# Bridging the Gap between Multiplexing and Diversity in Finite SNR Multiple Antenna Coded Caching

Eleftherios Lampiris, Petros Elia, Giuseppe Caire
lampiris@tu-berlin.de, elia@eurecom.fr, caire@tu-berlin.de

***Abstract*— We study the multi-antenna Coded Caching setting in the finite Signal-to-Noise-Ratio (SNR) region, with the purpose to utilize the extra resources from the multiple antennas to increase the rate performance of the system. Specifically, assuming an $L$-antenna setting with $K$ cache-enabled users (each with normalized cache-size $\gamma$), previous works showed that at the high SNR region the performance is proportional to $L+K\gamma$ degrees-of-freedom while, if one considers the low SNR region this would be outperformed by a strategy that focuses on beamforming a (single) multicast message (of $K\gamma+1$ elements) to its recipients. In order to bridge these two extremes we use a recently proposed algorithm that can balance the amount of users served in order to maximize the achieved sum-rate.**

## I. Introduction

The seminal work of Maddah-Ali and Niesen [1] showed that equipping users with caches can provide multicasting opportunities and thus to significantly increase the performance of the wired, single-stream, bottleneck channel. The setting assumes a server with access to a library of $N$ files, tasked with satisfying the demands of a set of $K$ users, each equipped with a cache of normalized size $\gamma \in (0,1)$. The authors showed that a new caching policy can allow the transmission of a XORed messages, containing content for $K\gamma+1$ users, which can proceed to decode the desired message by using their cache to remove unwanted interference.

*Multi-antenna Coded Caching:* In an effort to combine Coded Caching gains with multiplexing gains attributed to multiple antennas the work in [2] showed that in a wired channel, with $L$ servers[1] and $K$ single-antenna users each equipped with a cache of normalized size $\gamma$, every transmission can satisfy the demands of $L+K\gamma$ users thus, adding an antenna can allows treating an additional user. Further works showed that the above gains persist when some users are not equipped with caches [3], with reduced channel state information (CSI) [4], with much reduced amount of required subpackets [5] and with reduced both CSI and subpacketization [6].

The main idea, that allowed these gains, is that each subfile can be "cached-out" by $K\gamma$ users and at the same time, using the multiple antennas, to be nulled-out at $L-1$ users.

*Finite SNR performance:* While the above approaches focused on asymptotic results in the very high Signal-to-Noise-Ratio (SNR) region, recently the attention has shifted to the study of caching methods in more practical SNR regions. As a notable example, the work in [7] showed that in the low SNR region, beamforming a single message (one XOR generated as in the algorithm of [1]) to $K\gamma+1$ users can provide much higher sum-rate performance than transmitting to the maximum $(L+K\gamma)$ users.

This rate increase was achieved due to two characteristics that have been exploited. First, transmitting only the XOR allows the allocation of the power budget in a single message. The second reason that beamforming a single multicast message outperforms – in the low SNR region – schemes which utilize both multicasting and multiplexing resources, has to do with one of the bottlenecks of Coded Caching, namely the worst-user effect (cf. [8]). In its essence, the worst-user effect forces the multicast message's rate to be the minimum among those of the receiving nodes. Since a message has many recipients, and in order for it to be decoded reliably by every one of these users, then the user with the worst rate among them will define the transmission rate and thus form the bottleneck of the communication.

Serving the maximum $(K\gamma+L)$ amount of users requires exploiting all spatial degrees of freedom, thus compensating for the worst-user rate through beamforming is not possible. In contrast, in the multicast case the beamforming vector exploits the spatial diversity to increase the worst rate between the $K\gamma+1$ recipient users (cf. [9]).

From the above we can discern a tradeoff as a function of the SNR. On one extreme, serving $K\gamma+L$ users is preferable in the high SNR region, while on the other extreme focusing on serving fewer users and compensating, through beamforming, for the worst rate is the best strategy for lower SNR values. In this work we seek to understand these two regions and to further extend the analysis in schemes that are serving an intermediary amount of users, which schemes may potentially achieve better performance in mid-SNR values. We will use an

Eleftherios Lampiris and Giuseppe Caire are with the Electrical Engineering and Computer Science Department, Technical University of Berlin, 10587, Germany. Petros Elia is with the Communication Systems Department at EURECOM, Sophia Antipolis, 06410, France. The work is supported by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929. (ERC project DUALITY) and by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant 789190 - CARENET. The code from this work can be found in https://github.com/elefteros/multi-antenna-coded-caching.

