This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

Multi-antenna Interference Management for Coded Caching

Antti Tölli, *Senior Member, IEEE,* Seyed Pooya Shariatpanahi, Jarkko Kaleva, *Member, IEEE* and Babak Khalaj, *Senior Member, IEEE*

***Abstract*—A multi-antenna broadcast channel scenario is considered where a base station delivers contents to cache-enabled user terminals. A joint design of coded caching (CC) and multigroup multicast beamforming is proposed to benefit from spatial multiplexing gain, improved interference management and the global CC gain, simultaneously. The developed general content delivery strategies utilize the multiantenna multicasting opportunities provided by the CC technique while optimally balancing the detrimental impact of both noise and inter-stream interference from coded messages transmitted in parallel. Flexible resource allocation schemes for CC are introduced where the multicast beamformer design and the receiver complexity are controlled by varying the size of the subset of users served during a given time interval, and the overlap among the multicast messages transmitted in parallel, indicated by parameters α and β, respectively. Degrees of freedom (DoF) analysis is provided showing that the DoF only depends on α while it is independent of β. The proposed schemes are shown to provide the same degrees-of-freedom at high signal-to-noise ratio (SNR) as the state-of-art methods and, in general, to perform significantly better, especially in the finite SNR regime, than several baseline schemes.**

I. Introduction

Video delivery will be responsible for about 80 percent of the mobile traffic by 2021 according to the Cisco traffic forecast report [1], which draws attention to the content caching technology as a key element of next generation networks. Content caching involves prefetching most popular contents at network edge during low-congested hours mitigating network overcrowding when the real requests of users will show up. This idea has been widely investigated in various wireless network scenarios such as using cache-enabled helpers [2], device-to-device collaboration [3], [4], small cell networks [5], multi-hop networks [6], and Cooperative Multi-Point (CoMP) [7].

While the above works clearly demonstrate the benefits of caching in wireless networks, the pioneering work of [8] 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, instead of simply replicating high-popularity contents near-or-at end-users (at the cache content placement phase), one should spread different contents at different caches. This way, at the content delivery phase, common coded messages could be broadcast to different users with different demands, that would benefit all of the users resulting in substantial gains in large networks. This *global caching gain* relies on the observation that almost in all communication scenarios, broadcasting is much simpler than unicasting. Also, as was proven later in [9], [10], the performance of this CC scheme is optimal under the assumption of uncoded prefetching, i.e. when coding is allowed only at the delivery phase. Follow-up works extend the coded caching scheme proposed in [8] to other setups such as online coded caching [11], hierarchical coded caching [12], and multi-server scenarios [13]. All these works suggest that the same kind of CC gain is achievable under various network models.

In order to examine the CC approach in wireless networks the specific characteristics of wireless medium (such as the broadcast nature, fading, and interference) must be investigated to be able to implement the original idea of [8] in mobile delivery scenarios. In order to achieve this goal, in this paper, we investigate the potentials of applying CC to a single-cell multiple-input single-output (MISO) broadcast channel (BC). In such a scenario a multi-antenna base station (BS) transmitter, which has access to the contents library, satisfies content requests of single-antenna users (mobile devices) via a shared wireless medium. The users are cache-enabled, and thus, before the delivery phase begins, they have cached relevant data from the library during off-peak hours. We focus on a joint design of the beamforming scheme used at the BS and the CC design of multicast messages such that the achievable delivery rate is maximized in finite SNR regime. The main goal of our paper is to employ the multiple antennas at the transmitter to manage the interaction between noise and interference between coded messages (i.e., inter-stream interference) and at the same time to benefit from the gains promised by the CC paradigm.

A. Related Work

In the context of benefiting from CC gains in wireless networks, the authors in [14] consider the effect of delayed channel state information at the transmitter (CSIT) and demonstrate a synergy between CSIT and caching. Moreover, the work [15] investigates wireless interference channels where

A. Tölli and J. Kaleva are with Centre for Wireless Communications, University of Oulu P.O. Box 4500, FIN-90014 University of Oulu, Finland {antti.tolli, jarkko.kaleva}@oulu.fi, S. P. Shariatpanahi is with the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran IR 1439957131, Iran (p.shariatpanahi@ut.ac.ir), and B. Khalaj is with Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran (khalaj@sharif.edu). Parts of this work has been published in 2018 IEEE International Symposium of Information Theory, 2018 International Workshop on Contenct Caching and Delivery in Wireless Networks and 2018 Asilomar Conference on Signals, Systems and Computers. This work was supported in part by the Academy of Finland grants No. 279101 and 319059, as well as 6Genesis Flagship grant No. 318927.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

both the transmitters and receivers are cache-enabled. They show that, considering one-shot transmission schemes, the caches at receive and transmit sides are of equal value in the sense of network DoF, which is also confirmed to be the case in cellular networks [16]. In contrast, [17] treats the same setup with mixed-CSIT and unveils the importance of receiver side memory in such a scenario. Cache-enabled interference channels are also investigated by other works such as [18]–[21] which do not restrict the schemes to be one-shot, and thus benefit from practically more complex interference alignment (IA) schemes. Also, the authors in [22] investigate CC schemes in wireless device-to-device networks and adapt the original CC scheme to a server-less setup, while [23] shows the benefit of device mobility in such scenarios. Furthermore, the cache-enabled cloud radio access networks (C-RAN) are studied in [24].

All the aforementioned papers consider wireless networks in the high signal-to-noise-ratio (SNR) regime, expressing their performance in terms of degrees-of-freedom (DoF). As high SNR analysis is not always a good indicator for practical implementations performance, there is still a gap which should be filled in with finite SNR analysis of the CC idea. The papers [25] and [26] propose different CC schemes in a wireless MISO-BC model, and provide a finite SNR analysis, in different system operating regimes. While the main idea in [25] is to use rate-splitting along with CC, the authors in [26] 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 [26] originally came from adapting the multi-server CC scheme of [13] (which is almost optimal in terms of DoF as shown in [15]) to a Gaussian MISO-BC, the interesting observations in [26] reveal that careful code and beamformer design modifications have significant effects on the finite SNR performance.

Moreover, it should be noted that [27] also considers using the rate-splitting along with CC and propose schemes benefiting from spatial multiplexing and CC gains in a MISO-BC setup. However, as shown in [28] the resulting DoF performance is worse than the zero-forcing proposal in [26], and, consequently, is inferior to our scheme as well. Although the works [29] and [30] consider the finite SNR performance of coded caching in broadcast channels, they assume a single-antenna transmitter, and thus in contrast to our paper, the interference management potentials of transmitter via its multiple antennas are not investigated. Finally, the authors in [31] addressed the subpacketization bottleneck in the multicast CC schemes [8], [13], [26], and proposed a simple ZF based multiantenna transmission scheme substantially reducing the required subpacketization, while providing the same high SNR DoF as [13], [26].

B. Main Contributions

In this paper, extending the joint interference nulling and CC concept originally proposed in [26], [28], a joint design of CC and generic multicast beamforming is introduced to simultaneously benefit from spatial multiplexing gain, improved management of inter-stream interference from coded messages transmitted in parallel, and the global caching gain. 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 such that $t = KM/N$ is an integer value. Due to the ZF beamforming constraint, the number of antennas is restricted to be strictly $L \leq K - t$ in [26], [28] as the null space beamformer is unique only when $L \leq K - t$. However, the generic multicast beamformer design introduced in this paper can be designed to fully benefit from the the interference free signal space even when $L > K - t$. 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) of non-convex SINR constraints, similarly to [32], are used to devise efficient iterative algorithms. In this paper, the multigroup multicast beamformer design is also generalized to any combination of *overlapping user groups* while assuming successive interference cancellation (SIC) receiver.

The generalized flexible multigroup multicast beamformer design proposed in this paper allows us to introduce innovative flexible resource allocation schemes for coded caching. Depending on the spatial degrees of freedom and the available SNR, a varying number of multicast messages can be transmitted in parallel to distinct subsets of users. Instead of always serving a group of $t + L$ users as in [26], [28], *the size of the user subset served during a given time interval* is controlled by a parameter α such that $t + \alpha \leq t + L$. Furthermore, the parameter β is introduced to control the *overlap among the multicast messages* transmitted in parallel. It defines how many parallel messages the receiver should be able to distinguish from each other using the SIC receiver. The benefits of new α- and β-schemes are twofold. First, the complexity of the beamformer design at the transmitter is managed by controlling the number of constraints and variables in the corresponding optimization problem. The receiver complexity depends on the number of parallel messages to be decoded by each user, which is controlled by the parameter β. In the least complex form of implementation with $\beta = 1$, the multicast messages do not overlap at all. This results in linear receiver implementation, which does not require SIC unlike in the general case. Second, by using $\alpha < L$, a better rate performance can be attained at low to medium SNR range by exploiting the transmit antennas to achieve multiplexing gain and at the same time compensating for the worst users channel effects. Thus, with a modest loss in performance only at high SNR region, the complexity of both the receiver and transmitter implementation can be significantly reduced, compared to the generalized multicast beamformer design with $\alpha = \beta = L$. Finally, DoF analysis of the proposed schemes is provided showing that the DoF only depends on α and it is independent of β.

Parts of this paper have been published in the conference publications [33]–[35]. Considering simple 3- and 4-user scenarios, the basic idea of combining multi-group multicast beamformer design and CC was first introduced in [33],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

while the reduced complexity multicast mode selection idea and the simple linear multicast beamforming strategy were introduced in [34] and [35], respectively. In this paper, in addition to simple scenarios considered in [33]–[35], a general content delivery scheme applicable for a wide range of the problem parameters values is provided along with a corresponding DoF analysis. Furthermore, all Matlab codes to regenerate the results of this paper are available online at https://github.com/kalesan/sim-cc-miso-bc.

In this paper we use the following notations. We use $(.)^{\mathrm{H}}$ to denote the Hermitian of a complex matrix. Let $\mathbb{C}$ and $\mathbb{N}$ denote the set of complex and natural numbers and $\|.\|$ be the norm of a complex vector. Also $[m]$ denotes the set of integer numbers $\{1, ..., m\}$, and $\oplus$ represents addition in the corresponding finite field. For any vector $\mathbf{v}$, we define $\mathbf{v}^{\perp}$ such that $\mathbf{v}^{\mathrm{H}}\mathbf{v}^{\perp} = 0$. Moreover, $\mathcal{A}$ and $|\mathcal{A}|$ denote a set of indexes and its cardinality, while a collection of sets and the number of such sets are indicated by B ans $|\mathsf{B}|$, respectively.

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. Each user k is equipped with a cache memory of MF bits and 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 $d_k \in [N]$ 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 $\mathbf{h}_k \in \mathbb{C}^L$, $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ is the multicast beamformer dedicated to users in subset $\mathcal{T}$ of set $\mathcal{S} \subseteq [K]$ 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 $t \triangleq KM/N$ [13], [28]. The main idea in CC is (by careful cache content placement) to provide multicasting opportunities to groups of size $t+1$, in which a common coded message would be useful for all the members of the multicast group. This is called the *Global Coded Caching Gain*, and is proportional to the total memory of the users, i.e., KM, normalized by the library size, i.e., N (for more details refer to [8]). 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 $\mathcal{S}$ 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 $\mathcal{S} \subseteq [K]$.

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 all users requesting for any file of the library. Multicasting opportunities due to the coded caching [8], [13], [26] are utilized to devise an efficient multiantenna multicast beamforming method that perform well over the entire SNR region.

