MISO Cache-Aided Communication with Reduced Subpacketization

Soheil Mohajer
Department of ECE
University of Minnesota
Email: soheil@umn.edu

Itsik Bergel
Faculty of Engineering
Bar-Ilan University
Email: itsik.bergel@biu.ac.il

***Abstract*—We present a novel low complexity scheme for cache aided communication, where a multi-antenna base station serves multiple single-antenna mobiles. The scheme is based on transmission of coded messages to disjoint groups of users simultaneously. Compared to the state-of-the-art, the proposed scheme significantly reduces the transmission and decoding complexity. Furthermore, it substantially relaxes the subpacketization level, and involves a transmission of much smaller number of packets in each time block. The proposed scheme achieves the same degrees of freedom (DoF) as the best known scheme, but, it suffers from a performance degradation of about 1.5dB due to a loss of diversity. Nevertheless, the loss is acceptable as the reduction in complexity allows a practical implementation of this scheme even for a large number of users.**

***Index Terms*—cache-aided communication, MISO, subpacketization level, complexity**

I. Introduction

In spite of recent improvements in wireless communication technologies and data delivery networks, the rates supported by these networks are not likely to keep up with the overwhelming growth in demand caused by the popularity of large data and high speed applications. A unique characteristic of these applications is their variation over time, which results in a heavy network traffic in peak-hours, compared to the average network traffic. Cache aided communication is a strategy that allows us to benefit from off-peak hours, and shift a part of the traffic from high traffic time to lower traffic time of the network. The gain of traditional caching is limited to the fraction of the database stored at each individual user, which is typically negligible in practice. Recent caching techniques introduced in [1] allow for two separate gains: (1) a *local gain* due to the fraction of the desired data of each user which is cached by that user and does not need to be retransmitted, and (2) a *global gain* due to the interference cancellation and broadcasting opportunity provides by caching the desired data of one user by other users. While the local gain is still small due to the orders of magnitude difference between the size of users' cache and the size of the data set, the global gain scales with the aggregate size of the cache distributed among all the users in the network, and can be substantial due to the growth and popularity of cache enabled communication devices.

The so called *coded caching* or *cache-aided communication* consists of placement phase and delivery phase. [1]. During the placement phase, prior to knowing the users' requests, we can pre-fetch and store at each users' memory some packets from the files in the database. Once the requests are revealed, the server generates a set of coded messages and transmits them to all the receivers during the delivery phase. All users should be able to decode their desired file from the received signal and their cache content. The key feature of coded caching is the fact that caching a packet at one user provides an opportunity for *multicasting combined packets*, even if it is only requested by another user. This leads to an achievable degrees of freedom (DoF), proportional to the number of cached copies of each piece of data among all the users' cache.

In a network where the transmitter is equipped by L transmit antennas, it can simultaneously serve a number of users, and L degrees of freedom can be achieved. Interestingly, the spatial diversity gain and caching gains can be simultaneously achieved. More precisely, it is shown by Shariatpanahi *et al.* that $L+M$ degrees of freedom can be achieved, where M is the aggregate cache size normalized by the size of the data set, i.e., we can distributedly store M complete copies of the database across the users [2].

A practical concern in adoption of cache aided communication for practical systems is the subpacketization level, which refers to the number of segments each file has to be divided to in order to implement the solution. A large subpacketization level leads to a complex and computationally heavy scheme, where huge number of short length file segments need to be individually encoded at the transmitter and decoded at the users. The state-of-the-art strategy for cache-aided communication in MISO networks required dividing each file into $\binom{U}{M}\binom{U-M-1}{L-1} = O\left(U^{M+L-1}\right)$ file segments, where U is the number of users in the network [2]. This is a practically infeasible number, specially since cache aided communication is attractive mostly for networks with large number of users. Recently, we proposed a scheme with subpacketization level of $\binom{U}{M}$ based on uncoded transmission [3], and achieves the same DoF as [2]. However, due to transmission of uncoded packets, the scheme of [3] requires a power that is $M+1$ times higher than that of [2].

The problem of subpacketization in cache-aided communication is widely studied for the single antenna setting. In particular, the problem is formulated as an optimization problem in [4], [5]. The trade-off between subpacketization and achievable rate is reported in [6]. Moreover, for specific

978-1-7281-5089-5/20/$31.00 ©2020 IEEE

range of parameters, combinatorial solutions are proposed [7], [8]. However, in the MISO setting, the exponent of subpacketization order not only increases by the cache size, but also by the number of antennas. In a seminal work [9], Lampiris and Elia proposed a placement and delivery scheme based on grouping and cache replication ideas. The scheme of [9] treats the network as if there are only U/L *effective* users, yet achieves the the optimum DoF of $M+L$. In this scheme it is required that the number of copies of the database distributedly cached across the users is divisible by the number of antennas, i.e., $L|M$. In practice, M is typically small, and hence such a constraint is not feasible. A failure in satisfaction of this requirement may lead to a multiplicative gap (of at most 2) in the achievable DoF compared to the optimal DoF.

