

# Multicast Mode Selection for Multi-antenna 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 wireless coded caching (CC) setup is considered, where a multi-antenna transmitter delivers contents to multiple cache-enabled users. Exploiting multicasting opportunities provided by the coded caching paradigm, novel interference management schemes are proposed by assigning carefully designed beamforming vectors to different multicast messages. Thereby, the proposed design benefits from spatial multiplexing gain, improved interference management and the global CC gain, simultaneously. In addition, a novel multicast mode selection scheme is proposed which determines the optimum multicast group sizes providing the best complexity-performance trade-off for a given SNR range. While the proposed scheme exhibits the same near-optimal degrees-of-freedom (DoF) performance as previously proposed methods, it will surpass them at the practical finite SNR regimes. In addition to reducing the complexity, the proposed mode selection feature also provides significantly better rate than previously proposed schemes.

## 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. In the coded caching scheme the idea is that, by carefully designing cache contents, common coded messages could be broadcast to benefit different users with different demands, i.e., resulting the so-called *global caching gain*.

Coded caching is an attractive proposal in mobile delivery scenarios in wireless networks. In order to investigate the specific characteristics of wireless networks in designing the coded caching schemes several papers have assumed different wireless network scenarios such as [2]–[6]. However, in these papers the analysis is for the high signal-to-noise-ratio (SNR) regime, and in terms of degrees-of-freedom (DoF), which is not always a good indicator for practical implementations performance. Thus, there is still a gap which should be filled in with finite SNR analysis of the CC idea. Cache-enabled broadcast channels with a single-antenna transmitter at the finite SNR regime was explored in [7], [8]. Moreover, [9] and [10] propose different CC schemes in a wireless multiple-input single-output broadcast channel (MISO-BC), and provide a finite SNR analysis, in different system operating regimes. The main idea in [9] is to use rate-splitting along with CC, however, the authors in [10] 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 [10] originally came from adapting the multi-server CC scheme of [11] (which is almost optimal in terms of DoF [3]) to a Gaussian MISO-BC, the interesting observations in [10] reveal that careful code and beamformer design modifications should be further considered having significant effects on the finite SNR performance.

In this paper, extending the joint interference nulling and CC concept originally proposed in [10], [12], 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. Our proposal results in a general content delivery scheme for any values of the problem parameters, i.e., the number of users  $K$ , library size  $N$ , cache size  $M$ , and number of transmit antennas  $L$ . 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 [13]. Moreover, by observing the complexity of the optimization problem, we propose a novel mode selection scheme which controls the size of multicasting groups. The benefits of this mode selection scheme are twofold: first it manages the complexity of the beamformers design by controlling the number of constraints in the corresponding optimization problem, second it results in a better rate performance by exploiting the transmit antennas to achieve multiplexing gain, and at the same time, compensating for the worst users channel effects, at each specific SNR.

## 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  $\{W_1, \dots, W_N\}$ , 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  $Z_k = Z_k(W_1, \dots, W_N)$  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  $k = 1, \dots, K$  at time instant  $i, i = 1, \dots, n$  can be written as

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

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 to users in subset  $\mathcal{T}$  of set  $\mathcal{S} \subseteq [1 : K]$  of users, and  $\tilde{X}_{\mathcal{T}}^{\mathcal{S}}(i)$  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. MULTICAST BEAMFORMING FOR CODED CACHING

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. Multicasting opportunities due to CC [1], [10], [11] are utilized to devise efficient generic multicast beamforming methods. In the following, for the sake of easy exposure, we introduce the basic multi-antenna multicast beamforming concept for a simple scenario 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, where a transmitter with  $L \geq 2$  antennas should deliver requests arising at  $K = 3$  users from a library  $\mathcal{W} = \{A, B, C\}$  of size  $N = 3$  files each of  $F$  bits. Suppose that in the cache content placement phase each user can cache  $M = 1$  files of  $F$  bits, 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.

At the content delivery phase 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 to users  $\mathcal{S} = \{1, 2, 3\}$  one after another

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

where  $\oplus$  represents summation in the corresponding finite field, and the superscript  $\mathcal{S}$  is omitted for ease of presentation. In this coding scheme of [1], each coded message is heard by all the three users, but is only beneficial to two users. For example,  $X_{1,2}$  is useful for the first and second user only,  $X_{1,3}$  is useful for the first and third user, and  $X_{2,3}$  is useful

for the second and third user. 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. Moreover, for every possible combination of the users requests the scheme works, with the same cache content placement, but with another set of coded delivery messages. Now, in the given *Scenario 1* we can combine the spatial multiplexing gain, and the global caching gain following the scheme from [10] (see also [3], [11]). In [10], 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 [10]. Although, this scheme is order-optimal in terms of DoF [3] it is suboptimal at low SNR regime [10], [12].