Before presenting a high-level description of our proposed scheme, let us first provide a brief review of the original coded caching scheme proposed in [8], and then describe the multi-server and multi-antenna extensions in [13], and [26], [28], respectively. Assuming $t = KM/N \in \mathbb{N}$, the scheme proposed in [8] first divides each file W_n into $\binom{K}{t}$ subfiles (i.e., $W_n = \{W_{n,\tau}, \tau \subset [K], |\tau| = t\}$), then user k caches all subfiles $W_{n,\tau}$ in which $k \in \tau,\ \forall\ n$. In the delivery phase, it can be easily seen that for each $t+1$ subset of users $\mathcal{T}$, a common coded multicast message $\oplus_{k \in \mathcal{T}} W_{d_k, \mathcal{T}\backslash k}$ would benefit all the users in the subset $\mathcal{T}$ with providing them one subfile of their requested file [8]. If all such multicast coded messages are successfully delivered, then it can be shown that the missing subfiles will be received at all the users. This will result into the total normalized delay of $\binom{K}{t+1}/\binom{K}{t} = (K-t)/(t+1)$ [8].

When L servers have access to the library, instead of a single server, they can collaboratively send coded messages each of which would benefit $t+L$ users [13]. This means the global caching gain (proportional to t) and the collaboration multiplexing gain (proportional to L) are additive. This will directly reduce the delay by a multiplicative factor of $(t+1)/(t+L)$, which results in the normalized delay of $(K-t)/(t+L)$. The scheme proposed in [13] combines zero-forcing gain and multicasting opportunities provided by coded caching as follows. Assume a subset of users of size $t+L$. Then according to the proposal in [8], a coded common message can be constructed to serve each $t+1$ subset of these users. Now by combining all $\binom{t+L}{t+1}$ such coded messages, with each message directed in the null space of $L-1$ undesired users, all of them can be sent simultaneously. However, any user k will have to decode $m_k = \binom{t+L-1}{t}$ different messages, which grows exponentially when K, L, N are increased with the same ratio (linearly if $t = KM/N = 1$).

The multiserver scheme [13] is adapted to a multiple-antenna transmitter delivering files to users via a wireless medium in [26], [28], where each coded message destined to $t+1$ users is nulled at $L-1$ non-desired users. For each user k, the received signal y_k contains m_k desired signals. Hence, from the receiver perspective, it appears as m_k-dimensional Gaussian multiple access channel (MAC) with a feasible rate region defined by $2^{m_k} - 1$ rate constraints. Thus, a SIC structure has to be used at each receiver to decode the intended messages [26], [28].

While adapting the multi-server coded caching scheme to the wireless multiple-antenna setup was shown to improve

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

the rate of the system also at finite SNR [28], compared to non-coded schemes, using the ZF vectors at finite SNR is a highly sub-optimal strategy in general. At finite SNR, the detrimental impact of both inter-stream interference and noise needs to be balanced to arrive at the best performance. Thus, in our proposed scheme we address the challenge of finding the optimum multigroup multicast beamforming vectors with any combination of overlapping user groups for minimizing the delivery time while allowing controlled inter-stream interference among multicast messages transmitted in parallel. Although this modification will significantly improve the system performance especially at low SNR, it introduces a beamformer optimization problem which has significantly higher computational complexity than the simple ZF scheme used in [26], [28].

The aforementioned generalized multigroup multicast beamformer design allows us to introduce innovative flexible resource allocation schemes for coded caching, depending on the available transmit power or computational complexity constraints. Instead of always serving a group of $t+L$ users as in [26], [28], *the size of the user subset served during a given time interval* can be controlled by a parameter α such that $t+\alpha \leq t+L$. Furthermore, the parameter β is introduced to control the *overlap among the multicast messages* transmitted in parallel. It defines how many parallel messages the receiver should be able to distinguish from each other using the SIC receiver. Thus, the complexity of both the receiver and transmitter implementation can be significantly reduced without significant , compared to the generalized multicast beamformer design with fully overlapping multicast messages, $\alpha=\beta=L$.

In the rest of this section, we first introduce the proposed concept and its variations in four simple scenarios and discuss the generalization of the proposed schemes in Section IV. The multigroup multicast beamformer design for the classical 3-user case [8], [13], [26] is first described in *Scenario 1*, which in turn is extended to 4-user case in *Scenario 2* to demonstrate how the size and complexity of the problem quickly increases for larger values of K. Reduced complexity beamformer design alternatives to *Scenario 2* are introduced in *Scenarios 3 and 4* by controlling the size α of the subset $\{\mathcal{S}\subseteq[K]\}$ served during a given time interval, and the overlap β among the multicast messages transmitted in parallel, respectively.

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 $K=3$ 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 [8] 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.

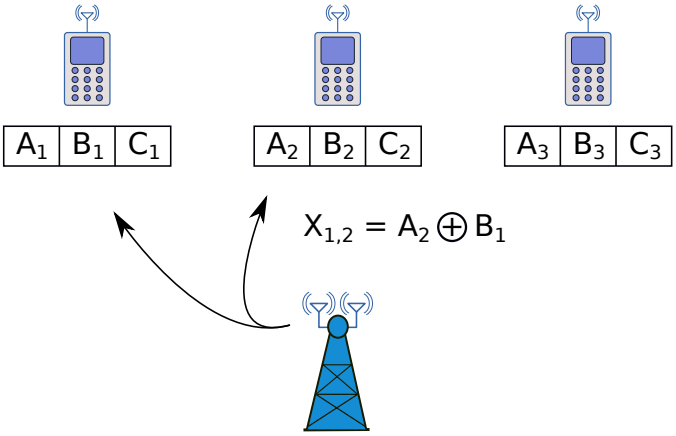


Fig. 1. *Scenario 1*: $L=2$, $K=3$, $N=3$ and $M=1$

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 [8], the following coded messages are sent to users $\mathcal{S}=\{1,2,3\}$ by the transmitter 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, and the superscript $\mathcal{S}$ is omitted for ease of presentation. In such coding scheme, each coded message is received by all 3 users, but is only beneficial to 2 of them. For example, $X_{1,2}$ is useful for the 1st and 2nd user only. It can be easily checked that after transmission is concluded, all 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.

Consequently, in *Scenario 1*, we can combine the spatial multiplexing gain and the global caching gain following the scheme in [26] (see also [13], [15]). In [26], 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 X, chosen from a unit power complex Gaussian codebook [26].

The key point here is to note that although this scheme is order-optimal in terms of DoF [15] it is suboptimal at low SNR regime [26], [28]. Therefore, in this paper, instead of nulling interference at unwanted users, general multicast beamforming vectors $\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}$ are defined 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 $\{1,...,K\}$ and the superscript $\mathcal{S}$ is omitted for simplicity. As a result, the received signals at users $1-3$ will be

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

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[1]

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

where $R^1_{Sum} = R^1_1 + R^1_2$ and the rates R^1_1 and R^1_2 correspond to $\tilde{X}_{1,2}$, and $\tilde{X}_{1,3}$, respectively. 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

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

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

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

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

$$\max_{\mathbf{w}_{2,3},\mathbf{w}_{1,3},\mathbf{w}_{1,2}} \min_{k=1,2,3} R^k_{MAC}. \tag{8}$$

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

$$\begin{aligned}
\max_{R^k,\gamma^k_i,\mathbf{w}_\mathcal{T},\forall k,i} \min_{k=1,2,3} \min &\left(\frac{1}{2}R^k_{\text{sum}}, R^k_1, R^k_2\right) \\
\text{s. t. } & R^k_1 \le \log(1+\gamma^k_1), R^k_2 \le \log(1+\gamma^k_2), \\
& R^k_{\text{sum}} \le \log(1+\gamma^k_1+\gamma^k_2),\ k=1,2,3, \\
& \gamma^1_1 \le \frac{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{1,2}|^2}{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2+N_0}, \gamma^1_2 \le \frac{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{1,3}|^2}{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2+N_0}, \\
& \gamma^2_1 \le \frac{|\mathbf{h}^{\text{H}}_2\mathbf{w}_{2,3}|^2}{|\mathbf{h}^{\text{H}}_2\mathbf{w}_{1,3}|^2+N_0}, \gamma^2_2 \le \frac{|\mathbf{h}^{\text{H}}_2\mathbf{w}_{1,2}|^2}{|\mathbf{h}^{\text{H}}_2\mathbf{w}_{1,3}|^2+N_0}, \\
& \gamma^3_1 \le \frac{|\mathbf{h}^{\text{H}}_3\mathbf{w}_{2,3}|^2}{|\mathbf{h}^{\text{H}}_3\mathbf{w}_{1,2}|^2+N_0}, \gamma^3_2 \le \frac{|\mathbf{h}^{\text{H}}_3\mathbf{w}_{1,3}|^2}{|\mathbf{h}^{\text{H}}_3\mathbf{w}_{1,2}|^2+N_0}, \\
& \textstyle\sum_{\mathcal{T}\in\{\{1,2\},\{1,3\},\{2,3\}\}} \|\mathbf{w}_\mathcal{T}\|^2 \le \text{SNR}
\end{aligned} \tag{9}$$

which can be equally presented in an epigraph from as

$$\begin{aligned}
\max_{r,\gamma^k,\mathbf{w}_{k,i}} \ & r \\
\text{s. t. } & r \le \frac{1}{2}\log(1+\gamma^k_1+\gamma^k_2),\ k=1,2,3, \\
& r \le \log(1+\gamma^k_1),\ r \le \log(1+\gamma^k_2) \\
& \text{The rest of the constraints as in (9).}
\end{aligned} \tag{10}$$

Problem (10) is non-convex due to the SINR constraints. Similarly to [32], 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

$$\gamma^1_1 \le \frac{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{1,2}|^2}{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2+N_0} \tag{11}$$

$$|\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2 + N_0 \le \frac{|\mathbf{h}^{\text{H}}_1\mathbf{w}_{1,2}|^2 + |\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2 + N_0}{1+\gamma^1_1} \tag{12}$$

where both sides of the inequality constraint are convex functions (quadratic or quadratic-over-linear) with respect to the optimization variables. In order to make it convex, the R.H.S of (12) is linearly approximated (lower bounded) as

