Linear Multicast Beamforming Schemes for Coded Caching

Antti Tölli⋆, Jarkko Kaleva⋆, Seyed Pooya Shariatpanahi∗ 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. Depending on the available degrees of freedom, several multicast messages are transmitted in parallel to distinct subsets of users. In the multicast beamformer design, we restrict ourselves to linear receiver implementation, which does not require successive interference cancellation unlike in the more optimal baseline scheme. With a small loss in performance, the complexity of both the receiver and transmitter implementation can be significantly reduced.**

I. Introduction

In Coded Caching (CC) [1], instead of simply replicating high-popularity contents near-or-at the end-users, the network should spread different contents to different caches such that common coded messages, broadcast during the network high-peak hours to different users with different demands, would benefit all the users simultaneously. CC has been extensively studied, in particular, for high SNR performance and achievable Degree of Freedom (DoF) gain [2]–[5]. Recent works on finite SNR regime have also shown CC to be greatly beneficial when the interference is properly accounted for [6]–[9]. However, the complexity of such schemes both at the transmitter and the receivers quickly becomes prohibitive as the number of simultaneously served users becomes large [8], [9].

In this paper, we consider a single cell downlink scenario similarly to [8], [9], where a multiple-antenna base station (BS) delivers contents to cache-enabled user terminals. Depending on the available DoF, several multicast messages are transmitted in parallel to distinct subsets of users. In the multicast beamformer design, we restrict ourselves to simple receiver implementation, which does not require successive interference cancellation (SIC) receiver structure as in the baseline scheme [9]. Therefore, with a small loss in performance, the complexity of both the receiver and transmitter implementation is significantly reduced

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 its own received signal from the wireless channel.

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

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

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. The size of $\mathcal{T}$ depends on the parameters K, M and N such that $|\mathcal{T}| = t+1$, where $t \triangleq KM/N$ [2], [10]. 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.

Note that (1) is defined for a given set of users $k \in \mathcal{S}$ served at time instant i. Depending on the chosen transmission strategy and parametrization, the delivery of the requested files $W_{d_k}\ \forall\ k$ may require multiple time intervals/slots carried out for all possible partitionings and subsets $\mathcal{S} \subseteq [1:K]$.

III. Multiantenna Coded Caching with Linear Receivers

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, we introduce the basic multiantenna multicast beamforming concept for two particular scenarios while the generalization of the proposed scheme along with all the necessary details is provided in [8].

A. Scenario 1: $L \geq 3$, $K = 4$, $N = 4$ and $M = 1$

We assume that the BS transmitter has $L \geq 3$ 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 each file is divided into four non-overlapping equal-sized subfiles.

At the content delivery phase, suppose that the users $1-4$ request files $A-D$, respectively. Following the approach of the CC-BF-SCA scheme introduced in [9], the transmit signal vector is

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

where

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

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.

For the *baseline reference scheme* introduced in [8], [9], 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, 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. Consequently, this user receives useful information with the rate $3R_{MAC}^k$, and the time required to fetch the entire file is $T_1 = \frac{3F}{4}\frac{1}{3R_{MAC}^k}$. In this case, the symmetric rate per user can be found as [8], [9]

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

where

$$R_{MAC}^k = \min\left(R_1^k, R_2^k, R_3^k, \frac{1}{2}R_4^k, \frac{1}{2}R_5^k, \frac{1}{2}R_6^k, \frac{1}{3}R_7^k\right) \quad (4)$$

and where the rate bounds R_1^1, R_2^1 and R_3^1 of user 1, for example, correspond to $\tilde{X}_{1,2}$, $\tilde{X}_{1,3}$ and $\tilde{X}_{1,4}$, respectively. The bounds R_4^1, R_5^1 and R_6^1 limit the sum rate of any combination of two transmitted multicast signals, and finally R_7^1 is the sum rate bound for all 3 messages.

