The expected power cost in the NCA $Q_{nc}(\mathbf{x}_1,\mathbf{x}_2)$ can be written as in (37). The optimization problem to be solved is the following

$$\underset{P^{(1)},P^{(2)},\mathbf{x}_1,\mathbf{x}_2}{\text{minimize}}\ Q_{\mathrm{nc}}(\mathbf{x}_1,\mathbf{x}_2), \qquad (38)$$

subject to: (9)-(11), (29)-(36)

$$x_{kg}\cdot\bar{x}_{\bar{k}g}\cdot\bar{x}_{\bar{k}\bar{g}}\cdot\bar{x}_{k\bar{g}}\cdot R_+\leq C\left(a_{+0}P_+\right), \qquad (39)$$

$$g=\{1,2\}\text{ if }g=1\text{ then }k=j\text{ else }k=i$$

$$x_{kg}\cdot\bar{x}_{\bar{k}g}\cdot\bar{x}_{\bar{k}\bar{g}}\cdot\bar{x}_{k\bar{g}}\cdot R_-\leq C\left(a_{-0}P_-/(1+a_{-0}P_+)\right), \qquad (40)$$

$$x_{i2}\cdot x_{j1}\cdot\bar{x}_{i1}\cdot\bar{x}_{j2}\cdot R_-\leq C\left(a_{-0}P_-/(1+a_{-0}P_+)\right), \qquad (41)$$

$$x_{i2}\cdot x_{j1}\cdot\cdot\bar{x}_{i1}\cdot\bar{x}_{j2}\cdot R_+\leq C\left(a_{+0}P_+\right), \qquad (42)$$

$$x_{lw}\cdot\bar{x}_{l\bar{w}}\cdot\bar{x}_{\bar{l}w}\cdot\bar{x}_{\bar{l}\bar{w}}\cdot R_{\bar{l}}\leq C\left(a_{\bar{w}0}\cdot P_{\bar{l}}^{(\bar{w})}\right), \qquad (43)$$

$$w=\{1,2\}\text{ if }w=1\text{ then }l=i\text{ else }l=j$$

$$x_{lw}\cdot\bar{x}_{l\bar{w}}\cdot\bar{x}_{\bar{l}w}\cdot\bar{x}_{\bar{l}\bar{w}}\cdot R_l\leq C\left(a_{ww}\cdot P_l^{(w)}\right), \qquad (44)$$

$$x_{tm}\cdot x_{\bar{t}m}\cdot\bar{x}_{t\bar{m}}\cdot\bar{x}_{\bar{t}\bar{m}}\cdot R_t\leq C(a_{tt}\cdot P_t^{(m)}), \qquad (45)$$

$$t=\{i,j\},m=\{1,2\},$$

$$x_{tm}\cdot x_{\bar{t}m}\cdot\cdot\bar{x}_{t\bar{m}}\cdot\bar{x}_{\bar{t}\bar{m}}\cdot R_{\bar{t}}\leq C(a_{\bar{t}0}\cdot P_{\bar{t}}^{(\bar{m})}), \qquad (46)$$

$$x_{tm}\cdot x_{t\bar{m}}\cdot\bar{x}_{\bar{t}m}\cdot\bar{x}_{\bar{t}\bar{m}}\cdot R_t\leq C(a_{mm}\cdot P_t^{(m)}), \qquad (47)$$

$$x_{tm}\cdot x_{t\bar{m}}\cdot\cdot\bar{x}_{\bar{t}m}\cdot\bar{x}_{\bar{t}\bar{m}}\cdot R_{\bar{t}}\leq C(a_{\bar{m}0}\cdot P_{\bar{t}}^{(\bar{m})}), \qquad (48)$$

$$\mathbb{1}_\alpha\cdot x_{i1}\cdot x_{j2}\cdot\bar{x}_{i2}\cdot\bar{x}_{j1}\cdot R_i\leq C\left(a_{11}P_i^{(1)}/(1+a_{12}P_j^{(2)})\right), \qquad (49)$$

$$\mathbb{1}_\alpha\cdot x_{i1}\cdot x_{j2}\cdot\bar{x}_{i2}\cdot\bar{x}_{j1}\cdot R_j\leq C\left(a_{22}P_j^{(2)}/(1+a_{21}P_i^{(1)})\right), \qquad (50)$$

$$\mathbb{1}_\alpha\cdot x_{i1}\cdot x_{i2}\cdot x_{j1}\cdot x_{j2}\cdot R_i\leq C\left(a_{11}P_i^{(1)}/(1+a_{12}P_j^{(2)})\right), \qquad (51)$$

$$\mathbb{1}_\alpha\cdot x_{i1}\cdot x_{i2}\cdot x_{j1}\cdot x_{j2}\cdot R_j\leq C\left(a_{22}P_j^{(2)}/(1+a_{21}P_i^{(1)})\right), \qquad (52)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{i2}\cdot x_{j1}\cdot x_{j2}\cdot R_i\leq C\left(a_{10}P_i^{(2)}\right),\quad i=j, \qquad (53)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{i2}\cdot x_{j1}\cdot x_{j2}\cdot R_j\leq C\left(a_{20}P_j^{(2)}\right),\quad i=j, \qquad (54)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{j2}\cdot\bar{x}_{i2}\cdot\bar{x}_{j1}\cdot R_i\leq C\left(a_{10}P_i^{(1)}\right),\quad i=j, \qquad (55)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{j2}\cdot\bar{x}_{i2}\cdot\bar{x}_{j1}\cdot R_j\leq C\left(a_{20}P_j^{(2)}\right),\quad i=j, \qquad (56)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{i2}\cdot x_{j1}\cdot x_{j2}\cdot R_-\leq C\left(a_{-0}P_-/(1+a_{-0}P_+)\right) \qquad (57)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{i2}\cdot x_{j1}\cdot x_{j2}\cdot R_+\leq C\left(a_{+0}P_+\right) \qquad (58)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{j2}\cdot\bar{x}_{i2}\cdot\bar{x}_{j1}\cdot R_-\leq C\left(a_{-0}P_-/(1+a_{-0}P_+)\right), \qquad (59)$$

$$\bar{\mathbb{1}}_\alpha\cdot x_{i1}\cdot x_{j2}\cdot\bar{x}_{i2}\cdot\bar{x}_{j1}\cdot R_+\leq C\left(a_{+0}P_+\right). \qquad (60)$$

