Multi-antenna Coded Caching at Finite-SNR: Breaking Down the Gain Structure

MohammadJavad Salehi and Antti Tölli

Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland
E-mail: {firstname.lastname}@oulu.fi

***Abstract*—Multi-antenna coded caching (CC) techniques are considered viable options for achieving higher data rates in future networks, especially for the prominent use case of multimedia-driven applications. However, despite their information-theoretic analyses, which are thoroughly studied in the literature, the research on the finite-SNR performance of multi-antenna CC techniques is not yet mature. In this paper, we try bridging this gap by breaking down, categorizing, and studying the effect of six crucial parameters affecting the finite-SNR performance of multi-antenna CC schemes. We also investigate the interaction of different parameters and clarify how they could affect the implementation complexity in terms of the necessary computation and subpacketization. Theoretical discussions are followed and verified by numerical analysis.**

***Index Terms*—coded caching; multi-antenna communications; multiplexing; multicasting; beamforming**

I. Introduction

Wireless communication networks are under mounting pressure to support higher data rates, and this trend will continue by emerging new services such as extended reality (XR) applications [1]. The coded caching (CC) technique [2] is considered a viable candidate to address this issue. The critical property of CC is that it enables the data storage memory in network devices to be efficiently used as a communication resource, especially for the prominent use case of multimedia-driven applications [3], [4]. Interestingly, the caching gain of CC scales with the cumulative cache size of all users in the network [2], and can be combined with the spatial multiplexing gain of incorporating multiple antennas at the transmitter [5], [6] or receivers [7]. However, this promising CC gain is followed by crucial practical bottlenecks hindering its applicability. Two important such bottlenecks are the subpacketization issue and the optimized beamformer design (for improved performance at the finite-SNR regime). Former stems from the fact that the original multi-antenna CC scheme in [6] required splitting files into exponentially-growing numbers of smaller parts, and the latter is due to the fact that designing optimized beamformers for this scheme requires solving complex non-convex optimization problems [8].

While the subpacketization issue seems to be fundamental in single-antenna systems [9], multi-antenna setups provide flexibility for reducing the required subpacketization without altering the theoretical degrees-of-freedom (DoF) gains [10]–[12]. The key to achieving this reduction is in using a new CC approach where the cache-aided interference cancellation is done *before* decoding the data and in the signal domain [10]. Following [4], we use the term *signal-level* while referring to CC schemes with this new approach and denote the traditional CC schemes (with cache-aided interference cancellation after decoding the received signal) as *bit-level*. Interestingly, even though signal-level CC schemes were initially developed for subpacketization reduction, the work in [11] showed that they also allow simplifying the optimized beamformer design by altering the underlying multicast structure of bit-level schemes.

However, achieving the same DoF and applicability of optimized beamformers does *not* reveal the whole story about the finite-SNR performance of signal-level CC schemes. For example, it was shown in [13] that with the same DoF, moving eventually from a pure signal-level approach (smallest subpacketization) to a pure bit-level one (largest subpacketization), the performance is improved at the finite-SNR regime. Also, the work in [8] revealed the fact that being DoF-optimal is not necessarily preferred in finite-SNR as the DoF can be traded-off with the more effective beamformer directivity parameter. Such results clarify a need to better understand all the parameters affecting the finite-SNR performance of CC schemes and to investigate the possibility of designing new CC schemes and beamforming techniques to better tune the performance for any given complexity constraint.

This paper breaks down and classifies all the parameters affecting the finite-SNR performance of multi-antenna CC schemes. Specifically, we identify six different parameters, as shown in Figure 1, and analyze how they affect the achievable rate of CC schemes in different SNR values. For each parameter, we provide the definition and a brief discussion on how it can be adjusted, followed by theoretical insights and numerical simulation results. As discussed, this paper is not the first to identify all these parameters and study their effects. However, it is the first to gather, organize, and study the interaction of all such performance-affecting parameters. We also carefully discuss how each parameter could affect the implementation complexity and provide guidelines for selecting the appropriate scheme following the available transmission power and acceptable complexity.

This work is supported by the Academy of Finland under grants no. 346208 (6G Flagship), 319059 (CCCWEE), and 343586 (CAMAIDE), and by the Finnish Research Impact Foundation under the project 3D-WIDE.

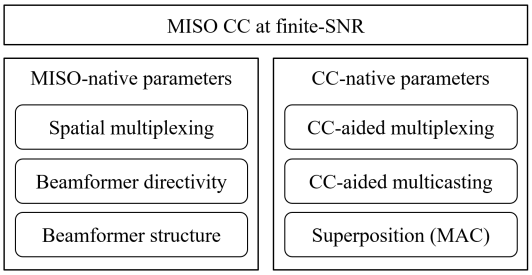

Fig. 1: Parameters affecting the finite-SNR performance of MISO CC schemes

In this paper, we have used the following notations. Boldface lower- and upper-case letters denote vectors and matrices, respectively. Calligraphic letters are used to represent sets. For an integer K, $[K]$ is the set of numbers $\{1, 2, \cdots, K\}$. The symbol $\oplus$ denotes the bit-wise XOR operation. For two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A}\backslash\mathcal{B}$ is the set of elements of $\mathcal{A}$ that are not in $\mathcal{B}$. $|\mathcal{A}|$ is the number of elements in $\mathcal{A}$. Set brackets and element separators are sometimes dropped for notational simplicity.