In this paper, instead of nulling interference at unwanted users, general multicast beamformers  $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$  are given as

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

where  $[K]$  denotes the set of integer numbers  $\{1, \dots, K\}$  and the superscript  $\mathcal{S}$  is omitted as all  $K = 3$  users are served in a single set  $\mathcal{S}$ . Then, the received signals at users 1 – 3 are

$$y_1 = \underline{(\mathbf{h}_1^H \mathbf{w}_{1,2})} \tilde{X}_{1,2} + \underline{(\mathbf{h}_1^H \mathbf{w}_{1,3})} \tilde{X}_{1,3} + \underline{(\mathbf{h}_1^H \mathbf{w}_{2,3})} \tilde{X}_{2,3} + z_1$$

$$y_2 = \underline{(\mathbf{h}_2^H \mathbf{w}_{1,2})} \tilde{X}_{1,2} + \underline{(\mathbf{h}_2^H \mathbf{w}_{1,3})} \tilde{X}_{1,3} + \underline{(\mathbf{h}_2^H \mathbf{w}_{2,3})} \tilde{X}_{2,3} + z_2$$

$$y_3 = \underline{(\mathbf{h}_3^H \mathbf{w}_{1,2})} \tilde{X}_{1,2} + \underline{(\mathbf{h}_3^H \mathbf{w}_{1,3})} \tilde{X}_{1,3} + \underline{(\mathbf{h}_3^H \mathbf{w}_{2,3})} \tilde{X}_{2,3} + z_3$$

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<sup>1</sup>

$$R_{MAC}^1 = \min\left(\frac{1}{2} R_{Sum}^1, R_1^1, R_2^1\right) \quad (5)$$

where the rate bounds  $R_1^1$  and  $R_2^1$  correspond to  $\tilde{X}_{1,2}$ , and  $\tilde{X}_{1,3}$ , respectively, and  $R_{Sum}^1$  is the sum rate of both messages. Thus, the total useful rate is  $2R_{MAC}^1$ . Since the user 1 must receive the missing  $2/3F$  bits ( $A_2$  and  $A_3$ ), the time needed to decode file  $A$  is  $T_1 = \frac{2F}{3} \frac{1}{2R_{MAC}^1}$ . As all the users decode their files *in parallel*, the decoding time is constrained by the worst user as

$$T = \frac{2F}{3} \frac{1}{\min_{k=1,2,3} 2R_{MAC}^k}. \quad (6)$$

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

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

which, when optimized with respect to the beamforming vectors, can be found as  $\max_{\mathbf{w}_{2,3}, \mathbf{w}_{1,3}, \mathbf{w}_{1,2}} \min_{k=1,2,3} R_{MAC}^k$ .

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

Finally, the symmetric rate for  $K = 3$  is given as

$$\begin{aligned}
& \max_{R^k, \gamma_i^k, \mathbf{w}_\mathcal{T}, \forall k, i} \min_{k=1,2,3} \min \left( \frac{1}{2} R_{\text{sum}}^k, R_1^k, R_2^k \right) \\
& \text{s. t. } R_1^k \leq \log(1 + \gamma_1^k), R_2^k \leq \log(1 + \gamma_2^k), \\
& \quad R_{\text{sum}}^k \leq \log(1 + \gamma_1^k + \gamma_2^k), k = 1, 2, 3, \\
& \quad \gamma_1^1 \leq \frac{|\mathbf{h}_1^H \mathbf{w}_{1,2}|^2}{|\mathbf{h}_1^H \mathbf{w}_{2,3}|^2 + N_0}, \gamma_2^1 \leq \frac{|\mathbf{h}_1^H \mathbf{w}_{1,3}|^2}{|\mathbf{h}_1^H \mathbf{w}_{2,3}|^2 + N_0}, \\
& \quad \gamma_1^2 \leq \frac{|\mathbf{h}_2^H \mathbf{w}_{2,3}|^2}{|\mathbf{h}_2^H \mathbf{w}_{1,3}|^2 + N_0}, \gamma_2^2 \leq \frac{|\mathbf{h}_2^H \mathbf{w}_{1,2}|^2}{|\mathbf{h}_2^H \mathbf{w}_{1,3}|^2 + N_0}, \\
& \quad \gamma_1^3 \leq \frac{|\mathbf{h}_3^H \mathbf{w}_{2,3}|^2}{|\mathbf{h}_3^H \mathbf{w}_{1,2}|^2 + N_0}, \gamma_2^3 \leq \frac{|\mathbf{h}_3^H \mathbf{w}_{1,3}|^2}{|\mathbf{h}_3^H \mathbf{w}_{1,2}|^2 + N_0}, \\
& \quad \sum_{\mathcal{T} \in \{\{1,2\}, \{1,3\}, \{2,3\}\}} \|\mathbf{w}_\mathcal{T}\|^2 \leq \text{SNR}.
\end{aligned} \tag{8}$$

