Multicast Beamformer Design for Coded Caching

Antti Tölli⋆, Seyed Pooya Shariatpanahi∗, Jarkko Kaleva⋆ and Babak Khalaj†
⋆ Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014, Finland
∗ School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
† Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.
firstname.lastname@oulu.fi, pooya@ipm.ir, khalaj@sharif.edu

***Abstract*—A single cell downlink scenario is considered where a multiple-antenna base station delivers contents to cache-enabled user terminals. Using the ideas from multi-server coded caching (CC) scheme developed for wired networks, a joint design of CC and general multicast beamforming is considered to benefit from spatial multiplexing gain, improved interference management and the global CC gain, simultaneously. The proposed multicast beamforming strategies utilize the multiantenna multicasting opportunities provided by the CC technique and optimally balance the detrimental impact of both noise and inter-stream interference from coded messages transmitted in parallel. The proposed scheme is shown to provide the same degrees-of-freedom at high SNR as the state-of-art methods and, in general, to perform significantly better than several baseline schemes including, the joint zero forcing and CC, max-min fair multicasting with CC, and basic unicasting with multiuser beamforming.**

I. Introduction

The pioneering work of [1] considers an information theoretic framework for the caching problem, through which a novel *coded caching* (CC) scheme is proposed. Instead of simply replicating high-popularity contents near-or-at end-users, one should spread different contents at different caches. At the content delivery phase, common coded messages could be then broadcast to different users with different demands, that would benefit all of the users, providing *global caching gain*. Follow-up works extend the coded caching scheme from [1] to other setups such as online coded caching [2], hierarchical coded caching [3], and multi-server scenarios [4].

The specific characteristics of wireless networks must be investigated to be able to implement the original idea of [1] in mobile delivery scenarios. To this end, the authors in [5] consider the effect of delayed channel state information at the transmitter (CSIT). Moreover, the work [6] considers a cache-enabled wireless interference channel while [7] treats the same setup with mixed-CSIT. In addition, the authors in [8] investigate coded caching schemes in wireless device-to-device networks. Furthermore, [9] assumes interference management in a cellular setup with coded caching, and [10] considers using the rate-splitting idea along with coded caching. In all these papers, the analysis is for the high signal-to-noise-ratio (SNR) regime, and in terms of degrees-of-freedom (DoF).

Different CC schemes for a wireless multiple-input single-output broadcast channel (MISO-BC) are proposed in [11] and [12] providing a finite SNR analysis in different system operating regimes. While the main idea in [11] is to use rate-splitting along with CC, the authors in [12], [13] propose a joint design of CC and zero-forcing (ZF) to benefit from the spatial multiplexing gain and the global gain of CC, at the same time. While the ideas in [12], [13] originally came from adapting the multi-server CC scheme of [4] (which is almost optimal in terms of DoF as shown in [6]) to a Gaussian MISO-BC, the interesting observations in [13] reveal that careful code and beamformer design modifications should be further considered for improving the finite SNR performance.

In this paper, extending the joint interference nulling and CC concept originally proposed in [12], [13], a joint design of CC and generic multicast beamforming is introduced to benefit from spatial multiplexing gain, improved management of inter-stream interference from coded messages transmitted in parallel, and the global caching gain, simultaneously. The general signal-to-interference-plus-noise ratio (SINR) expressions are handled directly to optimally balance the detrimental impact of both noise and inter-stream interference at low SNR. As the resulting optimization problems are not necessarily convex, successive convex approximation (SCA) methods are used to devise efficient iterative algorithms similarly to existing multicast beamformer design solutions [14].

II. System Model

Downlink transmission from a single L-antenna BS serving K cache enabled single-antenna users is considered. The BS is assumed to have access to a library of N files $\{W_1, \ldots, W_N\}$, each of size F bits. Every user is assumed to be equipped with a cache memory of MF bits. Furthermore, each user k has a message $Z_k = Z_k(W_1, \ldots, W_N)$ stored in its cache, where $Z_k(\cdot)$ denotes a function of the library files with entropy not larger than MF bits. This operation is referred to as the *cache content placement*, and it is performed once and at no cost, e.g. during network off-peak hours.

Upon a set of requests $d_k \in [1:N]$ at the *content delivery* phase, the BS multicasts coded signals, such that at the end of transmission all users can reliably decode their requested files. Notice that user k decoder, in order to produce the decoded file $\widehat{W}_{d_k}$, makes use of its own cache content Z_k as well as of its own received signal from the wireless channel.

