Rate-Memory Trade-Off for the Cache-Aided MISO Broadcast Channel with Hybrid CSIT

Antonio Bazco-Nogueras, Petros Elia
Communication Systems Department, EURECOM
Email: {antonio.bazco-nogueras; petros.elia}@eurecom.fr

Abstract—**One of the famous problems in communications was the so-called "PN" problem in the Broadcast Channel, which refers to the setting where a fixed set of users provide perfect Channel State Information (CSI) to a multi-antenna transmitter, whereas the remaining users only provide finite precision CSI or no CSI. The Degrees-of-Freedom (DoF) of that setting were recently derived by means of the Aligned Image Set approach. In this work, we resolve the cache-aided variant of this problem (i.e., the "PN" setting with side information) in the regime where the number of users providing perfect CSI is smaller than or equal to the number of transmit antennas. In particular, we derive the optimal rate-memory trade-off under the assumption of uncoded placement, and characterize the same trade-off within a factor of 2.01 for general placement. The result proves that the "PN" impact remains similar even in the presence of side information, but also that the optimal trade-off is not achievable through independently serving the two sets of users.**

I. Introduction

Coded caching has emerged as a promising tool for coping with the challenging increase of content demand in wireless networks. Initially, the benefit of coded caching was identified for the setting in which a server communicates to K users through an error-free single-stream shared link of fixed capacity [1]–[3]. In this setting, the server has access to a library of N files, and each user has access to a local memory (cache) of size equal to the size of M files, where it can store content from the library. In the aforementioned scenario, it was shown in [1]–[3] that coded caching provides a speed-up factor (or *coded-caching gain*) of $K\frac{M}{N}+1$ as compared to uncoded caching, since it allows us to simultaneously serve $K\frac{M}{N}+1$ users.

The promising gains of coded caching in this initial setting fostered the interest in understanding how these gains could be translated into wireless networks, which was analyzed e.g. by considering uneven link capacities [4], [5] or multi-antenna transmissions [6], [7]. The context of multi-antenna coded caching has recently received considerable attention, especially after the latest results in [8], which reveal that multi-antenna coded caching systems provide all the massive gains of coded caching, together with all the multiplexing gains of multi-antenna systems, and can do so without the hindrance of the subpacketization bottleneck that has previously kept such technologies from taking off.

In the multi-antenna context, it is crucial to characterize the impact of CSI availability. Toward this, among other works, [5] considered partial CSI at the transmitter (CSIT) on the cache-aided MISO Broadcast Channel (BC), [7] analyzed the CSIT required to maintain the ideal caching gains in multi-antenna settings, [9] studied cache-aided interference management with no CSIT, and [10] focused on the delayed CSIT setting.

In this work, we consider the cache-aided variant of a classical CSI-related problem: the MISO BC in which the L-antenna transmitter has perfect CSI only for the channel of a fixed set of K_P users, whereas the CSI of the other $K-K_P$ users is available only up to finite precision[1] [11]. This scenario is equivalent to the so-called "PN" BC setting [11], [13], [14], but with the non-trivial extension of considering side information at the users. The original "PN" BC remained an open problem for many years, and its DoF were finally derived by Davoodi and Jafar in [11] by means of the Aligned Image Set (AIS) approach. By incorporating the AIS approach into the derivation of the delivery time of coded caching, we obtain the optimal rate-memory trade-off of this setting under the assumption of uncoded placement when $K_P \leq L$, and characterize the same trade-off within a factor of 2.01 for general placement. Moreover, we show that the optimal trade-off is not achievable through separated transmission to the two sets of users.

II. System Model and Problem Definition

A. Communication Channel

We consider the $L\times K$ MISO BC in which a transmitter (TX) with L transmit antennas serves K single-antenna cache-aided users. The received signal at user i is written as

$$Y_i(t) = \sum\nolimits_{j=1}^{L} h_{i,j}(t)X_j(t) + \zeta_i(t), \qquad (1)$$

where $X_j(t)$ is the transmit signal from the j-th antenna of the transmitter, normalized such that $\mathbb{E}[|X_j(t)|^2] \leq P$, and where P is the nominal SNR value which in the DoF framework is allowed to approach infinity [11]. Furthermore, $\zeta_i(t)$ is the i.i.d. additive white Gaussian noise (AWGN), and $h_{i,j}$ denotes the fading coefficient between the j-th antenna of the TX and user i. We assume that $|h_{i,j}|$ is bounded away from zero and infinity. The channel vector from the L transmit antennas to user i is denoted by $\mathbf{h}_i$, and the global channel matrix by $\mathbf{H}$. We will make use of the notation $[n] = \{1, 2, \ldots, n\}$. We further define the vectors $\mathbf{X}(t) \triangleq (X_1(t), \ldots, X_L(t))$, $X_j^{[\tau]} \triangleq \{X_j(t)\}_{t\in[\tau]}$, $\mathbf{X}^{[\tau]} \triangleq \{\mathbf{X}(t)\}_{t\in[\tau]}$, and $Y_i^{[\tau]} \triangleq \{Y_i(t)\}_{t\in[\tau]}$.

