Coded caching for reducing CSIT-feedback in wireless communications

Jingjing Zhang
Mobile Communications Department
EURECOM, Sophia Antipolis, France
Email: jingjing.zhang@eurecom.fr

Felix Engelmann
Mobile Communications Department
EURECOM, Sophia Antipolis, France
Email: felix.engelmann@eurecom.fr

Petros Elia
Mobile Communications Department
EURECOM, Sophia Antipolis, France
Email: petros.elia@eurecom.fr

***Abstract*—The work explores the role of caching content at receiving users for the purpose of reducing the need for feedback in wireless communications. In the K-user broadcast channel (BC), we show how caching, when combined with a rate-splitting broadcast approach, can not only improve performance, but can also reduce the need for channel state information at the transmitter (CSIT), in the sense that the identified cache-aided optimal degrees-of-freedom performance, can in fact be achieved with reduced-quality CSIT. These CSIT savings can be traced back to an inherent relationship between caching, performance, and CSIT; caching improves performance by leveraging multicasting of common information, which automatically reduces the need for CSIT, by virtue of the fact that common information is not a cause of interference. At the same time though, too much multicasting of common information can be detrimental, as it does not utilize existing CSIT. Our caching method builds on the Maddah-Ali and Niesen coded caching scheme, by properly balancing multicast and broadcast opportunities, and by combing caching with rate-splitting communication schemes that are specifically designed to operate under imperfect-quality CSIT. The observed achievable CSIT savings here, are more pronounced for smaller values of K users and N files.**

I. Introduction

In the setting of communication networks, recent work in [1] has explored how caching content at receiving users, can increase the effective throughput and can reduce the load on the network, by utilizing *multicast gains*, i.e., by creating the need for common symbols that are simultaneously needed by more than one user. Our interest here is to explore this idea, not in the original multicast setting in [1], but in feedback-aided broadcast-type settings which utilize *broadcast gains* to communicate private messages. Our goal is to explore how caching, and the associated common messages that are generated as a result of this caching, can result in CSIT reductions, while at the same time, exploring how to properly balance the multicast and broadcast elements that emerge.

A. Cache-aided K-user broadcast channel

We consider a wireless communication setting where a K-antenna transmitter communicates information to K single-antenna receiving users. The entire information content at the transmitter consists of N distinct files $W_1, W_2, \dots, W_N$, each of size f bits. Each user $k = 1, 2, \dots, K$ has a cache Z_k of size Mf bits (Fig. 1), where $M < N$. Communication is split into two distinct phases; the *caching (or placement) phase* which happens before the users express their file demands[1], i.e., before the users inform the transmitter which file they want to receive, and the *delivery phase* that happens after the users' demands become known to the transmitter. During the placement phase, the caches $Z_1, Z_2, \dots, Z_K$ are pre-filled with information from the N files $W_1, W_2, \dots, W_N$. During the delivery phase, each user k requests a single file W_{F_k}, for some $F_k \in [1, 2, \dots, N]$. After each transmission, the corresponding received signals at receiver k, can be modeled as

$$y_k = \boldsymbol{h}_k^T \boldsymbol{x} + z_k, \quad k = 1, \dots, K \tag{1}$$

where $\boldsymbol{x} \in \mathbb{C}^{K\times 1}$ denotes the transmitted vector which satisfies a power constraint $\mathbb{E}(|\boldsymbol{x}|^2) \leq P$, where $\boldsymbol{h}_k \in \mathbb{C}^{K\times 1}$ denotes the vector fading coefficients, and where z_k represents unit power AWGN noise at user k. At the end of the delivery phase, each receiving user combines the observed y_k together with the side information in their cache Z_k, to reconstruct their desired file W_{F_k}. The effort here is to design an efficient caching-and-delivery method that minimizes the load on the delivery phase.

B. Measure of performance

As in [1], the measure of performance here is the duration T — in time slots, per file served per user — needed to complete the delivery process, for any request. Time is normalized such that one time slot corresponds to the amount of time it would take to communicate a single file to a single receiver, had their been no caching and no interference. As a result, in the high P setting of interest, with the capacity of a MISO channel scaling as $\log P$, we proceed to set $f = \log P$, which guarantees that the two measures of performance, here and in [1], now carry the same meaning, and can be meaningfully compared[2].

We here clarify that T is a 'worst-case' measure, as it corresponds to a duration that guarantees completion of delivery for *any* combination of requested files from the

[1]This placement phase may typically take place at a time of low network utilization, e.g. at nighttime, and is meant to ease the load during daytime.

[2]We note that the work in [1] tries to minimize T, which it refers to as the required achievable rate of the delivery link that guarantees completion of the delivery phase in a single time slot. This is the same as our measure here.

receiving users. As a result, we will henceforth assume without loss of generality that, *after caching takes place* (blindly, in terms of the requests), the delivery phase will respond to a request to send W_1 to user 1, W_2 to user 2, up to W_K to user K.