The received signal at user terminal $k = 1, \ldots, K$ at time instant $i, i = 1, \ldots, n$ can be written as

$$y_k = \mathbf{h}_k^{\mathrm{H}} \sum_{\mathcal{T} \subseteq \mathcal{S}} \mathbf{w}_{\mathcal{T}}^{\mathcal{S}} \tilde{X}_{\mathcal{T}}^{\mathcal{S}}(i) + z_k, \tag{1}$$

where the channel vector between the BS and UE k is denoted by $\mathbf{h}_k \in \mathbb{C}^L$, $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ denotes the multicast beamformer dedicated

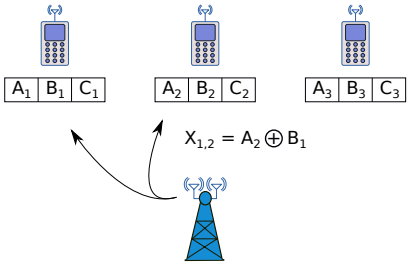

Fig. 1. *Scenario 1*: $L = 2$, $K = 3$, $N = 3$ and $M = 1$

to users in subset $\mathcal{T}$ of set $\mathcal{S} \subseteq [1 : K]$ of users, and $\tilde{X}^{\mathcal{S}}_{\mathcal{T}}(i)$ is the corresponding multicast message chosen from a unit power complex Gaussian codebook at time instant i. In the following, the time index i is ignored for simplicity. The receiver noise is assumed to be circularly symmetric zero mean $z_k \sim \mathcal{CN}(0, N_0)$. Finally, the CSIT of all K users is assumed to be perfectly known at the BS.

III. Multiantenna Coded Caching for Finite SNR

In this work, we focus on the worst-case (over the users) delivery rate at which the system can serve any users requesting any file of the library. In the following, for the sake of easy exposure, we introduce the basic multiantenna multicast beamforming concept for two simple scenarios and discuss the generalization of the proposed scheme afterwards.

A. Scenario 1: $L \geq 2$, $K = 3$, $N = 3$ and $M = 1$

Consider a content delivery scenario illustrated in Fig. 1, where a transmitter with $L \geq 2$ antennas should deliver requests arising at $\mathcal{S} = \{1, 2, 3\}$ users from a library $\mathcal{W} = \{A, B, C\}$ of size $N = 3$ files each of F bits. Now, each user can cache $M = 1$ files of F bits in the cache content placement phase, without knowing the actual requests beforehand. In the content delivery phase we suppose each user requests one file from the library. Following the same cache content placement strategy as in [1] the cache contents of users are as follows

$$Z_1 = \{A_1, B_1, C_1\}, Z_2 = \{A_2, B_2, C_2\}, Z_3 = \{A_3, B_3, C_3\}$$

where each file is divided into 3 equal-sized subfiles.

Suppose that the 1st, the 2nd, and the 3rd user request files A, B, and C, respectively. In the simple broadcast scenario in [1], the following coded messages are sent one after another

$$X_{1,2} = A_2 \oplus B_1, \;\; X_{1,3} = A_3 \oplus C_1, \;\; X_{2,3} = B_3 \oplus C_2 \quad (2)$$

where $\oplus$ represents summation in the corresponding finite field. In this coding scheme of [1], each coded message is heard by all the 3 users, but is only beneficial to 2 users. This multicasting gain is called as the *Global Caching Gain*. It can be easily checked that after the transmission is concluded all the users can decode their requested files.

Now, in the given *Scenario 1* we can combine the spatial multiplexing gain, and the global caching gain following the scheme from [12] (see also [4], [6]). In [12], the unwanted messages at each user are forced to zero by sending

$$\mathbf{h}_3^{\perp}\tilde{X}_{1,2} + \mathbf{h}_2^{\perp}\tilde{X}_{1,3} + \mathbf{h}_1^{\perp}\tilde{X}_{2,3} \quad (3)$$

where $\tilde{X}$ stands for the modulated version of X, chosen from a unit power complex Gaussian codebook [12].