II. System Model

We consider a multi-input single-output (MISO) cache-aided communication setup. A single, L-antenna transmitter with spatial multiplexing gain of $\alpha \leq L$ communicates with K single-antenna users over a shared wireless link. Every user has a cache memory of size MF bits and requests files from a library $\mathcal{F}$ of N files, each with the size of F bits. The CC gain is defined as $t \equiv \frac{KM}{N}$. For notational simplicity, we use a normalized data unit and drop F in our subsequent notations. We also use *CC*-(K, t, L, α) to refer to this setup.

The system operation consists of two phases, content placement and delivery. In the content placement phase, which is done at the low network traffic time, cache memories of the users are filled up by data chunks of files in $\mathcal{F}$. Then, in the delivery phase, after each user $k \in [K]$ reveals its requested file $W(k) \in \mathcal{F}$, the server creates and transmits a set of codewords such that all the users can decode their requested files using their cache contents and the received signals. The goal is to design content placement and delivery phases such that the symmetric rate, defined as the amount of data delivered to all the users over a defined time frame, is maximized. In this paper, we consider only CC schemes with uncoded data placement and single-shot data delivery, for which the MISO-CC scheme in [6] is shown to be DoF-optimal [14].

Data delivery is done using a set of transmission vectors, which are sent, for example, in consecutive time slots. A transmission vector $\mathbf{x}_\mathcal{S}$ delivers (parts of) the requested data to every user in the subset $\mathcal{S} \subseteq [K]$ of users, where $|\mathcal{S}| = t + \alpha$. This is done through the parallel transmission of codewords $X_\mathcal{T}$, where $\mathcal{T} \subseteq \mathcal{S}$. Every codeword $X_\mathcal{T}$ includes useful data for every user $k \in \mathcal{T}$ and is precoded with a beamforming vector $\mathbf{w}_{\mathcal{R}(\mathcal{T})}$ that nulls-out or suppresses the interference of $X_\mathcal{T}$ at every user $k \in \mathcal{R}(\mathcal{T})$. Using $\bar{\mathcal{T}}(\mathcal{S})$ to denote the set of subsets $\mathcal{T} \subseteq \mathcal{S}$ for which a codeword $X_\mathcal{T}$ is built, we can write

$$\mathbf{x}_\mathcal{S} = \sum_{\mathcal{T} \in \bar{\mathcal{T}}(\mathcal{S})} X_\mathcal{T} \mathbf{w}_{\mathcal{R}(\mathcal{T})}. \tag{1}$$

Then, after the transmission of $\mathbf{x}_\mathcal{S}$, the received signal at user $k \in \mathcal{S}$ can be modeled as

$$y_\mathcal{S}(k) = \sum_{\mathcal{T} \in \bar{\mathcal{T}}(\mathcal{S})} X_\mathcal{T} \mathbf{h}_\mathcal{S}(k)^H \mathbf{w}_{\mathcal{R}(\mathcal{T})} + z_\mathcal{S}(k), \tag{2}$$

where $\mathbf{h}_\mathcal{S}(k) \in \mathbb{C}^{L \times 1}$ and $z_\mathcal{S}(k)$ represent the channel vector and noise at user k during the transmission of $\mathbf{x}_\mathcal{S}$. In this paper, we use assume $\mathbf{h}_\mathcal{S}(k)$ does not change during the transmission of $\mathbf{x}_\mathcal{S}$, and full channel state information (CSI) is available at the transmitter. We also use $\mathbf{h}_k \equiv \mathbf{h}_\mathcal{S}(k)$ and $z_k \equiv z_\mathcal{S}(k)$ for notational simplicity.

Let us denote the maximum time required for every user k in $\mathcal{S}$ to decode $y_\mathcal{S}(k)$ as $T_\mathcal{S}$. Then, the total delivery time is $T = \sum_\mathcal{S} T_\mathcal{S}$, and the symmetric channel rate, used as the comparison metric here, can be defined as

$$R_{\text{sym}} = \frac{K - t}{T}. \tag{3}$$

Note that this metric reflects how well the wireless communication channel is used and excludes the effect of the local caching gain. This is unlike the total symmetric rate $\frac{K}{T}$ widely used in the literature [6]. Of course, the two metrics can be converted into each other with a simple coefficient.

III. MISO-stemmed parameters

As shown in Figure 1, we study three MISO-stemmed parameters affecting the finite-SNR performance of coded caching schemes. For this study, we choose the MISO scheme in [6] as the baseline. This is because this scheme enables the maximum performance for all CC-stemmed parameters and hence, provides a fair structure for capturing the effect of the MISO-stemmed parameters considered here.