C. The link between caching, communication, and imperfect-quality feedback

As we will see, feedback-quality is not only linked with the performance of the delivery phase (where more feedback allows for better interference management and thus for higher performance over the wireless link), but is also linked with the caching phase; after all, loosely speaking, the higher the value of M, the more side information the receivers have, the less interference one needs to handle, and the less feedback is potentially needed. Similarly one can note that the delivery phase will have a feedback-aided broadcast element (as each user can require a different file), but it will also be in the presence of feedback-reducing side information at each receiver's cache.

To capture feedback quality, we let $\hat{\boldsymbol{h}}_k$ denote the current CSIT estimates of channels $\boldsymbol{h}_k$, we let

$$\tilde{\boldsymbol{h}}_k = \boldsymbol{h}_k - \hat{\boldsymbol{h}}_k \tag{2}$$

denote the CSIT estimation error, and we let

$$\alpha = -\lim_{P\to\infty} \frac{\log \mathbb{E}[||\tilde{\boldsymbol{h}}_k||^2]}{\log P}, \quad k = 1, \ldots, K$$

denote the *current CSIT quality exponent*, describing the quality of current CSIT at any user.

It is easy to see that $\alpha = 0$ corresponds to finite-precision or no CSIT, while having $\alpha = 1$ has been shown by [2], [3] to correspond — in the high-P setting of interest here — to perfect-quality CSIT. As α varies, so does the overall performance. Interesting insights into the role of α (and of timeliness) on the performance of the MISO BC, have been found through degrees of freedom (DoF) characterizations under perfect CSIT [4], no CSIT [5]–[8], delayed CSIT [9], mixed CSIT [10]–[13], and alternating CSIT [14]. Other related work can be found in [15]–[29].

D. Combining multicast and broadcast gains to minimize T and the CSIT requirements

The existence of caching allows — by virtue of the fact that some data is already present at the receiver — for an automatic reduction in T. The fact that coded caching can lead to a delivery phase that involves common symbols, means that we can potentially achieve this reduced cache-aided T, in the presence of reduced CSIT. At the same time, a proper balance must be kept between private and common symbols, so that feedback is not under-utilized and thus performance is not reduced.

The general objective here is naturally to design caching and transmission schemes that jointly reduce T, under specific constraints on caching size Mf and under specific constraints on the CSIT-quality α. Our focus will be on describing the optimal $T^*(M)$ that is achievable with perfect CSIT ($\alpha = 1$), then to describe the effect of imperfect feedback by providing an achievable $T(M, \alpha)$ under CSIT-quality constraints, and then to translate this into an achievability bound on the CSIT threshold exponent

$$\alpha_{th} \triangleq \arg\min\{\alpha : T(M, \alpha) = T^*(M, \alpha = 1)\}$$

which, albeit corresponding to imperfect-quality CSIT, still guarantees the optimal $T^*(M) = T^*(M, \alpha = 1)$.

We first though proceed with motivating examples that provide insight on different elements of the problem.

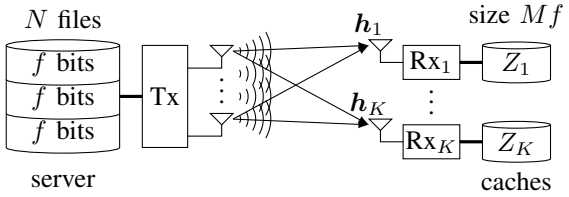

Fig. 1. System setup of cache-aided K-user MISO BC.

E. Motivating examples

1) Example - ($M = 0$): Let us first consider the no-caching case where $M = 0$, which — again without loss of generality — can be taken to correspond to a delivery phase that is strictly of a broadcast nature. In this broadcast channel, the sum DoF with perfect CSIT, is equal to K, which means that the K requested files[3] will take 1 time slot to deliver, corresponding to an optimal $T^*(M = 0, \alpha = 1) = T^*(0) = 1$. Applying the new result by Davoodi and Jafar [30], which states that the maximum DoF of K can only be achieved in the presence of $\alpha = 1$, immediately reveals that to achieve the optimal $T^*(0) = 1$, we need $\alpha = \alpha_{th} = 1$.

2) Example - ($N = K = 2$): Let us now offer a more involved example, which reveals an interesting achievable tradeoff between T, M and α. Specifically let us consider the case where $N = K = 2$, and let $0 \leq M \leq 1$. There are two files, which we relabel as $W_1 = A, W_2 = B$, each of size $f = \log P$ bits.

In this example, for the *placement phase*, we first split both files A and B into three subfiles, i.e, $A = (A_1, A_2, A_3), B = (B_1, B_2, B_3)$, where the subfiles $A_i, B_i, i = 1, 2$ are each of size $\frac{Mf}{2}$ bits, and where A_3 and B_3 are each of size $(1-M)f$ bits. We fill up the caches as follows $Z_1 = (A_1, B_1)$ and $Z_2 = (A_2, B_2)$, so that each user has in their cache, an equal part of clean information for each file.