In this paper, instead of nulling interference at unwanted users, general multicast beamforming vectors are defined as[1]

$$\sum_{\mathcal{T}\subseteq[3],|\mathcal{T}|=2} \mathbf{w}^{\mathcal{S}}_{\mathcal{T}}\tilde{X}^{\mathcal{S}}_{\mathcal{T}} = \mathbf{w}_{1,2}\tilde{X}_{1,2} + \mathbf{w}_{1,3}\tilde{X}_{1,3} + \mathbf{w}_{2,3}\tilde{X}_{2,3}. \quad (4)$$

Then, the received signals at users $1-3$ will be

$$\begin{aligned}
y_1 &= \underline{(\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2}} + \underline{(\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3}} + (\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3} + z_1\\
y_2 &= \underline{(\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2}} + (\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3} + \underline{(\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3}} + z_2\\
y_3 &= (\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2} + \underline{(\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3}} + \underline{(\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3}} + z_3
\end{aligned}$$

where the desired terms for each user are underlined. Let us focus on user 1 who is interested in decoding both $\tilde{X}_{1,2}$, and $\tilde{X}_{1,3}$ while $\tilde{X}_{2,3}$ appears as Gaussian interference. Thus, from receiver 1 perspective, y_1 is a Gaussian Multiple Access Channel (MAC). Suppose now user 1 can decode *both* of its required messages $\tilde{X}_{1,2}$ and $\tilde{X}_{1,3}$ with the equal rate[2]

$$R^1_{MAC} = \min(\frac{1}{2}R^1_{Sum}, R^1_1, R^1_2) \quad (5)$$

where the rate bounds R^1_1 and R^1_2 correspond to $\tilde{X}_{1,2}$, and $\tilde{X}_{1,3}$, respectively, and R^1_{Sum} is the sum rate of both messages. Thus, the total useful rate is $2R^1_{MAC}$. Since the user 1 must receive the missing $2/3F$ bits (A_2 and A_3), the time needed to decode file A is $T_1 = \frac{2F}{3}\frac{1}{2R^1_{MAC}}$. As all the users decode their files *in parallel*, the time needed to complete the decoding process is constrained by the worst user as

$$T = \frac{2F}{3}\frac{1}{\min\limits_{k=1,2,3} 2R^k_{MAC}}. \quad (6)$$

Then, the *Symmetric Rate (Goodput) per user* will be

$$R_{sym} = \frac{F}{T} = 3\min_{k=1,2,3} R^k_{MAC} \quad (7)$$

which, when optimized with respect to the beamforming vectors, can be found as

$$\begin{aligned}
&\max_{R^k,\gamma^k_i,\mathbf{w}_{\mathcal{T}}} \min_{k=1,2,3} \min\left(\frac{1}{2}R^k_{\text{sum}}, R^k_1, R^k_2\right)\\
&\text{s. t. } R^k_1 \leq \log(1+\gamma^k_1), R^k_2 \leq \log(1+\gamma^k_2),\\
&\quad R^k_{\text{sum}} \leq \log(1+\gamma^k_1+\gamma^k_2),\ k = 1,2,3,\\
&\quad \gamma^1_1 \leq \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2}|^2}{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2+N_0}, \gamma^1_2 \leq \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,3}|^2}{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2+N_0},\\
&\quad \gamma^2_1 \leq \frac{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{2,3}|^2}{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3}|^2+N_0}, \gamma^2_2 \leq \frac{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,2}|^2}{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3}|^2+N_0},\\
&\quad \gamma^3_1 \leq \frac{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{2,3}|^2}{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2+N_0}, \gamma^3_2 \leq \frac{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,3}|^2}{|\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2+N_0},\\
&\quad \textstyle\sum_{\mathcal{T}\in\{\{1,2\},\{1,3\},\{2,3\}\}} \|\mathbf{w}_{\mathcal{T}}\|^2 \leq \text{SNR}.
\end{aligned} \quad (8)$$

Problem (8) is non-convex due to the SINR constraints. Similarly to [14], successive convex approximation (SCA) approach can be used to devise an iterative algorithm that is able to converge to a local solution. To begin with, the SINR constraint for γ^1_1 can be reformulated as

$$|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + N_0 \leq \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + N_0}{1+\gamma^1_1}. \quad (9)$$

[1]Superscript $\mathcal{S}$ is omitted as all $K = 3$ users are served in a single set $\mathcal{S}$.

[2]Symmetric rate is imposed to minimize the time needed to receive both messages $\tilde{X}_{1,2}$, and $\tilde{X}_{1,3}$.