With the MISO scheme in [6] as the baseline, during the placement phase, we split every file $W \in \mathcal{F}$ into $\binom{K}{t}$ subfiles $W_\mathcal{U}$, where $\mathcal{U} \subseteq [K]$ and $|\mathcal{U}| = t$. Then, at the cache memory of user $k \in [K]$, we store $W_\mathcal{U}$ for every $W \in \mathcal{F}$ and $\mathcal{U} \ni k$. Subsequently, during the delivery phase, we create a separate transmission vector $\mathbf{x}_\mathcal{S}$ for *every* subset of users $\mathcal{S} \subseteq [K]$ with size $|\mathcal{S}| = t + \alpha$. Given one such set $\mathcal{S}$, the transmission vector $\mathbf{x}_\mathcal{S}$ includes a codeword for *every* user subset of $\mathcal{S}$ with $t + 1$ users, i.e., $\bar{\mathcal{T}}(\mathcal{S}) = \{\mathcal{T} \subseteq \mathcal{S} \mid |\mathcal{T}| = t + 1\}$.[1]

For better clarification, let us consider an example network of *CC*-$(6, 2, 4, \alpha)$. During the placement phase, we split each file in $\mathcal{F}$ into $\binom{6}{2} = 15$ subfiles and store five of them in the cache memory of each user. For example, a file $A \in \mathcal{F}$ is split into subfiles A_{12}, A_{13}, $\cdots$, A_{56}, from which A_{12}, A_{13}, A_{14}, A_{15}, and A_{16} (i.e., all subfiles $A_\mathcal{U}$ for which $1 \in \mathcal{U}$) are stored in the cache memory of user 1. Now, we review the delivery phase, assuming $\alpha = 1, 4$. For notational simplicity, let us denote the files requested by users 1-6 as $A, B, \cdots, F$, respectively.

• $\alpha = 1$. In this case, we have to create $\binom{K}{t+\alpha} = \binom{6}{3} = 20$ transmission vectors $\mathbf{x}_\mathcal{S}$. We also have $\bar{\mathcal{T}}(\mathcal{S}) = \{\mathcal{S}\}$, i.e., every

[1]It should be noted that during the delivery phase, we should further split every subfile $W_\mathcal{U}$ into $Q = \binom{K-t-1}{\alpha-1}$ smaller subpackets $W_\mathcal{U}^q$, $q \in [Q]$, to ensure new data chunks are delivered during every transmission. In this paper, we have ignored noting this second-level subpacketization as it does not affect the discussions. The effects are considered in numerical simulations, though.

vector $\mathbf{x}_{\mathcal{S}}$ includes only a single codeword $X_{\mathcal{T}}$. As a result, there is no need to null-out (or suppress) the interference from $X_{\mathcal{T}}$ at other users, and $\mathcal{R}(\mathcal{T}) = \varnothing$. For example, considering $\mathcal{S} = \{1, 2, 3\}$, the transmission vector $\mathbf{x}_{123}$ is built as

$$\mathbf{x}_{123} = X_{123}\mathbf{w}_{\varnothing}, \tag{4}$$

where the codeword $X_{123} = A_{23} \oplus B_{13} \oplus C_{12}$ includes data for all users 1, 2, and 3.

Now, let us consider the decoding process at user 1 after transmitting $\mathbf{x}_{123}$. Using (2), this user receives

$$y_{123}(1) = X_{123}\mathbf{h}_1^T\mathbf{w}_{\varnothing} + z_1, \tag{5}$$

and can extract X_{123} with a maximum rate of

$$r_{123}(1) \leq \log\big(1 + \frac{|\mathbf{h}_1^T\mathbf{w}_{\varnothing}|}{N_0}\big). \tag{6}$$

Then, user 1 has to use its cached contents to remove unwanted data terms B_{23} and C_{12} from X_{123}. Similarly, users 2 and 3 can also get their requested data terms.

• $\alpha = 4$. In this case, we create only a single transmission vector $\mathbf{x}_{1\cdots6}$. However, this transmission vector is comprised of $\binom{6}{3} = 20$ codewords, where each codeword delivers data to three users and its interference is suppressed by beamforming at the other three users. In other words,

$$\mathbf{x}_{1\cdots6} = X_{123}\mathbf{w}_{456} + X_{124}\mathbf{w}_{356} + \cdots + X_{456}\mathbf{w}_{123}. \tag{7}$$

The codewords are similar to the $\alpha = 1$ case. For example, we have $X_{123} = A_{23} \oplus B_{13} \oplus C_{12}$, $X_{124} = A_{24} \oplus B_{14} \oplus D_{12}$, and $X_{456} = D_{56} \oplus E_{46} \oplus F_{45}$.