Motivated by practically relevant parameters, in this work we focus on a specific range of system parameters where $M+L$ is divisible by $M+1$. We present a cache aided communication scheme which is much simpler than that of [2], yet achieves the same number of degrees of freedom. Our proposed scheme requires a subpacketization level of only $\binom{U}{M} = O\left(U^M\right)$ which is significantly smaller than $O\left(U^{M+L-1}\right)$ required by the scheme if [2]. Moreover, unlike the scheme of [2] in which users are simultaneously decoding multiple messages in a multiple access channel (MAC), in the proposed scheme each user receives and decode only one message at each transmission block. While the scheme is only presented for heterogeneous network where users experience fading channels with identical statistics, it can be extended to arbitrary network topologies. It is worth mentioning that a scheme with similar DoF and single message decoder at users is proposed in [10]. However, the subpacketization concern remain unsolved in that work.

The rest of this paper is organized as follows: the system model is introduced in Section II, our proposed scheme is presented in Section III, followed by a brief review of the scheme of [2] in Section IV. Then we analyze the two schemes and compare their performances in Section V. Finally, we conclude the paper in Section VI.

Notation. Throughout this paper we denote the set $\{1,2,\ldots,U\}$ by $[U]$ for an integer U. For two integers n and k we have $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. We use bold letters (e.g. $\mathbf{h}$) to denote vectors. The average is denoted by $\mathbb{E}[\cdot]$, and all the logarithms are in base 2.

II. System Model

We consider a network with U users each with one receive antenna, and a single base-station (BS) which is equipped wit L transmit antennas. We focus on a wideband communication scenario, in which the bandwidth, B, is divided into K frequency bins (e.g., OFDM), and each bin $k \in [K]$ carries one modulated symbol at a time without inter symbol interference. The received sample after matched filtering for the k-th frequency bin at User u is given by

$$y_{k,u}(t) = \mathbf{h}_{k,u}\mathbf{x}_k(t) + z_{k,u}(t), \tag{1}$$

where $\mathbf{x}_k(t) \in \mathbb{C}^{L\times 1}$ is the transmit vector at time slot t, $z_{k,u} \sim \mathbb{CN}\left(0, \frac{B}{K}\right)$ is the additive white Gaussian noise sample at User u in the frequency bin k, and $\mathbf{h}_{k,u} \in \mathbb{C}^{1\times L}$ is the channel vector from the BS to User u in the frequency bin k. We assume the BS has a total power constraint of P. We also assume the BS has perfect channel state information. Each link between two antennas experiences fading. Moreover, we consider a *homogeneous* system, in which all users experience a fading channel, where the channel statistics of all users are identical. More precisely, we assume $\mathbf{h}_{k,u} \sim \mathbb{CN}(0,1)$.

The BS has a dictionary of N files, namely $\{W_1, W_2, \ldots, W_N\}$, each of size F bits. Each user is interested in one of the files, chosen uniformly at random. In a cache aided communication system, each user $u \in [U]$ is equipped with a memory Z_u that can store up to MNF/U bits. That is, M copies of the entire dictionary of the files can be distributedly stored across the users. Cache *placement*, the process of filling the storage of the users with partial information from the dictionary, takes place before the users' demands are revealed. Later, each user u requests one file W_{d_u} from the dictionary, and the BS starts serving the users during the *delivery phase*. At the end of the delivery phase each user u should be able to decode W_{d_u} from Z_u and the signal received from the BS.

III. The Proposed Scheme

Let us define $\alpha = \frac{M+L}{M+1}$. Throughout this paper we assume that α is an integer. The main idea is to transmit α coded messages simultaneously at any given time, where each message serves a group of $M+1$ users. This approach gives a significant reduction in complexity compared to the state-of-the-art approach, which requires simultaneous transmission of much more messages at any given time. In the following we give a detailed description of the scheme. A short comparison to the state-of-the-art-scheme is given in the following section.

The placement strategy of this section is similar to that of [1]. To this end, we first split each file into $\binom{U}{M}$ segments, and label them with subsets of $[U]$ of size M. That is,

$$W_i = \{W_{i,\mathcal{S}} : \mathcal{S} \subseteq [U], |\mathcal{S}| = M\}, \quad i \in [N]. \tag{2}$$

The cache of user $u \in [U]$ will be filled by all file segments whose label contain u. More precisely,

$$Z_u = \{W_{i,\mathcal{S}} : u \in \mathcal{S}, i \in [N]\}. \tag{3}$$

This implies

$$|Z_u| = N\binom{U-1}{M-1}\frac{F}{\binom{U}{M}} = NMF/U, \tag{4}$$

which shows the cache size constraint is satisfied.

Given the fact that we serve α groups, each of size $M+1$, at any given time, we need a total of $T = \lceil \frac{1}{\alpha}\binom{U}{M+1} \rceil = \lceil \frac{M+1}{M+L}\binom{U}{M+1} \rceil$ time blocks to complete serving all users. Consider all subsets $\mathcal{T} \subseteq [U]$ of size $M+1$. In each time block m we serve α of such subsets of users, say $\mathcal{T}_1[m], \mathcal{T}_2[m], \ldots, \mathcal{T}_\alpha[m]$, such that

$$\mathcal{T}_i[m] \cap \mathcal{T}_j[m] = \varnothing, \qquad 1 \le i < j \le \alpha, m = 1, \ldots, T. \tag{5}$$

The set of users to be served at time block m is given by $\mathcal{U}[m] = \bigcup_{i=1}^{\alpha} \mathcal{T}_i[m]$, and since subsets $\mathcal{T}_j[m]$'s are pairwise disjoint, we have $|\mathcal{U}[m]| = \alpha(M+1) = M+L$, i.e, a total of $\alpha(M+1) = M+L$ users will be served in each time slot. To this end, we first generate a coded message $W_{\mathcal{T}}$ for each subset of users:

$$W_{\mathcal{T}} = \bigoplus_{u \in \mathcal{T}} W_{d_u, \mathcal{T}\setminus\{u\}}. \tag{6}$$