This work is supported by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929 (ERC project DUALITY).

[1] For the DoF metric, the results obtained assuming finite precision CSIT (see [11], [12] for a rigorous definition) also hold if there was no CSIT for the users with finite precision CSIT [11]. Thus, we omit the latter for concision.

B. Cached-aided Network

We consider a cached-aided scenario in which the TX has access to a content library of N files, each one of size equal to B bits. We assume that $N \geq K$, and we denote the n-th file of the library as W_n, $n \in [N]$. Each user has a local memory cache of size MB bits, with $M \in \{0, 1, \ldots, N\}$, in which it stores (coded or uncoded) data from the library. We denote the normalized cache size with respect to the library size as $\gamma \triangleq \frac{M}{N}$, and the cache available at user i as Z_i, $|Z_i| = MB$.

Coded caching systems operate in two phases: the *placement phase*, in which the users fill their cache with content from the library, and the *delivery phase*, in which each user requests a library file to be delivered by the TX. The file requested by user k is denoted as W_{d_k}, $d_k \in [N]$, and the vector of requested file indexes is denoted by $\boldsymbol{d} \triangleq (d_1, \ldots, d_K) \in [N]^K$. We allow coded cache placement, but we also analyze the particular case in which only uncoded placement is allowed.[2]

C. Hybrid CSIT

We focus on the MISO BC scenario[3] in which the TX only obtains perfect CSI for the channel of a subset of K_P users, whereas the CSI of the other $K_F \triangleq K - K_P$ users is available only up to finite precision. We assume that $K_P \leq L$.[4]

We remark that this setting is fundamentally different from that in [7], where it was shown that, in the $L \times K$ cache-aided MISO BC, we can achieve the optimal coded-caching gain under one-shot linear schemes ($K\gamma + L$) even if the TX has CSI for only L served users *at a time*. The difference resides in the fact that, in [7], the TX needs CSI from only L users at a time, but the users who are providing the CSI can (and do) change when the subset of served users changes. Hence, [7] derives the minimum instantaneous CSI requirements for a setting where all users must eventually provide CSI. In contrast, we consider in this work that a *fixed* subset of users provide CSI, and this subset does not change throughout the whole transmission. Thus, there are two classes of users according to the quality of the provided CSI.

D. Problem Definition

We present the formal definition of the rate-memory trade-off considered. Let $\{W_n\}_{n\in[N]}$ be N i.i.d. random variables, each uniformly distributed over $[2^{\lfloor B \rfloor}]$, and let us introduce the notation $\mathcal{W}_{[k]} = \{W_n\}_{n\in[k]}$, $k \leq N$. We scale the size of the files proportionally to the number of channel uses τ as $B \triangleq \tau R$, where R is the transmission rate in bits per channel use.

A code $\mathcal{C}(\tau, R, M)$ consists of a prefetching strategy, an encoding scheme and K decoding functions, which are explained in the following. We characterize a prefetching strategy $\boldsymbol{\phi}$ by its K caching functions ϕ_k, $\boldsymbol{\phi} \triangleq (\phi_1, \ldots, \phi_K)$, each of which is a function that maps the library content into the cache content of one user during the placement phase. Thus, $\phi_k : [2^{\lfloor B \rfloor}]^N \to [2^{\lfloor MB \rfloor}]$, and $Z_k \triangleq \phi_k(W_1, \ldots, W_N)$. The encoding scheme $\psi : [N]^K \times [2^{\lfloor B \rfloor}]^N \to \mathbb{C}^{L\times\tau}$ maps a demand vector $\boldsymbol{d}$ and N files into a codeword vector $\mathbf{X}^{[\tau]} \triangleq \psi(\boldsymbol{d}, \mathcal{W}_{[N]})$ satisfying the average power constraint $\mathbb{E}[|X_j(t)|^2] \leq P$. Finally, a decoding function $\mu_k : [N]^K \times \mathbb{C}^\tau \times [2^{\lfloor MB \rfloor}] \to [2^{\lfloor B \rfloor}]$ maps a requested demand, the signal received $Y_k^{[\tau]}$, and the cache content Z_k into an estimate $\hat{W}_{d_k} \triangleq \mu_k(\boldsymbol{d}, Y_k^{[\tau]}, Z_k)$ of W_{d_k}. The probability of error is defined as

$$P_{e,\tau} \triangleq \max\nolimits_{\boldsymbol{d}\in[N]^K} \max\nolimits_{k\in[K]} \Pr(\hat{W}_{d_k} \neq W_{d_k}). \qquad (2)$$

Note that (2) reflects a *worst-case* metric over all possible demands. Thus, we consider that each user requests a distinct file.