Now, let us consider the decoding process at user 1. The received signal at this user is

$$y_{1\cdots6}(1) = X_{123}\mathbf{h}_1^T\mathbf{w}_{456} + \cdots + X_{456}\mathbf{h}_1^T\mathbf{w}_{123} + z_1. \tag{8}$$

Unlike the $\alpha = 1$ case, user 1 now has to decode its requested data from a multiple access channel (MAC) of size 10, using a more complex successive interference cancellation (SIC) receiver, as there exist $\binom{5}{2} = 10$ codewords $X_{\mathcal{T}}$ for which $1 \in \mathcal{T}$. The maximum decoding rate in this case is determined by a rate region. Assume $\hat{\mathcal{V}}(1)$ and $\bar{\mathcal{V}}(1)$ include the user subsets $\mathcal{T} \subseteq \mathcal{S}$ of size $t + 1 = 3$, that include and not include user 1, respectively. In other words

$$\hat{\mathcal{V}}(1) = \{123, 124, \cdots, 156\}, \tag{9}$$

$$\bar{\mathcal{V}}(1) = \{234, 235, \cdots, 456\}. \tag{10}$$

Then, the SINR term $\gamma_{\hat{\mathcal{V}}}$ for a desired term $A_{\hat{\mathcal{V}}}$, $\hat{\mathcal{V}} \in \hat{\mathcal{V}}(1)$, is calculated as

$$\gamma_{\hat{\mathcal{V}}} = \frac{|\mathbf{h}_1^T\mathbf{w}_{[6]\setminus\hat{\mathcal{V}}}|^2}{\sum_{\bar{\mathcal{V}}\in\bar{\mathcal{V}}(1)}|\mathbf{h}_1^T\mathbf{w}_{[6]\setminus\bar{\mathcal{V}}}|^2 + N_0}, \tag{11}$$

and the decoding rate $r_{1\cdots6}(1)$ satisfies

$$r_{1\cdots6}(1) \leq \frac{1}{|\mathcal{W}|}\log\big(1+\sum_{\hat{\mathcal{V}}\in\mathcal{W}}\gamma_{\hat{\mathcal{V}}}\big),\ \forall\mathcal{W} \subseteq \hat{\mathcal{V}}(1), |\mathcal{W}| > 0. \tag{12}$$

Finding optimal beamformers maximizing the rate in this case is difficult as it requires solving a non-convex optimization problem with the number of constraints growing exponentially with the MAC size [8]. We will discuss the effect of different beamforming strategies in Section III-C.

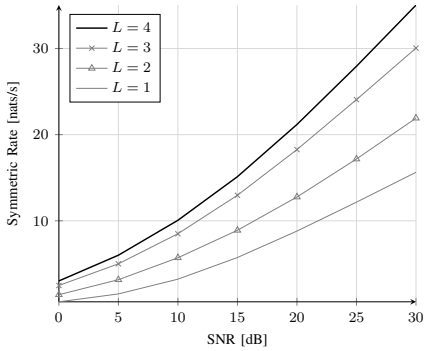

Fig. 2: Spatial multiplexing effect, $K = 6$, $t = 2$, $\alpha = L$

SNR (dB)	0	5	10	15	20	25	30
$L = 4$	405	297	212	164	140	130	124
$L = 3$	317	231	164	126	107	98	92
$L = 2$	138	109	77	55	45	41	40

TABLE I: Rate benefit (%) over $L = 1$ case, $K = 6$, $t = 2$, $\alpha = L$

A. Spatial multiplexing effect

Spatial multiplexing gain $\alpha \leq L$ determines the number of users at which we can null-out or suppress the interference caused by each term. As a larger α means serving more users in parallel (and hence, a larger DoF), most works in the literature have assumed $\alpha = L$. However, as discussed in [8], the DoF is not the best metric at the finite-SNR regime; setting $\alpha < L$ results in better performance due to an increased *beamformer directivity* gain.

In this section, to have a fair analysis of the effect of the spatial multiplexing gain, we set $\alpha = L$ to remove the effect of beamformer directivity. We perform numerical simulations for *CC*-$(6, 2, L, L)$ setup, where $L \in \{1, 2, 3, 4\}$. Simulation results are shown in Figure 2 and Table I. In all simulations, we use optimized beamformers for maximizing performance. Highlights of the results are:

1) If $\alpha = L$, a larger spatial multiplexing gain always improves the performance. This is because the DoF is increased, and beamformer directivity has no effect;
2) Performance improvement (over the $\alpha = L = 1$ case) is more prominent in the finite-SNR regime. This is because the $\alpha = L = 1$ case represents omni-casting data in all directions, which is very inefficient in finite-SNR;
3) Using multi-antenna transmission techniques is important in finite-SNR, even when they are used only for spatial multiplexing and not to improve beamformer directivity.

B. Beamformer directivity effect

As discussed, although choosing $\alpha < L$ decreases the DoF, it may improve the finite-SNR performance by enhancing beamformer directivity. In fact, as we decrease α (compared with L), the ratio of the variables to constraints in the beamformer optimization problem grows larger, enabling designing narrower beams that better direct data signals to end users [8], thus improving the performance, especially in finite-SNR.