Now, the R.H.S of (9) is a convex quadratic-over-linear function and it can be linearly approximated (lower bounded) as

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{2,3},\mathbf{w}_{1,2},\gamma_1^1) \triangleq\ & |\mathbf{h}_1^{\mathrm{H}}\bar{\mathbf{w}}_{1,2}|^2 + |\mathbf{h}_1^{\mathrm{H}}\bar{\mathbf{w}}_{2,3}|^2 + N_0 \\ & - 2\mathbb{R}\left(\bar{\mathbf{w}}_{1,2}^{\mathrm{H}}\mathbf{h}_1\mathbf{h}_1^{\mathrm{H}}\left(\mathbf{w}_{1,2}-\bar{\mathbf{w}}_{1,2}\right)\right) \\ & - 2\mathbb{R}\left(\bar{\mathbf{w}}_{2,3}^{\mathrm{H}}\mathbf{h}_1\mathbf{h}_1^{\mathrm{H}}\left(\mathbf{w}_{2,3}-\bar{\mathbf{w}}_{2,3}\right)\right) \\ & + \frac{|\mathbf{h}_1^{\mathrm{H}}\bar{\mathbf{w}}_{1,2}|^2 + |\mathbf{h}_1^{\mathrm{H}}\bar{\mathbf{w}}_{2,3}|^2 + N_0}{1+\bar{\gamma}_1^1}\left(\gamma_1^1 - \bar{\gamma}_1^1\right) \end{aligned} \quad (10)$$

where $\bar{\mathbf{w}}_{k,i}$ and $\bar{\gamma}_1^1$ denote the fixed values (points of approximation) for the corresponding variables from the previous iteration. Using (10) and reformulating the objective in the epigraph form, the approximated problem is written as the approximated problem is written as

$$\begin{aligned}\max_{t,\gamma^k,\mathbf{w}_{\mathcal{T}}}\ & t \\ \text{s. t.}\ & t \le 1/2\log(1+\gamma_1^k+\gamma_2^k),\ k=1,2,3, \\ & t \le \log(1+\gamma_1^k),\ t \le \log(1+\gamma_2^k)\ \forall\ k, \\ & \mathcal{L}(\mathbf{w}_{2,3},\mathbf{w}_{1,2},\gamma_1^1) \ge |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + N_0, \\ & \ \ \vdots \text{ In total 6 SINR constraints} \\ & \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{1,3},\gamma_2^3) \ge |\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + N_0, \\ & \textstyle\sum_{\mathcal{T}\in\{\{1,2\},\{1,3\},\{2,3\}\}} \|\mathbf{w}_{\mathcal{T}}\|^2 \le \text{SNR}\end{aligned} \quad (11)$$

This is a convex problem that can be readily solved using existing convex solvers. However, the logarithmic functions require further approximations to be able to apply the convention of convex programming algorithms. Problem (11) can be equally formulated as computationally efficient second order cone problem (SOCP), as shown in the extended version [15]. Finally, a solution for the original problem (8) can be found by solving (11) in an iterative manner using SCA, i.e, by updating the points of approximations $\bar{\mathbf{w}}_{k,i}$ and $\bar{\gamma}_j^l$ in (10) after each iteration. As each difference-of-convex constraint in (9) is lower bounded by (10), the monotonic convergence of the objective of (11) is guaranteed. Note that the final symmetric rates are achieved by time sharing (or rate splitting) between the rate allocations corresponding to different points (decoding orders) in the sum rate region of the MAC channel.

As a lower complexity alternative, a zero forcing solution, denoted as *CC with ZF*, is also proposed[3]. By assigning $\mathbf{w}_{1,2} = \mathbf{h}_3^{\perp}/\|\mathbf{h}_3^{\perp}\|\sqrt{p_{1,2}}$, $\mathbf{w}_{1,3} = \mathbf{h}_2^{\perp}/\|\mathbf{h}_2^{\perp}\|\sqrt{p_{1,3}}$, $\mathbf{w}_{2,3} = \mathbf{h}_1^{\perp}/\|\mathbf{h}_1^{\perp}\|\sqrt{p_{2,3}}$, the interference terms are canceled and (8) becomes:

$$\max_{R^k,\gamma^k,p_{\mathcal{T}}} \min_{k=1,2,3} \min\left(\frac{1}{2}R_{\text{sum}}^k, R_1^k, R_2^k\right) \quad (12)$$