Problem (8) is non-convex due to the SINR constraints. Similarly to [13], 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  $\gamma_1^1$  can be reformulated as

$$|\mathbf{h}_1^H \mathbf{w}_{2,3}|^2 + N_0 \leq \frac{|\mathbf{h}_1^H \mathbf{w}_{1,2}|^2 + |\mathbf{h}_1^H \mathbf{w}_{2,3}|^2 + N_0}{1 + \gamma_1^1}. \tag{9}$$

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

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

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

$$\begin{aligned}
& \max_{t, \gamma^k, \mathbf{w}_\mathcal{T}} t \\
& \text{s. t. } t \leq 1/2 \log(1 + \gamma_1^k + \gamma_2^k), k = 1, 2, 3, \\
& \quad t \leq \log(1 + \gamma_1^k), t \leq \log(1 + \gamma_2^k) \forall k, \\
& \quad \mathcal{L}(\mathbf{w}_{2,3}, \mathbf{w}_{1,2}, \gamma_1^1) \geq |\mathbf{h}_1^H \mathbf{w}_{2,3}|^2 + N_0, \\
& \quad \mathcal{L}(\mathbf{w}_{2,3}, \mathbf{w}_{1,3}, \gamma_2^1) \geq |\mathbf{h}_1^H \mathbf{w}_{2,3}|^2 + N_0, \\
& \quad \mathcal{L}(\mathbf{w}_{1,3}, \mathbf{w}_{2,3}, \gamma_1^2) \geq |\mathbf{h}_2^H \mathbf{w}_{1,3}|^2 + N_0, \\
& \quad \mathcal{L}(\mathbf{w}_{1,3}, \mathbf{w}_{1,2}, \gamma_2^2) \geq |\mathbf{h}_2^H \mathbf{w}_{1,3}|^2 + N_0, \\
& \quad \mathcal{L}(\mathbf{w}_{1,2}, \mathbf{w}_{2,3}, \gamma_1^3) \geq |\mathbf{h}_3^H \mathbf{w}_{1,2}|^2 + N_0, \\
& \quad \mathcal{L}(\mathbf{w}_{1,2}, \mathbf{w}_{1,3}, \gamma_2^3) \geq |\mathbf{h}_3^H \mathbf{w}_{1,2}|^2 + N_0, \\
& \quad \sum_{\mathcal{T} \in \{\{1,2\}, \{1,3\}, \{2,3\}\}} \|\mathbf{w}_\mathcal{T}\|^2 \leq \text{SNR}
\end{aligned} \tag{11}$$

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). To this end, we note that the sum rate constraint can be bounded as

$$t \leq \frac{1}{2} \log(1 + \gamma_1^k + \gamma_2^k) = \log(\sqrt{1 + \gamma_1^k + \gamma_2^k}) \leq \sqrt{1 + \gamma_1^k + \gamma_2^k}$$

Now, the equivalent SOCP reformulation follows as

$$\begin{aligned}
& \max_{\tilde{t}, \gamma^k, \mathbf{w}_k} \tilde{t} \\
& \text{s. t. } \tilde{t}^2 \leq 1 + \gamma_1^k + \gamma_2^k, k = 1, 2, 3, \\
& \quad \tilde{t} \leq 1 + \gamma_1^k, \tilde{t} \leq 1 + \gamma_2^k \forall k, \\
& \quad \text{The rest of the constraints as in (11)}.
\end{aligned} \tag{12}$$