For the *delivery phase* — and again focusing on the request $W_{F_1} = W_1 = A, W_{F_2} = W_2 = B$ — we see that to complete the task, user 1 needs subfile A_2 (which is available in the cache of user 2) as well as A_3, and user 2 needs subfile B_1 (available at the cache of user 1), as well as B_3. As a result, $A_2 \oplus B_1$ (containing $\frac{Mf}{2}$ bits) has information that

[3]Recall that the request considered over the delivery phase, is one where each user requests a different file.

can be useful to both users, while A_3 and B_3 has private information for user 1 and 2 respectively. The challenge will be to communicate this information, as efficiently as possible, over a channel with imperfect feedback, corresponding to some $\alpha < 1$. Towards this, let us consider a scheme that sends a single transmission of the form

$$\boldsymbol{x} = \boldsymbol{w}c + \hat{\boldsymbol{h}}_2^{\perp} a_1 + \hat{\boldsymbol{h}}_1^{\perp} a_2 \tag{3}$$

where $\boldsymbol{x} \in \mathbb{C}^{2\times 1}$, where $\boldsymbol{w} \in \mathbb{C}^{2\times 1}$ is a randomly chosen precoder, and where $\hat{\boldsymbol{h}}_k^{\perp} \in \mathbb{C}^{2\times 1}$ is a precoder that is orthogonal to the estimate $\hat{\boldsymbol{h}}_k$ for $\boldsymbol{h}_k$. Additionally, the above symbols c, a_1, a_2 are respectively allocated power as follows[4] given by

$$P^{(c)} \doteq P,\ P^{(a_1)} \doteq P^{(a_2)} \doteq P^{\alpha}$$

and are allocated rate as follows

$$r^{(c)} = (1-\alpha)f, \quad r^{(a_1)} = r^{(a_2)} = \alpha f.$$

In particular, a_1 is loaded with $\alpha \log P$ bits from A_3, and a_2 is loaded with $\alpha \log P$ bits from B_3, while c is loaded with the $\frac{Mf}{2}$ bits of $A_2 \oplus B_1$, as well as with the $\max\{2((1-M)f - T\alpha f), 0\}$ bits of A_3, B_3 that did not fit inside a_1 and a_2. The precoders, power-allocation and rate-allocation are known to all nodes. As a result the received signals at the two users, take the form

$$\begin{aligned} y_1 &= \underbrace{\boldsymbol{h}_1^T \boldsymbol{w} c}_{P} + \underbrace{\boldsymbol{h}_1^T \hat{\boldsymbol{h}}_2^{\perp} a_1}_{P^{\alpha}} + \underbrace{\boldsymbol{h}_1^T \hat{\boldsymbol{h}}_1^{\perp} a_2}_{P^0} + \underbrace{z_1}_{P^0} \\ y_2 &= \underbrace{\boldsymbol{h}_2^T \boldsymbol{w} c}_{P} + \underbrace{\boldsymbol{h}_2^T \hat{\boldsymbol{h}}_2^{\perp} a_1}_{P^0} + \underbrace{\boldsymbol{h}_2^T \hat{\boldsymbol{h}}_1^{\perp} a_2}_{P^{\alpha}} + \underbrace{z_2}_{P^0} \end{aligned} \tag{4}$$

where

$$\mathbb{E}|\boldsymbol{h}_1^T \hat{\boldsymbol{h}}_1^{\perp} a_2|^2 = \mathbb{E}|(\hat{\boldsymbol{h}}_1^T + \tilde{\boldsymbol{h}}_1^T)\hat{\boldsymbol{h}}_1^{\perp} a_2|^2 = \mathbb{E}|\tilde{\boldsymbol{h}}_1^T \hat{\boldsymbol{h}}_1^{\perp} a_2|^2 \doteq P^0$$

$$\mathbb{E}|\boldsymbol{h}_2^T \hat{\boldsymbol{h}}_2^{\perp} a_1|^2 = \mathbb{E}|(\hat{\boldsymbol{h}}_2^T + \tilde{\boldsymbol{h}}_2^T)\hat{\boldsymbol{h}}_2^{\perp} a_1|^2 = \mathbb{E}|\tilde{\boldsymbol{h}}_2^T \hat{\boldsymbol{h}}_2^{\perp} a_1|^2 \doteq P^0.$$

At this point, user 1 can decode common symbol c by treating all other signals as noise. Consequently, user 1 removes $\boldsymbol{h}_1^T \boldsymbol{w} c$ from y_1 and decodes private symbol a_1. From c, user 1 can recover $A_2 \oplus B_1$, which combined with Z_1 allows for recovery of A_2. Finally from c and a_1, user 1 can recover A_3. Given that A_1 is already available in its cache, user 1 can thus reconstruct A. User 2 similarly obtains B.

To calculate the achieved T for this example, we note that the total of $\frac{Mf}{2} + (1-M)f + (1-M)f = (\frac{4-3M}{2}) \log P$ bits in $A_2 \oplus B_1, A_3, B_3$, was sent at an achievable rate (provided by this specific rate-splitting scheme) of $(1-\alpha+\alpha+\alpha)f = (1+\alpha)\log P$ bits per time slot. Hence the corresponding achievable duration is

$$T(M,\alpha) = \frac{4-3M}{2(1+\alpha)}. \tag{5}$$

[4]We here use $\doteq$ to denote *exponential equality*, i.e., we write $f(P) \doteq P^B$ to denote $\lim_{P\to\infty} \frac{\log f(P)}{\log P} = B$.