To investigate the effect of beamforming directivity, we have provided simulation results for a *CC*-$(6, 2, 4, \alpha)$ setup, where $\alpha \in \{1, 2, 3, 4\}$, in Figure 3 and Table II. Again,

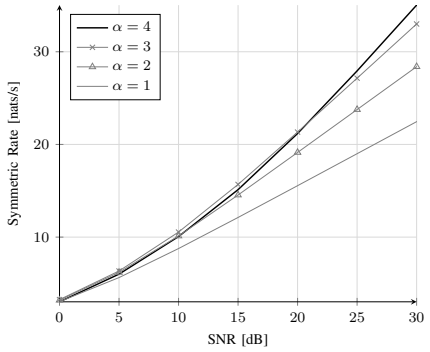

Fig. 3: Beamformer directivity effect, $K = 6$, $t = 2$, $L = 4$

SNR (dB)	0	5	10	15	20	25	30
$\alpha = 4$	-1	7	15	25	36	47	56
$\alpha = 3$	7	13	20	29	37	43	47
$\alpha = 2$	5	10	15	20	23	25	26

TABLE II: Rate benefit (%) over $\alpha = 1$ case, $K = 6$, $t = 2$, $L = 4$

optimized beamformers are used to maximize the performance. Highlights of the results are:

1) Beamformer directivity has a very strong effect in finite-SNR. It can even fully compensate for the performance loss due to decreased DoF. These results comply with [8];
2) The best choice seems to be choosing α to be slightly smaller than L. This choice also slightly simplifies the beamformer design problem. Of course, in smaller SNR values (the negative range, which is not considered here), this recommendation may change.

C. Beamformer structure effect

The beamforming vectors $\mathbf{w}_{\mathcal{R}(\mathcal{T})}$ in (1) can be designed in different ways. The simplest strategy is zero-forcing, i.e., to design $\mathbf{w}_{\mathcal{R}(\mathcal{T})}$ such that $\mathbf{h}_k^T \mathbf{w}_{\mathcal{R}(\mathcal{T})} = 0$ for every user $k \in \mathcal{R}(\mathcal{T})$. In other words, $\mathbf{w}_{\mathcal{R}(\mathcal{T})}$ should lie in the null-space of the matrix $\mathbf{H} = [\mathbf{h}_{k_1}, ..., \mathbf{h}_{k_{\alpha-1}}]$ formed by concatenating channel vectors of users in set $\mathcal{R}(\mathcal{T}) = \{k_1, ..., k_{\alpha-1}\}$. This is straightforward if $\alpha = L$, as in this case, the null-space is of dimension one. However, if $\alpha < L$, the null-space has higher dimensions and the beamformer vector can be any vector in that space. In this paper, we assume the best vector (for maximizing the symmetric rate) is found in the null-space by solving an optimization problem. The details are removed due to the lack of space.

Of course, zero-forcing is not the optimum strategy in the finite-SNR regime [8]. Instead, one needs to use optimized beamformers by maximizing the symmetric rate given the SNR constraints. Indeed, this can result in non-convex optimization problems, necessitating solutions such as successive convex approximation (SCA) [8]. Of course, the underlying scheme can be tweaked to reduce the optimized beamformer design complexity (e.g., using the signal-level scheme in [15]). However, we still use the baseline MISO scheme in [6] here, as it maximizes the positive effects of all CC-stemmed parameters.

To investigate the effect of the beamformer structure, we have provided numerical simulation results for zero-force and optimized beamforming strategies for a *CC*-$(6, 2, 4, \alpha)$ setup,

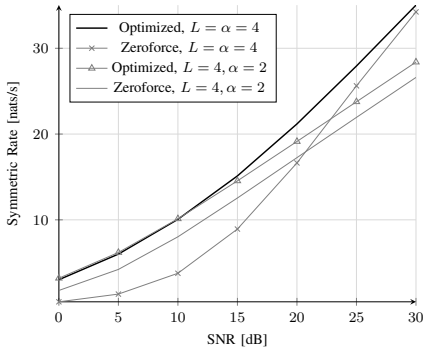

Fig. 4: Beamformer structure effect, $K = 6$, $t = 2$, $L = 4$

SNR (dB)	0	5	10	15	20	25	30
$\alpha = 4$	582	347	167	69	27	9	2
$\alpha = 2$	81	47	26	16	11	8	7

TABLE III: Rate benefit (%) over ZF beamforming, $K = 6$, $t = 2$, $L = 4$

$\alpha \in \{2, 4\}$, in Figure 4 and Table III. Highlights of the results are:

1) Performance gap of the two strategies is very big in low-SNR but narrows down as the SNR grows;
2) Performance gap at low-SNR gets smaller if $\alpha < L$. This is because with $\alpha < L$, zero-force beamformers are selected from a larger null-space [8];
3) If computation capability is a bottleneck, choosing $\alpha < L$ and zero-force beamforming is a wise choice for finite-SNR communications, as we also benefit from improved beamformer directivity.