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 (the proof follows similar lines as in [13]). Note that the final symmetric rates are achieved by time sharing 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<sup>2</sup>. 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:

$$\begin{aligned}
& \max_{R^k, \gamma^k, p_\mathcal{T}} \min_{k=1,2,3} \left( \frac{1}{2} R_{\text{sum}}^k, R_1^k, R_2^k \right) \\
& \text{s. t. } R_{\text{sum}}^k \leq \log(1 + \gamma_1^k + \gamma_2^k) \forall k, \\
& \quad R_1^k \leq \log(1 + \gamma_1^k), R_2^k \leq \log(1 + \gamma_2^k) \forall k, \\
& \quad \gamma_1^1 \leq u_{1,3} p_{1,2}, \gamma_2^1 \leq u_{1,2} p_{1,3}, \gamma_1^2 \leq u_{2,1} p_{2,3}, \\
& \quad \gamma_2^2 \leq u_{2,3} p_{1,2}, \gamma_1^3 \leq u_{3,1} p_{2,3}, \gamma_2^3 \leq u_{3,2} p_{1,3}, \\
& \quad \sum_{\mathcal{T} \in \{\{1,2\}, \{1,3\}, \{2,3\}\}} p_\mathcal{T} \leq \text{SNR}
\end{aligned} \tag{13}$$

where  $u_{k,i} = |\mathbf{h}_k^H \mathbf{h}_i^\perp|^2 / \|\mathbf{h}_i^\perp\|^2 N_0$ . 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) [10]:*

If the multicast transmit powers are made equal,  $p_{1,2} = p_{1,3} = p_{2,3} = \text{SNR}/3$ , the resulting scheme is the same as originally published in [10].

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:<sup>3</sup>

$$T_{1,2} = \frac{F/3}{\max_{\|\mathbf{w}\|^2 \leq \text{SNR}} \min \left( \log(1 + \frac{|\mathbf{h}_1^H \mathbf{w}|^2}{N_0}), \log(1 + \frac{|\mathbf{h}_2^H \mathbf{w}|^2}{N_0}) \right)} \tag{15}$$

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

$$R_{\text{maxmin}} = F / (T_{1,2} + T_{1,3} + T_{2,3}). \tag{16}$$

<sup>2</sup>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 V).

<sup>3</sup>This multicast maxmin problem is NP-hard in general, but near-optimal solutions can be obtained by a semidefinite relaxation (SDR) approach, see [10] and the references therein.

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  $\min(K, L)$  parallel independent streams to the users at each time instant. All the users can be served in parallel if  $L \geq K$ . On the other hand, if  $L < K$ , the users need to be divided into subsets of size  $L$  served in distinct time slots.

Now, let us consider a case  $L = 2$  and  $K = 3$ , and focus on 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

$$T_{1,2} = \frac{F/3}{\max_{\sum_{k=1,2} \|\mathbf{w}_k\|^2 \leq SNR} \min(R_1, R_2)} \quad (17)$$

where

$$R_k = \log \left( 1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{i \neq k} |\mathbf{h}_k^H \mathbf{w}_i|^2 + N_0} \right). \quad (18)$$

The minimum delivery time in (16) 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 (16).

### B. General $K, L, N$ and $M$

The guidelines for constructing general beamformed multicast messages for multi-antenna coded caching with any  $K, L, N$  and  $M$  are presented in Algorithm 1 included in the extended version [14, Algorithm 1] while the resulting symmetric rate is given in Theorem 1.

**Theorem 1.** [14, Algorithm 1] will result in the following symmetric rate

$$R_{\text{sym}} = c \left[ \sum_{\substack{\mathcal{S} \subseteq [K] \\ |\mathcal{S}| = \min(t+L, K)}} (R_C^*(\mathcal{S}))^{-1} \right]^{-1}, \quad (19)$$

where<sup>4</sup>

$$c = \binom{K}{t} \binom{K-t-1}{\min(L-1, K-t-1)} \quad (20)$$

and  $R_C^*(\mathcal{S})$  is the symmetric rate of user subset  $\mathcal{S}$  optimized over a set of multicast beamformers  $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = t+1$

$$R_C^*(\mathcal{S}) = \max_{\substack{\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T} \subseteq \mathcal{S}, k \in \mathcal{S} \\ |\mathcal{T}| = t+1, \\ \sum_{\mathcal{T} \subseteq \mathcal{S}} \|\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}\|^2 \leq SNR\}}} \min R_{MAC}^k(\mathcal{S}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = t+1\}) \quad (21)$$