Conditions (9)-(11) and (29)-(36) are in common with the CA. Conditions (39)-(42) denote the capacity region of the Gaussian broadcast channel, (43)-(48) define the capacity region of the orthogonal channels, while conditions (49)-(52) represent the constraints of the GIN channel. Constraints (53)-

$$Q_{\text{nc}}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{f_i\in\mathcal{F}}\sum_{\substack{f_j\in\mathcal{F}\\ f_i\neq f_j}} q_iq_j\Big\{c_{\text{GIN}}(i,j)\big[\mathbb{1}_\alpha x_{i1}x_{j2}\big]+c_{\perp}(i,j)\big[x_{i1}\bar{x}_{j2}+\bar{x}_{i1}x_{j2}\big]+c^{\text{MBS}}_{\text{BC}}(i,j)\big[\bar{\mathbb{1}}_\alpha x_{i1}x_{j2}$$
$$+\bar{x}_{i1}\bar{x}_{j2}\big]\Big\}+\sum_{\substack{f_j\in\mathcal{F}\\ f_i=f_j}} q_i^2\Big\{c_{\text{GIN}}(i,j)\big[\mathbb{1}_\alpha x_{i1}x_{i2}\big]+c^{\text{MBS}}_{\text{MC}}(i,j)[\bar{x}_{i1}\bar{x}_{i2}+\bar{\mathbb{1}}_\alpha x_{i1}x_{i2}\big]+c_{\perp}(i,j)\big[\bar{x}_{i2}x_{i1}+\bar{x}_{i1}x_{i2}\big]\Big\}. \tag{37}$$

Algorithm 2 $[\mathbf{x}_1 \quad \mathbf{x}_2]$ = low_complexity_allocation($\mathbf{R},\mathbf{q}$)

```
$\mathbf{x}_1 = \mathbf{0}$ and $\mathbf{x}_2 = \mathbf{0}$    ▷ $\mathbf{0}$ is an $N$-dimensional all-zero vector
$W = \{w_1, \dots, w_k, \dots, w_N\}$    ▷ $w_k = 2^{2R_k} \cdot q_k$ is the $k^{th}$ element of the set
for $l = 1, \dots, M$ do
    $w_m$ = $\arg\max_l\{w_l\}$
    $x_{m1} = 1$
    $W = W \setminus \{w_m\}$
    $w_m$ = $\arg\max_l\{w_l\}$
    $x_{m2} = 1$
    $W = W \setminus \{w_m\}$
end for
return $[\mathbf{x}_1 \quad \mathbf{x}_2]$
```

(56) and (57)-(60) denote the capacity region of the Gaussian multicast channel considered when the SBSs are not capable to deliver the files at the requested rates.

C. Implementation

The optimal cache allocation can be derived by minimizing the cost function in (7) independently along the power dimension and the cache allocation dimension. First, the minimum average power for a given cache allocation is calculated. Let us recall that N is the number of files. For each of the N^2 possible requests the power has to be minimized. Each request, according to whether the file is present or not in a cache, requires to solve one of the power minimization subproblems presented in Section III. The optimal cache allocation is obtained by exhaustive search within the set of possible allocations. The complexity for minimizing the average power for a given cache allocation is $O(N^2)$ while the cost for finding the optimal cache allocation is $O(N!)$.

Although such algorithm is useful to evaluate the performance limit of the considered setup, its complexity makes it unsuited for a practical application when the number of files is large. Therefore, we propose in the following a suboptimal algorithm with reduced complexity.

Let us consider equation (7). The cost of each channel depends in a non-linear way on the channel coefficients a_{nm} and on the rate of both requested files R_i, R_j. It also depends linearly on the probabilities of the files requests q_i, q_j. Let us now consider genie-aided receivers having access to the message destined to the other user. In this case, the interference can be entirely removed and each user sees an interference-free channel. In such channel, the cost of transmitting f_i is equal to its popularity ranking q_i times a term proportional to 2^{R_i}. Since we consider a setup in which the SBS-user channel is on average better than that between MBS-user one, having files with higher cost transmitted by SBSs rather than the MBS minimizes the overall power expenditure.

Let us define the utility function $w_k(R_k, q_j) = 2^{2R_k} \cdot q_k$ associated to f_k. The proposed algorithm, described in Algorithm 2, caches the files with highest $w(R,q)$. The complexity of the algorithm grows linearly with N. In section V, we compare this algorithm with the optimal one and with two benchmarks.

V. Numerical Results

In this section we compare CA and NCA in two different scenarios, namely: (i) static scenario and (ii) mobile transmitters scenario. In the static scenario no fading is present, i.e., all channel coefficients are fixed. In the mobile transmitters scenario the transmitter-user links experience fading while the master node-user links are AWGN channels. This models a heterogeneous satellite network in which two LEO satellites act as transmitters while a GEO satellite acts as master node. Due to the LEOs orbital motion and the presence of obstacles such as tall buildings, fading is present in the LEO-user channels[6] while the link to the GEO, assumed to be line of sight, does not change significantly over time, which is the case in clear-sky conditions. This setup also models a hybrid network with a GEO satellite acting as master node while hovering UAVs play the role of transmitters. Also in this case, the UAV-user links are affected by fading due to UAVs motion [34].

Table I: Rate and popularity rankings for each file in the direct probability and inverse probability setups.

$\mathcal{F}$	R_i [bits/sec/Hz]	q_i *direct*	q_i *inverse*
f_1	1.2	0.45	0.05
f_2	1.0	0.20	0.15
f_3	0.5	0.15	0.15
f_4	0.4	0.15	0.20
f_5	0.2	0.05	0.45

We model the channel coefficients in the links affected by fading as Lognormal random variables, i.e., given a channel coefficient a, we have $\log(a) \sim \mathcal{N}\left(\mu, \sigma^2\right)$, $\log(.)$ being the Napierian logarithm. The parameters μ and σ have been chosen so that in the three scenarios we have $\mathbb{E}\{a_{11}\} = \mathbb{E}\{a_{22}\} = 1$ and $\mathbb{E}\{a_{10}\} = \mathbb{E}\{a_{20}\} = 0.01$, $\mathbb{E}\{.\}$ being the expectation operator. We plot our results against the mean value of the interfering link coefficients as presented in the GIc equivalent model (c_{12}, c_{21}), where direct channel coefficients have unit mean. The parameters for such coefficients are chosen so that

[6] In practice the equivalent of an SBS (MBS) would be a LEO (GEO) satellite together with the respective ground station, where the cache would be physically located. This is because on-board storage in satellites is not common practice.