IV. CC-stemmed parameters

In this section, we review the three CC-stemmed parameters in Figure 1 that affect the finite-SNR performance of cache-aided MISO communications.

A. CC-aided multiplexing effect

CC-aided multiplexing is determined by the CC gain t and indicates how much interference can be removed by the cache contents of target users. As the value of t directly affects DoF, we expect the performance to improve as t grows larger.

To investigate the effect of the CC-aided multiplexing gain, we use numerical simulations for a *CC*-$(6, t, 4, 4)$ setup, where $t \in \{0, 1, 2\}$. We assume the MISO-CC scheme of [6] is used as the baseline, and optimized beamformers are applied to maximize the performance. The results are provided in Figure 5 and Table IV. Highlights of the results are:

1) As expected, better performance is attained with larger t;
2) Performance improvement by cC-aided multiplexing is independent of the SNR value. This is because cache-aided interference removal is also independent of SNR;
3) Cache-aided communications is a strong tool at finite-SNR as it does not impose similar complexities of beamforming in that regime.

B. CC-aided multicasting effect

So far, we have only considered bit-level MISO-CC schemes where the cache-aided interference cancellation is performed

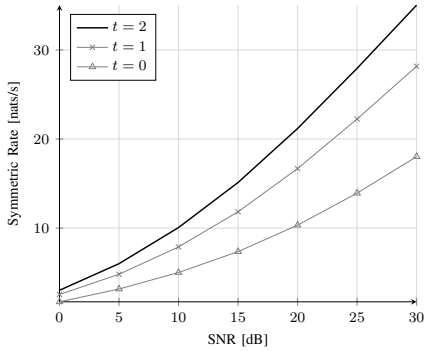

Fig. 5: Cache-aided multiplexing effect, $K = 6$, $\alpha = L = 4$

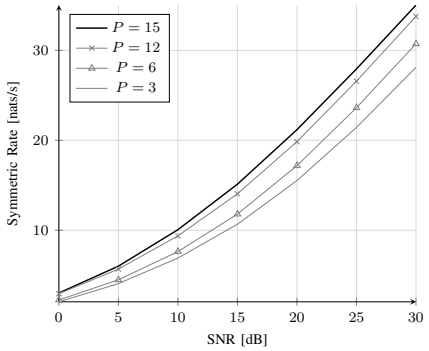

Fig. 6: Cache-aided multicasting effect, $K = 6$, $t = 2$, $\alpha = L = 4$

SNR (dB)	0	5	10	15	20	25	30
$t = 2$	75	90	101	105	105	101	94
$t = 1$	48	53	57	61	62	60	56

TABLE IV: Rate benefit (%) over $t = 0$ case, $K = 6$, $\alpha = L = 4$

SNR (dB)	0	5	10	15	20	25	30
$P = 15$	48	48	46	42	36	30	25
$P = 12$	44	40	36	32	28	24	20
$P = 6$	10	10	10	11	11	10	9

TABLE V: Rate benefit (%) over $P = 3$ case, $K = 6$, $t = 2$,$\alpha = L = 4$

after the received signal is decoded. As mentioned earlier, another class of MISO-CC schemes with cache-aided interference cancellation *before* decoding the received signal has recently gained popularity due to nice properties such as reduced subpacketization [10], [15], simpler beamformer design [15], and applicability to dynamic setups [11]. However, it is also discussed in [4], [13] that these nice properties are accompanied by penalties; most notably in reduced finite-SNR performance and more complex signaling requirements in the control plane. In this paper, we investigate the former issue and show that it is related to a reduction in the coded multicasting gain. In [13], this gain is referred to as *efficiency index*.

Let's consider a *CC*-$(6, 2, 4, 4)$ setup and review how the subpacketization can be reduced from 15 to 3 using a signal-level CC scheme.

Placement phase. We split each file $W \in \mathcal{F}$ into $P = 3$ subpackets W_1, W_2, W_3, and store W_1 in users 1 and 2, W_2 in users 3 and 4, and W_3 in users 4 and 5.

Delivery phase. Let us denote the files requested by users 1-6 as $A, B, \cdots, F$, respectively. We only need the following transmission vector:

$$\begin{aligned}\mathbf{x}_{1\cdots6} =& A_2\mathbf{w}_{256} + A_3\mathbf{w}_{234} + B_2\mathbf{w}_{156} + B_3\mathbf{w}_{134} \\ &+ C_1\mathbf{w}_{456} + C_3\mathbf{w}_{412} + D_1\mathbf{w}_{356} + D_3\mathbf{w}_{312} \\ &+ E_1\mathbf{w}_{634} + E_2\mathbf{w}_{612} + F_1\mathbf{w}_{534} + F_2\mathbf{w}_{512}.\end{aligned} \quad (13)$$