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{2,3},\mathbf{w}_{1,2},\mathbf{h}_1,\gamma^1_1) \triangleq\ & \frac{|\mathbf{h}^{\text{H}}_1\bar{\mathbf{w}}_{1,2}|^2 + |\mathbf{h}^{\text{H}}_1\bar{\mathbf{w}}_{2,3}|^2 + N_0}{1+\bar{\gamma}^1_1} \\
& - 2\mathbb{R}\left(\frac{\bar{\mathbf{w}}^{\text{H}}_{1,2}\mathbf{h}_1\mathbf{h}^{\text{H}}_1}{1+\bar{\gamma}^1_1}(\mathbf{w}_{1,2}-\bar{\mathbf{w}}_{1,2})\right) \\
& - 2\mathbb{R}\left(\frac{\bar{\mathbf{w}}^{\text{H}}_{2,3}\mathbf{h}_1\mathbf{h}^{\text{H}}_1}{1+\bar{\gamma}^1_1}(\mathbf{w}_{2,3}-\bar{\mathbf{w}}_{2,3})\right) \\
& + \frac{|\mathbf{h}^{\text{H}}_1\bar{\mathbf{w}}_{1,2}|^2 + |\mathbf{h}^{\text{H}}_1\bar{\mathbf{w}}_{2,3}|^2 + N_0}{(1+\bar{\gamma}^1_1)^2}\left(\gamma^1_1 - \bar{\gamma}^1_1\right)
\end{aligned} \tag{13}$$

where $\mathbb{R}(\cdot)$ is the real part of the complex valued argument, $\bar{\mathbf{w}}_{k,i}$ and $\bar{\gamma}^1_1$ denote the fixed values (points of approximation) for the corresponding variables from the previous iteration. Using (13) and reformulating the objective in the epigraph form, the approximated problem is written as

$$\begin{aligned}
\max_{r,\gamma^k,\mathbf{w}_\mathcal{T}} \ & r \\
\text{s. t. } & r \le 1/2\log(1+\gamma^k_1+\gamma^k_2), \\
& r \le \log(1+\gamma^k_1),\ r \le \log(1+\gamma^k_2)\ k=1,2,3, \\
& \mathcal{L}(\mathbf{w}_{2,3},\mathbf{w}_{1,2},\mathbf{h}_1,\gamma^1_1) \ge |\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2 + N_0, \\
& \mathcal{L}(\mathbf{w}_{2,3},\mathbf{w}_{1,3},\mathbf{h}_1,\gamma^1_2) \ge |\mathbf{h}^{\text{H}}_1\mathbf{w}_{2,3}|^2 + N_0, \\
& \mathcal{L}(\mathbf{w}_{1,3},\mathbf{w}_{2,3},\mathbf{h}_2,\gamma^2_1) \ge |\mathbf{h}^{\text{H}}_2\mathbf{w}_{1,3}|^2 + N_0, \\
& \mathcal{L}(\mathbf{w}_{1,3},\mathbf{w}_{1,2},\mathbf{h}_2,\gamma^2_2) \ge |\mathbf{h}^{\text{H}}_2\mathbf{w}_{1,3}|^2 + N_0, \\
& \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{2,3},\mathbf{h}_3,\gamma^3_1) \ge |\mathbf{h}^{\text{H}}_3\mathbf{w}_{1,2}|^2 + N_0, \\
& \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{1,3},\mathbf{h}_3,\gamma^3_2) \ge |\mathbf{h}^{\text{H}}_3\mathbf{w}_{1,2}|^2 + N_0, \\
& \textstyle\sum_{\mathcal{T}\in\{\{1,2\},\{1,3\},\{2,3\}\}} \|\mathbf{w}_\mathcal{T}\|^2 \le \text{SNR}
\end{aligned} \tag{14}$$

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 (14) 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

$$r \le \frac{1}{2}\log(1+\gamma^k_1+\gamma^k_2) = \log(\sqrt{1+\gamma^k_1+\gamma^k_2}) \le \sqrt{1+\gamma^k_1+\gamma^k_2}$$

Now, the equivalent SOCP reformulation follows as

$$\begin{aligned}
\max_{\tilde{r},\gamma^k,\mathbf{w}_k} \ & \tilde{r} \\
\text{s. t. } & \tilde{r}^2 \le 1+\gamma^k_1+\gamma^k_2, \tilde{r} \le 1+\gamma^k_1,\ \tilde{r} \le 1+\gamma^k_2\ k=1,2,3, \\
& \text{The rest of the constraints as in (14) .}
\end{aligned} \tag{15}$$

Finally, a solution for the original problem (9) can be found by solving (14) in an iterative manner using SCA, i.e, by updating the points of approximations $\bar{\mathbf{w}}_{k,i}$ and $\bar{\gamma}^l_j$ in (13) after each

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

iteration. As each difference-of-convex constraint in (12) is lower bounded by (13), the monotonic convergence of the objective of (14) is guaranteed. 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[2]. 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 (9) becomes:

$$\begin{aligned}\max_{R^k,\gamma^k,p_{\mathcal{T}}} \min_{k=1,2,3} \min\left(\frac{1}{2}R^k_{\text{sum}}, R^k_1, R^k_2\right) \quad (16)\\ \text{s. t.} \quad R^k_{\text{sum}} \le \log(1+\gamma^k_1+\gamma^k_2),\\ R^k_1 \le \log(1+\gamma^k_1), R^k_2 \le \log(1+\gamma^k_2)\ \forall\ k,\\ \gamma^1_1 \le u_{1,3}p_{1,2}, \gamma^1_2 \le u_{1,2}p_{1,3}, \gamma^2_1 \le u_{2,1}p_{2,3},\\ \gamma^2_2 \le u_{2,3}p_{1,2}, \gamma^3_1 \le u_{3,1}p_{2,3}, \gamma^3_2 \le u_{3,2}p_{1,3},\\ \textstyle\sum_{\mathcal{T}\in\{\{1,2\},\{1,3\},\{2,3\}\}} p_{\mathcal{T}} \le \text{SNR}\end{aligned}$$

where $u_{k,i} = |\mathbf{h}_k^{\text{H}}\mathbf{h}_i^{\perp}|^2/\|\mathbf{h}_i^{\perp}\|^2 N_0$. 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) [26]: 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 [26].

2) 2nd Baseline Scheme: MaxMinSNR Multicasting: In this case, a single multicast stream is transmitted at a time *without any inter-stream interference*. Thus, three time slots are required to deliver messages $X_{1,2}$, $X_{1,3}$ and $X_{2,3}$. In timeslot 1, the message $X_{1,2}$ is delivered to the users 1 and 2 by sending the signal $\mathbf{w}_{1,2}\tilde{X}_{1,2}$. A single transmit beamformer $\mathbf{w}_{1,2}$ is found to minimize the time needed for multicasting the common message:[3]

$$T_{1,2} = \frac{F/3}{\max\limits_{\|\mathbf{w}_{1,2}\|^2\le SNR} \min\left(\log(1+\frac{|\mathbf{h}_1^{\text{H}}\mathbf{w}_{1,2}|^2}{N_0}), \log(1+\frac{|\mathbf{h}_2^{\text{H}}\mathbf{w}_{1,2}|^2}{N_0})\right)}. \quad (17)$$

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

$$R_{\text{maxmin}} = F/(T_{1,2}+T_{1,3}+T_{2,3}). \quad (18)$$

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

[2]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).

[3]This multicast maxmin problem is NP-hard in general, but near-optimal solutions can be obtained by a semidefinite relaxation (SDR) approach, see [26] and the references therein.

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 $L \ge K$. On the other hand, if $L < K$, the users need to be divided into subsets of size L served at distinct time slots.

Now, let us consider the 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

$$T_{1,2} = \frac{F/3}{\max\limits_{\sum_{k=1,2}\|\mathbf{w}_k\|^2\le SNR} \min(R_1, R_2)} \quad (19)$$

where

$$R_k = \log\left(1 + \frac{|\mathbf{h}_k^{\text{H}}\mathbf{w}_k|^2}{\sum_{i\ne k}|\mathbf{h}_k^{\text{H}}\mathbf{w}_i|^2 + N_0}\right). \quad (20)$$

The minimum delivery time in (18) 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 (18).

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

In this scenario, the number of users K and files N is further increased in order to demonstrate how the size and complexity of the problem quickly increases for larger values of K. 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 [8] the cache contents of users are as follows

$$\begin{aligned}Z_1 = \{A_1, B_1, C_1, D_1\}, Z_2 = \{A_2, B_2, C_2, D_2\},\\ Z_3 = \{A_3, B_3, C_3, D_3\}, Z_4 = \{A_4, B_4, C_4, D_4\} \quad (21)\end{aligned}$$

where here 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. Here, we have $t \triangleq KM/N = 1$ and the subsets $\mathcal{S}$ and $\mathcal{T}$ will be of size 4 and $t+1 = 2$, respectively (for details see [13], [28] and Section IV). Following the approach of *Scenario 1*, the transmit signal vector is

$$\begin{aligned}\sum_{\mathcal{T}\subseteq\mathcal{S},|\mathcal{T}|=2} \mathbf{w}_{\mathcal{T}}\tilde{X}_{\mathcal{T}} = \mathbf{w}_{1,2}\tilde{X}_{1,2} + \mathbf{w}_{1,3}\tilde{X}_{1,3} + \mathbf{w}_{1,4}\tilde{X}_{1,4}\\ + \mathbf{w}_{2,3}\tilde{X}_{2,3} + \mathbf{w}_{2,4}\tilde{X}_{2,4} + \mathbf{w}_{3,4}\tilde{X}_{3,4} \quad (22)\end{aligned}$$

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

The received signal at user $k = 1, 2, 3, 4$ is written as

$$\begin{aligned}y_k = \underline{(\mathbf{h}_k^{\text{H}}\mathbf{w}_{1,2})\tilde{X}_{1,2}} + \underline{(\mathbf{h}_k^{\text{H}}\mathbf{w}_{1,3})\tilde{X}_{1,3}} + \underline{(\mathbf{h}_k^{\text{H}}\mathbf{w}_{1,4})\tilde{X}_{1,4}}\\ + (\mathbf{h}_k^{\text{H}}\mathbf{w}_{2,3})\tilde{X}_{2,3} + (\mathbf{h}_k^{\text{H}}\mathbf{w}_{2,4})\tilde{X}_{2,4} + (\mathbf{h}_k^{\text{H}}\mathbf{w}_{3,4})\tilde{X}_{3,4} + z_1\end{aligned}$$

where, as an example, the desired terms of user 1 are underlined. As in *Scenario 1*, each user faces a MAC channel, now with three desired signals, three Gaussian interference terms,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

and one noise term. Suppose that user k can decode each of its desired signals with the rate R^k_{MAC}. Consequently, this user receives useful information with the rate $3R^k_{MAC}$, and the time required to fetch the entire file is $T_1 = \frac{3F}{4}\frac{1}{3R^k_{MAC}}$. Following the same steps as in (6)–(7), the symmetric rate per user can be found as

$$R_{sym} = \frac{F}{T} = 4 \max_{\mathbf{w}_\mathcal{T}, \mathcal{T}\subseteq[4], |\mathcal{T}|=2} \min_{k=1,2,3,4} R^k_{MAC} \tag{23}$$

where

$$R^k_{MAC} = \min\left(R^k_1, R^k_2, R^k_3, \frac{1}{2}R^k_4, \frac{1}{2}R^k_5, \frac{1}{2}R^k_6, \frac{1}{3}R^k_7\right) \tag{24}$$

and where the rate bounds R^1_1, R^1_2 and R^1_3 of user 1, for example, correspond to $\tilde{X}_{1,2}$, $\tilde{X}_{1,3}$ and $\tilde{X}_{1,4}$, respectively. The bounds R^1_4, R^1_5 and R^1_6 limit the sum rate of any combination of two transmitted multicast signals, and finally R^1_7 is the sum rate bound for all 3 messages. The SCA method is again used to solve (23), similarly to (9)-(14).

C. Scenario 3: $L \geq 3$, $K = 4$, $N = 4$, $M = 1$ and $\alpha = 2$

In this example, a reduced complexity alternative for *Scenario 2* is considered. Instead of fixing the size of the served user set to $|\mathcal{S}| = \min(t + L, K) = 4$ as in *Scenario 2*, we set $\alpha = 2$ and restrict the size of the subsets $\mathcal{S} \subset [4]$ benefiting from a common transmitted signal to $|\mathcal{S}| = t + \alpha = 3$. Thus, the size of the MAC channel for each user is reduced from 3 to 2 and each user needs to decode just 2 multicast streams. This in turn, reduces the complexity of the problem for determining the beamforming vectors for each subset $\mathcal{S} \subset [4]$. As will be shown later, besides complexity reduction, controlling the size of each subset allows us to handle the trade-off between the multiplexing and multicast beamforming gains due to multiple transmit antennas, resulting in even better rate performance at certain SNR values. Note that for $|\mathcal{S}| = 2$ ($\alpha = 1$), the beamformer design for each subset $\mathcal{S}$ reduces to the baseline max-min SNR scheme (see (17) for $K = 3$).