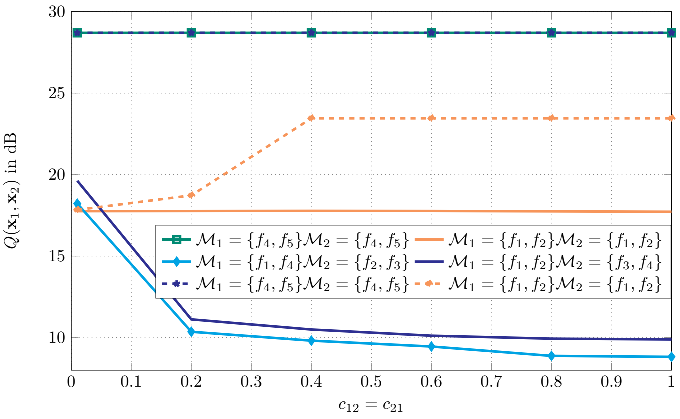


Figure 4: Minimum average power cost in dB in the static case plotted against the interference coefficient with direct file probabilities. The result for the CA is represented with solid curves while dashed curves are used for the the NCA. The optimization is done for each average interference coefficient. In the simulations we set $a_{10} = a_{20} = 0.01$ and $a_{11} = a_{22} = 1$.

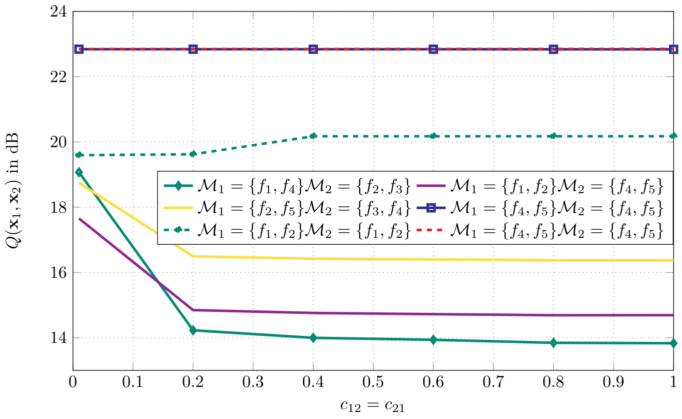


Figure 5: Minimum average power cost in dB in the static scenario plotted against the interference coefficient with inverse probabilities. The optimization is done for each average interference coefficient. In the simulations we set $a_{10} = a_{20} = 0.01$ and $a_{11} = a_{22} = 1$.

$\mathbb{E}\{c_{12}\} = \mathbb{E}\{c_{21}\}$ take values in $[0, 1]$. This allows to compare different scenarios against a single parameter. In order to do a fair comparison, this is done for both the non-cooperative as well as the cooperative scheme.

We consider the system depicted in Fig. 2 for the NCA and Fig. 3 for the CA where each SBS has a memory of size $M = 2$ and the library has cardinality $N = 5$. Two relevant setups are considered. In the first one, files with higher rate have a higher popularity ranking while in the second one files with higher rate have a lower popularity ranking. We refer to these setups as *direct probability* and *inverse probability*, respectively. The corresponding values for R_i and q_i are shown in Table I. We plot for each scenario the minimum power cost $Q(\mathbf{x}_1, \mathbf{x}_2)$ of the most significant cache allocations in the CA (solid curves) and in the NCA (dashed curves) versus the mean value of the interference coefficients, assumed to be the same for both users[7]. Note that, by the symmetry of the problem, the results do not change if the content of the cache of SBS$_1$ and that of SBS$_2$ are switched, i.e., the performance of the solution $\mathcal{M}_1 = \{f_i, f_j\}$, $\mathcal{M}_2 = \{f_k, f_l\}$ and $\mathcal{M}_1 = \{f_k, f_l\}$, $\mathcal{M}_2 = \{f_i, f_j\}$ coincide.

In Fig. 4 the most significant cache allocations for the static scenario with direct probabilities are plotted. When no cooperation is in place the optimal caching solution consists in each SBS storing the files with highest probabilities/transmission rate ($\mathcal{M}_1 = \mathcal{M}_2 = \{f_1, f_2\}$) which is indicated with the orange dashed curve. In this case, when mutual interference is low the MBS only transmits files with low probability and low transmission rate. On the contrary, when the level of interference increases the SBSs are not able to serve their own users even if they have the requested files and transmissions are delegated to the MBS. Thus, the usage of the MBS becomes more frequent and more power is consumed due to the high rate of the files and lower channel gain. This explains the step of roughly 4 dB from $c_{12} = c_{21} = 0.2$ to $c_{12} = c_{21} = 0.4$. For $c_{12} = c_{21}$ larger than or equal to 0.4 the curve flattens out because, from that point on, the mutual interference is so strong that the required rates cannot be achieved by the SBSs and the MBS takes over all transmissions. In the CA the best performance is obtained when one of the SBS stores the files with highest probability/transmission rate and the two caches store different files. The result is shown in the light blue curve with diamonds. A slight increase in power (from 0.5 dB up to 1 dB) is observed when the caches store different files but the two most probable files are in the same cache ($\mathcal{M}_1 = \{f_1, f_2\}$, $\mathcal{M}_2 = \{f_3, f_4\}$). In both CA and NCA approaches, the most power consuming solution coincides and corresponds to each SBS storing the two files with lower probabilities/rates ($\mathcal{M}_1 = \mathcal{M}_2 = \{f_4, f_5\}$). The flatness of the curves in both the CA and the NCA are due to the fact that the power consumed by the MBS has a strong weight in the overall mean, since transmissions at high data rate from MBS are the most probable.

Let us now consider the low interference regime. In the CA, all the caching schemes at the lowest interference level consume at least 5 dB more with respect to the other levels. This is because, when a single SBS serves both users, it applies either the broadcast or the multicast channel and such channels are penalized when one of the two channel coefficients are relatively small. Later on in this section we show that, despite the higher power consumed in some cases, cooperation has an advantage in terms of MBS resources sparing.