Let us review the decoding process at user 1, which is interested in the first and second data terms in (13). Using (2), this user receives

$$y_{1\cdots6}(1) = A_2\mathbf{h}_1^T\mathbf{w}_{256} + A_3\mathbf{h}_1^T\mathbf{w}_{234} + \mathbf{h}_1^T I_{BF} + \mathbf{h}_1^T I_{CC} + z_1,$$

where the interference terms are

$$\begin{aligned}I_{BF} =& B_2\mathbf{w}_{156} + B_3\mathbf{w}_{134} + C_3\mathbf{w}_{412} \\ &+ D_3\mathbf{w}_{312} + E_2\mathbf{w}_{612} + F_2\mathbf{w}_{512}, \\ I_{CC} =& C_1\mathbf{w}_{456} + D_1\mathbf{w}_{356} + E_1\mathbf{w}_{634} + F_1\mathbf{w}_{534}.\end{aligned}$$

However, the interference terms in I_{BF} are nulled-out or suppressed by beamforming, and the ones in I_{CC} could be regenerated and removed from $y_{1\cdots6}(1)$ using the cached contents of user 1. So, user 1 can decode A_2 and A_3 using a SIC receiver after cache-aided interference cancellation is done in the signal domain. Comparing $\mathbf{x}_{1\cdots6}$ in (13) with its bit-level counterpart in (7), it is seen that in the signal-level scheme, we have fewer codewords and hence, need less multicasting.

As shown in [13], for our example *CC*-$(6, 2, 4, 4)$ setup, different subpacketization levels of $P \in \{3, 6, 9, 12, 15\}$ are possible. Moreover, as we move from a pure signal-level scheme (smallest P) to a pure bit-level one (largest P), we can use more multicasting in the transmission. Here, we have used numerical simulations to compare the system performance for $P \in \{3, 6, 12, 15\}$. The results are provided in Figure 6 and Table V. Highlights of the results are:

1) Performance is improved as subpacketization is increased and more multicasting is supported. This is because multicasting is more power-efficient than unicasting;
2) Performance improvement is more prominent in finite-SNR, as the rate is power-limited in this regime.

C. Superposition (MAC) effect

In the transmission vector model in (1), each codeword $X_{\mathcal{T}}$, $\mathcal{T} \in \bar{\mathcal{T}}(\mathcal{S})$, includes data for every user $k \in \mathcal{T}$. Let us consider a specific user $\bar{k} \in \mathcal{S}$. After the transmission of $\mathbf{x}_{\mathcal{S}}$, if there exist multiple sets $\mathcal{T} \in \bar{\mathcal{T}}(\mathcal{S})$ that include $\bar{k}$, this user has to decode its requested data from a multiple access channel using a SIC receiver. However, SIC receivers are complex to implement, and hence, more effort has been recently put into designing CC schemes without a SIC requirement [12], [15]. Interestingly, it is also shown in [16] that removing the SIC requirement could greatly simplify optimized beamformer design through iterative optimization methods.

Removing the SIC requirement (or reducing the MAC size) is possible in both bit- and signal-level schemes. However, signal-level schemes generally provide more flexibility as they are less constrained by multicasting [12], [15]. Here, we only consider controlling the MAC size in bit-level schemes, as they enable CC-aided multicasting gain to be achieved at its full capacity. For example, let us consider a *CC*-$(6, 2, 4, 4)$ setup and its transmission vector $\mathbf{x}_{1\cdots6}$ in 7. As discussed in Section III, every user needs to decode its requested data from

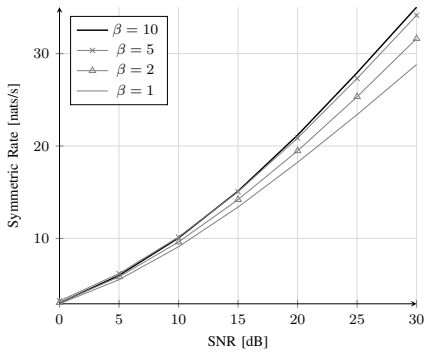

Fig. 7: Superposition (MAC) effect, $K = 6$, $t = 2$, $\alpha = L = 4$

SNR (dB)	0	5	10	15	20	25	30
$\beta = 10$	3	8	11	13	16	19	22
$\beta = 5$	11	12	12	12	14	17	18
$\beta = 2$	4	6	6	6	7	8	10

TABLE VI: Rate benefit (%) over $\beta = 1$ case, $K = 6$, $t = 2$,$\alpha = L = 4$

a MAC of size 10 after $\mathbf{x}_{1\dots6}$ is transmitted. However, for this network, we may control the MAC size (and even avoid it altogether) by scheduling the codewords sent by $\mathbf{x}_{1\dots6}$ into more transmissions. For example, to avoid the MAC, we can use the following ten transmission vectors instead:

$$\begin{aligned} \mathbf{x}^1_{1\dots6} &= X_{123}\mathbf{w}_{456} + X_{456}\mathbf{w}_{123}, \\ \mathbf{x}^2_{1\dots6} &= X_{124}\mathbf{w}_{356} + X_{356}\mathbf{w}_{124}, \\ &\dots \\ \mathbf{x}^{10}_{1\dots6} &= X_{156}\mathbf{w}_{234} + X_{234}\mathbf{w}_{156}. \end{aligned} \tag{14}$$