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:

$$\begin{aligned}
&\max_{r,\mathbf{w}_{\mathcal{T}},\mathcal{T}\subseteq[4],|\mathcal{T}|=2} r \quad (5)\\
&\text{s.t.} \quad r \le \log(1+\gamma_1^k), t \le \log(1+\gamma_2^k), t \le \log(1+\gamma_3^k) \\
&\quad r \le 1/2\log(1+\gamma_1^k+\gamma_2^k), \\
&\quad r \le 1/2\log(1+\gamma_1^k+\gamma_3^k), \\
&\quad r \le 1/2\log(1+\gamma_2^k+\gamma_3^k), \qquad \forall\ k = 1,2,3,4 \\
&\quad r \le 1/3\log(1+\gamma_1^k+\gamma_2^k+\gamma_3^k) \\
&\quad \gamma_1^1 \le \frac{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{1,2}|^2}{|\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{2,4}|^2 + |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{3,4}|^2 + N_0} \\
&\quad \vdots \text{ In total 12 SINR constraints} \\
&\quad \gamma_3^4 \le \frac{|\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{3,4}|^2}{|\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + |\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,3}|^2 + |\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{2,3}|^2 + N_0} \\
&\quad \textstyle\sum_{\mathcal{T}\subseteq[4],|\mathcal{T}|=2} \|\mathbf{w}_{\mathcal{T}}\|^2 \le \text{SNR}
\end{aligned}$$

In order to solve the above non-convex problem, the SCA method is used and the SINR constraints are approximated similarly to [8], [9].

In the reference scheme above, each user receives three useful streams which need to be decoded via SIC. In this paper, instead, we consider a simpler TX-RX strategy for this particular scenario where the transmission is split into three time slots, as illustrated in Fig. 1. In time slot 1–3, the multicast beamforming vectors are generated as

$$\mathbf{w}_{1,2}(A_2 \oplus B_1) + \mathbf{w}_{3,4}(C_4 \oplus D_3) \quad (6)$$
$$\mathbf{w}_{1,3}(A_3 \oplus C_1) + \mathbf{w}_{2,4}(B_4 \oplus D_2) \quad (7)$$
$$\mathbf{w}_{1,4}(A_4 \oplus D_1) + \mathbf{w}_{2,3}(B_3 \oplus C_2) \quad (8)$$

In each time slot, all four users are served with two parallel multicast streams. Each stream causes inter-stream interference to two other users not included in the given multicast group. Therefore, the BS, equipped at least with 3 antennas, has enough spatial degrees of freedom to manage the inter-stream interference between the multicast streams. The beamforming vectors are optimized separately to maximize the symmetric rate $R_C(i)$ for each transmission interval i. Thus, the corresponding time to deliver the multicast messages containing $F/4$ fractions of the files in time slot i is $T(i) = \frac{F}{4}\frac{1}{R_C(i)}$. Since these transmissions are done in three different time slots, the overall *Symmetric Rate Per User* of this scheme is

$$\frac{F}{\sum_{i=1,2,3} T(i)} = 4\left(\sum_{i=1,2,3}\frac{1}{R_C(i)}\right)^{-1}. \quad (9)$$

As will be numerically demonstrated in Section IV, the scheme provides the same overall DoF (slope) as the original scheme in [8], [9], but with a constant gap at high SNR due to simplified TX-RX processing.

As no overlap is allowed, each user decodes a single multicast message in a given time slot. Therefore, neither SIC receiver nor MAC rate region constraints are needed in the problem formulation unlike in [9]. As a result, the achievable rate is uniquely defined by the SINR of the received data stream. Let us define $\gamma_C(i)$ to be the common symmetric SINR for all users served in time slot i such that $R_C(i) = \log(1+$