$$\begin{aligned}\text{s. t.}\ & R_{\text{sum}}^k \le \log(1+\gamma_1^k+\gamma_2^k)\ \forall\ k, \\ & R_1^k \le \log(1+\gamma_1^k), R_2^k \le \log(1+\gamma_2^k)\ \forall\ k, \\ & \gamma_1^1 \le u_{1,3}p_{1,2}, \gamma_2^1 \le u_{1,2}p_{1,3}, \gamma_1^2 \le u_{2,1}p_{2,3}, \\ & \gamma_2^2 \le u_{2,3}p_{1,2}, \gamma_1^3 \le u_{3,1}p_{2,3}, \gamma_2^3 \le u_{3,2}p_{1,3}, \\ & \textstyle\sum_{\mathcal{T}\in\{\{1,2\},\{1,3\},\{2,3\}\}} p_{\mathcal{T}} \le \text{SNR}\end{aligned} \quad (13)$$

where $u_{k,i} = |\mathbf{h}_k^{\mathrm{H}}\mathbf{h}_i^{\perp}|^2/\|\mathbf{h}_i^{\perp}\|^2 N_0$. This is readily a convex power optimization problem with three real valued variables, and hence it can be solved in an optimal manner.

In the following, three baseline reference cases for the proposed multiantenna caching scheme are introduced.

1) 1st Baseline Scheme: CC with ZF (equal power) [12]: If the multicast transmit powers are made equal, $p_{1,2} = p_{1,3} = p_{2,3} = SNR/3$, the resulting scheme is the same as originally published in [12].

2) 2nd Baseline Scheme: MaxMinSNR Multicasting: The message $X_{1,2}$ is multicast to the users 1 and 2, *without any interference* (orthogonally), by sending the signal $\mathbf{w}\tilde{X}_{1,2}$. A single transmit beamformer is found to minimize the time needed for multicasting the common message:[4]

$$T_{1,2} = \frac{F/3}{\max\limits_{\|\mathbf{w}\|^2\le SNR}\min\left(\log(1+\frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}|^2}{N_0}), \log(1+\frac{|\mathbf{h}_2^{\mathrm{H}}\mathbf{w}|^2}{N_0})\right)} \quad (14)$$

Similarly, the messages $X_{1,3}$ and $X_{2,3}$ should be delivered to the users with corresponding times $T_{1,3}$ and $T_{2,3}$. Finally the resulting symmetric rate (Goodput) per user will be

$$R_{\text{maxmin}} = F/(T_{1,2}+T_{1,3}+T_{2,3}). \quad (15)$$

Note that, in this scheme, only the coded caching gain is exploited, while the multiple transmit antennas are used just for the beamforming gain.

3) 3rd Baseline Scheme: MaxMinRate Unicast: In this scheme, only the local caching gain is exploited and the CC gain is ignored altogether. The BS simply sends two parallel independent streams to the users at each time instant.

Now, let us consider users 1 and 2 in time slot 1. The transmitted signal to deliver A_2 and B_1 to users 1 and 2, respectively, is given as $\mathbf{w}_1\tilde{A}_2 + \mathbf{w}_2\tilde{B}_1$. Thus the delivery time of $F/3$ bits is

$$T_{1,2} = \frac{F/3}{\max\limits_{\sum_{k=1,2}\|\mathbf{w}_k\|^2\le SNR}\min(R_1,R_2)} \quad (16)$$

where

$$R_k = \log\left(1+\frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k|^2}{\sum_{i\neq k}|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_i|^2+N_0}\right). \quad (17)$$

The minimum delivery time in (15) can be equivalently formulated as a maxmin SINR problem and solved optimally. By repeating the same procedure for the subsets $\{1,3\}$ and $\{2,3\}$, the symmetric rate expression is equivalent to (15).

B. Scenario 2: $L \ge 3$, $K = 4$, $N = 4$ and $M = 1$

In this scenario, we assume that the BS transmitter has $L \ge 3$ antennas, and there are $K = 4$ users each with cache size $M = 1$, requesting files from a library $\mathcal{W} = \{A, B, C, D\}$ of $N = 4$ files. Following the same cache content placement strategy as in [1] the cache contents of users are as follows

$$Z_1 = \{A_1, B_1, C_1, D_1\}, Z_2 = \{A_2, B_2, C_2, D_2\}$$
$$Z_3 = \{A_3, B_3, C_3, D_3\}, Z_4 = \{A_4, B_4, C_4, D_4\}$$

where here each file is divided into four non-overlapping equal-sized subfiles.

[3]Note that the null space beamformer is unique only when $L = 2$. Generic multicast beamformers can be designed within the interference free signal space when $L > 2$ (See Section IV).

[4]This multicast maxmin problem is NP-hard in general, but near-optimal solutions can be obtained by a semidefinite relaxation (SDR) approach, see [12] and the references therein.