Controlling the MAC size is first introduced in [8] using a β parameter for reducing beamformer design complexity. Accordingly, we also use β to denote the MAC size. Simulation results for the example *CC*-$(6, 2, 4, 4)$ setup are provided in Figure 7 and Table VI. It is assumed that $\beta \in \{1, 2, 5, 10\}$, and optimized beamformers are used to maximize the performance. Highlights of the results are:

1) Increasing the MAC size generally improves performance (we suspect the deviations are due to numerical errors). This is due to improved superposition coding [8]. However, compared with other parameters, the effect is small;
2) Given the complexity of SIC receivers and the small gain of superposition coding, expanding the literature on MAC-avoiding bit-level CC schemes is a promising direction.

V. Conclusion and Future Work

In this paper, we studied six parameters affecting the finite-SNR performance of coded caching schemes in MISO setups. Out of the six parameters, three stemmed from the multi-antenna transmission part and the rest relied on cache-aided communications. For each parameter, we provided a brief explanation, followed by simulation results and insights.

As a quick summary, both multi-antenna and cache-aided communication techniques play important roles in finite-SNR. However, being DoF-optimal is not always preferred; one can slightly reduce the spatial multiplexing gain and still get better results due to improved beamformer directivity. Deviation from the optimal point also narrows the performance gap of simple zero-force beamformers with more complex optimized ones. On the other hand, CC gain is independent of the SNR regime, bit-level schemes have better performance at finite-SNR due to improved multicasting gain, and designing bit-level schemes without SIC requirement is of interest.

Future research directions include expanding the results to multi-input multi-output (MIMO) communication setups where the users are also equipped with antenna arrays, a more thorough mathematical analysis of the effect of the parameters, and expanding the results considering a more diverse set of baseline CC schemes.

References

[1] N. Rajatheva and et al., "White Paper on Broadband Connectivity in 6G," *arXiv preprint arXiv:2004.14247*, 2020. [Online]. Available: http://arxiv.org/abs/2004.14247

[2] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[3] H. B. Mahmoodi, M. J. Salehi, and A. Tolli, "Non-Symmetric Coded Caching for Location-Dependent Content Delivery," *IEEE International Symposium on Information Theory - Proceedings*, vol. 2021-July, pp. 712–717, 2021.

[4] M. Salehi, K. Hooli, J. Hulkkonen, and A. Tolli, "Enhancing Next-Generation Extended Reality Applications with Coded Caching," *arXiv preprint arXiv:2202.06814*, 2022. [Online]. Available: http://arxiv.org/abs/2202.06814

[5] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7253–7271, 2016.

[6] 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, 2019.

[7] M. J. Salehi, H. B. Mahmoodi, and A. Tölli, "A Low-Subpacketization High-Performance MIMO Coded Caching Scheme," in *WSA 2021 - 25th International ITG Workshop on Smart Antennas*, 2021, pp. 427–432.

[8] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2091–2106, 2020.

[9] Q. Yan, X. Tang, Q. Chen, and M. Cheng, "Placement Delivery Array Design Through Strong Edge Coloring of Bipartite Graphs," *IEEE Communications Letters*, vol. 22, no. 2, pp. 236–239, 2018.

[10] 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, 2018.

[11] M. J. Salehi, E. Parrinello, H. B. Mahmoodi, and A. Tolli, "Low-Subpacketization Multi-Antenna Coded Caching for Dynamic Networks," *2022 Joint European Conference on Networks and Communications and 6G Summit, EuCNC/6G Summit 2022*, pp. 112–117, 2022.

[12] S. Mohajer and I. Bergel, "MISO Cache-Aided Communication with Reduced Subpacketization," in *IEEE International Conference on Communications*, vol. 2020-June. IEEE, 2020, pp. 1–6.

[13] M. Salehi, A. Tolli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-rate trade-off in multi-antenna coded caching," in *2019 IEEE Global Communications Conference, GLOBECOM 2019 - Proceedings*. IEEE, 2019, pp. 1–6.

[14] E. Lampiris and P. Elia, "Resolving a Feedback Bottleneck of Multi-Antenna Coded Caching," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1811.03935

[15] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tolli, "Low-Complexity High-Performance Cyclic Caching for Large MISO Systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3263–3278, 2022.

[16] H. B. Mahmoodi, B. Gouda, M. Salehi, and A. Tolli, "Low-complexity Multicast Beamforming for Multi-stream Multi-group Communications," in *2021 IEEE Global Communications Conference, GLOBECOM 2021 - Proceedings*, 2022, pp. 01–06.