[1]Between the servers and the users is a set of intermediary nodes that perform network coding operations, which create a full rank channel.

algorithm that serves $K\gamma+s$ users, where $s \in \{1, ..., L\} \triangleq [L]$ is a parameter of choice. In Fig. 1 we display the per-user rate for a Multiple-Input-Multiple-Output (MISO) Broadcast Channel (BC) with 4 transmit antennas for each value of parameter $s$, where the stream number signifies the amount of spatial degrees of freedom used to serve "additional" users.

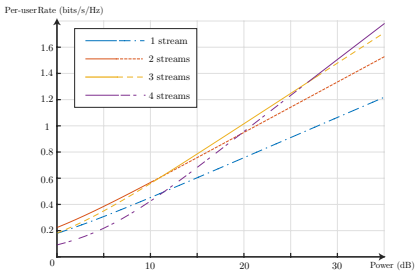


Fig. 1. Simulation results of the per-user rate of the MISO BC setting with 30 users, cache-redundancy $K\gamma = 3$ and $L = 4$ antennas. Solid part of a line reveals higher performance for the particular stream compared to all others.

In the low SNR region, transmission with 2 streams outperforms any other stream selection, while in the mid SNR region the best option is to transmit 3 streams and finally in the high SNR region the full stream scheme becomes the optimal.

### A. Paper Overview

The main idea is that we will exploit is that as we reduce the amount of streams, i.e. the number of users served through precoding, we allow more flexibility in shaping the beamforming vector, thus tackling the worst-user effect. From Fig. 1 we can see that the SNR region is divided into sub-regions, inside which the best strategy is to choose some specific stream value. Initially, we seek to understand how these sub-regions behave as we vary the two problem parameters, namely the number of antennas $L$ and the cache redundancy $K\gamma$.

Further, we will use the curve that is formed as the maximum among the rates achieved by any stream (solid parts in Fig. 1) in order to explore how varying parameters $L$ and $K\gamma$ will affect this rate. We will employ a recently proposed multi-antenna algorithm (see [10]), which has the desired ability to easily extend a scheme that uses $s < L$ streams to a scheme with $s+1$ streams by adding one more user, instead of selecting a new set of users. This will allow us to show how much a new user is affecting the system's performance.

### B. State-of-Art

The effect of channel variability in the context of Coded Caching has been investigated in a variety of works (see [8] and references therein). Specifically, the work in [8] investigated the single-antenna setting with each user having, potentially, different channel capacity and proposed a superposition coding scheme that achieves an order-optimal performance. The work in [11] used a different approach to tackle the worst-user effect, which involved employing multiple transmitters in order to provide diversity and thus lift the worst-user rate. Further, the work in [7] studied the finite SNR performance of the multi-antenna cache-aided setting, while the work in [12] (and subsequently the work in [13]) investigated efficient beamforming designs that can improve the rate performance as well as the complexity of a multi-antenna coded caching algorithm.

## II. Setting

We consider the cache-aided MISO BC where an $L$-antenna Base Station, with access to a library of $N$ files $\{W^n\}_{n=1}^N$, serves $K$ single antenna receivers. Each receiver is able to store fraction $\gamma \in (0, 1)$ of the library and we further assume that each user will request a unique and different file.

In order to satisfy the users' demands, the base station transmits an $L$-element vector with $s \in [L]$ messages, which collectively contain $K\gamma + s$ subfiles. Choosing parameter $s$ we can control the number of streams or else the number of different messages that will be transmitted. Specifically, a transmitted vector takes the form

$$\mathbf{x} = \mathbf{h}_\lambda^\perp X_\sigma + \sum_{i=1}^{s-1} \mathbf{h}_{\lambda_i}^\perp W_\tau^{d_k} \tag{1}$$