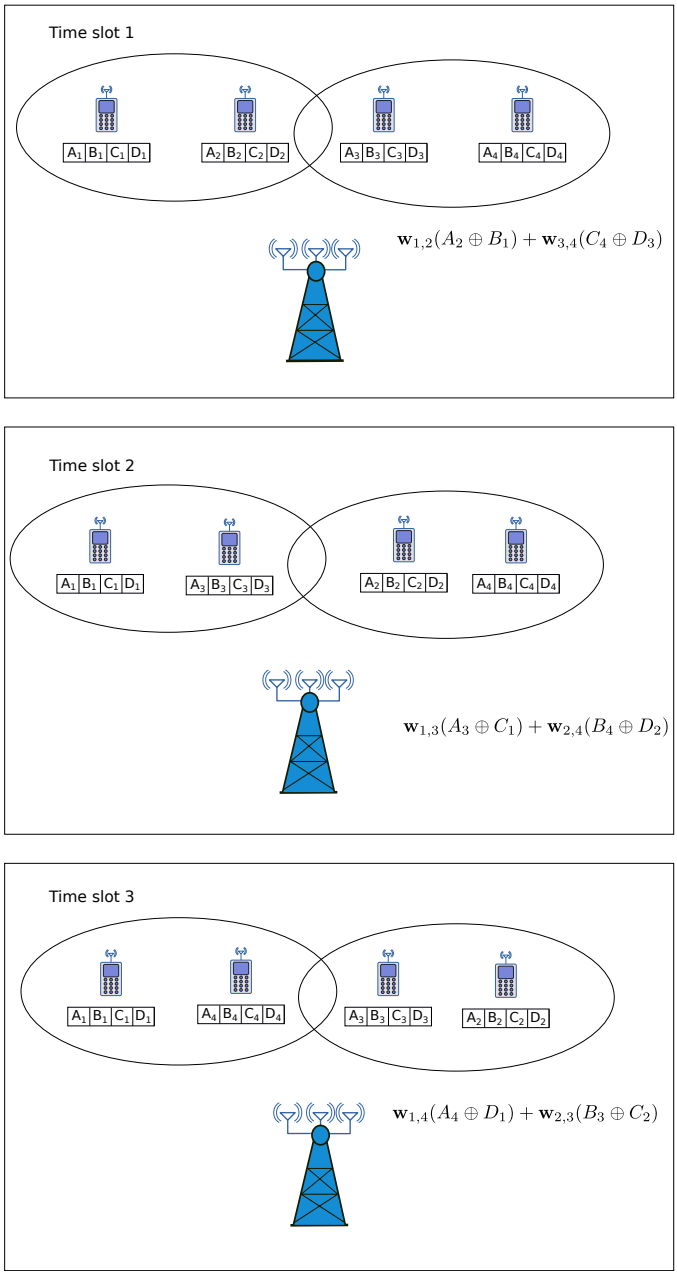


Fig. 1. Proposed simple multicast beamforming scheme, $L = 3$ and $K = 4$

$\gamma_C(i)$). The multigroup multicast beamformer optimization problem for ith timeslot can be then expressed as the following common SINR maximization problem:

$$\begin{aligned} \max_{\gamma_C(i),\mathbf{w}_\mathcal{T}} \quad & \gamma_C(i) \\ \text{s. t.} \quad & \gamma_C(i) \le \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_\mathcal{T}(i)|^2}{|\sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)\setminus\mathcal{T}} \mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\bar{\mathcal{T}}}(i)|^2 + N_0}, \\ & \forall k \in \mathcal{T}, \mathcal{T} \in \mathsf{P}(i), \\ & \sum_{\mathcal{T}\in\mathsf{P}(i)} \|\mathbf{w}_\mathcal{T}(i)\|^2 \le \text{SNR}. \end{aligned} \tag{10}$$

where $\mathsf{P}(1) = \{\{1,2\},\{3,4\}\}$, $\mathsf{P}(2) = \{\{1,3\},\{2,4\}\}$ and $\mathsf{P}(3) = \{\{1,4\},\{2,3\}\}$. The resulting problem is a multigroup multicast beamforming for common SINR maximization and several solutions exist, for example via semidefinite relaxation (SDR) of beamformers and solving (iteratively via bisection) as a semidefinite program (SDP) [11]. Here, instead, we adopt the SCA solution from [12], based on which (10) can be solved efficiently as a series of second order cone programs. Unlike the SDP based designs, the SCA technique solves for beamformers directly, thereby avoiding the need for any randomization procedure if rank-1 beamformers are to be recovered from the SDR solutions [12].

To begin with, the SINR constraint for $\gamma_C(i)$ and a given $k \in \mathcal{T}, \mathcal{T} \in \mathsf{P}(i), \bar{\mathcal{T}} \in \mathsf{P}(i) \setminus \mathcal{T}$ can be reformulated as

$$\gamma_C(i) \le \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_\mathcal{T}(i)|^2}{\sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)\setminus\mathcal{T}} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\bar{\mathcal{T}}}(i)|^2 + N_0}, \tag{11}$$

$$\sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)\setminus\mathcal{T}} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\bar{\mathcal{T}}}(i)|^2 + N_0 \le \frac{\sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\bar{\mathcal{T}}}(i)|^2 + N_0}{1+\gamma_C(i)}. \tag{12}$$