As we will show later using basic cut-set bound arguments, the optimal $T^*(M)$ — associated to perfect CSIT — takes the form $T^*(M) = 1 - \frac{M}{N} = 1 - \frac{M}{2}$. Hence equating

$$T(M,\alpha) = \frac{4-3M}{2(1+\alpha)} = T^*(M) = 1 - \frac{M}{2}$$

and solving for α, gives that any α bigger than

$$\alpha_{th} = 1 - \frac{M}{2-M}, \quad 0 \le M \le 1 \tag{6}$$

suffices to achieve the optimal $T^*(M, \alpha = 1) = 1 - \frac{M}{2}$.

A few simple observations include the fact that, as expected, α_{th} reduces with M (Fig. 2), as well as the fact that for $M = 0$, we have $\alpha_{th} = 1$, which correctly reflects the discussion in the previous example, which reminded us that in the broadcast channel, in the limit of high P, perfect CSIT is necessary for DoF optimality, i.e., perfect CSIT is necessary to transmit one file to each user within a time duration that is asymptotically optimal. On the other hand, we see that having $M \ge 1$, leads to $\alpha_{th} = 0$ because, as we have seen in [1], when $M \ge 1$ ($N = K = 2$), a simple transmission of a common message, suffices to achieve $T^*(M) = 1 - \frac{M}{2}$. Since the transmission is limited to a common symbol, it does not require CSIT.

3) Example - ($N = K = 2, M = 1/2$). Modifying coded caching for the BC: Let us now look at the interesting instance of $N = K = 2, M = 1/2$, which showcases some of the differences between the multicast case in [1] and the broadcast approach here, and which motivates caching specifically for the multi-antenna wireless setting, as compared to caching for the multicast case with a single shared medium with only serial multicast possibilities, that was explored in [1]. Towards this, we recall that in [1], the *optimal* $T = 1$ was achieved by splitting files A and B into two halves, i.e., as $A = (A_1, A_2)$ and $B = (B_1, B_2)$, by setting $Z_k = A_k \oplus B_k, k = 1, 2$, and by sequentially transmitting[5] two common messages, B_1 and then A_2, to achieve the aforementioned optimal $T = 1$. What we point out here is that this caching would not work for the MISO-BC case (where the optimal T is $T^* = 1 - \frac{M}{2} = \frac{3}{4}$) because it leads to a delivery phase that only transmits common information, and thus does not leverage existing CSIT to improve performance.

We proceed with the description of the main results.

II. Main Results

We now describe the optimal $T^*(M)$ that is achievable with perfect CSIT ($\alpha = 1$), and then provide an achievability bound on $T(M,\alpha)$, and thus an achievability bound on the smallest α_{th} that achieves the optimal $T^*(M)$ above. We recall that the following results hold for $f = \log P$, in the limit of large P.

Lemma 1: In the cache-aided K-user MISO BC, with N files of size f, and with caches of size Mf, the optimal

[5]For the worst-case request that is assumed without loss of generality.

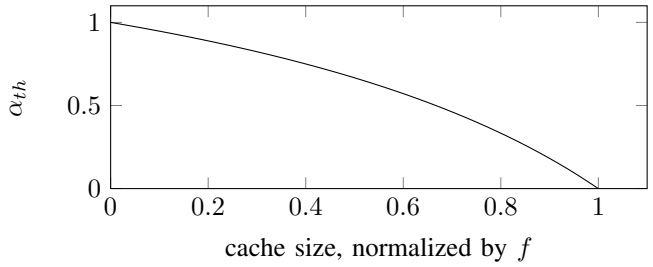

Fig. 2. Required α_{tr} to achieve the optimal $T^*(M)$ in the $K = N = 2$ cache-aided MISO BC.

$T^*(M, \alpha = 1)$ takes the form

$$T^*(M) = 1 - \frac{M}{N}. \tag{7}$$

Proof:

Let us first create the outer (lower) bound on T, using basic cut-set bound arguments in a manner that is similar to that in [1], [31]. Consider a simplified setting, where there are s users ($s \in \{1, \ldots, min\{\lfloor \frac{N}{M} \rfloor, K\}\}$). During the placement phase, the users' corresponding caches $Z_1, \ldots, Z_s$ are filled, while during the delivery phase, each of the s users makes $\lfloor \frac{N}{s} \rfloor$ sequential requests (one after the other), corresponding to a total of $s\lfloor \frac{N}{s} \rfloor$ requested files $W_1, \ldots, W_{s\lfloor \frac{N}{s} \rfloor}$ by all the users together. Note that for integer $\frac{N}{s}$, these requests span all N files. We now consider a total of $\lfloor \frac{N}{s} \rfloor$ sequential transmissions $X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}$, such that X_1 and $Z_1, \ldots, Z_s$ can reconstruct $W_1, \ldots, W_s$, such that similarly X_2 and $Z_1, \ldots, Z_s$ can reconstruct $W_{s+1}, \ldots, W_{2s}$, and so on, until we have that $X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}$ and $Z_1, \ldots, Z_s$ can reconstruct all the requested files $W_1, \ldots, W_{s\lfloor \frac{N}{s} \rfloor}$.