where $\mathbf{h}_\mu^\perp$ denotes a normalized beamforming vector designed to zero force the symbol's interference to users of set $\mu$, $|\mu| = s - 1$, where this set could potentially be empty in the case of $s = 1$. Symbol $X_\sigma$ denotes the XORed message (described in the Appendix), that is intended to convey information to the users of set $\sigma \subset [K]$, $|\sigma| = K\gamma + 1$. Finally, $d_k \in [N]$ denotes the file request of user $k$, while $\tau \subset [K]$, $|\tau| = K\gamma$ denotes the subfile index (described in the Appendix). We relegate all information regarding file segmentation, creation of the XORs and how subfiles are selected in each transmission to the Appendix (for a more detailed exposition of the algorithm, the reader is also referred to [10]). The received message at user $k \in [K]$ takes the form

$$y_k = \mathbf{h}_k^\dagger \mathbf{x} + w_k \tag{2}$$

where $\mathbf{h}_k \in \mathbb{C}^L$ represents the channel vector between the $L$-antenna transmitter and user $k$, signal $\mathbf{x}$ satisfies a power constraint $\mathbb{E}(\|\mathbf{x}\|^2) = P$, while $w_k \sim \mathbb{C}\mathcal{N}(0, 1)$ represents the noise at user $k$. We notice that, in this work, we will consider only Zero-Forcing (ZF) beamforming, which will allow for a first understanding of the performance of the above systems in the finite SNR and, unless otherwise stated, we are employing power control among the streams.

*Notation:* For sets $A, B$ we will use $A \setminus B$ to denote the difference set. Further, for integers $n, k$ we will denote the binomial coefficient with $\binom{n}{k}$.

## III. Beamformer Design

As we see from Eq. (1), in order for all $K\gamma + s$ involved users to be able to decode, apart from removing part of the unwanted interference via caching, we also need to design each beamforming vector $\mathbf{h}_\mu^\perp$ to satisfy

$$\mathbf{h}_k^\dagger \cdot \mathbf{h}_\mu^\perp \begin{cases} = 0, & \text{if } k \in \mu \\ \neq 0, & \text{else.} \end{cases} \tag{3}$$

It is easy to see that if $s = L$, the design of the beamformer vectors is explicitly defined through the normalized inverse of the channel matrix between the $L$ antenna transmitter and the precoding-assisted users. On the other extreme case, where $s = 1$, the transmitted message takes the form of a single message (a XOR intended for $K\gamma+1$ users), which is designed such as to maximize the worst-rate. Hence, the optimization problem takes the form

$$\max_{\mathbf{h}_{\emptyset}^{\perp} \in \mathbb{C}^L} \min_{k \in \sigma} \|\mathbf{h}_k^{\dagger} \cdot \mathbf{h}_{\emptyset}^{\perp}\|^2. \tag{4}$$

The above solution has been first proposed in the multicast messages setting (see [9]) and in the context of multi-antenna coded caching in [7], [12].

By allowing parameter $s$ to take an intermediary value, we seek to maintain some of the beamforming abilities of multiple antennas, as in Eq. (4), and at the same time to use some of the Zero-Forcing capabilities of the system. Thus, the design of a beamformer is dictated by the following optimization problem

$$\max_{\mathbf{h}_{\lambda}^{\perp} \in \mathbb{C}^L} \min_{k \in \sigma} \|\mathbf{h}_k^{\dagger} \cdot \mathbf{h}_{\lambda}^{\perp}\|^2 \tag{5}$$

$$\mathbf{s.t.} \;\; \mathbf{h}_k^{\dagger} \cdot \mathbf{h}_{\lambda}^{\perp} = 0, \;\; \forall k \in \lambda.$$

In essence, beamformer $\mathbf{h}_{\lambda}^{\perp}$ is required to belong in the null-space of the channel of all users of set $\lambda$, $|\lambda| < L$ and at the same time to increase the worst-user rate of the XOR message.

## IV. Results

As one can imagine (as also attested in Fig. 1) the SNR value defines how many streams one should use. One of our goals is to understand how changing the problem parameters, namely $L$ and $K\gamma$ would affect these sub-regions. Further, using the curve that presents the maximum over the rates achieved by any stream, we will explore i) how the achieved rates compare between systems that share the same $K\gamma + L$, but have different parameters $K\gamma$ and $L$, and ii) the multiplicative rate performance of systems when we increase either of the two parameters.