The cache content placement works similarly, except that each subfile is split into 2 mini-files (indicated by superscripts) in order to allow different contents to be transmitted in each subset $\mathcal{S}$. As a result, the following content is stored in user cache memories

$$\begin{aligned}
Z_1 &= \{A^1_1, A^2_1, B^1_1, B^2_1, C^1_1, C^2_1, D^1_1, D^2_1\},\\
Z_2 &= \{A^1_2, A^2_2, B^1_2, B^2_2, C^1_2, C^2_2, D^1_2, D^2_2\}\\
Z_3 &= \{A^1_3, A^2_3, B^1_3, B^2_3, C^1_3, C^2_3, D^1_3, D^2_3\},\\
Z_4 &= \{A^1_4, A^2_4, B^1_4, B^2_4, C^1_4, C^2_4, D^1_4, D^2_4\}
\end{aligned}$$

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

$$\mathbf{w}_{1,2}\tilde{X}_{1,2} + \mathbf{w}_{1,3}\tilde{X}_{1,3} + \mathbf{w}_{2,3}\tilde{X}_{2,3} \tag{25}$$

where $X_{1,2} = A^1_2 \oplus B^1_1, X_{1,3} = A^1_3 \oplus C^1_1, X_{2,3} = B^1_3 \oplus C^1_2$. This transmission should be such that $X_\mathcal{T}$ is received correctly at all users in $\mathcal{T} \subset \{1, 2, 3\}, |\mathcal{T}| = 2$. 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

$$\mathbf{w}_{1,2}\tilde{X}_{1,2} + \mathbf{w}_{1,4}\tilde{X}_{1,4} + \mathbf{w}_{2,4}\tilde{X}_{2,4} \tag{26}$$

where $X_{1,2} = A^2_2 \oplus B^2_1, X_{1,4} = A^1_4 \oplus D^1_1, X_{2,4} = B^1_4 \oplus D^1_2$ 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 $\{1, 2\}$ 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

$$\mathbf{w}_{1,3}\tilde{X}_{1,3} + \mathbf{w}_{1,4}\tilde{X}_{1,4} + \mathbf{w}_{3,4}\tilde{X}_{3,4} \tag{27}$$

$$\mathbf{w}_{2,3}\tilde{X}_{2,3} + \mathbf{w}_{2,4}\tilde{X}_{2,4} + \mathbf{w}_{3,4}\tilde{X}_{3,4} \tag{28}$$

respectively, where $X_{1,3} = A^2_3 \oplus C^2_1$, $X_{1,4} = A^2_4 \oplus D^2_1$, $X_{3,4} = C^1_4 \oplus D^1_3$ 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 $X_{2,3} = B^2_3 \oplus C^2_2, X_{2,4} = B^2_4 \oplus D^2_2, X_{3,4} = C^2_4 \oplus D^2_3$ 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

$$\begin{aligned}
&\frac{F}{T_{1,2,3} + T_{1,2,4} + T_{1,3,4} + T_{2,3,4}}\\
&= 8\left(\frac{1}{R_{1,2,3}} + \frac{1}{R_{1,2,4}} + \frac{1}{R_{1,3,4}} + \frac{1}{R_{2,3,4}}\right)^{-1}.
\end{aligned} \tag{29}$$

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 in this scenario we have potentially more antennas available ($L \geq 3$) allowing for further improved multicast beamforming performance.

D. Scenario 4: Simple Linear TX-RX strategy, $\alpha = 3$ and $\beta = 1$

In *Scenarios 1–3*, each user is allocated with a number of parallel streams that need to be decoded using SIC receiver structure. In this example, in contrast, we consider the same setting as in *Scenarios 2-3* with $L \geq 3$, $K = 4$, $N = 4$, $M = 1$ but no overlap is allowed among user groups served by multiple multicast messages transmitted in parallel, i.e., we use parameters $\beta = 1$ and $\alpha = 3$. This leads to a simpler TX-RX strategy where all 6 multicast streams introduced in *Scenario 2* are delivered across three orthogonal time intervals/slots, instead of transmitting all in parallel as in (22). In time slots 1–3, the multicast beamforming vectors are generated as $\mathbf{w}_{1,2}(A_2 \oplus B_1) + \mathbf{w}_{3,4}(C_4 \oplus D_3)$, $\mathbf{w}_{1,3}(A_3 \oplus C_1) + \mathbf{w}_{2,4}(B_4 \oplus D_2)$ and $\mathbf{w}_{1,4}(A_4 \oplus D_1) + \mathbf{w}_{2,3}(B_3 \oplus C_2)$, respectively.

In each time slot, all 4 users are served with 2 parallel multicast streams. Each stream causes inter-stream interference to 2 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 multicast streams. The beamforming vectors are optimized separately to maximize the symmetric

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

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 3 different time slots, the overall *Symmetric Rate Per User* of this scheme is

$$\frac{F}{\sum_{i=1,2,3} T(i)} = 4\left(\sum\nolimits_{i=1,2,3} R_C(i)^{-1}\right)^{-1}. \quad (30)$$

As will be shown in Section V, the scheme provides the same overall DoF (slope) as the original scheme in *Scenario 2*, 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 *Scenario 2*. 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 $R_C(i) = \log(1+\gamma_C(i))$. The multigroup multicast beamformer optimization problem for ith timeslot can be then expressed as the following common SINR maximization problem:

$$\begin{aligned} \max_{\gamma_C(i),\mathbf{w}_\mathcal{T}} \quad & \gamma_C(i) \\ \text{s. t.} \quad & \gamma_C(i) \le \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_\mathcal{T}(i)|^2}{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\bar{\mathcal{T}}}(i)|^2 + N_0}, \\ & \forall k \in \mathcal{T}, \mathcal{T} \in \mathsf{P}(i), \bar{\mathcal{T}} \in \mathsf{P}(i) \setminus \mathcal{T}, \\ & \textstyle\sum_{\mathcal{T}\in\mathsf{P}(i)} \|\mathbf{w}_\mathcal{T}(i)\|^2 \le \text{SNR}. \end{aligned} \quad (31)$$

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) [36]. Here, instead, we adopt the SCA solution from [32], based on which (31) 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 [32].

For example, by approximating the SINR constraints as in (12)–(13), 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

$$\begin{aligned} \max_{\gamma_C,\mathbf{w}_{1,2},\mathbf{w}_{3,4}} \quad & \gamma_C \\ \text{s. t.} \quad & \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{3,4},\mathbf{h}_1,\gamma_C) \ge |\mathbf{h}_1^{\mathrm{H}}\mathbf{w}_{3,4}|^2 + N_0, \\ & \mathcal{L}(\mathbf{w}_{1,2},\mathbf{w}_{3,4},\mathbf{h}_2,\gamma_C) \ge |\mathbf{h}_2^{\mathrm{H}}\mathbf{w}_{3,4}|^2 + N_0, \\ & \mathcal{L}(\mathbf{w}_{3,4},\mathbf{w}_{1,2},\mathbf{h}_3,\gamma_C) \ge |\mathbf{h}_3^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + N_0, \\ & \mathcal{L}(\mathbf{w}_{3,4},\mathbf{w}_{1,2},\mathbf{h}_4,\gamma_C) \ge |\mathbf{h}_4^{\mathrm{H}}\mathbf{w}_{1,2}|^2 + N_0, \\ & \|\mathbf{w}_{1,2}\|^2 + \|\mathbf{w}_{3,4}\|^2 \le \text{SNR}. \end{aligned} \quad (32)$$

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

IV. General Case Formulation, Algorithm, and Rate Analysis

As mentioned in Section III, in the multiserver scheme [13], [26], [28], each user k will have to decode $m_k = \binom{t+L-1}{t}$ different messages, which grows exponentially when K, L, N are increased with the same ratio (linearly if $t = KM/N = 1$). Thus, the total number of rate constraints in the beamformer optimization problem is $(t+L)(2^{m_k}-1)$. 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 to be approximated would be 5×4. 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 multicast messages transmitted in parallel as in *Scenario 3* or limit the overlap among the multicast messages as in *Scenario 4*, reflected in parameters α and β, respectively.

In the following, the general algorithm for the delivery phase for any set of parameters K, L, N and M in Algorithm 1 is described. Let us first provide a light description of the algorithm. Algorithm's inputs contains library contents $W_1,\ldots,W_N$, user requests indices $d_1,\ldots,d_K$, and the channel gain matrix $\mathbf{H} = [\mathbf{h}_1,\ldots,\mathbf{h}_K]$. In addition, the algorithm takes two design parameters α and β which should be tuned based on the working SNR, and the required complexity of the involved optimization problem.

The cache content placement phase is similar to the scheme [8] introduced in Section III, where each file is split into $\binom{K}{t}$ subfiles. The only difference is that here we further need to split each subfile in [8] into smaller fragments such that the total number of minifiles is

$$\binom{K}{t}\binom{K-t-1}{\alpha-1}\frac{(\alpha-1)!}{(\delta-1)!(\beta-1)!(t+\beta)!^{\delta-1}} \quad (33)$$

where $\delta := \frac{t+\alpha}{t+\beta} \in \mathbb{N}$. This further splitting is needed in order to allow different content to be transmitted in each additional time interval introduced due to parameters α and β, similarly to *Scenario 3*. More thorough justification for the second and third terms of (33) and their dependence of α and β is given in the latter part of this section. Note that (33) is reduced to [13] if $\alpha=\beta=L$, and [8] if $\alpha=\beta=1$.

As shown in [13], [28], coded caching/multicasting and spatial multiplexing gains are additive and the maximum DoF is upper bounded by $t+L$ (or by $\min(t+L,K)$ if $L>K-t$). In the generalized scheme, instead of fixing the size of the subsets $\{\mathcal{S}\subseteq[K]\}$ to be $\min(t+L,K)$ as in *Scenario 2*, we introduce a new integer parameter α bounded by

$$1 \le \alpha \le \min(L, K-t) \quad (34)$$

and define the size of subsets $\{\mathcal{S}\subseteq[K]\}$ to be $t+\alpha \le t+\min(L,K-t)$. The parameter α controls the available spatial multiplexing gain and has two main roles. First, it manages the trade-off between the spatial multiplexing and multicast beamforming/diversity gains due to optimized use of multiple transmit antennas, and thus should be designed carefully at each SNR to result in the maximum throughput. Second, it enables us to control the size of the MAC channel elements with respect to each user, and in turn, to control the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

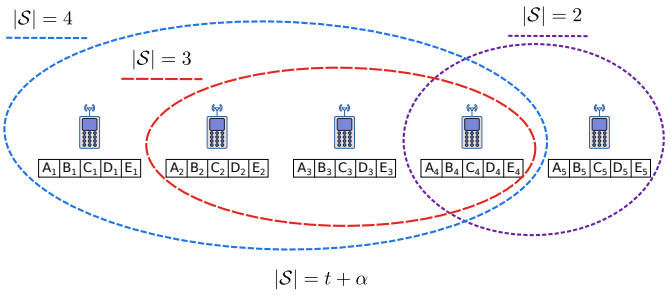
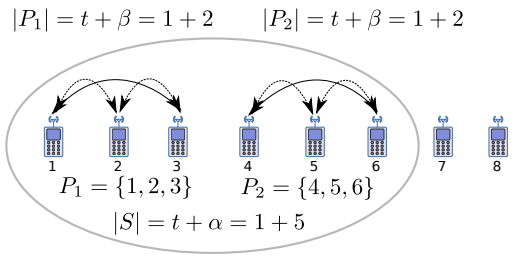


Fig. 2. Example partitioning of users into sets $\{\mathcal{S} \subseteq [K]\}$ in a scenario with $K = N = 5$, $L = 4$, $|\mathcal{S}| = t + \alpha = [2, 3, 4]$.