To apply the cut-set bound, we place the $\lfloor \frac{N}{s} \rfloor$ broadcasting signals $X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}$, each of duration T, on one side of the cut, together with all the caches $Z_1, \ldots, Z_s$, and then on the other side of the cut, we place all the requests of s users for a total of $s\lfloor \frac{N}{s} \rfloor$ files, each of size f. Hence it follows that

$$\begin{aligned}
&\lfloor \frac{N}{s} \rfloor sT + sM \geq H(Z_1, \ldots, Z_s, X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor}) \\
&\geq H(Z_1, \ldots, Z_s, X_1, \ldots, X_{\lfloor \frac{N}{s} \rfloor} | W_1, \ldots, W_{s\lfloor \frac{N}{s} \rfloor}) \\
&\quad + s\lfloor \frac{N}{s} \rfloor (1 - \epsilon_f) \\
&\geq s\lfloor \frac{N}{s} \rfloor (1 - \epsilon_f)
\end{aligned} \tag{8}$$

where we have used that the $K \times s$ interference-free MIMO channel provides s degrees of freedom (this is in the limit of $f \to \infty$), and where we have used Fano's inequality. In the same limit of $f \to \infty$, we have that $\epsilon_f \to 0$. Thus solving for T, and optimizing over all possible choices of s, we obtain

$$T \geq \max_{s \in \{1, \ldots, min\{\lfloor \frac{N}{M} \rfloor, K\}\}} (1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}) \tag{9}$$

which obviously gives that $T \geq 1 - \frac{M}{N}$.

To achieve this with perfect CSIT, is very easy. First let each user cache any $f\frac{M}{N}$ bits from each file, which leaves for $(1 - \frac{M}{N})f$ bits, per user, that must be delivered during the delivery phase. For the worst case where each user requests a different file, this corresponds to a broadcast transmission, which can be handled in $T = 1 - \frac{M}{N}$ time slots, in the presence of perfect CSIT. This completes the proof. ■

Having established the optimal $T^*(M, \alpha = 1) = 1 - \frac{M}{N}$, let us now establish an inner (achievability) bound on $T(M, \alpha)$, and translate that onto a bound on

$$\alpha_{th} = \arg\min\{\alpha : T(M, \alpha) = 1 - \frac{M}{N}\}.$$

The result will be presented for the simpler case where $K = N$.

Proposition 1: In the cache-aided K-user MISO BC, with $N = K$ files of size f, and with caches of size Mf, an achievable $T(M, \alpha)$ takes the form

$$T(M, \alpha) = \frac{K - \frac{M(1+K)}{K}}{1 + (K-1)\alpha} \tag{10}$$

which implies that the optimal $T^* = 1 - \frac{M}{N}$ can be achieved with an α that need not be bigger than

$$\alpha_{th} = \frac{N - 1 - M}{\frac{M}{K} + (N - 1 - M)}.$$

Proof: The proof is presented in the following subsection, by presenting the caching and delivery scheme that achieves the above performance in the presence of imperfect CSIT. ■

A. Coded caching and delivery with imperfect CSIT

To design the caching, each of the N files W_n, $n = 1, 2, \ldots, N$ is first divided into two parts,

$$W_n = (W_n^c, W_n^p)$$

where the information in W_n^p is never cached. For $p = \frac{M}{N-1}$, W_n^c has size pf, and W_n^p has size $(1-p)f$. The main idea is to apply the caching method of [1], but to restrict this to the subfiles $\{W_n^c\}_{n=1}^N$, rather than applying it on the whole $\{W_n\}_{n=1}^N$. Towards this, let us first split each subfile W_n^c into N subfiles $W_{n,\tau}, \tau \in \Omega$, where $\Omega = \{\tau \subset [K],\ s.t.\ |\tau| = N-1\}$, and where we have used the notation $[K] \triangleq \{1, 2, \ldots, K\}$. We note that the union of the above subfiles forms W_n^c, and that each subfile $W_{n,\tau}$ has size $\frac{Pf}{N}$. Based on the above, and following in the footsteps of [1], we form the caches as follows

$$Z_k \leftarrow W_{n,\tau}, \forall n = \{1, 2, \ldots, N\}, \forall \tau \in \Omega, \text{ such that } k \in \tau.$$

It is easy to see that each cache Z_k has $\frac{Mf}{N}$ bits originating from any specific W_n^c.