### A. Evolution of SNR sub-regions

The first result we will analyze concerns the SNR sub-regions inside which, choosing one of the $s \in [L]$ streams would outperform any other option. Specifically, we are interested in understanding if these SNR regions are changing as we increase the cache redundancy. Contrarily, if one increases the number of antennas it is easy to see that these SNR sub-regions will be achieved with much smaller power. This is a direct consequence of the fact that more antennas signify better ability to tackle the worst-user effect and in turn to increase the achieved rate at any SNR point.

We display the results in Fig. 2-4 where each figure assumes constant number of antennas equal to $L = 2, 3, 4$, respectively.

Fig. 2, corresponding to the case with $L = 2$ antennas, illustrates the SNR values where the transition happens from 1 stream to 2 streams. As we can see, most SNR values are centered around $20 - 30$ dB, while the trend shows a slight increase as the cache-redundancy increases.

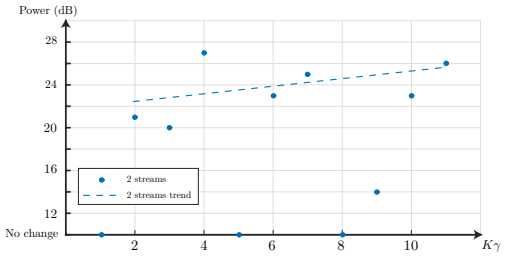


Fig. 2. Evolution of the optimal SRN region for settings with $K = 30$ users and same number of antennas $L = 2$. We can see that the 2 stream region begins approximately in the same power, even if the size of the caches increases.

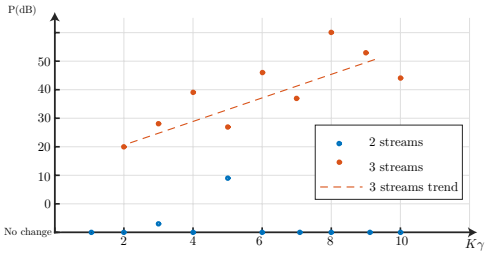


Fig. 3. Evolution of the optimal SRN region for settings with $K = 30$ users and same number of antennas $L = 3$. While choosing only one stream is not optimal in any SNR region, in fact transitioning from $s = 2$ to $s = 3$ is dependent on the the cache-redundancy.

The second figure of interest (Fig. 3) corresponds to settings with $L = 3$ antennas. We notice that employing 1 stream would always lead to a subpar performance compared to choosing 2 streams. Further, when examining the evolution of the SNR points that correspond to the 3-stream case, we can see a higher increasing trend than in the settings with $L = 2$ antennas.

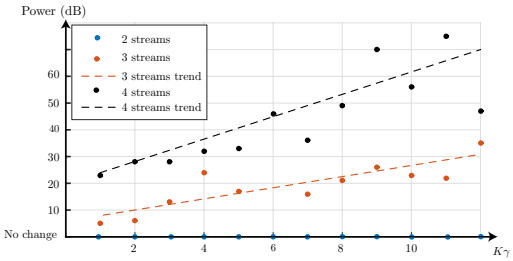


Fig. 4. Evolution of the optimal SRN region for settings with $K = 30$ users and same number of antennas $L = 4$. While choosing only one stream is not optimal in any SNR region, in fact transitioning from $s = 2$ to $s = 3$ (and subsequently from $s = 3$ to $s = 4$) is dependent on the the cache-redundancy.

Finally, in Fig. 4 we display the evolution of the sub-regions for settings with $L = 4$ antennas. In this case as well as in the previous ($L = 3$) we can see that choosing $s = 1$ stream would always lead to a subpar performance. Moreover, looking at the evolution of points in the case where $s = 3$ we can see a slight upward trend as the cache-redundancy is increasing, while in the case where $s = 4$ we can see that this trend is even steeper.

### B. Rate Comparison for constant $K\gamma + L$

In this section we will compare systems that share the same $K\gamma+L$, but differ on the number of antennas $L$ and the cache redundancy $K\gamma$. Quantity $K\gamma + L$ signifies the sum degrees-of-freedom performance (DoF) which is given as

$$D_\Sigma = \lim_{P\to\infty} \frac{\mathcal{R}}{\log(1+P)} \tag{6}$$

with $\mathcal{R}$ denoting the rate performance and $P$ the power. Thus, in theory, systems that share the same DoF should achieve the same rate performance in high SNR.