Then, this coded message will be modulated to a codeword $\overline{\mathcal{W}}_{\mathcal{T}}$ of length τK. We further split this codeword into K chunks, each of length τ, and send each chunk in one frequency bin. Let us denote the chunk of $\overline{\mathcal{W}}_{\mathcal{T}}$ corresponding to the frequency bin k by $\overline{\mathcal{W}}_{k,\mathcal{T}}$. The transmit signal for time block m will be then formed by beamforming of $\overline{\mathcal{W}}_{k,\mathcal{T}}$ for all the active subsets $\mathcal{T}$ in time block m. The appropriate beamforming will guarantee that the signal corresponding to a coded message $W_{\mathcal{T}}$ is zero-forced at all the users except those in $\mathcal{T}$. More precisely, we have

$$\mathbf{X}_k[m] = \sum_{i=1}^{\alpha} \sqrt{\frac{P}{\alpha K}} \mathbf{h}^{\perp}_{k,\mathcal{U}[m]\setminus\mathcal{T}_i[m]} \overline{\mathcal{W}}_{k,\mathcal{T}_i[m]}, \tag{7}$$

where $\mathbf{X}_k[m] \in \mathbb{C}^{L\times\tau}$ is given by

$$\mathbf{X}_k[m] = \begin{bmatrix} \mathbf{x}_k((m-1)\tau+1) & \cdots & \mathbf{x}_k(m\tau) \end{bmatrix} \tag{8}$$

and denotes the sequence of transmit signals $\mathbf{x}_k(t)$'s which are sent during the m-th time block at frequency bin k. Note that $\mathbf{h}^{\perp}_{k,\mathcal{A}} \in \mathbb{C}^{L\times 1}$ is a unit length beamforming vector which is orthogonal to the channel vectors of all users in $\mathcal{A}$ in frequency bin k. That is,

$$\begin{aligned} \mathbf{h}_{k,u}\mathbf{h}^{\perp}_{k,\mathcal{A}} &= \mathbf{0}, \qquad \forall u \in \mathcal{A}, k \in [K] \\ \|\mathbf{h}^{\perp}_{k,\mathcal{A}}\| &= 1. \end{aligned} \tag{9}$$

Note that $|\mathcal{U}[m] \setminus \mathcal{T}_i[m]| = (\alpha-1)(M+1) = \frac{L-1}{M+1}(M+1) = L-1$, and hence there is unique direction for the vector $\mathbf{h}^{\perp}_{\mathcal{U}[m]\setminus\mathcal{T}_i[m]}$.

Note that the optimal performance would require optimization of the power allocated to each frequency bin of each user, subject to the BS average power constraint. However, we assume the power is equally allocated between the frequency bins and the data streams.

The received vector at user $u \in \mathcal{T}_i[m] \subset \mathcal{U}[m]$ is given by

$$\begin{aligned} \mathbf{y}_{k,u}[m] &= \mathbf{h}_{k,u}\mathbf{X}_k[m] + \mathbf{z}_{k,u}[m] \\ &= \sqrt{\frac{P}{\alpha K}} \mathbf{h}_{k,u}\mathbf{h}^{\perp}_{k,\mathcal{U}[m]\setminus\mathcal{T}_i[m]} \mathcal{W}_{\mathcal{T}_i[m]} + \mathbf{z}_{k,u}[m], \end{aligned} \tag{10}$$

where the second equality is due to the fact that $\mathbf{h}_{k,u}$ is orthogonal to all $\mathbf{h}^{\perp}_{k,\mathcal{U}[m]\setminus\mathcal{T}_j[m]}$ for all $j \neq i$. Note that $\mathbf{y}_{k,u}[m] \in \mathbb{C}^{1\times\tau}$ and $\mathbf{z}_{k,u}[m] \in \mathbb{C}^{1\times\tau}$ denote the vectors of received signal and the additive Gaussian noise during time block m, respectively.

Upon receiving $\mathbf{y}_{k,u}[m]$ for all $k \in [K]$, user u has to decode the coded message $W_{\mathcal{T}_i[m]} = \bigoplus_{v\in\mathcal{T}_i[m]} W_{d_v,\mathcal{T}_i[m]\setminus\{v\}}$. Note that for every $v \neq u$ we have $u \in \mathcal{T}_i[m] \setminus \{v\}$, and hence a copy of $W_{d_v,\mathcal{T}_i[m]\setminus\{v\}}$ is stored in the cache of user u. Therefore, it can remove all the interfering segments and retrieve its desired file segment $W_{d_u,\mathcal{T}_i[m]\setminus\{u\}}$.

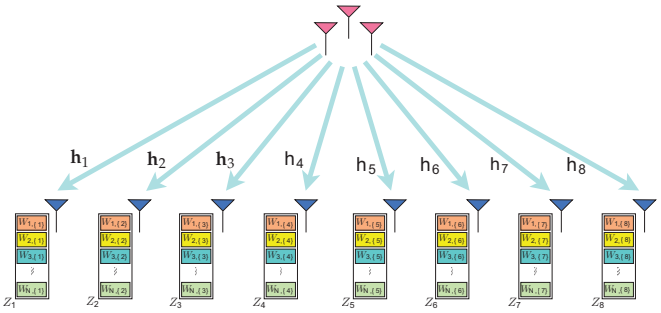

Fig. 1. A homogeneous MISO cache-aided communication system with $U = 8$ users and $L = 3$ transmit antennas. The normalized aggregate cache size is $M = 8$, so each user can store $M/U = 1/8$ of each file.