At the content delivery phase, suppose that the users $1-4$ request files $A-D$, respectively. Here, we have $t \triangleq KM/N = 1$ and the subsets $\mathcal{S}$ and $\mathcal{T}$ will be of size $t+L=4$, and $t+1=2$, respectively (for details see [4]). Following the approach of *Scenario 1*, the transmit signal vector is

$$\sum_{\mathcal{T}\subseteq\mathcal{S},|\mathcal{T}|=2} \mathbf{w}_{\mathcal{T}}\tilde{X}_{\mathcal{T}} = \mathbf{w}_{1,2}\tilde{X}_{1,2} + \mathbf{w}_{1,3}\tilde{X}_{1,3} + \mathbf{w}_{1,4}\tilde{X}_{1,4} + \mathbf{w}_{2,3}\tilde{X}_{2,3} + \mathbf{w}_{2,4}\tilde{X}_{2,4} + \mathbf{w}_{3,4}\tilde{X}_{3,4} \quad (18)$$

where

$$\begin{aligned} X_{1,2} &= A_2 \oplus B_1,\ X_{1,3} = A_3 \oplus C_1, X_{1,4} = A_4 \oplus D_1, \\ X_{2,3} &= B_3 \oplus C_2,\ X_{2,4} = B_4 \oplus D_2,\ X_{3,4} = C_4 \oplus D_3 \end{aligned}$$

It can be easily verified that if each multicast message $X_{\mathcal{T}}$ is delivered to all the members of $\mathcal{T}$ then all the users can decode their requested files.

The received signals at each user $k=1,2,3,4$ are

$$\begin{aligned} y_1 &= \underline{(\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2}} + \underline{(\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3}} + \underline{(\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,4})\tilde{X}_{1,4}} \\ &\quad + (\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3} + (\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,4})\tilde{X}_{2,4} + (\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{3,4})\tilde{X}_{3,4} + z_1 \\ y_2 &= \underline{(\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2}} + (\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3} + (\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{1,4})\tilde{X}_{1,4} \\ &\quad + \underline{(\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3}} + \underline{(\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{2,4})\tilde{X}_{2,4}} + (\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{3,4})\tilde{X}_{3,4} + z_2 \\ y_3 &= (\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2} + \underline{(\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3}} + (\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,4})\tilde{X}_{1,4} \\ &\quad + \underline{(\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3}} + (\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{2,4})\tilde{X}_{2,4} + \underline{(\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{3,4})\tilde{X}_{3,4}} + z_3 \\ y_4 &= (\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2} + (\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3} + \underline{(\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,4})\tilde{X}_{1,4}} \\ &\quad + (\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3} + \underline{(\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{2,4})\tilde{X}_{2,4}} + \underline{(\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{3,4})\tilde{X}_{3,4}} + z_4 \end{aligned}$$

where the desired terms are underlined. Thus, as in *Scenario 1*, each user faces a MAC channel with three desired signals, three Gaussian interference terms, and one noise term. Suppose that user k can decode each of its desired signals with the rate R_{MAC}^k. Then this user receives useful information with the rate $3R_{MAC}^k$, and the time required to fetch the entire file is $T_1 = \frac{3F}{4}\frac{1}{3R_{MAC}^k}$. Following the same steps as in (6)–(7), the symmetric rate per user can be found as

$$R_{sym} = \frac{F}{T} = 4 \max_{\mathbf{w}_{\mathcal{T}},\mathcal{T}\subseteq[4],|\mathcal{T}|=2} \min_{k=1,2,3,4} R_{MAC}^k \quad (19)$$

where we have used the notation $[i] \triangleq \{1,\ldots,i\}$. As the 3-dimensional MAC rate region for each user is formed by 7 rate constraints, the following optimization problem is solved to find the symmetric rate per stream:

$$\begin{aligned} \max_{\mathbf{w}_{\mathcal{T}},\mathcal{T}\subseteq[4],|\mathcal{T}|=2} \quad & t \\ \text{s.t.} \quad & t \le \log(1+\gamma_1^k), t \le \log(1+\gamma_2^k), t \le \log(1+\gamma_3^k) \\ & t \le 1/2\log(1+\gamma_1^k+\gamma_2^k), t \le 1/2\log(1+\gamma_1^k+\gamma_3^k), \\ & t \le 1/2\log(1+\gamma_2^k+\gamma_3^k), \qquad \forall\ k=1,2,3,4 \\ & t \le 1/3\log(1+\gamma_1^k+\gamma_2^k+\gamma_3^k) \\ & \gamma_1^1 \le \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2}|^2}{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,4}|^2 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{3,4}|^2 + N_0} \\ & \vdots \text{ In total 12 SINR constraints} \\ & \gamma_3^4 \le \frac{|\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{3,4}|^2}{|\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + |\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,3}|^2 + |\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + N_0} \\ & \sum_{\mathcal{T}\subseteq[4],|\mathcal{T}|=2} \|\mathbf{w}_{\mathcal{T}}\|^2 \le SNR \end{aligned}$$