The results are illustrated in Fig. 5 where we can see a comparison of the achieved rates of all the systems, and in Fig. 6 where we display the normalized rate of three of the schemes with the rate of the system with $L = 7$ antennas. The maximal DoF performance that can be achieved in all systems is 9, while the number of antennas varies from 2 up to 7.

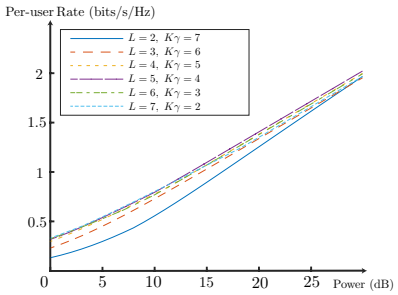


Fig. 5. Simulation results of achieved Rate as a function of Power, for settings with $K = 30$ users and same $K\gamma + L = 9$ value. We can see that for higher number of antennas the achieved rate remains relatively unchanged. On the other hand, the rate achieved by the settings with lower number of antennas is much smaller in the low SNR region.

*Caches vs Antennas:* An interesting outcome from this comparison can be seen by considering which of the two resources, either antennas or cache-redundancy, is more impactful in increasing a system's rate. Examining Fig. 5 we can see that all systems achieve approximately the same rate in the very high SNR region, something that is also displayed in Fig. 6. While it is expected that each setting would eventually achieve the same slope (i.e. DoF performance), at the same time these systems also approximately achieve the same rate performance in the high SNR region, thus hinting that in higher SNR there is no loss in rate, regardless of how the resources are distributed.

In contrast, focusing in the low and mid-SNR regions we can see a much different behaviour. Specifically, in Fig. 6 we compare the rate of the setting with most antennas ($L = 7$) with the rate of three settings with less amount of antennas. The $y$-axis displays the rate of the 7-antenna system *over* the rate of some other setting and reveals a multiplicative difference, in favour of the high antenna setting as compared to the 2-antenna setting, of approximately 2.4 in the very low SNR, which is vanishing as we approach the high SNR region. Furthermore, the system with 4 antennas achieves a smaller rate difference, while the 6-antenna setting achieves approximately the same performance as the 7-antenna system.

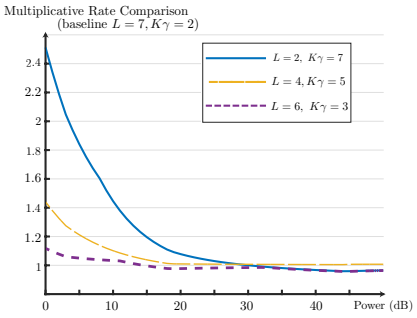


Fig. 6. Simulation results for settings with $K = 30$ users and fixed $K\gamma+L = 9$. In $y$-axis we display the rate achieved by the setting with 7 antennas *over* some of the remaining settings. In the lower SNR region the rate achieved by the setting with 7 antennas is approximately 2.4 times higher than in the setting with $L = 2$. Contrarily, as the number of antennas increases, this rate is significantly reduced.

*Conclusions:* Two interesting observations can be made at this point. First, that all rates are converging to the same values (high SNR) and second that the low SNR rates are higher for systems with more antennas. As we discussed above, the fact that rates are converging is particularly important, showing that the asymptotic behaviour is the same regardless of choosing a system with more antennas and less cache-redundancy and vice versa. This becomes more important when we see that in the low SNR region systems with less antennas exhibit a subpar performance.

Combining the above we can say that having more antennas helps by utilizing the stream that exhibits the highest performance. That is due to the ability to partially compensate for the worst-user effect by handling users with bad channels (through efficient beamforming). This comes in contrast to systems with fewer antennas that, even though they serve the same amount of users, cannot compensate for a worst user.

### C. Multiplicative Performance

In this section we investigate how the (multiplicative) performance of a system varies as we change the two resources, i.e. the number of antennas and the cache-redundancy. The outcome of this comparison would allow us to understand the importance of each of the two parameters in the finite SNR region. We will explore the connections between number of antennas and cache-redundancy with rate by first considering systems with fixed cache-redundancy and then systems with fixed number of antennas.

*Increasing the number of antennas:* We begin by comparing systems that share the same cache-redundancy but have different number of antennas. The results are presented in Fig. 7 for systems with cache-redundancy $K\gamma = 2$.