It is worth noting that since $\mathcal{T}_i[m]$ and $\mathcal{T}_j[m]$ are disjoint, each user decodes at most one message per time block. Moreover, following the transmission scheme explained above for all time blocks $m \in [T]$, each user u will retrieve all its file segments $W_{d_u,\mathcal{S}}$ which are not cached in its memory, i.e., $u \notin \mathcal{S}$. Concatenating these segments with those in the cache, user u will be able to decode its desired file. The overall delay of service for all users in T time blocks, each of duration τ will be $T\tau$. We evaluate this delay in Section V and compare it against that of the state-of-the-art scheme in the literature.

The following example demonstrates the delivery scheme explained in this section.

Example 1. Consider the wireless network depicted in Fig. 1, in which $U = 8$ users are served by a base station which is equipped with $L = 3$ antennas. Each user has a memory in which it can cache $M/U = 1/8$ of each file. Note that we have $\alpha = \frac{L+M}{M+1} = \frac{4}{2} = 2$, which is an integer. In the placement phase we divide each file W_i into $\binom{U}{M} = 8$ equal size segments, and label them as $W_{i,\{1\}}, W_{i,\{2\}}, \ldots, W_{i,\{8\}}$. Then user $u \in \{1,\ldots,8\}$ will cache all the file segments whose includes u, that is, $C_u = \{W_{i,\{u\}} : i = 1,\ldots,N\}$.

Without loss of generality, let us assume User u requests file W_u, i.e., $d_u = u$ for $u = 1,\ldots,8$. In the delivery phase, we need to serve a $\binom{U}{M+1} = \binom{8}{2} = 28$ groups (pairs) of users. In each time block we can serve $\alpha = 2$ groups, and hence we need $28/2 = 14$ blocks to complete the communication. To this end, we choose

$$\begin{array}{ll} \mathcal{T}_1(1) = \{1,2\} & \mathcal{T}_2(1) = \{3,4\} \\ \mathcal{T}_1(2) = \{1,3\} & \mathcal{T}_2(2) = \{5,6\} \\ \mathcal{T}_1(3) = \{1,4\} & \mathcal{T}_2(3) = \{7,8\} \\ \mathcal{T}_1(4) = \{1,5\} & \mathcal{T}_2(4) = \{2,3\} \\ \mathcal{T}_1(5) = \{1,6\} & \mathcal{T}_2(5) = \{4,5\} \\ \mathcal{T}_1(6) = \{1,7\} & \mathcal{T}_2(6) = \{6,8\} \\ \mathcal{T}_1(7) = \{1,8\} & \mathcal{T}_2(7) = \{3,5\} \\ \mathcal{T}_1(8) = \{2,4\} & \mathcal{T}_2(8) = \{5,7\} \\ \mathcal{T}_1(9) = \{2,5\} & \mathcal{T}_2(9) = \{6,7\} \\ \mathcal{T}_1(10) = \{2,6\} & \mathcal{T}_2(10) = \{4,8\} \\ \mathcal{T}_1(11) = \{2,7\} & \mathcal{T}_2(11) = \{3,8\} \end{array}$$

$$\mathcal{T}_1(12)=\{2,8\} \qquad \mathcal{T}_2(12)=\{4,6\}$$
$$\mathcal{T}_1(13)=\{3,6\} \qquad \mathcal{T}_2(13)=\{4,7\}$$
$$\mathcal{T}_1(14)=\{3,7\} \qquad \mathcal{T}_2(14)=\{5,8\}.$$

Consider a block, say $m=3$, during which Users in $\mathcal{U}(3)=\mathcal{T}_1(3)\cup\mathcal{T}_2(3)=\{1,4,7,8\}$ will be served. The base station first computes

$$W_{\{1,4\}}=W_{1,\{4\}}\oplus W_{4,\{1\}}$$
$$W_{\{7,8\}}=W_{7,\{8\}}\oplus W_{8,\{7\}},$$

and encodes them to $\overline{\mathcal{W}}_{\{1,4\}}$ and $\overline{\mathcal{W}}_{\{7,8\}}$, respectively. Each codeword will be divided into K sub-codewords, e.g.,

$$\overline{\mathcal{W}}_{\{1,4\}}=\left[\overline{\mathcal{W}}_{1,\{1,4\}} \quad \overline{\mathcal{W}}_{2,\{1,4\}} \quad \cdots \quad \overline{\mathcal{W}}_{K,\{1,4\}}\right].$$

Then we send a linear combination of the associated sub-codewords over each frequency band. The transmit signal in frequency bin k in time block $m=3$ is given by

$$\mathbf{X}_k[3]=\sqrt{\frac{P}{2}}\left(\mathbf{h}^{\perp}_{k,\{7,8\}}\overline{\mathcal{W}}_{k,\{1,4\}}+\mathbf{h}^{\perp}_{k,\{1,4\}}\overline{\mathcal{W}}_{k,\{7,8\}}\right).$$