Now, the R.H.S of (12) is a convex quadratic-over-linear function and it can be linearly approximated or lower bounded (ignoring the time index i) as

$$\begin{aligned} \mathcal{L}(\mathbf{w}_\mathcal{T},\mathbf{w}_{\bar{\mathcal{T}}},\mathbf{h}_k,\gamma_C) \triangleq & \sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)} |\mathbf{h}_k^{\mathrm{H}}\bar{\mathbf{w}}_{\bar{\mathcal{T}}}|^2 + N_0 \\ & - 2\sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)} \mathbb{R}\left(\bar{\mathbf{w}}_{\bar{\mathcal{T}}}^{\mathrm{H}}\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}\left(\mathbf{w}_{\bar{\mathcal{T}}} - \bar{\mathbf{w}}_{\bar{\mathcal{T}}}\right)\right) \\ & + \frac{\sum_{\bar{\mathcal{T}}\in\mathsf{P}(i)} |\mathbf{h}_k^{\mathrm{H}}\bar{\mathbf{w}}_{\bar{\mathcal{T}}}|^2 + N_0}{1+\bar{\gamma}_C}(\gamma_C - \bar{\gamma}_C) \end{aligned} \tag{13}$$

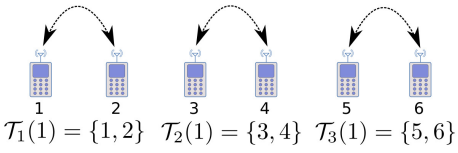
where $\bar{\mathbf{w}}_{\bar{\mathcal{T}}}$, $\bar{\mathbf{w}}_\mathcal{T}$ and $\bar{\gamma}_C$ denote the fixed values (points of approximation) for the corresponding variables from the previous iteration. For example, the common SINR for time slot 1, $\gamma_C(1)$ can be solved (for a given approximation point $\bar{\mathbf{w}}_{1,2}, \bar{\mathbf{w}}_{3,4}, \bar{\gamma}_C(1)$ and by omitting the slot index i) as

$$\begin{aligned} \max_{\gamma_C,\mathbf{w}_{1,2},\mathbf{w}_{3,4}} \quad & \gamma_C \\ \text{s. t.} \quad & \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{3,4},\mathbf{h}_1,\gamma_C) \ge |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{3,4}|^2 + N_0, \\ & \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{3,4},\mathbf{h}_2,\gamma_C) \ge |\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{3,4}|^2 + N_0, \\ & \mathcal{L}(\mathbf{w}_{3,4},\mathbf{w}_{1,2},\mathbf{h}_3,\gamma_C) \ge |\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + N_0, \\ & \mathcal{L}(\mathbf{w}_{3,4},\mathbf{w}_{1,2},\mathbf{h}_4,\gamma_C) \ge |\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + N_0, \\ & \|\mathbf{w}_{1,2}\|^2 + \|\mathbf{w}_{3,4}\|^2 \le \text{SNR}. \end{aligned} \tag{14}$$

where $\mathcal{L}(\mathbf{w}_\mathcal{T},\mathbf{w}_{\bar{\mathcal{T}}},\mathbf{h}_k,\gamma_C)$ is given in (13).

B. Scenario 2: $L \ge 5$, $K = 6$, $N = 6$ and $M = 1$

In general, the number of parallel multicast streams to be decoded at each user grows linearly when K, L, N are increased with the same ratio. In the baseline approach in [8], the number of rate constraints in the user specific MAC region grows exponentially, i.e., by $2^{(K-1)} - 1$ per user if $L \ge N = K$. The case $L \ge 5$, $K = 6$, $N = 6$ and $M = 1$ would require altogether $\binom{6}{2} = 15$ multicast messages and each user should be able to decode 5 multicast messages. Thus, the total number of rate constraints would be $K \times (2^{(K-1)} - 1) = 6 \times 31$ while the number of SINR constraints to be approximated would be 6×5. As an efficient way to reduce the complexity of the problem both at the



i	P		
1	**1,2**	**3,4**	**5,6**
2	1,2	3,5	4,6
3	1,2	3,6	4,5
4	1,3	2,4	5,6
5	1,3	2,5	4,6
6	1,3	2,6	4,5
7	1,4	2,3	5,6
8	1,4	2,5	3,6
9	1,4	2,6	3,5
10	1,5	2,3	4,6
11	1,5	2,4	3,6
12	1,5	2,6	3,4
13	1,6	2,3	4,5
14	1,6	2,4	3,5
15	1,6	2,5	3,4

Fig. 2. Proposed multicast scheme for $K = 6$, $N = 6$, $M = 1$

transmitter and the receivers (with a certain performance loss at high SNR), we may limit the overlap among the multicast messages as in *Scenario 1*.