P_1	P_2
1, 2, 3	**4, 5, 6**
1, 2, 4	3, 5, 6
1, 2, 5	3, 4, 6
1, 2, 6	3, 4, 5
1, 3, 4	2, 5, 6
1, 3, 5	2, 4, 6
1, 3, 6	2, 4, 5
1, 4, 5	2, 3, 6
1, 4, 6	2, 3, 5
1, 5, 6	2, 3, 4

Fig. 3. Example partitioning of users in $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ into $(t + \beta)$-groups in a scenario with $K = N = 8$, $L \geq 5$, $t = 1$, $\alpha = 5$ and $\beta = 2$.

optimization problem complexity for determining the beamforming vectors, as will be explained later. This generalization reduces to the baseline max-min SNR beamforming scheme if $\alpha = 1$ (see (17) for $K = 3$).

Fig. 2 illustrates a possible partitioning of users into sets $\{\mathcal{S} \subseteq [K]\}$ in a scenario with $K = N = 5$, $L = 4$ and $M = 1$, and with $|\mathcal{S}| = t + \alpha = \{2, 3, 4\}$ ($\alpha = \{1, 2, 3\}$). 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 $\{\mathcal{S}, |\mathcal{S}| = 3\}$ or $\{\mathcal{S}, |\mathcal{S}| = 4\}$ corresponds to *Scenario 1* or *Scenario 2*, respectively, and the optimal multicast beamformers can be found by solving (14) or (23) (for corresponding $k \in \mathcal{S}$).

After the required initializations, the algorithm contains an outer loop which goes over all the $(t + \alpha)$-subsets of all the users $[K]$. Let us now consider a scenario with $K = 8$ users with $t = 1$ and $\alpha = 5$ depicted in Fig. 3 and focus on one particular realization of these $(t + \alpha)$-subsets $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. For this specific set $\mathcal{S}$, the second loop goes over all possible partitionings of $\mathcal{S}$ into $(t + \beta)$-groups, which are collected in P. Here, β, bounded by $1 \leq \beta \leq \alpha$, is another design parameter which controls the overlap among the multicast messages, i.e., the complexity of the beamformer design problem.

In the example of Fig. 3, $\beta = 2$ is selected showing one possible partitioning of users in to $\delta = 2$ groups $\mathsf{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$ (of all 10 possible partitionings shown also in the table), where $\mathcal{P}_1 = \{1, 2, 3\}$ and $\mathcal{P}_2 = \{4, 5, 6\}$. Then, for a specific partitioning P, we form the coded messages for all $(t+1)$-subsets of each group, for all the groups $\mathcal{P}_i \in \mathsf{P}, i = 1, ..., \delta$. In this paper, the number of $t+\beta$ sets within each $t+\alpha$ set is restricted to integer values $\delta := \frac{t+\alpha}{t+\beta} \in \mathbb{N}$.[4] In this example, there are 3 parallel coded messages for every pair of users inside $\mathcal{P}_1$, and 3 coded messages for every pair in $\mathcal{P}_2$, resulting in a total of 6 coded messages. It should be noted that common messages for users in different groups are not allowed, which is the main ingredient behind controlling complexity of the beamformer design. In general, assuming $\delta \in \mathbb{N}$, there will be $\delta\binom{t+\beta}{t+1}$ coded messages involved for a fixed partitioning P, while these $(t+1)$-subsets for multicast beamforming are collected in the collection of sets $\Omega^{\mathcal{S},\mathsf{P}} := \bigcup_{i=1,...,\delta}\{\mathcal{T} \subseteq \mathcal{P}_i, |\mathcal{T}| = t + 1\}$ for a specific $\mathcal{S}$ and P. The transmit vector $\underline{\mathbf{X}}(\mathcal{S}, \mathsf{P})$ consists of all these coded messages multiplied by their corresponding beamformers. In the example shown in Fig. 3, the transmit vector $\underline{\mathbf{X}}(\mathcal{S}, \mathsf{P})$ for $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{P}_1 = \{1, 2, 3\}$ and $\mathcal{P}_2 = \{4, 5, 6\}$ is generated as

$$\underline{\mathbf{X}}(\mathcal{S}, \mathsf{P}) = \mathbf{w}_{1,2}^{\mathcal{S},\mathcal{P}_1}\tilde{X}_{1,2}^{\mathcal{S},\mathcal{P}_1} + \mathbf{w}_{1,3}^{\mathcal{S},\mathcal{P}_1}\tilde{X}_{1,3}^{\mathcal{S},\mathcal{P}_1} + \mathbf{w}_{2,3}^{\mathcal{S},\mathcal{P}_1}\tilde{X}_{2,3}^{\mathcal{S},\mathcal{P}_1} + \mathbf{w}_{4,5}^{\mathcal{S},\mathcal{P}_2}\tilde{X}_{4,5}^{\mathcal{S},\mathcal{P}_2} + \mathbf{w}_{5,6}^{\mathcal{S},\mathcal{P}_2}\tilde{X}_{5,6}^{\mathcal{S},\mathcal{P}_2} + \mathbf{w}_{4,6}^{\mathcal{S},\mathcal{P}_2}\tilde{X}_{4,6}^{\mathcal{S},\mathcal{P}_2}. \quad (35)$$

Note that here β can control the number of coded messages aimed at each user. For example, if we allow $\beta = \alpha = 5$, then there will be a total of $\binom{6}{2} = 15$ coded messages transmitted in parallel, of which every user would need to decode 5. By contrast, in the example scenario for $\beta = 2$, there are in total 6 parallel coded messages of which every user needs to decode 2.

Finally, the beamformers are optimized to deliver each coded message to its intended users at the highest common rate, considering interference from other terms as well as noise. The optimum beamformers are denoted by $\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}\}^*$ for a specific partitioning P of the set $\mathcal{S}$. The inner loop in the algorithm (line 8) ensures that the above procedure is repeated for all possible partitionings of a given $\mathcal{S}$ in a TDMA manner (for example in Fig. 3, all the 10 possible partitionings in the table should be considered), and finally the outer loop repeats this process for all possible $(t + \alpha)$-subsets $\mathcal{S}$.

The following theorem characterizes the achievable delivery rate of this algorithm. A detailed analysis of the algorithm elements, and the corresponding performance analysis is provided in the proof that follows.

Theorem 1. *Algorithm 1 will result in the following symmetric rate*

$$R_{\text{sym}} = \frac{F}{\sum\limits_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+\alpha}} \sum\limits_{\substack{\mathsf{P}=\{\mathcal{P}_i\}\\ \dot{\bigcup}\mathcal{P}_i=\mathcal{S}\\ |\mathcal{P}_i|=t+\beta}} T_C^*(\mathcal{S}, \mathsf{P})} \quad (36)$$

where $T_C^*(\mathcal{S}, \mathsf{P})$ is the optimized transmission time to the subset $\mathcal{S}$ for a specific partitioning P. The outer sum is over all possible $t + \alpha$ sets $\mathcal{S}$ while the inner sum collects all disjoint unions $\dot{\bigcup}\mathcal{P}_i$ of $\mathcal{S}$ such that $|\mathcal{P}_i| = t + \beta$, and given $\delta := \frac{t+\alpha}{t+\beta} \in \mathbb{N}$. Each $T_C^*(\mathcal{S}, \mathsf{P})$ is optimized over a set of

[4]The generalization of δ is discussed in Remark 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

Algorithm 1 Interference aware Multi-Antenna Coded Caching

procedure DELIVERY($W_1,\dots,W_N$, $d_1,\dots,d_K$, $\mathbf{H}=[\mathbf{h}_1,\dots,\mathbf{h}_K]$, α, β)
  $t \leftarrow MK/N, \quad \delta \leftarrow \frac{t+\alpha}{t+\beta} \in \mathbb{N}$
  for all $\mathcal{S} \subseteq [K], |\mathcal{S}| = t+\alpha$ **do**
    for all $\mathsf{P} = \{\mathcal{P}_i\}_{i=1,\dots,\delta}$: $\bigcup_{i=1,\dots,\delta} \mathcal{P}_i = \mathcal{S}, |\mathcal{P}_i| = t+\beta$ **do**
      $\Omega^{\mathcal{S},\mathsf{P}} \leftarrow \bigcup_{i=1,\dots,\delta} \{\mathcal{T} \subseteq \mathcal{P}_i, |\mathcal{T}| = t+1\}$
      for all $\mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}$ **do**
        $X_{\mathcal{T}} \leftarrow \oplus_{k\in\mathcal{T}} \mathrm{NEW}(W_{d_k,\mathcal{T}\backslash\{k\}})$
      end for
      $\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}\}^* = \arg \max_{\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S}}, \mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}\}} \min_{k\in\mathcal{S}} R_{MAC}^k \left(\mathcal{S}, \mathsf{P}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}\}\right)$
      $\underline{\mathbf{X}}(\mathcal{S},\mathsf{P}) \leftarrow \sum_{\mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}} \mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}} \tilde{X}_{\mathcal{T}}$
      transmit $\underline{\mathbf{X}}(\mathcal{S},\mathsf{P})$ with the rate $\min_{k\in\mathcal{S}} R_{MAC}^k \left(\mathcal{S}, \mathsf{P}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}\}^*\right)$
    end for
  end for
end procedure

multicast beamformers $\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}$

$$T_C^*(\mathcal{S},\mathsf{P}) = \frac{F}{\binom{K}{t}\binom{K-t-1}{\alpha-1}\frac{(\alpha-1)!}{(\delta-1)!(\beta-1)!(t+\beta)!^{\delta-1}}} \min_{\substack{\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}, \\ \sum_{\mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}} \|\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}\|^2 \le SNR\}}} \left[\min_{k\in\mathcal{S}} R_{MAC}^k \left(\mathcal{S}, \mathsf{P}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}\}\right)\right]^{-1} \tag{37}$$

where R_{MAC}^k *is the generalized stream specific rate expression for user* k *and given as*

$$R_{MAC}^k \left(\mathcal{S}, \mathsf{P}, \{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}, \mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}}\}\right) = \min_{\mathsf{B}\subseteq\Omega_k^{\mathcal{S},\mathsf{P}}} \left[\frac{1}{|\mathsf{B}|} \log\left(1 + \sum_{\mathcal{T}\in\mathsf{B}} \gamma_{\mathcal{T}}^k(\mathcal{S},\mathsf{P})\right)\right] \tag{38}$$

and where

$$\gamma_{\mathcal{T}}^k(\mathcal{S},\mathsf{P}) = \frac{|\mathbf{h}_k^H \mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}|^2}{N_0 + \sum_{\bar{\mathcal{T}}\in\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}} |\mathbf{h}_k^H \mathbf{w}_{\bar{\mathcal{T}}}^{\mathcal{S},\mathsf{P}}|^2} \tag{39}$$

and

$$\Omega^{\mathcal{S},\mathsf{P}} := \bigcup_{i=1,\dots,\delta} \{\mathcal{T} \subseteq \mathcal{P}_i, |\mathcal{T}| = t+1\},$$
$$\Omega_k^{\mathcal{S},\mathsf{P}} := \{\mathcal{T} \in \Omega^{\mathcal{S},\mathsf{P}} |\ k \in \mathcal{T}\},\ \bar{\Omega}_k^{\mathcal{S},\mathsf{P}} := \Omega^{\mathcal{S},\mathsf{P}} \backslash \Omega_k^{\mathcal{S},\mathsf{P}}. \tag{40}$$