Then the received signal at user 4 will be

$$\begin{aligned}\mathbf{y}_{k,4}[3]&=\sqrt{\frac{P}{2}}\left(\mathbf{h}_{k,4}\mathbf{h}^{\perp}_{k,\{7,8\}}\overline{\mathcal{W}}_{k,\{1,4\}}+\mathbf{h}_{k,4}\mathbf{h}^{\perp}_{k,\{1,4\}}\overline{\mathcal{W}}_{k,\{7,8\}}\right)\\&\quad+\mathbf{z}_{k,4}[3]\\&=\sqrt{\frac{P}{2}}\mathbf{h}_{k,4}\mathbf{h}^{\perp}_{k,\{7,8\}}\overline{\mathcal{W}}_{k,\{1,4\}}+\mathbf{z}_{k,4}[3].\end{aligned}\tag{11}$$

Collecting $\{\mathbf{y}_{k,4}[3]\}_{k=1}^{K}$, user 4 will decode $W_{\{1,4\}}=W_{1,\{4\}}\oplus W_{4,\{1\}}$. Then it can subtract $W_{1,\{4\}}$ using its cache content, and obtain $W_{4,\{1\}}$. All other segments $W_{4,\{v\}}$ can be decoded when $\{4,v\}$ is an active group for transmission. Finally user 4 can decode W_4 from delivered segments $\{W_{4,\{v\}}\}_{v\neq 4}$ and the segment $W_{4,\{4\}}$ which is cached in its memory. All other users will be served in a similar manner. $\diamond$

We present a larger example in the Appendix to further describe the scheme.

IV. The SCK Scheme

This section is dedicated to reviewing the communication scheme proposed by Shariatpanahi, Caire, and Khalaj in [2], which we refer to as the SCK scheme. The labeling and placement phase of the SCK schemes is identical to those explained above. After the placement phase is completed, the users reveal their requests. As before, and without loss of generality, we assume User u requests file W_u, i.e., $d_u=u$, for $u\in[U]$. Before the delivery phase gets started, the file segments in the cache of the users are further split into $\binom{U-M-1}{L-1}$ subsegments of equal size. The new subsegments will be labeled according to the revealed requests. Consider a file segment $W_{u,\mathcal{S}}$ which is cached by all users v with $v\in\mathcal{S}$, and requested by user u (since $d_u=u$). Each subsegment of this file segment will be labeled by a subset $\mathcal{A}\subseteq[U]\backslash(\mathcal{S}\cup\{u\})$ of size $|\mathcal{A}|=L-1$. More precisely, we have

$$W_{u,\mathcal{S}}=\left\{W^{\mathcal{A}}_{u,\mathcal{S}}:\mathcal{A}\subseteq[U]\setminus(\mathcal{S}\cup\{u\}),|\mathcal{A}|=L-1\right\},\tag{12}$$

for every $u,\mathcal{S}$ satisfying $u\notin\mathcal{S}$. Recall that the size each file segment is $F/\binom{U}{M}$, and hence, the size of each file subsegment is $F/\binom{U}{M}\binom{U-M-1}{L-1}$.

In the delivery phase, similar to our proposed scheme, the SCK scheme serves a total of $M+L$ users in each time block. However, the delivery phase here will be completed in $T^{\mathsf{SCK}}=\binom{U}{M+L}$ blocks, each associated with one subset $\mathcal{Q}\subset[U]$ of $|\mathcal{Q}|=M+L$. Indeed, $\mathcal{Q}$ identifies the set of users to be served in the corresponding time block.

During a time block $\mathcal{Q}$ one coded file subsegment will be sent for *every* subset $\mathcal{V}$ of $\mathcal{Q}$ with $|\mathcal{V}|=M+1$. More precisely, we define

$$W^{\mathcal{Q}}_{\mathcal{V}}=\bigoplus_{u\in\mathcal{V}}W^{\mathcal{Q}\backslash\mathcal{V}}_{u,\mathcal{V}\backslash\{u\}},\qquad \mathcal{V}\subseteq\mathcal{Q},|\mathcal{V}|=M+1.\tag{13}$$

Similar to Section III, a coded subsegments $W^{\mathcal{Q}}_{\mathcal{V}}$ will be encoded by a channel code to a codeword $\mathcal{W}^{\mathcal{Q}}_{\mathcal{V}}$ of length $K\tau^{\mathsf{SCK}}$, and then the codeword will be divided into K chunks $\overline{\mathcal{W}}^{\mathcal{Q}}_{k,\mathcal{V}}$ for $k\in[K]$, each of length τ^{SCK} to be modulated and sent in frequency bin k. The transmission block in time block $\mathcal{Q}$ and frequency bin k will be then formed as

$$\mathbf{X}_k[\mathcal{Q}]=\sum_{\substack{\mathcal{V}:\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\sqrt{\frac{P}{\binom{M+L}{M+1}}}\mathbf{h}^{\perp}_{k,\mathcal{Q}\backslash\mathcal{V}}\overline{\mathcal{W}}^{\mathcal{Q}}_{k,\mathcal{V}}.\tag{14}$$