In the given scenario, there are $K = 6$ users each with cache size $M = 1$, requesting files from a library $\mathcal{W} = \{A, B, C, D, E, F\}$ of $N = 6$ files. Following the same cache content placement strategy as in [1] the cache contents of users are as follows

$$\begin{aligned} Z_1 &= \{A_1, B_1, C_1, D_1, E_1, F_1\}, \\ Z_2 &= \{A_2, B_2, C_2, D_2, E_2, F_2\}, \\ Z_3 &= \{A_3, B_3, C_3, D_3, E_4, F_4\}, \\ Z_4 &= \{A_4, B_4, C_4, D_4, E_4, F_4\}, \\ Z_5 &= \{A_5, B_5, C_5, D_5, E_5, F_5\}, \\ Z_6 &= \{A_6, B_6, C_6, D_6, E_6, F_6\} \end{aligned}$$

where here each file is divided into 6 non-overlapping equal-sized subfiles.

Similarly to *Scenario 1*, the multicast transmission is split into orthogonal time intervals. For the given scenario, in total 15 time slots are required to cover all disjoint unions $\dot{\bigcup}\mathcal{T}$ of $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ such that $|\mathcal{T}| = t + 1 = 2$. In time slot i, the multicast beamforming vectors are generated as

$$\sum_{\mathcal{T} \in \mathsf{P}(i)} \mathbf{w}_{\mathcal{T}} \tilde{X}_{\mathcal{T}} \tag{15}$$

where the set of multicast beamformer indexes $\mathsf{P}(i) = \{\mathcal{T}_1(i), \mathcal{T}_2(i), \mathcal{T}_3(i)\}$ for all time slots i are given in Fig. 2.

Unlike in *Scenario 1*, each subfile is split into 3 mini-files in *Scenario 2* in order to allow different contents to be transmitted in each subset $\mathsf{P}(i)$. This is due to the fact that each user index pair $\mathcal{T}$ (e.g. $\mathcal{T} = \{1, 2\}$) is repeated 3 times, as seen from Fig. 2.

In each time slot i, all 6 users are served with 3 multicast streams transmitted in parallel. Similarly to *Scenario 1*, the BS, equipped at least with 5 antennas, is able to manage the inter-stream interference between multicast streams. One spatial degree of freedom is used for delivering the multicast message to a user pair $\mathcal{T}$ while four degrees of freedom are needed to control the interference towards users $\bar{\mathcal{T}} \in \mathsf{P}(i) \setminus \mathcal{T}$. The beamforming vectors are optimized separately to maximize the symmetric rate $R_C(i)$ for each transmission interval i.

The time to deliver the multicast messages containing

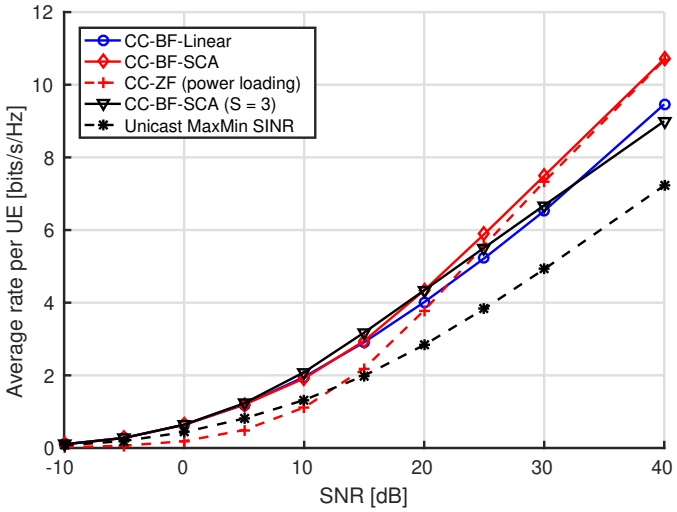


Fig. 3. Average 'Goodput' per user, $L = 3$ and $K = 4$

$F/(6\times3)$ fractions of the files in time slot i is $T(i) = \frac{F}{18}\frac{1}{R_C(i)}$. Since these transmissions are done in 15 different time slots, the overall *Symmetric Rate Per User* of *Scenario 2* is

$$\frac{F}{\sum_{i=1}^{15} T(i)} = 18\left(\sum_{i=1}^{15} \frac{1}{R_C(i)}\right)^{-1}. \tag{16}$$

The symmetric rate $R_C(i) = \log(1 + \gamma_C(i))$ is found by solving the generic optimization problem (10) while using $\mathsf{P}(i)$ given in Fig. 2.