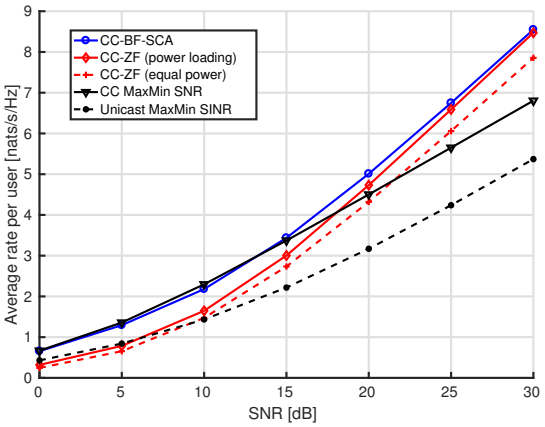

Fig. 2. Coded caching with multiantenna transmission, $L=2$ and $K=3$

In order to solve the above non-convex problem, the SCA method is again used and the SINR constraints are approximated similarly to (9)–(10).

The baseline schemes for this scenario are detailed in [15].

C. General K, L, N and M

The general guidelines for constructing multicast messages for multi-server/antenna coded caching with any K, L, N and M are given in [15], [13]. For example, the case $L=4$, $K=5$, $N=5$ and $M=1$ would require altogether $\binom{5}{2}=10$ multicast messages and each user should be able to decode 4 multicast messages. Thus, the total number of rate constraints would be 5×15 while the number of SINR constraints would be 20. In general, if K, L, N are increased with the same ratio, the number of rate constraints grows exponentially, i.e., $2^{(K-1)}-1$ per user [15]. As an efficient way to limit the complexity of the problem (with a certain performance loss at high SNR), we may limit the size of subsets benefiting from a common transmitted signal to *three*, for example. Consequently, the number of decoded multicast messages at each user would be two, as in (8). This can be achieved by splitting each subfile into minifiles and using slotted transmission to serve different user combinations in each slot. More detailed development of the reduced complexity scheme is included in the journal version of this work [15, Section IV].

IV. Numerical Examples

The numerical examples are simulated for *Scenarios 1* and *2*. The channels are considered to be i.i.d. complex Gaussian. The average performance is attained over 500 independent channel realizations. The SNR is defined as $\frac{P}{N_0}$, where P is the power budget and $N_0=1$ is the fixed noise floor.

Fig. 2 shows the performance of the interference coordination with CC in *Scenario 1*, with $K=3$ users and $L=2$ antennas. It can be seen that the proposed CC-BF-SCA scheme achieves $3-5$ dB gain at low SNR as compared to the ZF with equal power loading [12]. At high SNR, the ZF with optimal power loading in (12) achieves comparable performance while other schemes have significant performance gap. At low SNR regime, the simple MaxMin SNR multicasting with CC has similar performance as the proposed CC-BF-SCA scheme. This is due to the fact that, at low SNR, an efficient strategy for