We want to characterize the channel uses required to transmit a single bit of content to each user [1], [5]. This delay will be referred to as the *delivery time*, and it is given by $T \triangleq \frac{1}{R}$. More rigorously, for a given prefetching strategy $\boldsymbol{\phi}$, a delivery time $T_{\varepsilon,\phi}(M)$ is said to be ε-achievable if and only if, for every $\varepsilon > 0$ and big enough file size B, there exists a code $\mathcal{C}(\tau, R, M)$ with probability of error less than ε.

For a given memory constraint M, the rate-memory trade-off $T^\star(M)$ is defined as the minimum delivery time that can be achieved by *any* prefetching scheme with vanishing error probability and for sufficiently large file size. Thus, it is defined as $T^\star(M) \triangleq \sup_{\varepsilon>0} \limsup_{B\to\infty} \min_\phi T^\star_{\epsilon,\phi}(M)$. We omit hereinafter the dependence on M due to space constraints.

In pursuit of highlighting the impact of the multi-user interference, we consider the fundamental limit where $P \to \infty$. Therefore, we consider the optimal *Normalized Delivery Time* (NDT), which is defined as the ratio of the optimal delivery time $T^\star$ over the time required to deliver a single bit to a cache-less user, in the absence of interference, as P approaches infinity [4], [15]. Consequently, it is defined as

$$\mathrm{NDT} \triangleq \lim_{P\to\infty} \frac{T^\star}{1/\log P} = \lim_{P\to\infty} T^\star \log P. \qquad (3)$$

Let us introduce also the Degrees-of-Freedom [11] (DoF) metric, which are defined as $\mathrm{DoF} \triangleq \lim_{P\to\infty} \frac{C(P)}{\log P}$, where $C(P)$ denotes the capacity of the setting. Hence, the NDT can also be written as $\mathrm{NDT} = \frac{K(1-\gamma)}{\mathrm{DoF}}$.

The metrics presented above are similarly defined when considering only uncoded prefetching, in which case we will just insert a sub-index u (e.g. $\mathrm{NDT}_u, T_{u,\varepsilon,\phi}$). Furthermore, when necessary, we will use the full notation $\mathrm{NDT}(K, L, \gamma, K_P)$ to reflect the particular network configuration. For example, the result for the well-known K-user SISO BC with uncoded prefetching would correspond to $\mathrm{NDT}_u(K, 1, \gamma, 0)$.

III. Main results

We will henceforth make use of the term (K, L, γ, K_P) MISO BC to denote the $L \times K$ cache-aided MISO BC in which only a fixed set of $K_P \leq L$ users provide perfect CSIT, whereas the other $K_F = K - K_P$ users provide finite precision CSIT. We begin with the characterization of the optimal rate-memory trade-off for the particular case of uncoded placement.

[2] A prefetching scheme is called an *uncoded prefetching scheme* if each user stores MB bits from the database without coding.

[3] Although [5] studied also the cache-aided MISO BC, in [5] all users provide the same level of (possibly imperfect) CSI, while here we consider different CSIT level for different users.

[4] Cases where $K_P = 0$ or $K_F = 0$ are not comprised within the cases with *hybrid* CSIT, but we include them for completeness.

Theorem 1. *Under the assumption of uncoded prefetching, the optimal normalized delivery time of the* (K, L, γ, K_P) *MISO BC when* $K_P \leq L$ *is given by*[5]

$$\mathrm{NDT}_u(K, L, \gamma, K_P) = \mathrm{Conv}_{(K_F+1)\gamma}\Big(\frac{(K_F+1)(1-\gamma)}{(K_F+1)\gamma+1}\Big) \\ = \mathrm{NDT}_u(K_F+1, 1, \gamma, 0). \tag{4}$$

where $K_F = K - K_P$ *and* $\mathrm{Conv}_A\big(f(A)\big)$ *denotes the lower convex envelope of the points* $\big\{\big(A,\ f(A)\big) \,|\, A \in \{0, 1, ..., K\}\big\}$.