The size of set  $\mathcal{S}$  is bounded by  $|\mathcal{S}| = \min(t+L, K)$ . Thus, only a single set with all  $K$  users is considered in (19) if  $L \geq K-t$ . In (21), we have used

$$R_{MAC}^k(\mathcal{S}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = t+1\}) = \min_{\mathcal{B} \subseteq \Omega_k^{\mathcal{S}}} \left[ \frac{1}{|\mathcal{B}|} \log \left( 1 + \frac{\sum_{\mathcal{T} \in \mathcal{B}} |\mathbf{h}_k^H \mathbf{w}_{\mathcal{T}}^{\mathcal{S}}|^2}{N_0 + \sum_{\mathcal{T} \in \Omega_k^{\mathcal{S}} \setminus \mathcal{B}} |\mathbf{h}_k^H \mathbf{w}_{\mathcal{T}}^{\mathcal{S}}|^2} \right) \right] \quad (22)$$

<sup>4</sup>If  $t = K-1$ , then  $c = \binom{K}{t}$ .

where

$$\begin{aligned} \Omega^{\mathcal{S}} &:= \{\mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = t+1\} \\ \Omega_k^{\mathcal{S}} &:= \{\mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| = t+1 \mid k \in \mathcal{T}\} \end{aligned} \quad (23)$$

For a specific subset  $\mathcal{S}$ ,  $\Omega_k^{\mathcal{S}}$  includes the set of desired terms for each user's MAC channel and  $\Omega^{\mathcal{S}} \setminus \Omega_k^{\mathcal{S}}$  shows the set of interference terms for each user  $k$ .

Theorem 1 is a generalization of (8) for general  $K, L, N$  and  $M$ . If the beamforming vectors are chosen as the ZF vectors and  $L \leq K-t$ , the interference terms vanish, and this theorem reduces to [12, Theorem 2].

*Proof.* The proof can be found in the extended version of this work [14]  $\square$

## IV. MULTICAST MODE SELECTION FOR CC WITH REDUCED COMPLEXITY

In general, the number of parallel multicast streams to be decoded grows linearly when  $K, L, N$  are increased with the same ratio. As a result, the number of rate constraints in the user specific MAC region grows exponentially, i.e., by  $2^{(K-1)} - 1$  per user if  $L \geq N = K$  as can be seen from (22). 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  $K \times (2^{(K-1)} - 1) = 5 \times 15$  while the number of SINR constraints to be approximated would be  $5 \times 4$ . As an efficient way to reduce the complexity of the problem both at the transmitter and the receivers (with a certain performance loss at high SNR), we may limit the size of user subsets benefiting from transmitted parallel multicast messages.

### A. Reduced complexity approach with $|\mathcal{S}| < \min(t+L, K)$

In this section, a reduced complexity alternative to Theorem 1 is proposed. In this scheme, instead of fixing the size of the subsets  $\{\mathcal{S} \subseteq [K]\}$  to be  $\min(t+L, K)$ , we introduce a new integer parameter

$$1 \leq \alpha \leq \min(L, K-t) \quad (24)$$

and define the size of subsets  $\{\mathcal{S} \subseteq [K]\}$  to be  $t+\alpha$ . Then, a common signal is transmitted to each  $(t+\alpha)$ -subset  $\mathcal{S}$ , which contains a coded multicast message, beamformed to one of the  $(t+1)$ -subsets  $\{\mathcal{T} \subseteq \mathcal{S}\}$ . This generalization reduces to the baseline max-min SNR beamforming scheme if  $\alpha = 1$  (see (15) for  $K = 3$ ), and to Theorem 1 if  $\alpha = \min(L, K-t)$ . Also, it enables us to control the size of the MAC channel elements with respect to each user, and in turn, to control the optimization problem complexity for determining the beamforming vectors in (21) and (22). As will be shown later, besides complexity reduction, this generalization enables us to handle the trade-off between the multiplexing and multicast beamforming gains, resulting in even better rate performance at certain SNR values.

For this general setting, the cache content placement and content delivery phases are as before with slight modifications described below. First, instead of splitting each subfile into

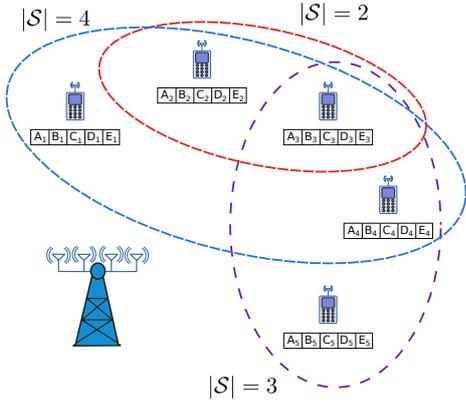


Fig. 1. Mode selection scenario,  $K = N = 5$ ,  $L = 4$ ,  $|\mathcal{S}| = \{2, 3, 4, 5\}$  ( $\alpha = \{1, 2, 3, 4\}$ )