Let us focus on user $u\in\mathcal{Q}$ and analyze its received signal during time block $\mathcal{Q}$. We have

$$\begin{aligned}\mathbf{y}_{k,u}[\mathcal{Q}]&=\mathbf{h}_{k,u}\mathbf{X}_k[\mathcal{Q}]+\mathbf{z}_{k,u}[\mathcal{Q}]\\&=\mathbf{h}_{k,u}\sum_{\substack{\mathcal{V}:\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\sqrt{\frac{P}{\binom{M+L}{M+1}}}\mathbf{h}^{\perp}_{k,\mathcal{Q}\backslash\mathcal{V}}\overline{\mathcal{W}}^{\mathcal{Q}}_{k,\mathcal{V}}+\mathbf{z}_{k,u}[\mathcal{Q}]\\&\overset{(a)}{=}\sum_{\substack{\mathcal{V}:u\in\mathcal{V}\\\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\sqrt{\frac{P}{\binom{M+L}{M+1}}}\mathbf{h}_{k,u}\mathbf{h}^{\perp}_{k,\mathcal{Q}\backslash\mathcal{V}}\overline{\mathcal{W}}^{\mathcal{Q}}_{k,\mathcal{V}}+\mathbf{z}_{k,u}[\mathcal{Q}].\end{aligned}\tag{15}$$

Note that (a) follows the fact that if $\mathcal{V}$ does not include u, then $u\in\mathcal{Q}\setminus\mathcal{V}$, and hence (9) implies that $\mathbf{h}_u\mathbf{h}^{\perp}_{\mathcal{Q}\backslash\mathcal{V}}=\mathbf{0}$.

Upon receiving $\{\mathbf{y}_{k,u}[\mathcal{Q}]\}_{k=1}^{K}$, user u decodes all coded subsegments intended for user u in block $\mathcal{Q}$, that is,

$$\left\{W^{\mathcal{Q}}_{\mathcal{V}}:u\in\mathcal{V}\subseteq\mathcal{Q}\right\}.\tag{16}$$

Once a coded subsegments $W^{\mathcal{Q}}_{\mathcal{V}}$ is obtained by user u, it can retrieve $W^{\mathcal{Q}\backslash(\mathcal{V}\cup\{u\})}_{u,\mathcal{V}\backslash\{v\}}$ by subtracting all interfering subsegments from its cache. This leads to a set of decoded subsegments given by

$$\begin{aligned}&\bigcup_{\substack{\mathcal{Q}\subseteq[U]\\|\mathcal{Q}|=M+L}}\bigcup_{\substack{\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\left\{W^{\mathcal{Q}\backslash\mathcal{V}}_{u,\mathcal{V}\backslash\{v\}}:u\in\mathcal{V}\right\}\\&=\bigcup_{\substack{\mathcal{S}\subseteq[U]\\|\mathcal{S}|=M\\u\notin\mathcal{S}}}\bigcup_{\substack{\mathcal{A}\subseteq[U]\\|\mathcal{Q}|=L-1\\\mathcal{A}\cap(\mathcal{S}\cup\{u\})=\varnothing}}\{W^{\mathcal{A}}_{u,\mathcal{S}}\}=\bigcup_{\substack{\mathcal{S}\subseteq[U]\\|\mathcal{S}|=M\\u\notin\mathcal{S}}}\{W_{u,\mathcal{S}}\},\end{aligned}$$

which together with $\{W_{u,\mathcal{S}}:u\in\mathcal{S}\}\subseteq Z_u$ can fully recover W_u, which is the desired file of user u.

V. Performance Comparison

In this section we analyze the performance of the two schemes presented in Sections III and IV, and compare them. The main factors of interest in our comparison are the overall delay of delivery, the complexity, and the subpacketization level, which is the number of pieces each file need to be divided into. The comparison is also summarized in Table I below.

A. The Subpacketization Level

Recall that in the scheme proposed in this paper, each file is split into $\binom{U}{M}$ segments. The delivery phase is solely based on combining packets of size $F/\binom{U}{M}$, encoding them using error correction codes, and modulating and sending them from the BS. In the SCK scheme, even though each file is only divided into $\binom{U}{M}$ segments in the placement phase, further splitting is required for the delivery phase. More precisely, each segment will be split into $\binom{U-M-1}{L-1}$ subsegments. Therefore, the overall subpacketization level is $\binom{U}{M}\binom{U-M-1}{L-1}$, which is substantially larger than that of the scheme proposed in this paper.

On the other hand, the number of codewords broadcasted by the proposed scheme is $\binom{U}{M+L}$, while this number is $\binom{U}{M+L}\binom{M+L}{M+1}$ for the proposed scheme.

B. The Overall Delay

In the scheme proposed in this paper, in order for user u to be able to decode $W_{\mathcal{T}_i[m]}$, the rate of information sent through the channel should not exceed the rate supported by that channel. More precisely, we need

$$\frac{1}{\tau}\frac{F}{\binom{U}{M}} \leq \sum_{k=1}^{K}\frac{B}{K}\log_2\left(1+\frac{P}{\alpha B}\left|\mathbf{h}_{k,u}\mathbf{h}^{\perp}_{\mathcal{U}[m]\setminus\mathcal{T}_i[m],k}\right|^2\right). \tag{17}$$

Let us define $\eta_{k,u,m} = \left|\mathbf{h}_{k,u}\mathbf{h}^{\perp}_{k,\mathcal{U}[m]\setminus\mathcal{T}_i[m]}\right|^2$. Due to the Gaussian distribution of $\mathbf{h}_{k,u}$ and the unit norm of $\mathbf{h}^{\perp}_{k,\mathcal{T}}$ for all $\mathcal{T}$'s, the random variables $\eta_{k,u,m}$'s admit an exponential distribution with a mean of 1, for every frequency band $k\in[K]$, every user $u\in[U]$, and every time block $m\in[T]$. We also define

$$R = B\mathbb{E}\left[\log_2\left(1+\frac{P}{\alpha B}\eta_{k,u,m}\right)\right]. \tag{18}$$