The SINR expressions in (39) *are non-convex, and hence, they need to be relaxed and approximated in a successive manner, similarly to* (12)–(13)*. First,* (39) *is relaxed as*

$$N_0 + \sum\nolimits_{\bar{\mathcal{T}}\in\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}} |\mathbf{h}_k^H \mathbf{w}_{\bar{\mathcal{T}}}^{\mathcal{S},\mathsf{P}}|^2 \le \frac{\sum_{\bar{\mathcal{T}}\in\mathcal{T}\bigcup\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}} |\mathbf{h}_k^H \mathbf{w}_{\bar{\mathcal{T}}}^{\mathcal{S},\mathsf{P}}|^2 + N_0}{1 + \gamma_{\mathcal{T}}^k(\mathcal{S},\mathsf{P})} \tag{41}$$

Now, the R.H.S of (41) *is a convex quadratic-over-linear function and it can be linearly approximated and lower bounded as (sets* $\mathcal{S}, \mathsf{P}$ *omitted)*

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{\mathcal{T}}, \mathbf{w}_{\bar{\mathcal{T}}}, \mathbf{h}_k, \gamma_{\mathcal{T}}^k) \triangleq \Bigg(&\sum_{\bar{\mathcal{T}}\in\mathcal{T}\bigcup\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}} |\mathbf{h}_k^H \bar{\mathbf{w}}_{\bar{\mathcal{T}}}|^2 + N_0 \\ &- 2\sum_{\bar{\mathcal{T}}\in\mathcal{T}\bigcup\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}} \mathbb{R}\left(\bar{\mathbf{w}}_{\bar{\mathcal{T}}}^H \mathbf{h}_k \mathbf{h}_k^H (\mathbf{w}_{\bar{\mathcal{T}}} - \bar{\mathbf{w}}_{\bar{\mathcal{T}}})\right) \\ &+ \frac{\sum_{\bar{\mathcal{T}}\in\mathcal{T}\bigcup\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}} |\mathbf{h}_k^H \bar{\mathbf{w}}_{\bar{\mathcal{T}}}|^2 + N_0}{1+\bar{\gamma}_{\mathcal{T}}^k} \left(\gamma_{\mathcal{T}}^k - \bar{\gamma}_{\mathcal{T}}^k\right)\Bigg) \frac{1}{1+\bar{\gamma}_{\mathcal{T}}^k}\end{aligned} \tag{42}$$

where $\bar{\mathbf{w}}_{\mathcal{T}}$ *and* $\bar{\gamma}_{\mathcal{T}}^k$ *denote the fixed values (points of approximation) for the corresponding variables from the previous iteration.*

Before going to the proof of Theorem 1, let us revisit the simple scenarios introduced in Section III and relate each of them to the generic algorithm above. By inserting the parameters listed below into (36)–(38), the corresponding scenario specific symmetric rate expressions given in Section III can be recovered.

- *Scenario 1*: $\alpha = 2$, $\beta = 2$, $\delta = 1$, $\mathcal{S} = \mathsf{P} = \{1,2,3\}$
- *Scenario 2*: $\alpha = 3$, $\beta = 3$, $\delta = 1$, $\mathcal{S} = \mathsf{P} = \{1,2,3,4\}$
- *Scenario 3*: $\alpha = 2$, $\beta = 2$, $\delta = 1$, $\mathcal{S} = \mathsf{P} \subset [4], |\mathcal{S}| = \alpha + 1 = 3$
- *Scenario 4*: $\alpha = 3$, $\beta = 1$, $\delta = 2$, $\mathcal{S} = \{1,2,3,4\}$, $\mathsf{P}(1) = \{\mathcal{P}_1(1), \mathcal{P}_2(1)\} = \{\{1,2\},\{3,4\}\}$, $\mathsf{P}(2) = \{\{1,3\},\{2,4\}\}$ and $\mathsf{P}(3) = \{\{1,4\},\{2,3\}\}$

The proof of Theorem 1 is given in the following.

Proof. In the cache content placement phase, each file is divided into $\binom{K}{t}$ subfiles as follows

$$W_n = \{W_{n,\tau}, \tau \subset [K], |\tau| = t\}, \tag{43}$$

and each subfile is further divided into mini-files

$$W_{n,\tau} = \{W_{n,\tau}^j, j = 1,\dots,\Gamma\} \tag{44}$$

where

$$\Gamma = \binom{K-t-1}{\alpha-1} \frac{(\alpha-1)!}{(\delta-1)!(\beta-1)!(t+\beta)!^{\delta-1}}. \tag{45}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

In the original coded caching scheme of [8], there are $\binom{K}{t+1}$ coded messages (called *coded sub-files*, each of size equal to a sub-file) which should be delivered to all $(t+1)$-subsets of users $[K]$, i.e., $X_{\mathcal{T}} := \oplus_{k\in\mathcal{T}} W_{d_k,\mathcal{T}\setminus\{k\}}$ should be delivered to all members of $\mathcal{T}$ for all $\mathcal{T}\subseteq[K], |\mathcal{T}|=t+1$. Since in our construction (inner and outer loops in Algorithm 1, each $(t+1)$-subset appears multiple times, we need to transmit smaller coded messages (called *coded mini-files*, each of size equal to a mini-file) in each appearance, which ensures that delivering each coded mini-file provides the targeted users with *fresh* (not transmitted before) mini-files they require. This is the main reason behind dividing each subfile into Γ mini-files. In order to do this, we define the operator $\mathrm{NEW}(.)$ which when operated on each sub-file returns the next *fresh* mini-file of that sub-file, which then will be used in forming coded mini-files. More specifically we have

$$\mathrm{NEW}(W_{n,\tau}) = W_{n,\tau}^{j+1} \tag{46}$$

if the last application of NEW on the sub file $W_{n,\tau}$ had returned $W_{n,\tau}^{j}$. Next, we describe how these tasks are fulfilled with the help of multi-antenna interference management.

Let us focus on a specific $(t+\alpha)$-subset of the users, namely $\mathcal{S}$, and a specific partitiong of this subset, namely $\mathsf{P}=\{\mathcal{P}_i\}_{i=1,\ldots,\delta}$: $\dot{\bigcup}_{i=1,\ldots,\delta}\mathcal{P}_i=\mathcal{S}, |\mathcal{P}_i|=t+\beta$. Then, $\Omega^{\mathcal{S},\mathsf{P}}$ is the collection of all $(t+1)$-subsets of $\mathcal{S}$, such that each subset is contained inside a group $\mathcal{P}_i$ of the partition. Then, sum of coded mini-files of these $(t+1)$-subsets with the corresponding beamformers will be transmitted to users in $\mathcal{S}$ in the form of the transmit signal

$$\underline{\mathbf{X}}(\mathcal{S},\mathsf{P}) \leftarrow \sum_{\mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}} \mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}} \tilde{X}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}} \tag{47}$$

where $\tilde{X}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}}$ is ensured to be a coded mini-file combined of fresh mini-files for each involved user. Assume that all the involved coded mini-files are successfully received at their intended users. Then, all the subsets $\mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}$ will receive one coded mini-file, containing a fresh mini-file for each user in $\mathcal{T}$. It can be easily verified that, if we go over all the possible $(t+\alpha)$-subsets and their corresponding partitionings, each $(t+1)$-subset of $[K]$ will appear Γ times (given in (45)), and due to the appropriate mini-file indexing, each user will be able to decode a fresh mini-file in each transmission shot. Thus, these coded mini-files constitute the whole coded subfile. As this is true for all the $(t+1)$-subset of $[K]$, all the original tasks of [8] are fulfilled.

It just remains to be proven that by transmitting $\underline{\mathbf{X}}(\mathcal{S},\mathsf{P})$ with the rate stated in Theorem 1, all the users in $\mathcal{S}$ will be able to decode their desired coded mini-files. Consider a user $k\in\mathcal{S}$, which happens to be in the group $\mathcal{P}_i$ of the partitioning P. Then, it is clear that this user will be interested in the coded mini-files $X_{\mathcal{T}}^{N\tau}$ such that $\mathcal{T}\in\Omega_k^{\mathcal{S},\mathsf{P}}$, and all the remaining coded mini-files $X_{\mathcal{T}}^{N\tau}, \mathcal{T}\in\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}$ will appear as interference to this user. Thus, this user faces a Gaussian MAC with $|\Omega_k^{\mathcal{S},\mathsf{P}}|$ desired terms, $|\bar{\Omega}_k^{\mathcal{S},\mathsf{P}}|$ interference terms, and a noise term. Clearly, by restricting the transmission rate to the achievable Gaussian MAC rate in (38), this user can decode all the desired terms with an equal rate. Since we are transmitting the common message of size $\frac{F}{\binom{K}{t}\Gamma}$ to the users in $\mathcal{S}$ at the rate of the worst user, all of them will be able to decode the file within the minimum delivery time given in (37).

Finally, since each user decodes one requested file at the end, the symmetric (per-user) rate of the proposed scheme will be $R_{\mathrm{sym}}=F/T$ where the total time T can be derived as

$$T=\sum_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+\alpha}}\sum_{\substack{\mathsf{P}=\{\mathcal{P}_i\}\\\dot{\bigcup}\mathcal{P}_i=\mathcal{S}\\|\mathcal{P}_i|=t+\beta}} T_C^*(\mathcal{S},\mathsf{P}) \tag{48}$$

where $T_C^*(\mathcal{S},\mathsf{P})$ is the transmission time for the given subset $\mathcal{S}$ and partitioning P. □

The following Degrees of Freedom (DoF) analysis of the proposed scheme shows that the DoF only depends on α and it is independent of β. By choosing $\alpha=L$, we achieve a DoF shown to be order-optimal among one-shot linear schemes in [15] (For an information theoretic optimality analysis based on interference alignment techniques see [20]).