Proof. The achievable scheme and the converse are presented in Section IV and Section V-B, respectively. □

Remark 1. *The NDT required to serve* $K_F + K_P$ *users* ($K_P \leq L$) *from a multi-antenna TX is the same as the one required to serve* $K_F + 1$ *users from a single-antenna TX. Thus, starting from a cache-aided setting with* K_F *users and finite precision CSIT, we can add as many as* L *users who provide perfect CSI at the cost of a single finite-precision-CSIT user.*

Remark 2. *The* NDT *in* (4) *is not achievable through separate transmission to the two classes of users, since that approach attains a* NDT *of* $\frac{K_F(1-\gamma)}{1+K_F\gamma} + (1-\gamma)$, *which is strictly bigger.*

Remark 3. *Theorem 1 also implies that perfect CSIT for a single user does not improve the* NDT *with uncoded prefetching w.r.t. the case where all users provide only finite precision CSIT. This is analogous to the collapse of DoF proved in [11] for the cacheless setting with perfect CSIT for only one user.*

Next, we remove the assumption of uncoded prefetching, and we show that the proposed scheme is within a factor of 2.01 from the optimal.

Theorem 2. *The optimal normalized delivery time of the* (K, L, γ, K_P) *MISO BC when* $K_P \leq L$ *satisfies that*

$$\mathrm{NDT}_u(K_F+1,\ 1,\ \gamma,\ 0) \geq \mathrm{NDT}(K, L, \gamma, K_P) \\ \geq {}^{1}\!/_{2.00884}\ \mathrm{NDT}_u(K_F+1,\ 1,\ \gamma,\ 0). \tag{5}$$

Proof. The achievable scheme is the same as for Theorem 1, whereas the converse is presented in Section V-C. □

While the achievable scheme can be directly generalized for the case where $K_P > L$, the generalization of the bounds is not straightforward. Nevertheless, it is expected that the insight in Remark 1 will hold (i.e., that we can add K_P perfect-CSIT users at the cost of $\lceil K_P/L \rceil$ finite-precision-CSIT users).

IV. Achievability of Theorem 1 and Theorem 2

We denote the set of users providing perfect CSI as $\mathcal{K}_P \subseteq [K]$. In the scheme, our aim will be to provide the users in $\mathcal{K}_P$ with both full spatial multiplexing gains and full coded caching gains, which will naturally surpass the performance of a simple separate transmission to the two user sets. As it turns out, the way to achieve these gains is allocating identical caches at all users in $\mathcal{K}_P$, and thus the scheme assumes only $\Lambda \triangleq K - K_P + 1$ different cache states. This is clarified below.

A. Placement

We consider Λ cache states of size MB bits, and denote them as $Z^{(c)}_i$, $i \in [\Lambda]$. Let $\Lambda\gamma$ be an integer and $\mathcal{T}$ the set of $\binom{\Lambda}{\Lambda\gamma}$ subsets in $[\Lambda]$ of size $\Lambda\gamma$, i.e., $\mathcal{T} \triangleq \{\tau \subset [\Lambda] : |\tau| = \Lambda\gamma\}$. First, we split each message W_n, $n \in [N]$, into $|\mathcal{T}| \triangleq \binom{\Lambda}{\Lambda\gamma}$ non-overlapping subfiles of equal size, such that $W_n \to \{W_{n,\tau} : \tau \in \mathcal{T}\}$, and we assign the content of each of the Λ caches as $Z^{(c)}_i = \{W_{n,\tau} : i \in \tau, \tau \in \mathcal{T}\}_{n=1}^{N}$. Although the subpacketization is analogous to the one developed by Maddah-Ali and Niesen in [1], in our case the cache assignment is different: Let us first assume that the users providing perfect CSI are the last ones, i.e., $\mathcal{K}_P = \{K_F + 1, \ldots, K\}$. Then, all the users in $\mathcal{K}_P$ store the content of the last cache state, $Z_i = Z^{(c)}_{\Lambda}$, $\forall i \in \mathcal{K}_P$, while each of the other users stores a different cache state, such that $Z_i = Z^{(c)}_i$, $\forall i \in [\Lambda - 1]$. Hereinafter, we refer to the set of users that share the same cache as a *cache group*, even if only one of these groups has more than one user.

B. Transmit Signal

During the delivery phase, the TX simultaneously serves $\Lambda\gamma + 1$ cache groups. Let $\mathcal{X} \triangleq \{\chi \subseteq [\Lambda] : |\chi| = \Lambda\gamma + 1\}$ be defined as the set of $|\mathcal{X}| = \binom{\Lambda}{\Lambda\gamma+1}$ subsets of size $\Lambda\gamma + 1$ in $[\Lambda]$. The delivery consists in $|\mathcal{X}|$ sequential transmissions, one for each $\chi \in \mathcal{X}$. Let $W_{d_j,\chi\backslash c_j}$ denote the part of the file requested by user j that is stored in all the caches in χ except in the cache allocated at user j (c_j). Note that $c_j = j$ if $j \notin \mathcal{K}_P$, and $c_j = \Lambda$ if $j \in \mathcal{K}_P$. For each such χ, the transmit signal takes a different expression depending on whether $\Lambda \in \chi$ or not (recall that all users in $\mathcal{K}_P$ store the Λ-th cache state).