Then, as K grows, we have

$$\frac{B}{K}\sum_{k=1}^{K}\log_2\left(1+\frac{P}{\alpha B}\eta_{k,u,m}\right)\to R. \tag{19}$$

Therefore, the decodability constraint in (17) reduces to

$$\tau \geq \frac{F}{\binom{U}{M}R}. \tag{20}$$

The number[1] of time blocks is $T=\frac{M+1}{M+L}\binom{U}{M+1}$. Therefore, the overall delay will be

$$D = T\tau = \frac{M+1}{M+L}\frac{\binom{U}{M+1}}{\binom{U}{M}}\frac{F}{R} = \frac{U-M}{M+L}\frac{F}{R}. \tag{21}$$

Now, let us consider the SCK scheme. It is evident that the received signal in (15) models a multiple access channel (MAC), in which user u should decode a total of $\binom{M+L-1}{M}$ messages simultaneously (including all $W^{\mathcal{Q}}_{\mathcal{V}}$'s for which $\mathcal{V}\subseteq\mathcal{Q}$ satisfies $u\in\mathcal{V}$ and $|\mathcal{V}|=M+1$). It is easy to verify that the number of messages user u should decode in a time block is $\binom{M+L-1}{M}$. The total signal power received at user u in frequency bin k is

$$\begin{aligned}
&\frac{P}{\binom{M+L}{M+1}K}\sum_{\substack{\mathcal{V}:u\in\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\left|\mathbf{h}_{k,u}\mathbf{h}^{\perp}_{k,\mathcal{Q}\setminus\mathcal{V}}\right|^2\\
&=\frac{\binom{M+L-1}{M}P}{\binom{M+L}{M+1}K}\frac{1}{\binom{M+L-1}{M}}\sum_{\substack{\mathcal{V}:u\in\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\eta_{k,u,\mathcal{Q}}\\
&=\frac{P}{\alpha K}\frac{1}{\binom{M+L-1}{M}}\sum_{\substack{\mathcal{V}:u\in\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\eta_{k,u,\mathcal{Q}}
\end{aligned}$$

where $\eta_{k,u,\mathcal{Q},\mathcal{V}} = \left|\mathbf{h}_{k,u}\mathbf{h}^{\perp}_{k,\mathcal{Q}\setminus\mathcal{V}}\right|^2$ admits an exponential distribution with parameter 1. Therefore, the rate of each coded message intended for u is upper bounded by the maximum *symmetric* decodable rate of the user, that is,

$$\begin{aligned}
&\frac{1}{\tau^{\mathsf{SCK}}}\frac{F}{\binom{U}{M}\binom{U-M-1}{L-1}}\\
&\quad\leq\frac{1}{\binom{M+L-1}{M}}\sum_{k=1}^{K}\frac{B}{K}\log\left(1+\frac{P}{\alpha B}\frac{\sum_{\mathcal{V}:u\in\mathcal{V}\subseteq\mathcal{Q}}\eta_{k,u,\mathcal{Q}}}{\binom{M+L-1}{M}}\right)\\
&\quad\to\frac{1}{\binom{M+L-1}{M}}R^{\mathsf{SCK}}.
\end{aligned} \tag{22}$$

where

$$R^{\mathsf{SCK}} = B\mathbb{E}\left[\log\left(1+\frac{P}{\alpha B}\frac{1}{\binom{M+L-1}{M}}\sum_{\substack{\mathcal{V}:u\in\mathcal{V}\subseteq\mathcal{Q}\\|\mathcal{V}|=M+1}}\eta_{k,u,\mathcal{Q}}\right)\right] \tag{23}$$

This implies that the length of each time block should be at least

$$\tau^{\mathsf{SCK}}\geq\frac{\binom{M+L-1}{M}}{\binom{U}{M}\binom{U-M-1}{L-1}}\frac{F}{R^{\mathsf{SCK}}}.$$

[1]We assume T is an integer for the sake of comparison simplicity.

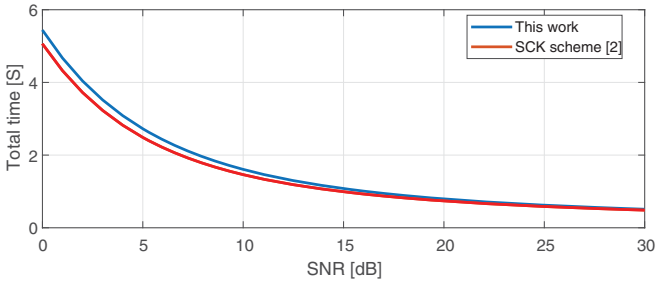

Fig. 2. Comparison of the (average) total delivery time for the proposed scheme vs. that of the SCK scheme [2].