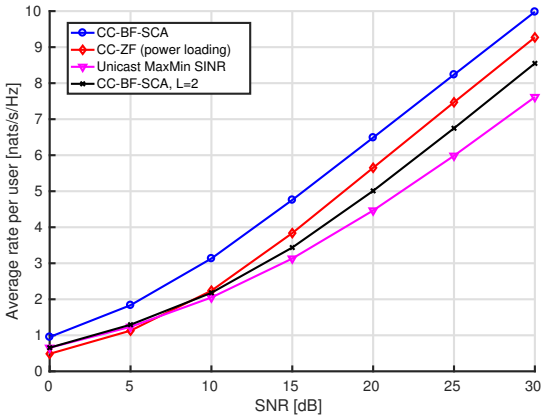

Fig. 3. Coded caching with multiantenna transmission, $L = 3$ and $K = 3$

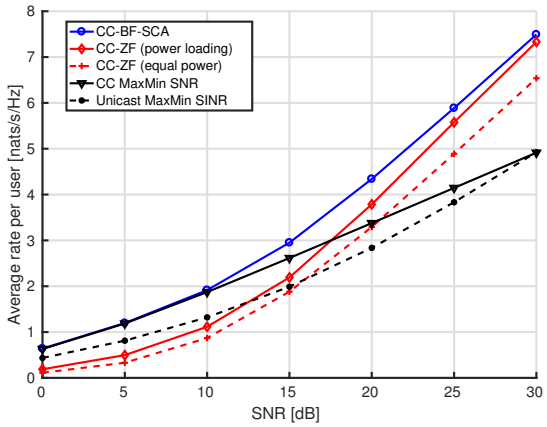

Fig. 4. Coded caching with multiantenna transmission, $L = 3$ and $K = 4$

beamforming is to concentrate all available power to a single (multicast) stream at a time and to serve different users/streams in TDMA fashion. Due to simultaneous global CC gain and inter-stream interference handling, both CC-BF-SCA and CC-ZF schemes achieve an additional DoF, which was already shown (for high SNR) in [4], [12]. The unicasting scheme does not perform well in this scenario as it does not utilize the global caching gain (only the local cache) and the spatial DoF is limited to two.

In Fig. 3, the number of transmit antennas is increased to $L = 3$. This provides more than 3dB additional gain for the CC-BF at low SNR, when compared to the $L = 2$ antenna scenario, while the DoF is the same for all the compared schemes. The optimal ZF multicast beamformer solution is no longer trivial, as the additional antenna makes the interference free signal space two-dimensional for the ZF schemes. A heuristic solution is used where orthogonal projection is first employed to get interference free signal space and then the strongest eigenvector of the stacked user channel matrix, projected to null space, is used to get a sufficiently good direction within the interference free signal space. It can be seen that the ZF scheme does achieve the same DoF as CC-BF method, but there is a constant performance gap at high SNR. Interestingly, the CC-BF scheme with $L = 2$ antennas has better performance than MaxMin SINR unicast with $L = 3$ antennas. Both schemes have the same DoF, but the global caching gain is more beneficial than the additional spatial DoF of the unicast method.

The performance of different schemes in *Scenario 2* is illustrated in Fig. 4. The CC-BF-SCA achieves $5 - 7$dB gain as compared to CC-ZF scheme at low SNR, which is considerably more than in the less complex *Scenario 1*. Again, at high SNR, the CC-ZF with optimal power loading provides comparable performance with lower computational complexity. Similarly, the performance of the MaxMin SNR multicasting overlaps with CC-BF-SCA at low SNR but there is a large gap at high SNR due to the DoF difference.

V. Conclusions

Multicasting opportunities provided by caching at user terminal were utilized to devise an efficient multiantenna transmission with CC. General multicast beamforming strategies were employed with CC, optimally balancing the detrimental impact of both noise and inter-stream interference from coded messages transmitted in parallel. The scheme was shown to perform significantly better than several baseline schemes over the entire SNR region.

References

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

[2] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, Apr 2016.

[3] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3212–3229, Jun 2016.

[4] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inform. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec 2016.

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

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

[7] M. A. T. Nejad, S. P. Shariatpanahi, and B. H. Khalaj, "On storage allocation in cache-enabled interference channels with mixed CSIT," in *2017 IEEE ICC Workshops*, May 2017, pp. 1177–1182.

[8] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inform. Theory*, vol. 62, no. 2, pp. 849–869, Feb 2016.

[9] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Cache-aided interference management in wireless cellular networks," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–7.

[10] E. Piovano, H. Joudeh, and B. Clerckx, "On coded caching in the overloaded MISO broadcast channel," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun 2017, pp. 2795–2799.

[11] K. H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–1, 2017.

[12] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna coded caching," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun 2017, pp. 2113–2117.

[13] ——, "Physical-layer schemes for wireless coded caching," *CoRR*, vol. abs/1711.05969, 2017. [Online]. Available: http://arxiv.org/abs/1711.05969

[14] G. Venkatraman, A. Tölli, M. Juntti, and L. N. Tran, "Multigroup multicast beamformer design for MISO-OFDM with antenna selection," *IEEE Trans. Signal Processing*, vol. 65, no. 22, pp. 5832–5847, Nov 2017.

[15] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *CoRR*, vol. abs/1711.03364, 2017. [Online]. Available: http://arxiv.org/abs/1711.03364