First, we can observe that all the curves converge to the theoretical values in the high SNR region. The second observation has to with the low and mid-SNR regions where we can see a high performance boost when adding more antennas to a system. While all systems provide a higher performance in the finite SNR region, nonetheless systems with more antennas produce even higher rate boost in finite SNR.

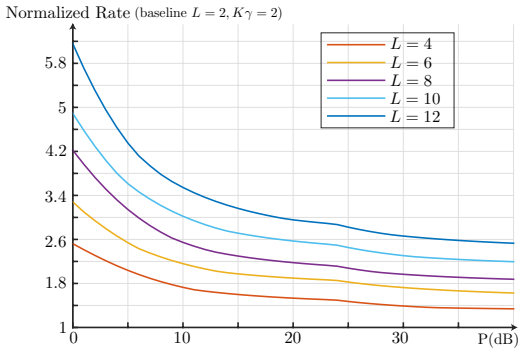


Fig. 7. Simulation results for settings with $K = 30$ users and same number of antennas $L = 9$. We compare the rate of the setting with $L = 2$ antnenas to the rates of the settings with more antennas. While in the high SNR region the systems converge to the theoretical values nevertheless, in the low and mid SNR regions more antennas provide significant rate increase.

This complements the results from the previous section which showed that systems with more antennas demonstrate higher multiplicative-rate performance in the finite SNR region. Moreover, as we also concluded in the previous section, having more antennas allows to more efficiently increase the worst rate and at the same time it provides a way to increase the number of users served by incurring a penalty in the beamformer design.

*Increasing the Cache Redundancy:* Further, we will look at the rate's behaviour by increasing the cache redundancy at the users, while retaining the same number of antennas. We display the results in Fig. 8 where we compare systems with 2 antennas against a system with 2 antennas and $K\gamma = 2$.

First, we can see that, as in the previous case, the asymptotic behaviour is in par with the theoretical predictions. On the other hand, the low and mid-SNR regions exhibit a much different behaviour, where in the lower SNR region the systems perform worse than the asymptotic trend, while in the mid-SNR region each system provides a slightly higher performance than the asymptotic trend.

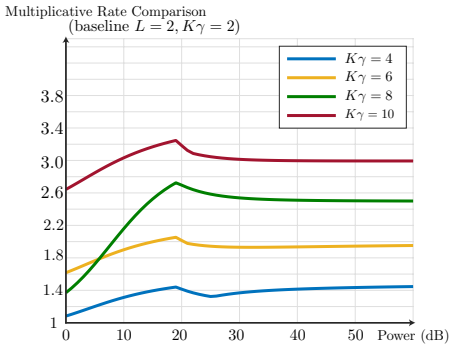


Fig. 8. Simulation results for settings with $K = 30$ users and same number of antennas $L = 9$. We compare the rate of the setting with $K\gamma = 2$ to the rates of the settings with higher amount of cache. While in the high SNR region the systems converge to the theoretical values nevertheless, in the low and mid SNR regions that the gain compared to a system with smaller cache is even higher.

*Conclusions:* From varying the number of antennas or the cache-redundancy, we saw that in the high SNR regime the simulation results converge to the theoretical predictions thus, the two resources can provide equal performance improvements. Comparing, though, the performances in the finite SNR region we have seen that adding antennas to a system provides much higher rate improvement compared to the asymptotic behaviour. This comes to a stark difference with the effect of increased cache-redundancy, where in the low SNR region remains below the theoretical value, while in the mid-SNR region it provides a small advantage than the asymptotic behaviour.

## APPENDIX

In this section we describe the caching and delivery processes of the algorithm of [10] that we use in this work.