Recall that there is a total of $\binom{U}{M+L}$ time blocks, the overall minimum delay is given by

$$D^{\mathsf{SCK}} = T^{\mathsf{SCK}} \tau^{\mathsf{SCK}} = \binom{U}{M+L} \frac{\binom{M+L-1}{M}}{\binom{U}{M}\binom{U-M-1}{L-1}} \frac{F}{R^{\mathsf{SCK}}} = \frac{U-M}{M+L} \frac{F}{R^{\mathsf{SCK}}} \tag{24}$$

Comparing (24) and (21), it is clear that both schemes achieve the same number of degrees of freedom. Nevertheless, due to the spatial diversity of the multiple access channel, the user rate achieved by the SCK scheme (i.e., R^{SCK}) is higher than that or ours.

Note that there are two levels of averaging to obtain the ergodic capacity, one is averaging over the achievable rates of different frequency bands, and the other is due to the randomness of η which is an exponential random variable. In the evaluation of the rate of the proposed scheme in (18), both expectations are taken outside of the logarithm. However, due to the multiple access effect, the rate of the SCK scheme in (23), we first take (empirical) average with respect to η inside the logarithm, and then averaging over k is taken outside. Due to the concavity of the $\log(\cdot)$ function, the latter is always greater than the the former. When the number of messages in the MAC, i.e. $\binom{M+L-1}{M}$, is large, we have

$$R^{\mathsf{SCK}} = B\mathbb{E}\left[\log\left(1 + \frac{P}{\alpha B}\frac{\|\mathbf{h}_{k,u}\|^2}{\binom{M+L-1}{M}} \sum_{\substack{\mathcal{V}: u\in\mathcal{V}\subseteq\mathcal{Q} \\ |\mathcal{V}|=M+1}} \frac{\eta_{k,u,\mathcal{Q}}}{\|\mathbf{h}_{k,u}\|^2}\right)\right] \to B\mathbb{E}\left[\log\left(1 + \frac{P}{\alpha B}\frac{\|\mathbf{h}_{k,u}\|^2}{L}\right)\right] \tag{25}$$

In an (hypothetical) extreme scenario that both averages are taken inside the logarithm, the achievable rate will be

$$R^{\mathsf{G}} = B\log\left(1 + \frac{P}{\alpha}\mathbb{E}_{k,\eta}\left[\eta_{k,u}\right]\right) = B\log\left(1 + \frac{P}{\alpha}\right)$$

which is the capacity of a Gaussian channel. We have $R \leq R^{\mathsf{SCK}} \leq R^{\mathsf{G}}$. A theoretical analysis for the gap between R and R^{G} is fairly standard, and skipped for the sake of brevity. Our numerical results in Fig. 2 shows that the SCK scheme offers a maximum of 1.5dB improvement compared to the proposed scheme. We leave more precise gap analysis between R and R^{SCK} for the extended version of this work.

	the SCK scheme [2]	the proposed scheme
Number of Time Slots	$\binom{U}{M+L}$	$\frac{M+1}{M+L}\binom{U}{M+1}$
Subpacketization	$\binom{U}{M}\binom{U-M-1}{L-1}$	$\binom{U}{M}$
Duration of Delivery	$\frac{U-M}{M+L}\frac{F}{R^{\mathsf{SCK}}}$	$\frac{U-M}{M+L}\frac{F}{R}$
Diversity Gain	≈ 1.5dB	0dB
Decoding	MAC decoding	Single message decoding
Parameters	all L, M	$M+1$ divides $M+L$

TABLE I
A GENERAL COMPARISON BETWEEN THE SCK SCHEME AND THE ONE PROPOSED IN THIS WORK.

Example 2. Recall the network of Example 1, with $U = 8$, $M = 1$, and $L = 3$. The subpacketization levels of the proposed scheme and the SCK scheme for this network are 8 and 120, respectively. Moreover, the course of communication is divided into only 14 blocks for the proposed scheme, while 70 blocks are needed for the SCK scheme; this yields to shorter codewords and larger overhead, which reduce the effective communication rate. Finally, in SCK scheme each user has to decode 3 received message over a MAC simultaneously in each active block, while only one message will be decoded at a time by the proposed scheme. Even though this leads to a diversity gain for the SCK scheme, it requires a more sophisticated decoding algorithm. ◇

VI. Conclusion

We studied the subpacketization problem for a cache-aided MISO communication system. For a system with parameters satisfying $\frac{M+L}{M+1} \in \mathbb{N}$, we proposed a scheme that achieves the optimum DoF, while its subpacketization level is identical to that of the single antenna case. The proposed scheme also offers a low-complexity decoding, since each user has to decode at most one message per transmission block.

References

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

[2] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna coded caching," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2113–2117.

[3] I. Bergel and S. Mohajer, "Cache aided communications with multiple antennas at finite SNR," *IEEE J. Sel. Areas Commun.*, 2018.

[4] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5821–5833, Sept 2017.

[5] S. Jin, Y. Cui, H. Liu, and G. Caire, "Uncoded placement optimization for coded delivery," in *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. IEEE, 2018, pp. 1–8.

[6] M. Salehi, A. Tölli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-rate trade-off in multi-antenna coded caching," *arXiv preprint arXiv:1905.04349*, 2019.

[7] K. Shanmugam, A. M. Tulino, and A. G. Dimakis, "Coded caching with linear subpacketization is possible using ruzsa-szeméredi graphs," in *IEEE ISIT*, 2017, pp. 1237–1241.

[8] L. Tang and A. Ramamoorthy, "Low subpacketization schemes for coded caching," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2790–2794.

[9] 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.

[10] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Transactions on Wireless Communications*, 2020.