1) Case $\Lambda \notin \chi$*:* The transmit signal when the Λ-th cache group is not in χ is given by $\mathbf{X}(t) = \mathbf{v}_0(t)\sum_{g\in\chi} W_{d_g,\chi\backslash g}$, where $\mathbf{v}_0(t)$ is a randomly chosen $L \times 1$ precoding vector.

2) Case $\Lambda \in \chi$*:* In this case, we have that

$$\mathbf{X}(t) = \mathbf{v}_0(t)\sum_{g\in\{\chi\backslash\Lambda\}} W_{d_g,\chi\backslash g} + \sum_{p\in\mathcal{K}_P} \mathbf{v}_{\mathcal{K}_P\backslash p}(t) W_{d_p,\chi\backslash\Lambda}, \tag{6}$$

where $\mathbf{v}_{\mathcal{K}_P\backslash p}(t)$ is an $L \times 1$ precoding vector designed to belong to the null space of the $K_P - 1$ users in $\{\mathcal{K}_P\backslash p\}$.

C. Received Signal

1) Case $\Lambda \notin \chi$*:* The received signal at user i is given by

$$Y_i(t) = \mathbf{h}_i(t)\mathbf{v}_0(t)\Big(\sum\nolimits_{g\in\chi} W_{d_g,\chi\backslash g}\Big) + \zeta_i(t). \tag{7}$$

Note that, for any $i \in \chi$, user i can remove all the undesired messages because they are cached in Z_i. Thus, user i obtains

$$Y'_i(t) = \mathbf{h}_i(t)\mathbf{v}_0(t)W_{d_i,\chi\backslash i} + \zeta_i(t), \tag{8}$$

and it can decode the intended subfile $W_{d_i,\chi\backslash i}$.

2) Case $\Lambda \in \chi$*:* Any user not in $\mathcal{K}_P$ can decode its subfile as in the previous case, since all the non-intended subfiles are available in its local cache. For users in $\mathcal{K}_P$, they can remove the messages intended by the finite-precision-CSI users, which are available in their local cache, thus obtaining

$$Y'_i(t) = \mathbf{h}_i(t)\big(\sum\nolimits_{p\in\mathcal{K}_P} \mathbf{v}_{\mathcal{K}_P\backslash p}(t)W_{d_p,\chi\backslash\Lambda}\big) + \zeta_i(t). \tag{9}$$

[5] For sake of readability and concision, we omit the fact that the NDT is naturally upper bounded by $(1-\gamma)$; for example, in (4), it holds that $\mathrm{NDT}_u(K, L, \gamma, K_P) = \min\big((1-\gamma), \mathrm{Conv}_{(K_F+1)\gamma}\big(\frac{(K_F+1)(1-\gamma)}{(K_F+1)\gamma+1}\big)\big)$.

By definition of $\mathbf{v}_{\mathcal{K}_P\backslash p}(t)$, it holds that $\mathbf{h}_i(t)\mathbf{v}_{\mathcal{K}_P\backslash p}(t) = 0$ for any $i \in \mathcal{K}_P : i \neq p$. From that, it follows that

$$Y'_i(t) = \mathbf{h}_i(t)\mathbf{v}_{\mathcal{K}_P\backslash i}(t) W_{d_i,\chi\backslash\Lambda} + \zeta_i(t), \tag{10}$$

and thus, user i can decode its intended subfile $W_{d_i,\chi\backslash\Lambda}$.

Remark 4. *The users must decide whether they will provide perfect CSIT at the placement phase to be able to correctly assign the cache states. Yet, the optimal* NDT *is also achieved if only a subset $\mathcal{K}'_P \subset \mathcal{K}_P$ of users in $\mathcal{K}_P$ appear for delivery.*

For the case where $K_P > L$, the scheme is generalized by having $\lceil K_P/L \rceil$ cache groups with several users.