For the *delivery phase*, the transmitter sends

$$\boldsymbol{x} = \boldsymbol{w}c + \boldsymbol{g}_1 a_1 + \cdots + \boldsymbol{g}_k a_k + \cdots + \boldsymbol{g}_K a_K \tag{11}$$

where each a_k carries information from W_k^p, i.e., information that has not been cached anywhere, while c carries all the $\frac{pf}{N}$ bits of

$$X_c = \oplus_{k=1}^{K} W_{F_k,[K]\setminus\{k\}}$$

which can be seen as common information that is simultaneously useful to all receivers. Additionally, c carries the extra private information that could not fit in each a_k. In the above, $\boldsymbol{g}_k,\ k = 1, 2, \ldots, K$ are precoders that are designed to be orthogonal to the channel estimates of all users other than k. Finally the power and rate allocation was given by

$$\begin{aligned} &P^{(c)} \doteq P,\ P^{(a_k)} \doteq P^{\alpha} \\ &r^{(c)} = (1-\alpha)f,\ r^{(a_k)} = \alpha f, \quad k = 1, \ldots, K. \end{aligned} \tag{12}$$

As a result, the received signals $y_k,\ k = 1, 2, \ldots, K$ take the form

$$y_k = \underbrace{\boldsymbol{h}_k^T \boldsymbol{w} c}_{P} + \underbrace{\boldsymbol{h}_k^T \boldsymbol{g}_k a_k}_{P^{\alpha}} + \underbrace{\sum_{i=1, i\neq k}^{K} \boldsymbol{h}_k^T \boldsymbol{g}_i a_i}_{P^0} + \underbrace{z_k}_{P^0} \tag{13}$$

and we can see that, due to the power allocation and CSIT quality, symbols a_k do not cause interference to unintended users; at least not above the noise level. At this point, user k can decode the common symbol c by treating all other signals as noise. Consequently, user k removes $\boldsymbol{h}_k^T \boldsymbol{w} c$ from y_k, and decodes its private symbol a_k. Then it can recover $W_{F_k,[K]\setminus\{k\}}$ from Z_k and c, and $W_{F_k}^p$ from c and a_k. Since it already has $W_{F_k,\tau}$, user 1 reconstructs W_{F_k}. The same approach resolves the requests of the other users.

As before, we see that we were able to communicate

$$(K - \frac{MK}{N-1} + \frac{M}{K(N-1)})f = (K - \frac{MK}{N-1} + \frac{M}{K(N-1)}) \log P$$

bits of information. With the achievable rate of the communication scheme scaling as $(1 + (K-1)\alpha)\log P + o(\log P)$ per channel use, it becomes clear that as f increases, the transmission duration becomes

$$\begin{aligned} T(M, \alpha) &= \frac{K - \frac{MK}{N-1} + \frac{M}{K(N-1)}}{1 + (K-1)\alpha} \\ &= \frac{K - \frac{M(1+K)}{K}}{1 + (K-1)\alpha}. \end{aligned} \tag{14}$$

Setting $T(M, \alpha) = T^*(M) = 1 - \frac{M}{N}$, and solving for α, provides the achievable CSIT threshold[6] α_{th}.

Example: We here present a final example to offer some clarity on the designed scheme. We do this for the case of $N = K = 3, M = 1$. As before, we rename the files $(W_1, W_2, W_3) =: (A, B, C)$, and since $\tau \in \{\{1,2\},\{1,3\},\{2,3\}\}$, we split these as $A = (A_{12}, A_{13}, A_{23}, A^p)$, $B = (B_{12}, B_{13}, B_{23}, B^p)$ and $C = (C_{12}, C_{13}, C_{23}, C^p)$, where A^p, B^p, C^p will be private information for user 1, 2 and 3 respectively. In the above, the subfiles A_τ, B_τ, C_τ are each of size $\frac{f}{6}$, and they appear in Z_k for any $k \in \tau$ (see Fig. 3).

[6]It is interesting to note that for $\alpha = \alpha_{th}$, all private information is stored in symbols $a_k,\ k = 1, 2, \ldots, K$.

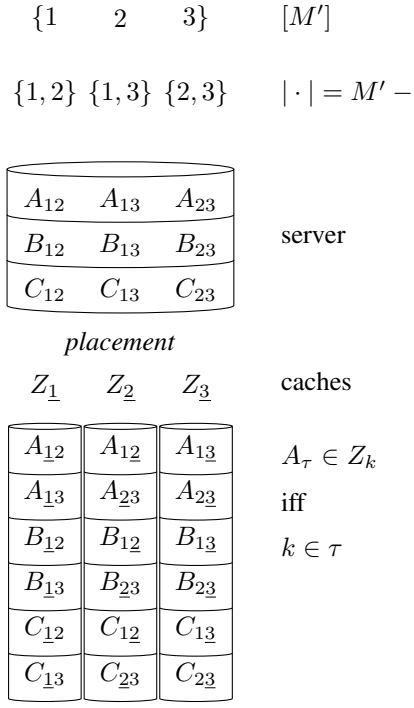

Fig. 3. Placement of parts into user caches for $N = K = 3$, $M' \triangleq N - 1 = 2$.