### A. Placement Phase

Initially, each file is subpacketized into $S = \binom{K}{K\gamma}$ subpackets, which are further split into $K\gamma + s$ smaller packets[2]. Each packet of file $W^n$, $n \in [N]$ is described by two indices namely $W^{n,p}_\tau$, where $\tau \subset [K]$, $|\tau| = K\gamma$, while $p \in [K\gamma + s]$. We will assume that $T_{MN} = \frac{K(1-\gamma)}{1+K\gamma}$ is an integer, while extending the scheme to non-integer values requires a little higher subpacketization. Users' caches are filled according to

$$\mathcal{Z}_{k\in[K]} = \big\{W^{n,p}_\tau : \tau \subset [K], |\tau| = K\gamma, k \in \tau, \quad \forall p \in [K\gamma + s], \forall n \in [N]\big\}. \tag{7}$$

The purpose of index $p$ is two-fold. On the one hand it assists with the combinatorial problem of matching XORs with subfile indices (cf. [10]), while on the other hand it allows to deliver every time a "fresh" subfile making a total of $K\gamma + s$ subfiles for each associated $W^n_\tau$. We will refrain from using this index in the following algorithm, in order to keep the notation more clear, but we will show that each subfile $W^n_\tau$ is transmitted $K\gamma + s$ times, thus each individual $p, \tau$ pair will be transmitted.

The main idea behind the algorithm is to transmit $s$ messages, which carry a total of $K\gamma + s$ subfiles. One of the messages is a XOR comprized of $K\gamma + 1$ subfiles, while each of the remaining $s - 1$ messages carries one subfile. In each transmission we will pick a set of $K\gamma + 1$ users, which we denote by set $\sigma$, and who will be receiving the multicast message. The remaining $s-1$ users included in this transmission form the – ordered – set $\lambda$. If $s > 1$ we denote $\lambda_i \triangleq \lambda \setminus \lambda(i)$, $i \in [s]$, where $\lambda(i)$ denotes the $i-$th element of ordered set $\lambda$, while $q \in \sigma$ along with users of set $\lambda$ will denote the precoding-assisted users, thus users of set $\tau \triangleq \sigma \setminus q$ are not-precoding assisted.

The transmitted message takes the following form

$$\mathbf{x} = \mathbf{h}^\perp_\lambda X_\sigma + \sum_{i=1}^{s-1} \mathbf{h}^\perp_{\lambda_i \cup q} W^{d_{\lambda(s)}}_\tau. \tag{8}$$

[2]This "extra" subpacketization can be performed after the requests and the SNR are revealed thus, no such knowledge is required during placement.

While the explicit design of beamforming vectors $\mathbf{h}_{\mu}^{\perp}$ has been discussed in the main part of this document, here we are interested only in their main attribute i.e., the ability to null-out interference at some users (see Eq. (3)).

Further, in this section we will discuss the explicit design of the multicast message $X_{\sigma}$ and the selection of the uncoded subfiles $\{W_{\mu}^{d_k}\}_{k\in\lambda}$ that can allow a successful decoding and delivery of all the requested subfiles to the users.

### *B. Delivery Phase*

First, we note that in the case $s = 1$ the delivery algorithm is based on the algorithm of [1], where in our case it is repeated $K\gamma+1$ times such as to account for the extra subpacketization. All the remaining cases are based on the algorithm of [10] and are given in the form of pseudo-code in Alg. 1.

**Algorithm 1:** Delivery Phase

**for** *all* $\sigma \subseteq [K], |\sigma| = K\gamma + 1$ *(pick XOR)* **do**
  **for** *all* $q \in \sigma$ *(pick precoded user)* **do**
    Set: $\tau = \sigma \setminus \{q\}$
    Set: $\lambda = \beta_{\tau,q}$.
    Transmit:

$$\mathbf{x}_{q,\tau} = \mathbf{h}_{\lambda}^{\perp} X_{\sigma} + \sum_{i=1}^{s-1} \mathbf{h}_{\lambda_i \cup \{q\}}^{\perp} W_{\tau}^{d_{\lambda(s)}}. \quad (9)$$

*Details of Algorithm 1:* The algorithm begins by selecting a subset $\sigma$ of the users of size $K\gamma + 1$. For these users, the algorithm will form XOR $X_{\sigma}$ in the same way as does the algorithm of [1] i.e., $X_{\sigma} = \bigoplus_{k\in\sigma} W_{\sigma\setminus\{k\}}^{d_k}$. Then, the algorithm selects one user, $q$, from the users of set $\sigma$, which user will be helped by precoding. It is easy to see that the subfile index that this user will receive is $\sigma \setminus \{q\} = \tau$.

The algorithm proceeds to select the remaining $s-1$ users. These users are described by set $\beta_{\tau,q}$, which is calculated by finding the $s-1$ consecutive elements of the ordered set $[K]\setminus\tau$ after element $q$.