D. Degrees-of-Freedom

This scheme always serves $\Lambda\gamma + 1$ groups simultaneously. However, the number of served users depends on whether $\Lambda \in \chi$ or not: the TX serves $\Lambda\gamma + 1$ users if $\Lambda \notin \chi$, but $\Lambda\gamma + K_P$ users if $\Lambda \in \chi$. Let $\mathcal{X}_\Lambda \triangleq \{\chi \in \mathcal{X} : \Lambda \in \chi\}$ denote the set of those $\chi \in \mathcal{X}$ that include Λ. Then, it follows that

$$|\mathcal{X}_\Lambda| = |\{\chi \in \mathcal{X} : \Lambda \in \chi\}| = \binom{\Lambda - 1}{\Lambda\gamma}. \tag{11}$$

Since all the subfiles have the same size, the time required for decoding each of them is the same, no matter which user-set χ is being served. Thus, from (11) it follows that

$$\begin{aligned} \mathrm{DoF} &= \frac{\binom{\Lambda-1}{\Lambda\gamma}(\Lambda\gamma + K_P) + \left(\binom{\Lambda}{\Lambda\gamma+1} - \binom{\Lambda-1}{\Lambda\gamma}\right)(\Lambda\gamma+1)}{\binom{\Lambda}{\Lambda\gamma+1}} \\ &= (1 + \Lambda\gamma) + (K_P - 1)\frac{\Lambda\gamma + 1}{\Lambda}. \end{aligned} \tag{12}$$

By applying that $\Lambda = K - K_P + 1$ into (12), we obtain that

$$\mathrm{DoF} = 1 + K\gamma + \frac{K_P - 1}{K - K_P + 1}. \tag{13}$$

Since the NDT can be written as $\mathrm{NDT} = \frac{K(1-\gamma)}{\mathrm{DoF}}$, by applying that $\Lambda = K - K_P + 1 = K_F + 1$ into (13) yields

$$\mathrm{NDT} = \frac{\Lambda(1-\gamma)}{1 + \Lambda\gamma} = \mathrm{NDT}_u(K_F + 1, 1, \gamma, 0), \tag{14}$$

which concludes the achievability proof for the cases in which $\Lambda\gamma = (K_F + 1)\gamma$ is an integer. In other cases, the NDT can be achieved through the usual memory-sharing approach [1]–[3].

V. Converse of Theorem 1 and Theorem 2

We first present two useful lemmas that are instrumental for the converse of both theorems. This will be followed first by the converse of Theorem 1 and later by that of Theorem 2.

A. Lower Bounding the Achievable Delivery Time

In the following, we derive a lower bound for any ε-achievable delivery time. Before presenting the result, we introduce the notations $\mathcal{Z}_{[j]} \triangleq \{Z_i\}_{i\in[j]}$, $\mathcal{W}_{[j]} \triangleq \{W_{d_i}\}_{i\in[j]}$, and $\mathcal{H} \triangleq \{h_{i,j}(t)\}_{i\in[K],j\in[L],t\in[\tau]}$. We start by focusing on the "PN" problem with side information. This important setting can be characterized by the following key lemma.

Lemma 1. *Consider the $(K, L, \gamma, K_P = 1)$ MISO BC where the TX has perfect CSI only for one user. Then, it holds that*

$$\begin{aligned} H(Y_k^{[\tau]}|\mathcal{Z}_{[k]}, \mathcal{W}_{[k]}, \mathcal{H}) - H(Y_{k+1}^{[\tau]}|\mathcal{Z}_{[k]}, \mathcal{W}_{[k]}, \mathcal{H}) \\ \geq \tau\, o(\log P) \quad \forall k \in [K]. \end{aligned} \tag{15}$$

Proof. The proof follows from a new application of the Aligned Image Set approach. Due to space constraints, the proof is relegated to the extended version [12]. □

The previous lemma is instrumental in the derivation of the next lower bound on the delivery time. Before presenting the lemma, let us define $K' \triangleq \min(N, K)$.

Lemma 2. *Consider the $(K, L, \gamma, K_P = 1)$ MISO BC where the TX has perfect CSI only for one user. For any prefetching scheme ϕ and any demand **d**, the ε-achievable delivery time $T_{\varepsilon,\phi}$ is lower-bounded by*

$$\begin{aligned} &T_{\varepsilon,\phi}\big(\log(1 + KP) + o(\log P)\big) \\ &\quad \geq \frac{1}{B}\sum\nolimits_{k=1}^{K'} H(W_{d_k} \mid \mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}) - K'\Big(\frac{1}{B} + \varepsilon\Big). \end{aligned} \tag{16}$$