$\binom{K-t-1}{\min(L-1, K-t-1)}$  minifiles, now each subfile is split into  $\binom{K-t-1}{\alpha-1}$  minifiles. Second, the size of subset  $\mathcal{S}$  will change to  $t + \alpha$ . Finally, the rate expression in (19) is changed to

$$R_{\text{sym}} = \binom{K}{t} \binom{K-t-1}{\alpha-1} \left[ \sum_{\substack{\mathcal{S} \subseteq [K] \\ |\mathcal{S}|=t+\alpha}} (R_C^*(\mathcal{S}))^{-1} \right]^{-1} \quad (25)$$

Fig. 1 illustrates a scenario with  $K = N = 5$ ,  $L = 4$  and  $M = 1$  and 3 possible subsets of size  $|\mathcal{S}| = \{2, 3, 4, 5\}$  ( $\alpha = \{1, 2, 3, 4\}$ ). In total, there can be  $T = \binom{5}{4} = 5$ ,  $T = \binom{5}{3} = 10$  and  $T = \binom{5}{2} = 10$  subsets of sizes  $|\mathcal{S}| = 4$ ,  $|\mathcal{S}| = 3$  and  $|\mathcal{S}| = 2$ , respectively. In this example, every subset in  $\{\mathcal{S}, |\mathcal{S}| = 3\}$  corresponds to *Scenario 1*, and the optimal multicast beamformers can be found by solving (11) (for corresponding  $k \in \mathcal{S}$ ).

The following example illustrates how the multicast transmission sets are constructed for the proposed reduced-complexity scheme.

### B. Scenario 3: $L \geq 3$ , $K = 4$ , $N = 4$ , $M = 1$ and $|\mathcal{S}| = 3$

In this example, we consider the parameters  $L \geq 3$ ,  $K = 4$ ,  $N = 4$ ,  $M = 1$ . This corresponds to *Scenario 2* in [14]. However, here we set the parameter  $\alpha$  defined in (24) to  $\alpha = 2$ . Therefore, we restrict the size of the subsets  $\mathcal{S} \subset [4]$  benefiting from a common transmitted signal to  $|\mathcal{S}| = 3$ , instead of serving the full set of four users in parallel (as in *Scenario 2* in [14]). The cache content placement works similarly, except for that we split each subfile into  $\binom{K-t-1}{\alpha-1} = 2$  mini-files (indicated by superscripts) resulting in the following contents in user cache memories

$$\begin{aligned} Z_1 &= \{A_1^1, A_1^2, B_1^1, B_1^2, C_1^1, C_1^2, D_1^1, D_1^2\} \\ Z_2 &= \{A_2^1, A_2^2, B_2^1, B_2^2, C_2^1, C_2^2, D_2^1, D_2^2\} \\ Z_3 &= \{A_3^1, A_3^2, B_3^1, B_3^2, C_3^1, C_3^2, D_3^1, D_3^2\} \\ Z_4 &= \{A_4^1, A_4^2, B_4^1, B_4^2, C_4^1, C_4^2, D_4^1, D_4^2\} \end{aligned}$$

Now we focus on the users  $\mathcal{S} = \{1, 2, 3\}$ . Let us send them the following transmit vector by the transmitter

$$\mathbf{w}_{1,2} \tilde{X}_{1,2} + \mathbf{w}_{1,3} \tilde{X}_{1,3} + \mathbf{w}_{2,3} \tilde{X}_{2,3} \quad (26)$$

where

$$X_{1,2} = A_2^1 \oplus B_1^1, \quad X_{1,3} = A_3^1 \oplus C_1^1, \quad X_{2,3} = B_3^1 \oplus C_2^1 \quad (27)$$