*Example:* Assume a MISO BC with $K = 5$, $K\gamma = 2$ and $s = 2$. Further, let $\sigma = \{1, 2, 3\}$ and $q = 1$ then $\tau = \sigma\setminus\{q\} = \{2, 3\}$ thus, $[K]\setminus\tau = \{1, 4, 5\}$, which means that $\beta_{\tau,q} = \{4\}$, since 4 is the consecutive element of element $q = 1$.

The users of set $\tau\cup\{q\}\cup\beta_{\tau,q}$ consist the $K\gamma+s$ users that will receive a subfile in this slot. Further, each user $k$ of set $\beta_{\tau,q}$ will receive subfile indexed by set $\tau$, i.e. $W_{\tau}^{d_k},\ k \in \beta_{\tau,q}$.

*Decoding Process:* We begin with user $k$ of set $\lambda\cup\{q\}$ i.e, the "precoding-assisted" users. Due to the design of the precoder (cf. Eq. (3)), we can see that these users will receive either the multicast message (user $q$) or each of the uncoded messages to the respective user i.e.,

$$y_k = \mathbf{h}_k^{\dagger}\mathbf{x}_{q,\tau} = \begin{cases} X_{\sigma}, & \text{if } k = q \\ W_{\tau}^{d_k}, & \text{if } k \in \lambda \end{cases} \quad (10)$$

where for simplicity we have removed the noise and the channel precoding product. It is easy to see that users in set $\lambda$ (assisted by precoding) will only "see" the uncoded subfile that they want. Further, user $q$ will receive no interference from any of the uncoded messages thus, will receive XOR $X_{\sigma}$ which can proceed to decode using its cached content.

On the other hand, users in set $\tau$ will be receiving a linear combination of all $s$ messages, which will proceed to decode using both CSIT knowledge and their cached subfiles. The received message at some user $k \in \tau$ takes the form

$$y_{k\in\tau} = \mathbf{h}_k^{\dagger}\left(\mathbf{h}_{\lambda}^{\perp} X_{\sigma} + \sum_{i=1}^{s-1} \mathbf{h}_{\lambda_i\cup\{q\}}^{\perp} W_{\tau}^{d_{\lambda(i)}}\right). \quad (11)$$

From Eq. (11), the subfiles that are included in the summation term have all been cached by all receivers of set $\tau$ and as such they can be removed from the equation. What remains is XOR $X_{\sigma}$ which, by design, is decodable by all users in $\tau$.

**Corollary 1.** *In Algorithm 1, each subfile* $W_{\tau}^{d_k},\ k \in [K]$ *is transmitted exactly* $K\gamma + s$ *times.*

*Proof.* Due to lack of space, we omit this proof. The interested reader is pointed to [10]. □

## REFERENCES


[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. on Information Theory*, vol. 60, pp. 2856–2867, May 2014.

[2] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Transactions on Information Theory*, vol. 62, pp. 7253–7271, Dec 2016.

[3] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," in *IEEE Information Theory Workshop (ITW)*, pp. 1–5, Nov 2018.

[4] E. Lampiris and P. Elia, "Achieving full multiplexing and unbounded caching gains with bounded feedback resources," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1440–1444, June 2018.

[5] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 36, pp. 1176–1188, June 2018.

[6] E. Lampiris and P. Elia, "Bridging two extremes: Multi-antenna coded caching with reduced subpacketization and CSIT," *SPAWC*, 2019.

[7] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Transactions on Information Theory*, vol. 65, pp. 2792–2807, May 2019.

[8] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, "Fundamental limits of wireless caching under uneven-capacity channels," *arXiv preprint arXiv:1908.04036*, 2019.

[9] N. D. Sidiropoulos, T. N. Davidson, and Zhi-Quan Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, pp. 2239–2251, June 2006.

[10] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *arXiv preprint arXiv:1806.07800*, 2018.

[11] K. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 548–562, Jan 2018.

[12] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multi-antenna interference management for coded caching," *arXiv preprint arXiv:1711.03364*, 2017.

[13] J. Zhao, M. M. Amiri, and D. Gündüz, "A low-complexity cache-aided multi-antenna content delivery scheme," in *IEEE 20th Int. Workshop on Sig. Proc. Advances in Wireless Comm. (SPAWC)*, pp. 1–5, July 2019.