In Fig. 5 the power cost versus interference level in the static scenario and for inverse probabilities (i.e. files with higher R_i have lower q_i) is plotted. The plot shows that for both approaches the most power-efficient cache allocation (solid and dashed green curves) is not the one including the file with the highest popularity but rather the one with the highest rate. The reason for this is that the weight of MBS transmissions on the overall power budget is higher with respect to the SBSs' due to the lower channel gains. For the considered rates and popularity ranking it is better to have the MBS transmitting low rate messages more often rather than transmitting higher rate messages with lower popularity. In the cooperative case, if the file with highest rate is not cached there is a loss larger

[7] Note that, while in the static case this is equivalent to having a symmetric channel matrix, this is not necessarily the case in the mobile transmitter case due to the independence of fading.

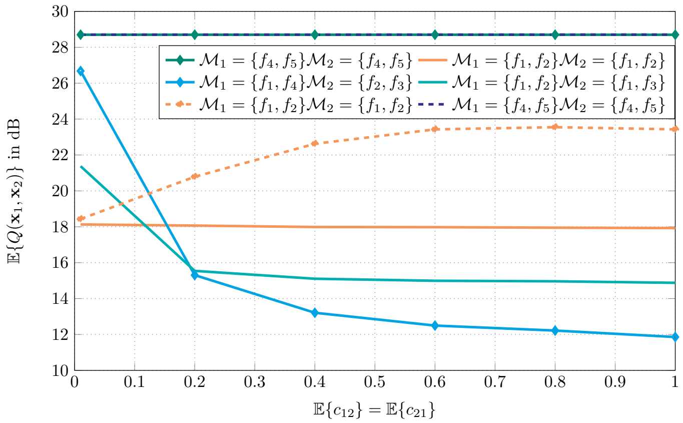


Figure 6: Average power cost in dB versus interference link coefficients for different memory allocations in the mobile transmitters setup in a fading environment with direct probabilities. In the simulations we set $a_{10} = a_{20} = 0.01$ and $\mathbb{E}\{a_{11}\} = \mathbb{E}\{a_{22}\} = 1$.

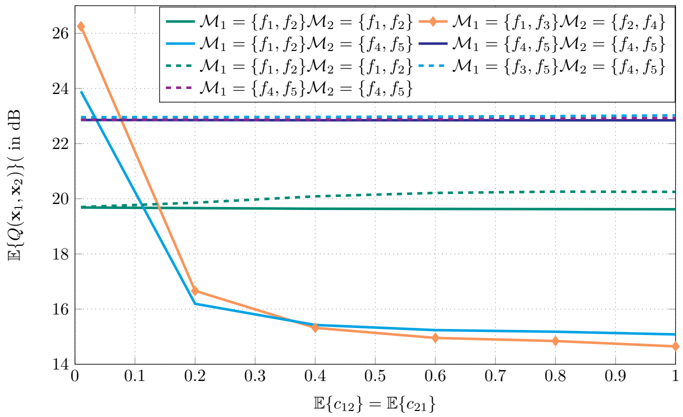


Figure 7: Avarege power cost in dB versus interference link coefficients for different memory allocations for the mobile transmitters setup in a fading environment with inverse probabilities. In the simulations we set $a_{10} = a_{20} = 0.01$ and $\mathbb{E}\{a_{11}\} = \mathbb{E}\{a_{22}\} = 1$.

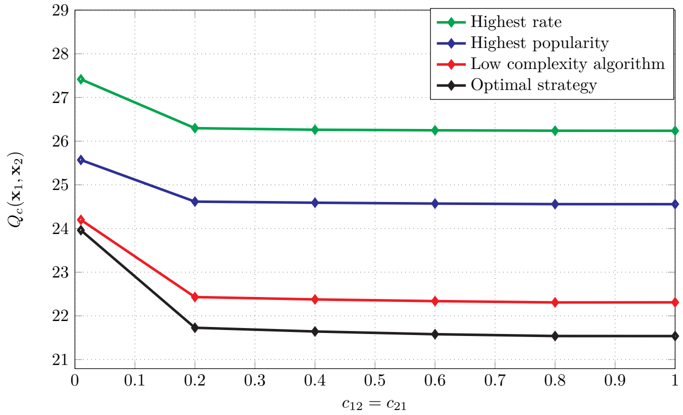


Figure 8: Minimum average power cost in dB in the static scenario plotted against the interference coefficient. For $N = 10$ and $M = 2$. R_k and q_k for each file are described on Table II. The optimization is done for each average interference coefficient. In the simulations we set $a_{10} = a_{20} = 0.01$ and $a_{11} = a_{22} = 1$.

than or equal to 2.5 dB with respect to the best curve. This holds also for the allocation including the most popular file (yellow curve), which confirms that it is better to store the file with highest rate rather than the one with highest popularity. As a matter of facts, the worst performance in the both the CA and the NCA case is obtained when each SBS stores the two files with largest probabilities/lowest data rate.

We examine now the mobile transmitters scenario (fading on the the SBS-user link, no fading on the the MBS-user link). Note that, due to the difference between two Lognormal random variables at the denominator of equation (3), the average transmission power in the GIN channel is not finite [35]. Comparing the CA and the NCA in such conditions always leads to an infinite power gain of the former with respect to the latter as long as the probability that the two SBSs transmit together is non-zero. In order to get some interesting insight in the presence of fading, in the simulations of the mobile transmitters scenario we introduced a constraint on the maximum transmit power. Such constraint is chosen so that the probability that the SBSs cannot transmit due to it is around 10^{-5} in the worst case scenario (i.e., GIN with maximum rates and $\mathbb{E}\{c_{12}\} = \mathbb{E}\{c_{21}\} = 0.4$[8]).

In Fig. 6 the minimum cost $\mathbb{E}\{Q(\mathbf{x}_1, \mathbf{x}_2)\}$ for different average interference levels in the mobile transmitters case for direct probabilities is plotted. Let us first consider the CA. Similarly to the static scenario, when interference is not close to zero the best performance is obtained when each SBS stores different files and the most probable ones are in $\mathcal{M}_1 \cup \mathcal{M}_2$. The power consumption is constant (18 dB) for all levels of interference when the caches of the SBSs are equal and each SBS stores the files with highest rate/popularity. Note that such caching solution is the best for very low levels of interference. In fact, when both SBSs cache the same files, broadcast and multicast transmissions, which are the most expensive ones, are not implemented. In the NCA the caching scheme which consumes less power is the one in which each SBS stores the most probable files.