This transmission should be such that  $X_{\mathcal{T}}$  is received at all users in  $\mathcal{T} \subset \mathcal{S}$ ,  $|\mathcal{T}| = 2$  correctly. Let us call the corresponding common rate for coding each  $X_{\mathcal{T}}$  as  $R_{1,2,3}$ . Then, since each minifile is of length  $F/8$ , the time needed for this transmission is  $T_{1,2,3} = \frac{F}{8} \frac{1}{R_{1,2,3}}$ . Now we consider the other 3-subsets (subsets of size 3) of users. For the subset  $\mathcal{S} = \{1, 2, 4\}$  the transmitter sends

$$\mathbf{w}_{1,2} \tilde{X}_{1,2} + \mathbf{w}_{1,4} \tilde{X}_{1,4} + \mathbf{w}_{2,4} \tilde{X}_{2,4} \quad (28)$$

where

$$X_{1,2} = A_2^2 \oplus B_1^2, \quad X_{1,4} = A_4^1 \oplus D_1^1, \quad X_{2,4} = B_4^1 \oplus D_2^1 \quad (29)$$

each coded with the rate  $R_{1,2,4}$  and the corresponding transmission time is  $T_{1,2,4} = \frac{F}{8} \frac{1}{R_{1,2,4}}$ . Please note that the subset  $\{1, 2\}$  here appears for the second time, and thus the second minifiles are used for the coding. The other subsets  $\{1, 4\}$ , and  $\{2, 4\}$  have not yet appeared and the first minifiles are still not transmitted. For the subsets  $\mathcal{S} = \{1, 3, 4\}$  and  $\mathcal{S} = \{2, 3, 4\}$  the transmitter sends

$$\mathbf{w}_{1,3} \tilde{X}_{1,3} + \mathbf{w}_{1,4} \tilde{X}_{1,4} + \mathbf{w}_{3,4} \tilde{X}_{3,4} \quad (30)$$

$$\mathbf{w}_{2,3} \tilde{X}_{2,3} + \mathbf{w}_{2,4} \tilde{X}_{2,4} + \mathbf{w}_{3,4} \tilde{X}_{3,4} \quad (31)$$

respectively, where

$$X_{1,3} = A_3^2 \oplus C_1^2, \quad X_{1,4} = A_4^2 \oplus D_1^2, \quad X_{3,4} = C_4^1 \oplus D_3^1 \quad (32)$$

are coded with the rate  $R_{1,3,4}$  with the corresponding transmission time  $T_{1,3,4} = \frac{F}{8} \frac{1}{R_{1,3,4}}$ , while

$$X_{2,3} = B_3^2 \oplus C_2^2, \quad X_{2,4} = B_4^2 \oplus D_2^2, \quad X_{3,4} = C_4^2 \oplus D_3^2 \quad (33)$$

are coded with the rate  $R_{2,3,4}$  and  $T_{2,3,4} = \frac{F}{8} \frac{1}{R_{2,3,4}}$ . Since these transmissions are done in different time slots, the *Symmetric Rate Per User* of this example is

$$\begin{aligned} & \frac{F}{T_{1,2,3} + T_{1,2,4} + T_{1,3,4} + T_{2,3,4}} \\ &= 8 \left( \frac{1}{R_{1,2,3}} + \frac{1}{R_{1,2,4}} + \frac{1}{R_{1,3,4}} + \frac{1}{R_{2,3,4}} \right)^{-1}. \end{aligned} \quad (34)$$

The beamforming vectors are optimized separately to maximize the symmetric rate for each transmission interval. For each subset  $\mathcal{S}$  the formulation is exactly the same as the one in *Scenario 1*. The difference is that here we have  $L = 3$  antennas, and hence, the beamforming vectors are 3-dimensional instead of 2-dimensional as in *Scenario 1*.

## V. NUMERICAL EXAMPLES

The numerical examples are generated for various combinations of parameters  $L, K, N, M$  and  $|\mathcal{S}|$ . 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. The performance of different schemes with  $L = 3$ ,  $K = 4$ ,  $N = 4$ ,  $M = 1$  is illustrated in Fig. 2. It can be seen that the proposed CC multicast beamforming scheme via SCA, denoted as CC-BF-SCA, achieves 5–7 dB gain at low SNR as compared to the CC-ZF with equal power loading [10]. At high SNR, the CC-ZF with optimal power loading in (13) achieves comparable performance while other schemes have significant performance gap. At low SNR