IV. Numerical Examples

Numerical examples are generated for both *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. 3 illustrates the performance of the proposed simple linear TX-RX multicasting scheme in comparison with the CC-BF-SCA and unicast (which only utilizes the local caching gain) beamforming baseline cases from [8], [9]. Furthermore, the performance of the 'reduced complexity scheme' from [8] is plotted for comparison. In this scheme, labelled as 'CC-BF-SCA (S=3)', the multicast transmission is split into 4 time slots, where the CC-BF-SCA strategy is applied to each subset of 3 users (with the corresponding loss in DoF). The proposed linear scheme labeled as 'CC-BF-Linear' is able to serve 4 users simultaneously in each time slot with 3 antennas. Thus, it can provide the same degrees of freedom ($= 4$) at high SNR as the baseline CC-BF-SCA scheme as well as its zero forcing (ZF) variant. However, there is about 3dB power penalty at high SNR due to less optimal TX-RX processing, but it still greatly outperforms the unicast reference case.

In Fig. 4 the performance of the linear scheme introduced in *Scenario 2* is compared with the more optimal but highly complex baseline scenario from [8]. Parameters α and β control the size of the subset $\{\mathcal{S} \subseteq [K]\}$ served during a given time interval, and the overlap among the multicast messages

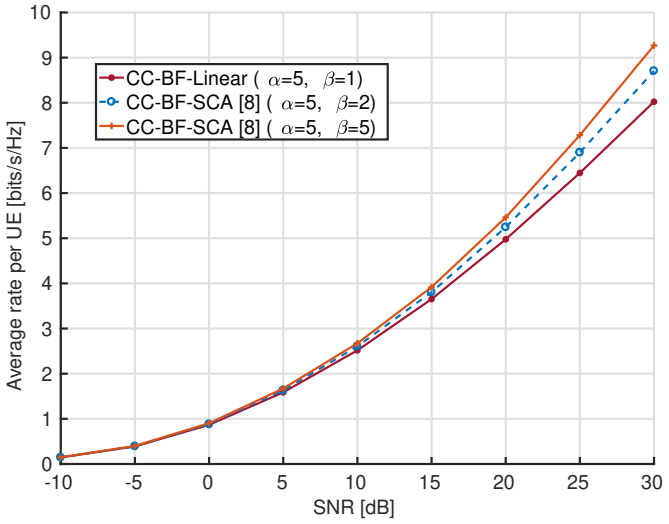


Fig. 4. Average 'Goodput' per user, $L = 5$ and $K = 6$

transmitted in parallel, respectively. The case labeled as 'CC-BF-SCA ($\alpha = 5$, $\beta = 2$) is an intermediate between the Linear scheme and the fully overlapping case ($\alpha = 5$, $\beta = 5$), allowing partial overlap among multicast messages transmitted in parallel, see more details in [8]. Similarly to *Scenario 1*, the results demonstrate that the simple linear scheme provides almost identical performance as the reference cases while having a minor SNR penalty at high SNR region.

V. Conclusions

A simple multicast beamforming strategy for CC was introduced. With a small loss in performance, the complexity of both the receiver and transmitter implementation can be significantly reduced as it avoids SIC structure altogether.

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] 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.

[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] ——, "Cache-aided interference management in wireless cellular networks," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–7.

[5] 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.

[6] 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.

[7] 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.

[8] 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

[9] ——, "Multicast beamformer design for coded caching," in *Proc. IEEE Int. Symp. Inform. Theory*, Vail, CO, USA, Jun 2018.

[10] 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

[11] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of Service and Max-Min Fair Transmit Beamforming to Multiple Cochannel Multicast Groups," *IEEE Trans. Signal Processing*, vol. 56, no. 3, pp. 1268–1279, 2008.

[12] 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.