Proof. Let us consider that a particular delay $T_{\varepsilon,\phi}$ is ε-achievable. Then, for any request $\boldsymbol{d} = \{d_1, d_2, \ldots, d_K\}$, there exists a transmitted signal vector $\mathbf{X}_d^{[\tau]}$ such that each user $k \in [K]$ can decode W_{d_k} from Z_k and $Y_k^{[\tau]}$ with probability of error at most ε. From Fano's inequality and the fact that conditioning reduces entropy, it follows that

$$H(W_{d_k} \mid Y_k^{[\tau]}, \mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}, \mathcal{H}) \leq 1 + \varepsilon B, \tag{17}$$

for any $k \in [K]$. From the above, it follows that

$$\begin{aligned} H(Y_k^{[\tau]} \mid \mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}, \mathcal{H}) \geq H(W_{d_k} \mid \mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}, \mathcal{H}) \\ + H(Y_k^{[\tau]} \mid \mathcal{Z}_{[k]}, \mathcal{W}_{[k]}, \mathcal{H}) - (1 + \varepsilon B), \end{aligned} \tag{18}$$

for any $k \in [K]$. Summing up the previous inequality for all $k \in [K']$ and re-ordering terms yields

$$\begin{aligned} H(Y_1^{[\tau]}|Z_1, \mathcal{H}) \geq &\sum_{k=1}^{K'} H(W_{d_k}|\mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}, \mathcal{H}) - K'(1 + \varepsilon B) \\ +&\sum_{k=1}^{K'-1} H(Y_k^{[\tau]}|\mathcal{Z}_{[k]}, \mathcal{W}_{[k]}, \mathcal{H}) - H(Y_{k+1}^{[\tau]}|\mathcal{Z}_{[k+1]}, \mathcal{W}_{[k]}, \mathcal{H}). \end{aligned} \tag{19}$$

Note that the LHS of (19) can be upper-bounded as $H(Y_1^{[\tau]} \mid Z_1, \mathcal{H}) \leq \tau \log(1 + KP)$. Moreover, from the fact that conditioning reduces entropy, the RHS of (19) can be further lower-bounded by applying $H(Y_{k+1}^{[\tau]}|\mathcal{Z}_{[k+1]}, \mathcal{W}_{[k]}, \mathcal{H}) \leq H(Y_{k+1}^{[\tau]} \mid \mathcal{Z}_{[k]}, \mathcal{W}_{[k]}, \mathcal{H})$. This, together with Lemma 1, leads to

$$\begin{aligned} \sum\nolimits_{k=1}^{K'} H(W_{d_k}|\mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}) - K'(1 + \varepsilon B) - \tau\, o(\log P) \\ \leq H(Y_1^{[\tau]} \mid Z_1, \mathcal{H}) \leq \tau \log(1 + KP), \end{aligned} \tag{20}$$

where we have removed the condition on $\mathcal{H}$ because both messages and caches are independent of the channel $\mathcal{H}$. Since $\tau = T_{\varepsilon,\phi}\, B$, we obtain Lemma 2 by rearranging the terms. □

Lemmas 1-2 are the key contribution of this work, and they are essential for proving both Theorem 1 and Theorem 2. Lemma 2 can be seen from two perspectives: First, it represents the non-trivial extension of [3, Lemma 2] from the single-server error-free shared-link setting of [3] to the (K, L, γ, K_P) MISO BC with $K_P = 1$. As explained in [3], these lemmas represent an enhanced cut-set bound that improves the compound cut-set bound previously used in other works [1]. Second, it extends the results of the "PN" BC setting to the important case with receiver side information.

B. Proof of Theorem 1

We now proceed to prove Theorem 1. Some steps, especially those that draw from the proof in [2, Section V], are omitted and can be found in the extended version of this work [12].

As a first step, let us restrict ourselves to the case $K_P = 1$, and let $K' \triangleq \min(N, K)$. Before applying the derivation, recall that the bits in the library are i.i.d. and uniformly distributed. Let $B_{n,b}$ denote the b-th bit of file n, and let $\mathcal{B}^{(k)}_{d_k,b}$ represent the event that $B_{d_k,b}$ is not cached by any user in the set $\{i\}_{i=1}^{k}$. Then, it follows (cf. [2]) that

$$\sum\nolimits_{k=1}^{K'} H(W_{d_k}|\mathcal{Z}_{[k]}, \mathcal{W}_{[k-1]}) \geq \sum\nolimits_{k=1}^{K'} \sum\nolimits_{b=1}^{B} \mathbb{1}\big(\mathcal{B}^{(k)}_{d_k,b}\big) \quad (21)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Thus, by applying the same steps as in [2, Appendix A], and from Lemma 2, the value of $T^{\star}_{u,\varepsilon,\phi}$ is lower bounded by

$$\begin{aligned} &T^{\star}_{u,\varepsilon,\phi}\big(\log(1+KP)+o(\log P)\big) \\ &\qquad \geq \mathrm{Conv}_{K\gamma}\left(\frac{K(1-\gamma)}{K\gamma+1}\right) - K'\Big(\frac{1}{B}+\varepsilon\Big). \quad (22)\end{aligned}$$