Moving from the static to the mobile transmitters scenario the minimum required power increases. This is not surprising since the number of links affected by fading increases. However, we note that in the direct file probability case, the optimal cache allocation in both the scenarios remains the same. A similar observation is also true in the inverse probability case. We performed other simulations for a third setup in which all links are affected by fading, which models a situation where the users are mobile, although we do not report here the plots for lack of space. Interestingly, also in this third scenario the optimal cache allocation is the same as in the other two. This may suggest that imperfect knowledge of channel statistics may be tolerated at least up to a certain extent. This aspect can have important practical implications but due to the lack of space we do not delve further into this topic and leave it for future work.

Let us compare the performance of the optimal algorithm with that of the low-complexity algorithm proposed in Section

[8] Note that $\mathbb{E}\{c_{12}\} = \mathbb{E}\{c_{21}\} = 1$ would lead to a higher outage probability for the SBSs in the NCA, but the outage in such case is due almost exclusively to the lack of a feasible solution to equation (3) rather than to the constraint on the maximum power.

IV-C. Two other benchmarks are also considered, namely, one in which the files with highest popularity ranking are cached and one in which the files with the highest rate requirements are cached. The probabilities and rate of each file are given in Table II.

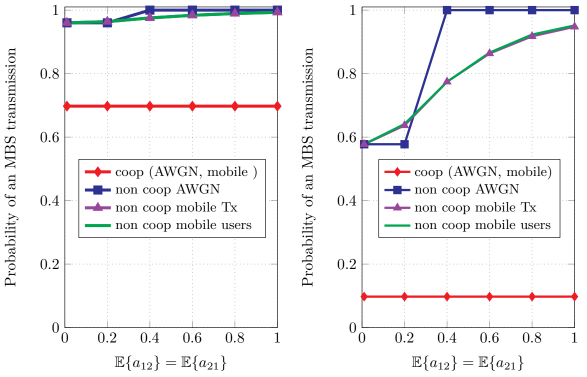
We consider $N = 10$ files and a memory size for each SBS of $M = 2$. The results are shown in Fig. 8 for the AWGN scenario. The black curve represents the optimal caching strategy. The result of the proposed low-complexity algorithm is plotted in red. As can be seen from the plot the suboptimal solution outperforms the benchmark strategies that store files with either the highest rates or the highest popularity rankings. We recall that this is achieved with a complexity that grows linearly with N. As a last remark, note that the low-complexity algorithm and the two benchmark allocate the cache independently of the average interference, while the optimal algorithm takes it into account. Although the low-complexity algorithm consumes more power than the optimal one, it has the advantage of not requiring previous knowledge of the average interference level.

So far we saw that in some scenarios and for low levels of interference the power gain deriving from cooperation is limited. This is because we are trying to spare the usage of the MBS as much as possible. Since from the plots showed so far the MBS utilization cannot be appreciated, in Fig. 9a and Fig. 9b we plot, for a given caching allocation and for different levels of interference, the probability that a transmission is served by the MBS for inverse and direct probabilities, respectively. For the CA we consider the cache allocation: $\mathcal{M}_1 = \{f_1, f_3\}, \mathcal{M}_2 = \{f_2, f_4\}$ while for the NCA: $\mathcal{M}_1 = \mathcal{M}_2 = \{f_1, f_2\}$.

In the the CA (red curve) the probability of having a transmission from the MBS is independent from the level of interference and in the static and mobile transmitters setups the probability coincides. This is because in the cooperative case the usage of the MBS only depends on the cache allocation of the SBSs while in the NCA it depends also on the mutual interference level. In fact, in the non-cooperative case the probability of MBS transmissions grows with the level of interference. This is due to the fact that, in the GIN channel, transmissions are forwarded to the MBS when a feasible solution does not exists and the probability of this happening increases with the interference level. In the inverse file probabilities case, using cooperation reduces the MBS usage of at least 25% with respect to the non-cooperative case. In Fig. 9b we see how the usage of the MBS is greatly reduced with cooperation if files have direct probabilities. In this case, since the files stored in the SBSs' caches are those with higher popularity ranking (and rates), transmissions from the MBS occur with very low probability. In fact, in the CA MBS transmissions occur only in 10% of the cases when the best caching allocation is considered, with a gain of up to 90% with respect to the best allocation of the non-cooperative case. However, as mentioned previously in the present section, such gain implies, in some case, a higher cost in terms of power.

Table II: Rate and popularity ranking.

$\mathcal{F}$	q_i	R_i [bits/sec/Hz]	$\mathcal{F}$	q_i	R_i [bits/sec/Hz]
f_1	0.5911	1.20	f_6	0.0235	1.00
f_2	0.1697	0.20	f_7	0.0178	1.60
f_3	0.0818	1.40	f_8	0.0140	0.60
f_4	0.0487	0.80	f_9	0.0113	2.00
f_5	0.0326	1.80	f_{10}	0.0094	0.40



(a) Inverse

(b) Direct

Figure 9: Probability of having an MBS transmission at different levels of interference for a given allocation memory. $\mathcal{M}_1 = \{f_1, f_3\}, \mathcal{M}_2 = \{f_2, f_4\}$ for the CA and $\mathcal{M}_1 = \{f_1, f_2\}, \mathcal{M}_2 = \{f_1, f_2\}$ for the NCA.

VI. Conclusions

We studied the optimal caching strategy for minimizing the power and limiting MBS use in heterogeneous networks with mutual interference under different per-file rate constraints and one-shot delivery phase. We formulated and analysed the caching problem both in case of no cooperation between SBSs as well as in the cooperative case. One among a set of different cooperative strategies is chosen by the transmitters depending on the cache allocation and file requests. We derived for each of the considered transmission strategies the minimum power needed for satisfying the rate constraints. An optimal cache allocation algorithm as well as a low-complexity suboptimal algorithm were proposed. Our results show that memorizing the most popular files is not always the most convenient strategy when rate requirements have to be respected. Furthermore, we showed that in the CA the probability of a transmission by the MBS only depends on the cache allocation and file popularity ranking while in the NCA such probability also depends on the mutual interference. Thanks to cooperation up to 90% of the MBS resources can be saved with respect to the non-cooperative case. This implies a trade-off in terms of SBSs power.