Corollary 1. *The DoF of the rate derived in the above theorem is*

$$DoF=\frac{t+\alpha}{K-t}=\frac{KM/N+\alpha}{K(1-M/N)}. \tag{49}$$

Proof. DoF is defined as

$$\begin{aligned}
DoF &= \lim_{SNR\to\infty}\frac{R_{\mathrm{sym}}}{\log SNR}\\
&= \frac{F}{\sum\limits_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+\alpha}}\sum\limits_{\substack{\mathsf{P}=\{\mathcal{P}_i\}\\\dot{\bigcup}\mathcal{P}_i=\mathcal{S}\\|\mathcal{P}_i|=t+\beta}}\lim\limits_{SNR\to\infty}(\log SNR\times T_C^*(\mathcal{S},\mathsf{P}))}\\
&\overset{(a)}{=} \frac{F}{\binom{K}{t+\alpha}\frac{(t+\alpha)!}{\delta!(t+\beta)!^\delta}\lim\limits_{SNR\to\infty}(\log SNR\times T_C^*(\mathcal{S},\mathsf{P}))}\\
&\overset{(b)}{=} \frac{\binom{K}{t}\binom{K-t-1}{\alpha-1}\frac{(\alpha-1)!}{(\delta-1)!(\beta-1)!(t+\beta)!^{\delta-1}}}{\binom{K}{t+\alpha}\frac{(t+\alpha)!}{\delta!(t+\beta)!^\delta}}\\
&\qquad\lim_{SNR\to\infty}\frac{R_{MAC}^k\left(\mathcal{S},\mathsf{P},\{\mathbf{w}_{\mathcal{T}}^{\mathcal{S},\mathsf{P}},\mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}\}\right)}{\log SNR}\\
&\overset{(c)}{=} \frac{\binom{K}{t}\binom{K-t-1}{\alpha-1}\frac{(\alpha-1)!}{(\delta-1)!(\beta-1)!(t+\beta)!^{\delta-1}}}{\binom{K}{t+\alpha}\frac{(t+\alpha)!}{\delta!(t+\beta)!^\delta}\binom{t+\beta-1}{t}}\\
&= \frac{t+\alpha}{K-t}=\frac{KM/N+\alpha}{K(1-M/N)}.
\end{aligned}$$

where (a) is due to that fact that the number of terms in the inner and outer summations are $\frac{(t+\alpha)!}{\delta!(t+\beta)!^\delta}$ and $\binom{K}{t+\alpha}$ respectively, and since $\lim_{SNR\to\infty}(\log SNR\times T_C^*(\mathcal{S},\mathsf{P}))$ does not depend on particular $\mathcal{S}$ and P, (a) is valid for any of $\mathcal{S}$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

P indexed in the summations. Also (b) follows from (37) and (c) is due to the fact that

$$\lim_{SNR\to\infty} \frac{R^k_{MAC}\left(\mathcal{S},\mathsf{P},\{\mathbf{w}^{\mathcal{S},\mathsf{P}}_{\mathcal{T}},\mathcal{T}\in\Omega^{\mathcal{S},\mathsf{P}}\}\right)}{\log SNR}$$
$$= \lim_{SNR\to\infty} \min_{\mathsf{B}\subseteq\Omega^{\mathcal{S},\mathsf{P}}_k} \left[\frac{1}{|\mathsf{B}|}\frac{\log\left(1+\sum_{\mathcal{T}\in\mathsf{B}}\gamma^k_{\mathcal{T}}(\mathcal{S},\mathsf{P})\right)}{\log SNR}\right]$$
$$= \frac{1}{\max\limits_{\mathsf{B}\subseteq\Omega^{\mathcal{S},\mathsf{P}}_k}|\mathsf{B}|} = \frac{1}{\binom{t+\beta-1}{t}}$$

which concludes the proof. □

Also, we characterize the results in [26] and the max-min SNR beamforming baseline scheme as a special cases of Theorem 1 in the following remark.

Remark 1. *In Theorem 1, if we set $\alpha=\beta=\min(L,K-t)$ and the beamforming vectors are chosen based on the zero forcing principle, the interference terms vanish, and it reduces to the results of [26]. Furthermore, if we set $\alpha=\beta=1$, the result reduces to the baseline maxmin SNR beamforming scheme.*

Moreover, the complexity of the optimization problem is characterized in the following remark.

Remark 2. *All of the constraints involved can be rewritten as second-order cones (SOCs). The SINR and transmit power constraints are readily in SOC form. However, the MAC sum rate constraints involving exponents (as seen, e.g., in* (15)*) require some additional steps for the complete SOC formulation [37]. In the general case, the complexity of the beamformer design* (36) *is largely dominated by the number of simultaneously transmitted messages, that is, the partitioning size $|\Omega^{\mathcal{S},\mathsf{P}}_k| = \binom{t+\beta-1}{t}$. The number of MAC rate region constraints increases exponentially with $\beta+t$. However, the size of each SOC constraint involved with the MAC region is fairly small. On the other hand, the complexity of the SINR constraints scales quadratically with $\alpha+1$ and L [38]. It should be noted that, the beamformer design can be split into $\binom{K}{\alpha+t}$ parallel problems, which greatly improves the optimization latency and individual problem complexity as α is decreased. The receiver complexity is mostly affected by parameter β, i.e., whether or not SIC is needed. From the receiver perspective, $\beta>1$ indicates the number of desired multicast messages decoded at each user using the SIC receiver structure.*

Remark 3. *The above discussion is for parameter values such that $t+\beta$ divides $t+\alpha$. In general, one can vary α and β such that this condition holds true, however if it is not possible to ensure, a readily available option is always to set $\beta=\alpha$, which by choosing $\alpha=L$ will achieve full DoF. For other cases where $t+\beta$ does not divide $t+\alpha$, extending the above techniques to arrive at satisfactory finite-SNR performance is challenging due to the asymmetries arising in the combinatorial nature of the problem. Therefore, this problem can be posed as an interesting topic for further research.*

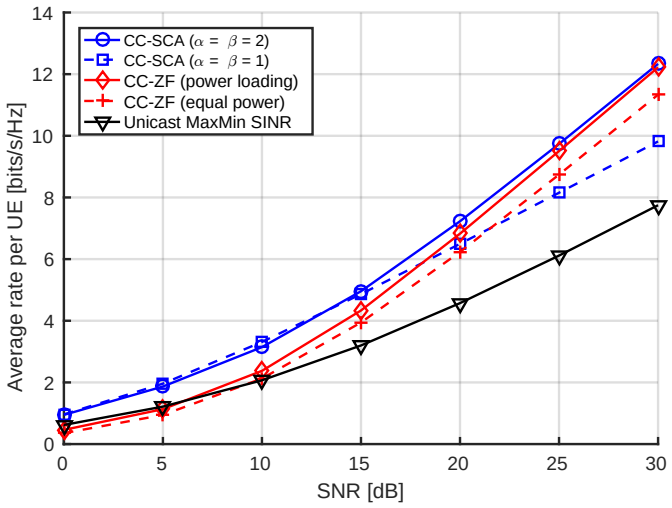


Fig. 4. *Scenario 1*: $L=2$, $K=3$, $N=3$ and $M=1$.

V. Numerical Examples

The numerical examples are generated for various combinations of parameters L, K, N, M and $|\mathcal{S}|$, including *Scenarios 1 – 4*. 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. All the Matlab codes are available online at https://github.com/kalesan/sim-cc-miso-bc.

Fig. 4 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 multicast beamforming scheme via SCA, denoted as CC-SCA, achieves $3-5$ dB gain at low SNR as compared to the ZF with equal power loading [26]. At high SNR, the ZF with optimal power loading in (16) achieves comparable performance while other schemes have significant performance gap. At low SNR regime, the simple MaxMin SNR multicasting with CC (labelled as 'CC-SCA ($\alpha=\beta=1$)') has similar performance as the CC-SCA scheme with full overlap between multicast streams ($\alpha=\beta=2$). 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-SCA and CC-ZF schemes achieve an additional DoF, which was already shown (for high SNR) in [13], [26]. The unicasting scheme does not perform well in this scenario as it does not utilize the global caching gain (only the local cache).

In Fig. 5, the number of transmit antennas is increased to $L=3$. This provides more than 3dB additional gain for the CC-SCA 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

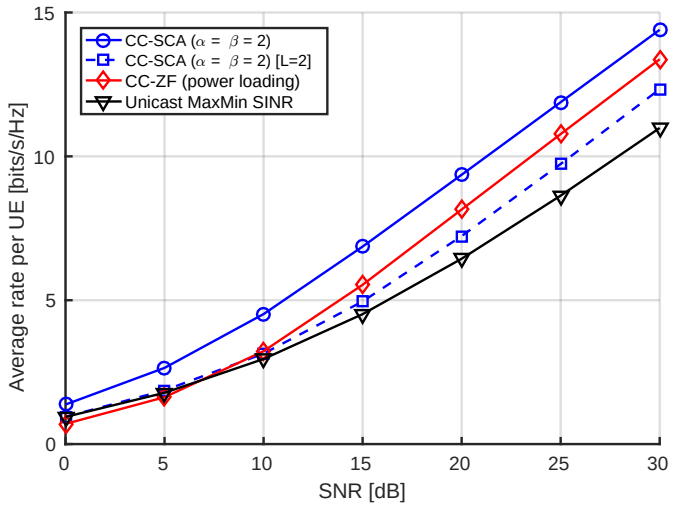


Fig. 5. *Scenario 1*: $L = 3$, $K = 3$, $N = 3$ and $M = 1$.

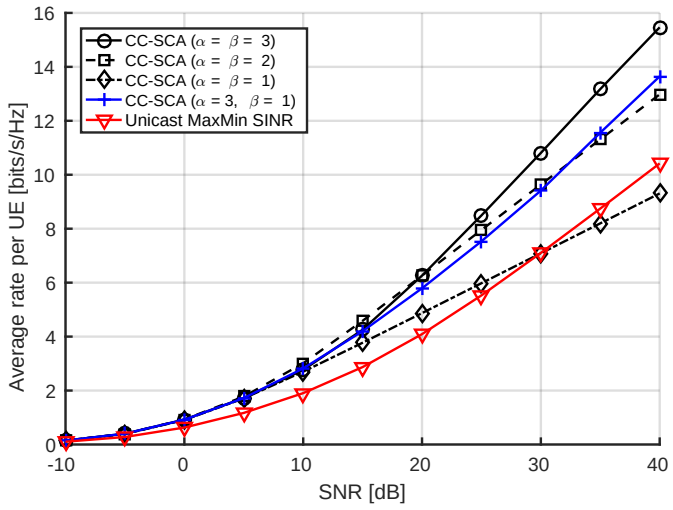


Fig. 6. Impact of parameters α and β in *Scenarios 2-4*: $L = 3$, $K = N = 4$, and $M = 1$.

strongest eigenvector of the stacked user channel matrix, projected to the 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-SCA method, but there is a constant performance gap at high SNR. Interestingly, the CC-SCA 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 introduced in *Scenarios 2–4* are illustrated in Fig. 6. The CC-ZF scheme is not shown in Fig. 6, but it is noted that the CC-SCA achieves $5 - 7$dB gain at low SNR, which is considerably more than in the less complex *Scenario 1*. At high SNR, however, the CC-ZF with optimal power loading provides a comparable performance. The performance of the reduced subset method for $K = 4$, $L = 3$ (*Scenario 3*) are also shown in Fig. 6. Similar to Fig. 4, $\alpha = \beta = 1(|\mathcal{S}| = 2)$ provides comparable

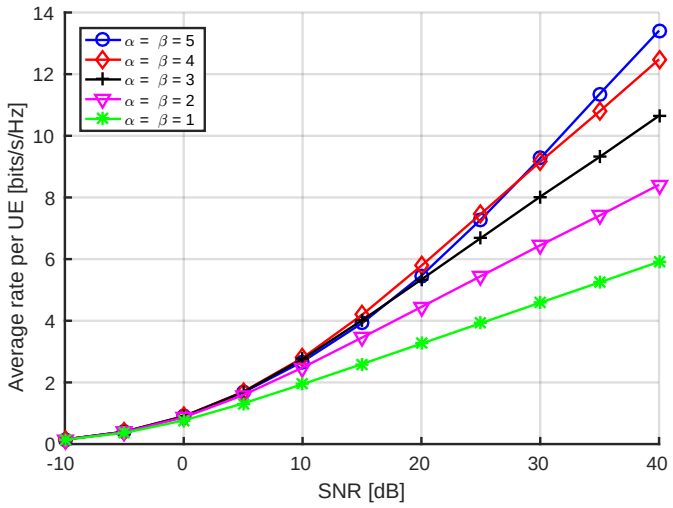


Fig. 7. CC-SCA performance with $K = 6$, $L = 5$, $|S| = \alpha + 1 = [2, 3, 4, 5, 6]$.

performance at the low SNR regime. Interestingly, as the SNR is increased, $|\mathcal{S}| = 3$ ($\alpha = \beta = 2$) subset size outperforms the case with $|\mathcal{S}| = 4$ ($\alpha = \beta = 3$). Again, 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 then in TDMA fashion.