For request A, B, C by user $1, 2, 3$ respectively, we can see that user 1 needs A_{23} which is available at Z_2, Z_3, user 2 needs B_{13} which is available at Z_1, Z_3, and user 3 needs C_{12} which is available at Z_1, Z_2. Hence considering the common information $A_{23} \oplus B_{13} \oplus C_{12}$, we see that upon decoding this common information, user 1 can automatically reconstruct A_{23} (by removing $B_{13} \oplus C_{12}$), and users 2 and 3 can act similarly to respectively reconstruct B_{13} and C_{12}.

During the delivery phase, the transmitter sends

$$\boldsymbol{x} = \boldsymbol{w} c + \boldsymbol{g}_1 a_1 + \boldsymbol{g}_2 a_2 + \boldsymbol{g}_3 a_3 \tag{15}$$

with power and rates set as

$$\begin{aligned} &P^{(c)} \doteq P,\ P^{(a_k)} \doteq P^{\alpha} \\ &r^{(c)} = 1 - \alpha,\ r^{(a_k)} = \alpha, \quad k = 1, 2, 3 \end{aligned} \tag{16}$$

where $A_{23} \oplus B_{13} \oplus C_{12}$ is carried exclusively by c, while A^p, B^p, C^p are respectively placed in a_1, a_2, a_3, and any leftover information is placed in c.

The received signals y_k take the form

$$y_k = \underbrace{\boldsymbol{h}_k^T \boldsymbol{w} c}_{P} + \underbrace{\boldsymbol{h}_k^T \boldsymbol{g}_k a_k}_{P^{\alpha}} + \underbrace{\sum_{i=1, i\neq k}^{3} \boldsymbol{h}_k^T \boldsymbol{g}_i a_i}_{P^0} + \underbrace{z_k}_{P^0} \tag{17}$$

and as before, user 1 can decode c and a_1 to reconstruct all of A, and similarly for user 2 and 3 which reconstruct B and C respectively.

To calculate T, we note that there is a total of $3 - \frac{4M}{3}$ bits of information to be communicated (total information in $A_{23} \oplus B_{13} \oplus C_{12}$, A^p, B^p, C^p). Since the rate-splitting scheme has an achievable rate of $(1 + 2\alpha) \log P$ bits per time slot, we have that

$$T(M = 1, \alpha) = \frac{3 - \frac{4M}{3}}{1 + 2\alpha}$$

which, when equated with $T^*(M, \alpha = 1) = 1 - \frac{M}{N} = \frac{2}{3}$, gives

$$\alpha_{th} = \frac{1}{\frac{1}{3} + 1} = \frac{3}{4}.$$

It is worth comparing the above scheme which achieves $\alpha_{th} = \frac{3}{4}$, to a scheme that uses the caching method in [1], which — for these values of M, N — would not immediately allow for the possibility to have a common symbol that is simultaneously useful to everyone, and would thus not allow for the CSIT savings presented above. This is because in [1], the files are divided as $A = (A_1, A_2, A_3), B = (B_1, B_2, B_3), C = (C_1, C_2, C_3)$, forming caches $Z_k = (A_k, B_k, C_k),\ k = 1, \ldots, K$, which means that (again for delivery of different files $W_{F_1} = A, W_{F_2} = B, W_{F_3} = C$), $A \setminus Z_1 = (A_2, A_3), B \setminus Z_2 = (B_1, B_3), C \setminus Z_3 = (C_1, C_2)$, which in turn implies transmission of two-pair XORs $(A_2 \oplus B_1, A_3 \oplus C_1, B_3 \oplus C_2)$ (rather than triplet XORs in our case) which allows for the optimal $T^* = 3/4$, but only under the condition of perfect CSIT.

III. Conclusions

Motivated by recent advances in caching content at users (cf. [1] [32] [31]), which utilize *multicast gains* to increase throughput and reduce the network load, and motivated by sophisticated transmission schemes in multiuser settings that utilize precoder-enabled *broadcast gains* to increase the overall capacity of the system (sometimes in the presence of imperfect feedback [13] [30] [10] [12]), we have here jointly treated these multicast and broadcast efforts, in a complementary manner that synergistically compensated for each approach's limitations. Particularly we focused on the synergistic effect of jointly treating caching and communication in broadcast-type communications, and explored the benefits of caching, not only in improving performance, but also in reducing the CSIT required to achieve this optimal performance.

Future work will include an effort to reduce the achievable bound $T(M, \alpha)$, as well as efforts to further reduce the CSIT-quality exponent α_{th} associated to the optimal performance.

Acknowledgment

The work was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) / grant agreement no.318306 (NEWCOM#).

References

[1] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *Information Theory, IEEE Transactions on*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[2] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045 – 5060, Nov. 2006.

[3] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845 – 2866, Jun. 2010.

[4] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691 – 1706, Jul. 2003.

[5] S. Jafar and A. Goldsmith, "Isotropic fading vector broadcast channels: The scalar upper bound and loss in degrees of freedom," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 848 – 857, Mar. 2005.

[6] C. Huang, S. A. Jafar, S. Shamai, and S. Vishwanath, "On degrees of freedom region of MIMO networks without channel state information at transmitters," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 849–857, Feb. 2012.