Appendix I

We can rewrite the minimization problem of the the GI$_c$ as

$$\begin{aligned}
\underset{P_i, P_j}{\text{minimize}} \quad & \frac{P_i^{(1)}}{a_{11}} + \frac{P_j^{(2)}}{a_{22}}, \\
\text{subject to} \quad & -\sqrt{2^{2R_i} - 1} + \sqrt{P_i^{(1)}} + \sqrt{c_{12} \cdot P_j^{(2)}} \geq 0, \\
& -\sqrt{2^{2R_i} - 1} + \sqrt{P_j^{(2)}} + \sqrt{c_{21} \cdot P_i^{(1)}} \geq 0, \\
& P_i^{(1)}, P_j^{(2)} \geq 0.
\end{aligned}$$

Let us indicate with $\mathbf{P}^* = [P_i^*, P_j^*]$ the solution and let L be the Lagrange function associated with the optimization problem

$$\begin{aligned}L(\mathbf{P},\lambda) = &-\frac{P_i^{(1)}}{a_{11}} - \frac{P_j^{(2)}}{a_{22}} + \\ &\lambda_1\left(-\sqrt{2^{2R_i}-1} + \sqrt{P_i^{(1)}} + \sqrt{c_{12}P_j^{(2)}}\right) + \\ &\lambda_2\left(-\sqrt{2^{2R_i}-1} + \sqrt{P_j^{(2)}} + \sqrt{c_{21}P_i^{(1)}}\right) + \\ &\lambda_3 \cdot P_i^{(1)} + \lambda_4 \cdot P_j^{(2)}. \end{aligned} \tag{61}$$

Let us consider the points where functions $g_1, ..., g_m$ are differentiable. Applying the KKT conditions [25] we have

$$\frac{\delta L(\mathbf{P},\lambda)}{\delta P_i} = -\frac{1}{a_{11}} + \frac{\lambda_1^*}{2}\frac{1}{\sqrt{P_i^*}} + \frac{\lambda_2^*}{2}\frac{\sqrt{c_{21}}}{\sqrt{P_i^*}} + \lambda_3 = 0, \tag{62}$$

$$\frac{\delta L(\mathbf{P},\lambda)}{\delta P_j} = -\frac{1}{a_{22}} + \frac{\lambda_1^*}{2}\frac{\sqrt{c_{12}}}{\sqrt{P_i^*}} + \frac{\lambda_2^*}{2}\frac{1}{\sqrt{P_j^*}} + \lambda_4 = 0, \tag{63}$$

$$g_1(P_i^*, P_j^*) = -\sqrt{2^{2R_i}-1} + \sqrt{P_j^*} + \sqrt{c_{12} \cdot P_i^*} \geq 0, \tag{64}$$

$$g_2(P_i^*, P_j^*) = -\sqrt{2^{2R_i}-1} + \sqrt{P_i^*} + \sqrt{c_{21} \cdot P_j^*} \geq 0, \tag{65}$$

$$g_3(P_i^*, P_j^*) = P_i^* \geq 0, \tag{66}$$

$$g_4(P_i^*, P_j^*) = P_j^* \geq 0, \tag{67}$$

$$\lambda_1^* \geq 0, \quad \lambda_2^* \geq 0, \quad \lambda_3^* \geq 0, \quad \lambda_4^* \geq 0, \tag{68}$$

$$\lambda_1^* g_1(P_i^*, P_j^*) = \lambda_1^*\left(-\sqrt{2^{2R_i}-1} + \sqrt{P_j^*} + \sqrt{c_{12}P_i^*}\right) = 0, \tag{69}$$

$$\lambda_2^* g_2(P_i^*, P_j^*) = \lambda_2^*\left(-\sqrt{2^{2R_i}-1} + \sqrt{P_i^*} + \sqrt{c_{21}P_j^*}\right) = 0, \tag{70}$$

$$\lambda_3^* g_3(P_i^*, P_j^*) = \lambda_3^* \cdot P_i^* = 0, \tag{71}$$

$$\lambda_4^* g_4(P_i^*, P_j^*) = \lambda_4^* \cdot P_j^* = 0. \tag{72}$$

From (62)-(63) and (71)-(72) follows that $\lambda_3 = 0$ and $\lambda_4 = 0$ then from (62) and (63) we find

$$P_i^* = \frac{1}{4}(\lambda_1^* \cdot a_{11} + \lambda_2^* \cdot a_{11} \cdot \sqrt{c_{21}})^2, \tag{73}$$

$$P_j^* = \frac{1}{4}(\lambda_1^* \cdot a_{22} \cdot \sqrt{c_{12}} + \lambda_2^* \cdot a_{22})^2. \tag{74}$$

By plugging in (69)-(70) the results of (73)-(74), we get the following system of equations

$$\begin{aligned}\lambda_1^*\big[\lambda_1^*(a_{11} + a_{22}c_{12}) + \lambda_2^*(a_{11}\sqrt{c_{21}} + a_{22}\sqrt{c_{12}}) - \\ 2\sqrt{2^{2R_i}-1}\big] = 0,\end{aligned} \tag{75}$$

$$\begin{aligned}\lambda_2^*\big[\lambda_1^*(a_{22}\sqrt{c_{12}} + a_{11}\sqrt{c_{21}}) - 2\sqrt{2^{2R_i}-1} + \\ \lambda_2^*(a_{22} + a_{11}c_{21})\big] = 0.\end{aligned} \tag{76}$$

Now we need to find the quadruples (λ_1, λ_2, λ_3, λ_4) which solve the system composed by equations (62), (63), (71), (72), (75) and (76). We obtain the following three solutions

$$S_1 = \{\frac{2\sqrt{2^{2R_i}-1} - \lambda_2(a_{11}\sqrt{c_{21}} + a_{22}\sqrt{c_{12}})}{a_{11} + c_{12}a_{22}}, \tag{77}$$

$$\frac{2\sqrt{2^{2R_i}-1}(1 - \frac{\sqrt{c_{12}}a_{22} + \sqrt{a_{21}}a_{11}}{a_{11}+a_{12}a_{22}})}{a_{22} + c_{21}a_{11} - \beta}, 0, 0\}, \tag{78}$$

$$S_2 = \left\{(2\sqrt{2^{2R_i}-1})/(a_{11} + a_{22}c_{12}), 0, 0, 0\right\}, \tag{79}$$

$$S_3 = \left\{0, (2\sqrt{2^{2R_i}-1})/(a_{22} + c_{21}a_{11}), 0, 0\right\}, \tag{80}$$