Since $T^{\star}_{u} = \sup_{\varepsilon>0}\limsup_{B\to\infty}\min_{\phi} T^{\star}_{u,\epsilon,\phi}$, it follows that

$$T^{\star}_{u}\big(\log(1+KP)+o(\log P)\big) = \mathrm{Conv}_{K\gamma}\Big(\frac{K(1-\gamma)}{K\gamma+1}\Big). \quad (23)$$

Now, applying $\mathrm{NDT}_u(K, L, \gamma, 1) = \lim_{P\to\infty} T^{\star}_{u} \log P$ yields

$$\mathrm{NDT}_u(K, L, \gamma, 1) = \mathrm{Conv}_{K\gamma}\left(\frac{K(1-\gamma)}{K\gamma+1}\right), \quad (24)$$

and thus yields that $\mathrm{NDT}_u(K, L, \gamma, 1) = \mathrm{NDT}_u(K, 1, \gamma, 0)$, which proves Theorem 1 for the case $K_P = 1$.

Let us now consider the general case in which $K_P \leq L$. We start by considering the reduced setting in which we only serve the $K - K_P$ users for which there is finite precision or no CSIT and only one additional user among the K_P users providing perfect CSIT. The optimal delay on this reduced scenario will clearly lower bound the optimal delay of the whole setting. This optimal delay of the reduced scenario is given by (24) after setting the total number of users to be $K - K_P + 1$. Then, it follows from (24) that

$$\begin{aligned}\mathrm{NDT}_u(K, L, \gamma, K_P) &\geq \mathrm{NDT}_u(K - K_P + 1, L, \gamma, 1) \quad (25)\\ &= \mathrm{NDT}_u(K - K_P + 1, 1, \gamma, 0), \quad (26)\end{aligned}$$

which concludes the proof of Theorem 1. □

C. Proof of Theorem 2

From the derivation of Lemma 2, the converse of Theorem 2 follows a similar path as the proof of Theorem 2 in [3] for the error-free symmetric setting (see [3, Section V]). Consequently, for the sake of space, its details will be found in the extended version of this work in [12].

VI. Conclusions

We have considered the cache-aided $L \times K$ MISO BC where only a fixed set of $K_P \leq L$ users provide perfect CSI to the transmitter, whereas the remaining users provide finite precision CSIT. This setting also corresponds to the side-information variant of the well-known "PN" BC setting. For this setting, we have derived the optimal rate-memory trade-off at high SNR under uncoded placement, and characterized the same trade-off within a factor of 2.01 for coded placement. The proposed scheme capitalizes on the fact that the optimal trade-off cannot be achieved through separate transmission between the users that provide different CSIT level. The derived limits clearly show that coded-caching and multi-antenna gains are synergistic and schemes should integrate both aspects, but also that the existence of users that provide only finite precision CSIT greatly reduces the performance, which is in line with the results of the "PN" setting without side information. Analyzing the case $K_P > L$ is a meaningful extension currently under investigation, and finally, considering partial or heterogeneous CSI settings are also interesting research directions.

References

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

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

[3] ——, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 647–663, Jan 2019.

[4] H. Joudeh, E. Lampiris, P. Elia, and G. Caire, "Fundamental limits of wireless caching under mixed cacheable and uncacheable traffic," in *2020 IEEE Int. Symp. on Inf. Theory (ISIT)*, 2020, pp. 1693–1698.

[5] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized DoF of the symmetric cache-aided MISO Broadcast Channel with partial CSIT," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5799–5815, Sep. 2019.

[6] 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.

[7] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," 2020. [Online]. Available: arxiv.org/abs/1811.03935

[8] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1176–1188, June 2018.

[9] E. Lampiris, J. Zhang, and P. Elia, "Cache-aided cooperation with no CSIT," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 2960–2964.

[10] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.

[11] A. G. Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling conjectures on the collapse of Degrees of Freedom under finite precision CSIT," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5603–5618, Oct 2016.

[12] A. Bazco-Nogueras and P. Elia, "Rate-memory trade-off for the cache-aided MISO Broadcast Channel with hybrid CSIT," 2020. [Online]. Available: arxiv.org/abs/2010.13109

[13] R. Tandon, S. A. Jafar, S. Shamai, and H. V. Poor, "On the synergistic benefits of alternating CSIT for the MISO Broadcast Channel," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4106–4128, July 2013.

[14] S. Lashgari, R. Tandon, and S. Avestimehr, "MISO Broadcast Channel with hybrid CSIT: Beyond two users," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7056–7077, Dec 2016.

[15] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, 2017.