[7] A. Lapidoth, S. Shamai, and M. A. Wigger, "On the capacity of fading MIMO broadcast channels with imperfect transmitter side-information," in *Proc. Allerton Conf. Communication, Control and Computing*, Sep. 2005.

[8] C. Vaze and M. Varanasi, "The degree-of-freedom regions of MIMO broadcast, interference, and cognitive radio channels with no CSIT," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5254 – 5374, Aug. 2012.

[9] M. A. Maddah-Ali and D. N. C. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418 – 4431, Jul. 2012.

[10] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 315–328, Jan. 2013.

[11] J. Chen and P. Elia, "Degrees-of-freedom region of the MISO broadcast channel with general mixed-CSIT," in *Proc. Information Theory and Applications Workshop (ITA)*, Feb. 2013.

[12] T. Gou and S. Jafar, "Optimal use of current and outdated channel state information: Degrees of freedom of the MISO BC with mixed CSIT," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1084 – 1087, Jul. 2012.

[13] J. Chen and P. Elia, "Toward the performance versus feedback tradeoff for the two-user miso broadcast channel," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8336–8356, Dec. 2013.

[14] R. Tandon, S. A. Jafar, S. Shamai, and H. V. Poor, "On the synergistic benefits of alternating CSIT for the MISO BC," Aug. 2012, to appear in *IEEE Trans. Inform. Theory*, available on arXiv:1208.5071.

[15] A. Ghasemi, A. S. Motahari, and A. K. Khandani, "On the degrees of freedom of X channel with delayed CSIT," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2011.

[16] J. Xu, J. G. Andrews, and S. A. Jafar, "Broadcast channels with delayed finite-rate feedback: Predict or observe?" *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1456 – 1467, Apr. 2012.

[17] Y. Lejosne, D. Slock, and Y. Yuan-Wu, "Degrees of freedom in the MISO BC with delayed-CSIT and finite coherence time: A simple optimal scheme," in *Proc. IEEE Int. Conf. on Signal Processing, Communications and Control (ICSPCC)*, Aug. 2012.

[18] J. Chen and P. Elia, "MISO broadcast channel with delayed and evolving CSIT," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2013.

[19] N. Lee and R. W. Heath Jr., "Not too delayed CSIT achieves the optimal degrees of freedom," in *Proc. Allerton Conf. Communication, Control and Computing*, Oct. 2012.

[20] J. Chen and P. Elia, "MIMO BC with imperfect and delayed channel state information at the transmitter and receivers," Jun. 2013.

[21] C. Hao and B. Clerckx, "Imperfect and unmatched CSIT is still useful for the frequency correlated MISO broadcast channel," Feb. 2013, to appear in *ICC13*, available on arXiv:1302.6521.

[22] J. Chen and P. Elia, "Can imperfect delayed CSIT be as useful as perfect delayed CSIT? DoF analysis and constructions for the BC," in *Proc. Allerton Conf. Communication, Control and Computing*, Oct. 2012.

[23] J. Chen, S. Yang, and P. Elia, "On the fundamental feedback-vs-performance tradeoff over the MISO-BC with imperfect and delayed CSIT," Feb. 2013, to appear in *ISIT13*, available on arXiv:1302.0806.
[24] J. Chen and P. Elia, "Symmetric two-user MIMO BC and IC with evolving feedback," Jun. 2013, available on arXiv:1306.3710.
[25] X. Yi, S. Yang, D. Gesbert, and M. Kobayashi, "The degrees of freedom region of temporally-correlated MIMO networks with delayed CSIT," Nov. 2012, submitted to *IEEE Trans. Inform. Theory*, available on arXiv:1211.3322.
[26] C. S. Vaze and M. K. Varanasi, "The degrees of freedom region of two-user and certain three-user MIMO broadcast channel with delayed CSI," Dec. 2011, submitted to *IEEE Trans. Inf. Theory*, available on arXiv:1101.0306.
[27] A. Vahid, M. A. Maddah-Ali, and A. S. Avestimehr, "Capacity results for binary fading interference channels with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6093 – 6130, Oct. 2014.
[28] G. Caire, N. Jindal, and S. Shamai, "On the required accuracy of transmitter channel state information in multiple antenna broadcast channels," in *Proc. Allerton Conf. Communication, Control and Computing*, Nov. 2007.
[29] J. Chen, S. Yang, A. Özgür, and A. Goldsmith, "Outdated CSIT can achieve full DoF in heterogeneous parallel channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2014.
[30] A. G. Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling a conjecture by lapidoth, shamai and wigger on the collapse of degrees of freedom under finite precision CSIT," *CoRR*, vol. abs/1403.1541, 2014. [Online]. Available: http://arxiv.org/abs/1403.1541
[31] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," *CoRR*, vol. abs/1504.01123, 2015. [Online]. Available: http://arxiv.org/abs/1504.01123
[32] M. A. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," *CoRR*, vol. abs/1301.5848, 2013. [Online]. Available: http://arxiv.org/abs/1301.5848