where $\beta = \frac{(a_{11}\sqrt{c_{21}}+a_{22}\sqrt{c_{12}})(\sqrt{c_{12}}a_{22}+\sqrt{c_{21}}a_{11})}{a_{11}+c_{12}a_{22}}$. By plugging S_1, S_2 or S_3 in (64)-(68) we obtain at most three feasible points (P_i^*, P_j^*). We also need to consider pairs of values $(P_i^{(1)}, P_j^{(2)})$ where the constraint functions are not differentiable. Such pairs are $A = (0, P_j^{(2)})$, $B = (P_i^{(1)}, 0)$ and $C = (0, 0)$. Note that (0,0) cannot give a feasible solution due to the fact that in (61) we would have $\sqrt{2^{2 \cdot R_i} - 1} < 0$. Let us calculate the value of $P_i^{(1)}$ and $P_j^{(2)}$ in A and B. In point A and with $P_i^{(1)} = 0$ we have $\frac{1}{2}\log_2\left(1 + c_{12} \cdot P_j^{(2)}\right) \geq R_i$ and $\frac{1}{2}\log_2\left(1 + P_j^{(2)}\right) \geq R_i$. Then: $A = \left(0, \max\left\{2^{2R_i} - 1, \frac{2^{2R_i}-1}{c_{12}}\right\}\right)$ and dually $B = \left(\max\left\{2^{2R_i} - 1, \frac{2^{2R_i}-1}{c_{21}}\right\}, 0\right)$. Thus, the minimum power will be among A, B and the feasible points (P_i^*, P_j^*) obtained by inserting S_1, S_2 and S_3 in (64)-(68). The cost of this channel is $c_{GIc}(i,j) = P_i^{(1)} + P_j^{(2)}$, where $P_i^{(1)} + P_j^{(2)}$ are derived from the previous results.

Appendix II

The left hand side of Han-Kobayashi [33] inequalities (14)-(18) can be written as

$$\begin{aligned}\rho_1 =& \sigma_1^* + I(Y_1; u_1|w_1w_2), \\ \rho_2 =& \sigma_2^* + I(Y_2; u_2|w_1w_2), \\ \rho_{12} =& \sigma_{12} + I(Y_1; u_1|w_1w_2) + I(Y_2; u_2|w_1w_2), \\ \rho_{10} =& 2\sigma_1^* + 2I(Y_1, u_1|w_1w_2) + I(Y_2; u_2|w_1w_2) + \\ &- [\sigma_1^* - I(Y_2; w_1|w_2)]^+ + \min\{I(Y_2; w_2|w_1), \\ &I(Y_2; w_2) + [I(Y_2; w_1|w_2) - \sigma_1^*]^+, I(Y_1; w_2|w_1), \\ &I(Y_1, w_1w_2) - \sigma_1^*\},\end{aligned}$$

where

$$\begin{aligned}\rho_{20} =& 2\sigma_2^* + 2I(Y_2, u_2|w_1w_2) + I(Y_1; u_1|w_1w_2) + \\ &- [\sigma_2^* - I(Y_1; w_2|w_1)]^+ + \min\{I(Y_1; w_1|w_2), \\ &I(Y_1; w_1) + [I(Y_1; w_2|w_1) - \sigma_2^*]^+, I(Y_2; w_1|w_1), \\ &I(Y_2, w_1w_2) - \sigma_1^*\} \\ \sigma_1^* =& \min\{I(Y_1; w_1|w_2), I(Y_2; w_1|u_2w_2)\} \\ \sigma_2^* =& \min\{I(Y_2; w_1|w_1), I(Y_1; w_1|u_1w_1)\} \\ \sigma_{12} =& \min\{I(Y_1; w_1w_2), I(Y_2; w_1w_2); I(Y_1; w_1|w_2) + \\ &I(Y_2; w_2|w_1), I(Y_2; w_1|w_2) + I(Y_1; w_2|w_1)\},\end{aligned}$$

and

$$
\begin{aligned}
&I(Y_1;u_1|w_1w_2) = C\left(\lambda_1P_1/(1+a_{12}\lambda_2P_2)\right),\\
&I(Y_2;u_2|w_1w_2) = C\left(\lambda_2P_2/(1+a_{21}\lambda_1P_1)\right),\\
&I(Y_1;w_1|w_2) = C\left(\bar{\lambda}_1P_1/(1+\lambda_1P_1+a_{12}\lambda_2P_2)\right),\\
&I(Y_1;w_2|w_1) = C\left(a_{12}\bar{\lambda}_2P_2/(1+\lambda_1P_1+a_{12}\lambda_2P_2)\right),\\
&I(Y_1;w_1w_2) = C\left((\bar{\lambda}_1P_1{+}a_{12}\bar{\lambda}_2P_2)/(1{+}\lambda_1P_1{+}a_{12}\lambda_2P_2)\right),\\
&I(Y_2;w_2|w_1) = C\left(\bar{\lambda}_2P_2/(1+\lambda_2P_2+a_{21}\lambda_1P_1)\right),\\
&I(Y_2;w_1|w_2) = C\left(a_{21}\bar{\lambda}_1P_1/(1+\lambda_2P_2+a_{21}\lambda_1P_1)\right),\\
&I(Y_2;w_1w_2) = C\left((\bar{\lambda}_2P_2{+}a_{21}\bar{\lambda}_1P_1)/(1{+}\lambda_2P_2{+}a_{21}\lambda_1P_1)\right),\\
&I(Y_1;w_1) = C\left(\bar{\lambda}_1P_1/(1+\lambda_2P_2+a_{21}\lambda_1P_1)\right),\\
&I(Y_2;w_2) = C\left(\bar{\lambda}_2P_2/(1+\lambda_1P_1+a_{12}\lambda_2P_2)\right),\\
&I(Y_1;w_2|u_1w_1) = C\left(a_{12}\bar{\lambda}_2P_2/(1+a_{12}\lambda_2P_2)\right),\\
&I(Y_2;w_1|u_2w_2) = C\left(a_{21}\bar{\lambda}_1P_1/(1+a_{21}\lambda_1P_1)\right).
\end{aligned}
$$

while $\lambda_n = 1 - \bar{\lambda}_n$ represents the splitting coefficient, v_n is the private message and w_n the common message. Finally, Y_1 and Y_2 indicate the output of the channels for u_1 and u_2, respectively, while $[x]^+ = \max\{0, x\}$.