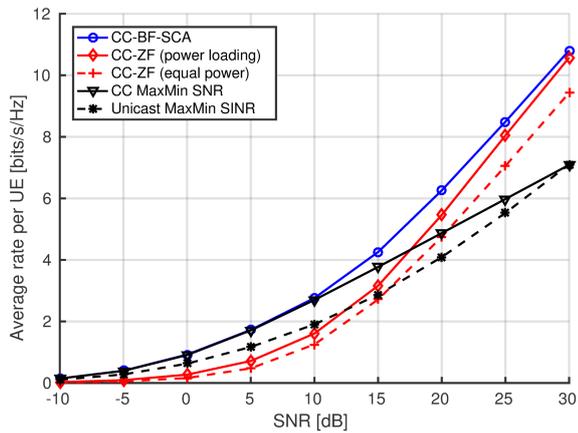


Fig. 2. Coded caching with multiantenna transmission,  $L = 3$  and  $K = 4$  regime, the simple MaxMin SNR multicasting with CC has similar performance as the proposed CC-BF-SCA scheme. This is due to the fact that, at low SNR, an efficient strategy for beamforming is to concentrate all available power 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 [10], [11]. The unicasting scheme does not perform well here as it does not utilize the global caching gain (only local) and the spatial DoF is limited to 3.

Fig. 1 illustrates the subset selection possibilities for  $K = 5$ ,  $L = 4$  scenario, and its performance is plotted in Fig. 3. In this case, there are three possible reduced subset sizes  $|\mathcal{S}| \in \{2, 3, 4\}$  that can be used to reduce the serving set size in  $\mathcal{S}$ . From Fig. 3, we can observe that, by reducing the subset size to  $|\mathcal{S}| = 4$ , and  $|\mathcal{S}| = 3$ , the average symmetric rate per user can be even improved at low to medium SNR as compared to the case where all users are served simultaneously, i.e.,  $|\mathcal{S}| = 5$ . This is due to similar reasoning as in Fig. 2. At lower SNR, it is better to focus the available power to fewer multicast streams transmitted in parallel. This will reduce the inter-stream interference and, at the same time, provide increased spatial degrees of freedom for multicast beamformer design. All distinct user subsets  $\mathcal{S} \subseteq [K]$  are served in TDMA fashion.

At high SNR region, however, the reduced subset cases become highly suboptimal as they do not utilize all spatial degrees of freedom for transmitting parallel streams. From complexity reduction perspective, the multicast mode with the smallest subset size providing close to optimal performance should be selected. In Fig. 3, for example, subset sizes  $|\mathcal{S}| = 2$ ,  $|\mathcal{S}| = 3$ ,  $|\mathcal{S}| = 4$  could be used up to 0 dB, 10 dB and 25 dB, respectively, for optimal performance-complexity trade-off.

## VI. CONCLUSIONS

Multicasting opportunities provided by caching at user terminal were utilized to devise an efficient multiantenna transmission with CC. General multicast beamforming strategies for content delivery with any values of the problem parameters, i.e., the number of users  $K$ , library size  $N$ , cache size  $M$ , and number of transmit antennas  $L$  were employed with CC,

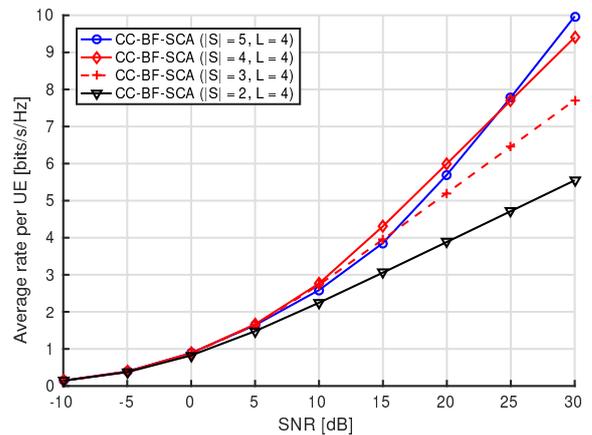


Fig. 3.  $K = 5$ ,  $L = 4$ ,  $|\mathcal{S}| = \{2, 3, 4\}$ .

optimally balancing the detrimental impact of both noise and inter-stream interference from coded messages transmitted in parallel. In addition, a novel multicast mode selection scheme was proposed where the optimum multicast group sizes providing the best complexity-performance trade-off were found for a given SNR range. The schemes were shown to perform significantly better than several base-line 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] 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.
- [3] 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.
- [4] J. Zhang and P. Elia, "Wireless Coded Caching: A Topological Perspective," *ArXiv e-prints*, Jun. 2016.
- [5] 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.
- [6] 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.
- [7] M. M. Amiri and D. Gunduz, "Caching and Coded Delivery over Gaussian Broadcast Channels for Energy Efficiency," *ArXiv e-prints*, Dec. 2017.
- [8] S. S. Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1222–1226.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *CoRR*, vol. abs/1711.05969, 2017. [Online]. Available: <http://arxiv.org/abs/1711.05969>
- [13] 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.
- [14] 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>