Fig. 6 illustrates also the performance of the proposed simple linear TX-RX multicasting scheme introduced in *Scenario 4*. The linear scheme labeled as 'CC-SCA ($\alpha = 3$, $\beta = 1$)' 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-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.

Fig. 2 illustrates some subset selection possibilities for $K = 5$. For a six user ($K = 6$) scenario shown in Fig. 7, there are four possible subset sizes $|\mathcal{S}| = [2, 3, 4, 5]$ that can be used to reduce the serving set size for multicast transmission in $\mathcal{S}$. From Fig. 7, we can observe again that, by reducing the subset size to $|\mathcal{S}| = 5$ or 4, the average symmetric rate per user can be even improved at medium SNR as compared to the case where all users are served simultaneously, i.e., $|\mathcal{S}| = 6$. At high SNR region, however, the reduced subset cases become highly suboptimal as the spatial DoF for transmitting parallel streams is limited by α. The high SNR slope for each curve in Fig. 7 is equivalent to the user specific DoF given in (**??**), ranging from $\frac{2}{5}$ ($\alpha = 1$) to $\frac{6}{5}$ ($\alpha = 5$). From complexity reduction perspective, the multicast mode with the smallest subset size providing close to optimal performance should be selected. In Fig. 7, for example, subset sizes $|\mathcal{S}| = 3$, $|\mathcal{S}| = 4$, $|\mathcal{S}| = 5$ could be used up to 0 dB, 10 dB and 30 dB, respectively, for optimal performance-complexity trade-off.

In Fig. 8, the impact of parameter $\beta = [1, 2, 5]$ controlling

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

the overlap among the parallel multicast messages is assessed with a fixed $\alpha = 5$ for both SCA and ZF methods. The CC-ZF plots are generated by imposing zero-interference constraint similarly to (16). The results with $\alpha = 5$ and $\beta = 1$ represent the case with no overlap, and hence, SIC is not required at the receivers. The multicast transmission is split into in total 15 time slots to cover all disjoint unions $\dot{\bigcup}\mathcal{T}$ of $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ such that $|\mathcal{T}| = t + 1 = 2$. Furthermore, each subfile is split into 3 mini-files in order to allow different contents to be transmitted in each subset 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 in distinct P_i. In each time slot, all 6 users are served with 3 multicast streams transmitted in parallel. Thus, 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 $\bar{\mathcal{T}} \in \mathsf{P}(i) \backslash \mathcal{T}$. The case labeled as 'CC-SCA ($\alpha = 5$, $\beta = 2$) is an intermediate case between the linear scheme ($\beta = 1$) and the fully overlapping case ($\beta = 5$), allowing partial overlap among multicast messages transmitted in parallel. All possible 10 partitionings of size $t + \beta = 3$ served in a TDMA fashion are shown in Fig. 3. In this case, each subfile must be further split into 4 minifiles as each user index pair gets repeated in 4 different subsets $\mathcal{P}_i$

The results in Fig. 8 verify the DoF analysis of Corollary 1 where the asymptotic DoF (slope) is shown to be independent of β at high SNR region. However, the lower complexity $\beta = 1$ case in Fig. 8 suffers from 5dB SNR penalty, which in turn can be alleviated by using a higher overlap ($\beta = 2$) among parallel multicast streams. In general, a non-linear SIC structure has to be used at the receiver for $\beta > 1$ in order to decode multiple parallel streams at each user. In Fig. 8, the case with $\beta = L = 5$ has the highest degree of flexibility for beamformer and power allocation since all 15 multicast streams are transmitted in parallel. On the other hand, the optimization space for $\beta < 5$ is much more constrained. For example, only 3 multicast messages are sent in parallel in each transmit interval for $\beta = 1$ and the SIC receiver is not needed at all. This translates to a constant penalty at high SNR when using $\beta < L$.

At low SNR, the performance loss from using highly suboptimal ZF criterion can be more than 10 dB at low SNR. Similarly, more than 150% rate gain can be achieved by using the CC-SCA design at 10 dB SNR, for example. Furthermore, the performance impact of β diminishes as the inter-stream interference is no longer dominant over the noise and all β parametrizations provide almost identical performance. Asymptotically, the optimal transmit beamformers are reduced to SNR maximizing multicast beamformers, which can be simply obtained as weighted superposition of conjugate beamformers matching to the channels of each $t+1$-sized user subset $\mathcal{T}$. Consequently, the multicast beamformers specific to given $\mathcal{T}$ are essentially equivalent for any β as they become independent of the inter-stream interference. Due to the maxmin rate objective, the beamformers are weighted such that the received signal level would be the same for all

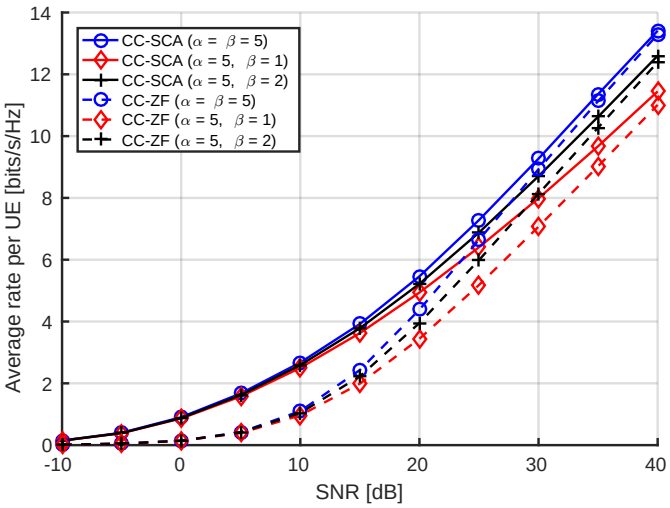


Fig. 8. CC-SCA performance with $K = 6$, $L = 5$, $\alpha = 5$ ($|S| = 6$), $\beta = [1, 2, 5]$.

$k \in \mathcal{T}$. As a result, the effective channel gains $\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_{\mathcal{T}}$ in the given symmetric scenario (equal path loss for all users) with $L = 5$ antennas become (nearly) equal for all k and $\mathcal{T}$. From the receiver perspective, this scenario resembles a classical symmetric downlink channel scenario where the superposition coding and SIC do not give any rate gain in comparison to a simple TDMA strategy [39, Section 6.2.1]. In general, β should be selected as small as possible, since there is only a minor impact in the performance and, at the same time, there is a significant reduction in the computational complexity.

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 antennas L, size of the user subset $t + \alpha$, and the overlap among the multicast messages β were employed, optimally balancing the detrimental impact of both noise and inter-stream interference from coded messages transmitted in parallel. Furthermore, the DoF was shown to only depend on α while being independent of β. The schemes were shown to perform significantly better than several base-line schemes over the entire SNR region.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021 white paper," Cisco, Tech. Rep., Feb. 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inform. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[3] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[4] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[5] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[6] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[7] A. Liu and V. K. N. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Processing*, vol. 62, no. 2, pp. 390–402, Jan. 2014.

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

[9] Kai Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *2016 IEEE Information Theory Workshop (ITW)*, Sep. 2016, pp. 161–165.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun. 2015, pp. 809–813.

[19] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.

[20] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5359–5380, July 2018.

[21] J. S. P. Roig, S. A. Motahari, F. Tosato, and D. Gündüz, "Fundamental limits of latency in a cache-aided 4x4 interference channel," in *2017 IEEE Information Theory Workshop (ITW)*, Nov. 2017, pp. 16–20.

[22] 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.

[23] A. Shabani, S. P. Shariatpanahi, V. Shah-Mansouri, and A. Khonsari, "Mobility increases throughput of wireless device-to-device networks with coded caching," in *Proc. IEEE Int. Conf. Commun.*, May 2016, pp. 1–6.

[24] J. Zhang and O. Simeone, "Fundamental limits of cloud and cache-aided interference management with multi-antenna base stations," in *Proc. IEEE Int. Symp. Inform. Theory*, June 2018, pp. 1425–1429.

[25] K. H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan 2018.

[26] 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.

[27] 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.

[28] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[29] M. M. Amiri and D. Gündüz, "Caching and coded delivery over gaussian broadcast channels for energy efficiency," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 8, pp. 1706–1720, Aug 2018.

[30] S. S. Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun. 2017, pp. 1222–1226.

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

[32] 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.

[33] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multicast beamformer design for coded caching," in *Proc. IEEE Int. Symp. Inform. Theory*, Vail, CO, USA, Jun 2018.

[34] ——, "Multicast mode selection for multi-antenna coded caching," in *The 2018 International Workshop on Content Caching and Delivery in Wireless Networks (CCDWN)*, Shanghai, China, May 2018.

[35] ——, "Linear multicast beamforming schemes for coded caching," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, Pacific Grove, CA, USA, Oct 2018.

[36] 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.

[37] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Mathematical Programming*, vol. 95, pp. 3–51, 2001.

[38] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second–order cone programming," *Linear Algebra and Applications*, vol. 284, pp. 193–228, Nov. 1998.

[39] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.

Antti Tölli (M'08, SM'14) received the Dr.Sc. (Tech.) degree in electrical engineering from the University of Oulu, Oulu, Finland, in 2008. Before joining the Centre for Wireless Communications (CWC) at the University of Oulu, he worked for 5 years with Nokia Networks as a Research Engineer and Project Manager both in Finland and Spain. Currently, he holds an Associate Professor position with the University of Oulu. In May 2014, he was granted a five year (2014-2019) Academy Research Fellow post by the Academy of Finland. During the academic year 2015-2016, he visited at EURECOM, Sophia Antipolis, France, while from August 2018 till June 2019 he visited University of California - Santa Barbara, USA. He has authored numerous papers in peer-reviewed international journals and conferences and several patents all in the area of signal processing and wireless communications. His research interests include radio resource management and transceiver design for broadband wireless communications with a special emphasis on distributed interference management in heterogeneous wireless networks. He is currently serving as an Associate Editor for IEEE Transactions on Signal Processing.

Seyed Pooya Shariatpanahi received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran, in 2006, 2008, and 2013, respectively. Currently, he is an assistant professor at the School of Electrical and Computer Engineering at University of Tehran. Before joining University of Tehran, he was a researcher with the Institute for Research in Fundamental Sciences (IPM), Tehran, Iran. His research interests include information theory, network science, wireless communications, and complex systems. He was a recipient of the Gold Medal at the National Physics Olympiad in 2001.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2962686, IEEE Transactions on Wireless Communications

Jarkko Kaleva (S'11-M'18) received his Dr.Sc. (Tech.) degree in communications engineering from University of Oulu, Oulu, Finland in 2018 with distinction. In 2010, he joined Centre for Wireless Communications (CWC) at University of Oulu, Finland. He is co-founder of Solmu Technologies, where he is working as the chief software architect. His main research interests are in deep learning, structural analysis and nonlinear programming.

Babak Hossein Khalaj received the B.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 1989, and M.Sc. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, USA, in 1993 and 1996, respectively. He has been with the pioneering team at Stanford University where he was involved in adoption of multi-antenna arrays in mobile networks. He joined KLA-Tencor in 1995, as a Senior Algorithm Designer, involved on advanced processing techniques for signal estimation. From 1996 to 1999, he was with Advanced Fiber Communications and Ikanos Communications. Since then, he has been a Senior Consultant in the area of data communications, and a Visiting Professor with CEIT, San Sebastian, Spain, from 2006 to 2007. He has co-authored many papers in signal processing and digital communications. He holds three U.S. patents and was the recipient of the Alexander von Humboldt Fellowship from 2007 to 2008 and Nokia Visiting Professor Fellowship in 2018.