References

[1] V.W.S. Wong, R. Schober, D. Kwan Ng and Li-Chun Wang, "Key techonologies for 5G wireless systems." Cambridge Univ. Press., Cambridge, U.K.: 2017.

[2] W. Han, A. Liu and V. Lau, "PHY-Caching in 5G wireless networks: design and analysis," *IEEE Comm. Mag.*, vol. 54, Aug. 2016.

[3] E. Bastug, M. Bennis and M. Debbah, "Living on the edge: the role of proactive caching in 5G wireless networks," *IEEE Comm. Mag., IEEE*, vol. 52, no. 8, pp. 82-89, Aug. 2014.

[4] M. Gregori, J. Gómez-Vilarbedó and J.Matamoros, "Joint transmission and caching policy design for energy minimization in the wireless backhaul link," *in Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Hong Kong, China, 2015.

[5] A. Can Güngör and D. Gündüz, "Proactive wireless caching at mobile user devices for energy efficiency," *in Proc. IEEE Int. Symp. Wireless Comm. Syst. (ISWCS)*, Brussels, Belgium, 2015.

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

[7] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 809-813.

[8] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092-3107, May 2017.

[9] J. S. P. Roig, D. Gündüz, and F. Tosato, "Interference networks with caches at both ends," *in Proc. IEEE Int. Conf. on Comm.*, Paris, France, May 2017.

[10] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, "Hypergraph-based analysis of clustered co-operative beamforming with application to edge caching," *IEEE Wireless Comm. Lett.*, vol. 5, no. 1, pp. 84-87, Feb. 2016.

[11] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Comm.*, vol. 16, no. 8, pp. 5061-5076, Aug. 2017.

[12] N. Zhao, F. Cheng, F. Richard Yu, "Caching UAV Assisted Secure Transmission in Hyper-Dense Networks Based on Interference Alignment", Jan. 2018, *IEEE Trans. on Comm.* (In Press)

[13] H. Wu, J. Li, H. Lu and P.Hong, "A two-layer caching model for control delivery services in satellite-terrestrial networks," *IEEE Globecom.*, Washington, DC USA, Dec. 2016.

[14] M. Chen, M. Mozzaffari, W. Saad, C. Yin, M. Debbah and C. S. Hong, "Caching in the sky: proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Comm.*, vol. 35, no. 5, pp. 1046-1061, May 2017.

[15] E. Bastug, J.-L. Guenego and M. Debbah, "Proactive small cell networks," *in Proc. of the IEEE Int. Conf. on Telecom*, Casablanca, Morocco, May 2013.

[16] J. Hachem, U. Niesen and S. Diggavi, "Degrees of freedom of cache aided wireless interference networks," *arXiv:1606.03175*, 2016.

[17] X. Huang, Z. Zhao and H. Zhang, "Latency analysis of cooperative caching with multicast for 5G wireless networks". *IEEE Utility and Cloud Computing*, Shanghai, China, Dec. 2016.

[18] E. Recayte, G. Cocco and A. Vanelli-Coralli, "Caching in Gaussian interference channel with QoS constraints," *IEEE Int. Conf. on Comm.*, Paris, France, May 2017.

[19] L. Shanmugan, N. Golrezaei, A. Dimakis, A. Molisch and G. Caire, "FemtoCaching: wireless video content delivery through distributed caching helpers," in *IEEE Trans. on Inf. Theory*, vol. 59, no. 12, pp. 2467-2472, Dec. 2013.

[20] M. Mohammadi Amiri and D. Gündüz, "Decentralized caching and coded delivery over Gaussian broadcast channels," *in Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017.

[21] Y. Tian, S. Lu and C. Yang, "Macro-Pico Amplitude-Space Sharing with Optimized Han-Kobayashi Coding," in IEEE Trans. on Commun., vol. 61, no. 10, pp. 4404-4415, Oct. 2013.

[22] L. Wei, R. Q. Hu, Y. Qian and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Comm.*, vol. 21, no. 6, pp. 136-143, Dec. 2014.

[23] A. Carleial, "Interference channels," *in IEEE Trans. on Inf. Theory*, vol. 24, no. 1, pp. 60-70, Jan. 1978.

[24] J. Jiang, Y. Xin and H. K. Garg, "Interference channels with common information," *IEEE Trans. on Inf. Theory*, vol. 54, no. 1, pp. 171-187, Jan. 2008.

[25] S. Boyd, *Convex optimization*, Eds. Cambridge, UK: Cambridge Univ. Press., 2004, ch. 5, pp. 241-247.

[26] European Space Agency, European Data Relay Satellite, [Online]. Available: https://artes.esa.int/edrs.

[27] A. Chaaban and A. Sezgin, "On the capacity of a class of multi-user interference channels," in *2011 International ITG Workshop on Smart Antennas*.

[28] T. M. Cover and J. A. Thomas, "Elements of information theory," 2nd ed., New Jersey, Canada, Wiley, 2006, ch. 15.1, pp. 515-516.

[29] M. Costa, "Writing on dirty paper," *IEEE Trans. on Inf. Theory*, vol. 29, no. 3, pp. 439-441, May 1983.

[30] J. Oh, S. J. Kim, R. Narasihan and J. M. Cioffi, "Transmit power optimization for Gaussian vector broadcast channels," in *Proc. IEEE Conf. Comm.,* Seoul, Corea, May 2005.

[31] S. Vishwanath, N. Jindal and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. on Inf. Theory,* vol. 50, no. 10, pp. 1875-1892, Sep. 2004.

[32] D. Tse, P. Viswanath, "MIMO II: capacity and multiplexing architectures," in *Fundamentals of Wireless Comm.*, Cambridge, UK, 2005, pp. 332-376.

[33] T. Han, K. Kobayashi. "A new achievable rate region for the interference channel," *IEEE Trans. on Inf. Theory,* vol. 27, no. 1 pp. 49-60, Jan. 1981.

[34] M. M. Azari, F. Rosas, K. C. Chen and S. Pollin, "Ultra reliable UAV communication using altitude and cooperation diversity," in *IEEE Trans. on Comm.*, vol. 66, no. 1, pp. 330-344, Jan. 2018.

[35] C.F. Lo, "The sum and difference of two Lognormal random variables," in *Hindawi Publishing Corporation Journal of Applied Mathematics*, vol. 2012, pp. 1-13., Jul. 2012.