# Multi-Antenna Coded Content Delivery with Caching: A Low-Complexity Solution

Junlin Zhao, *Member, IEEE,* Mohammad Mohammadi Amiri, *Member, IEEE,* and Deniz Gündüz, *Senior Member, IEEE*

***Abstract*—We study downlink beamforming in a single-cell network with a multi-antenna base station serving cache-enabled users. Assuming a library of files with a common rate, we formulate the minimum transmit power with proactive caching and coded delivery as a non-convex optimization problem. While this multiple multicast problem can be efficiently solved by successive convex approximation (SCA), the complexity of the problem grows exponentially with the number of subfiles delivered to each user in each time slot, which itself grows exponentially with the number of users. We introduce a low-complexity alternative through time-sharing that limits the number of subfiles received by a user in each time slot. We then consider the joint design of beamforming and content delivery with sparsity constraints to limit the number of subfiles received by a user in each time slot. Numerical simulations show that the low-complexity scheme has only a small performance gap to that obtained by solving the joint problem with sparsity constraints, and outperforms state-of-the-art results at all signal-to-noise ratio (SNR) and rate values with a sufficient number of transmit antennas. A lower bound on the achievable degrees-of-freedom (DoF) of the low-complexity scheme is derived to characterize its performance in the high SNR regime.**

## I. Introduction

The seminal work of Maddah-Ali and Niesen showed that by exploiting caches at the users in order to create and exploit multicasting opportunities we can reduce the delivery time, or equivalently increase the throughput in wireless networks [2]. With coded caching, uncoded contents can be proactively pushed at user devices without knowing users' demands, and a server can serve multiple users simultaneously by broadcasting specially designed coded combinations of the remaining parts of all the users' requests, to guarantee that all the users can recover their desired contents. This feature is particularly favorable in wireless medium due to its broadcast nature. However, [2] ignored the physical characteristics of the channel, and simply assumed error-free communication, and focused on minimizing the number of bits delivered error-free over this multicast channel.

Over the last years many follow-up works have studied coded delivery over noisy broadcast channels. When users may have different channel capacities, the user with the worst channel condition becomes the bottleneck limiting the performance of multicasting. The global caching gain promised in [2] is hence not straightforward in practice. Coded caching in erasure broadcast channels is studied in [3] and [4] by allocating cache memories at weak receivers to overcome this bottleneck. A simple binary Gaussian broadcast channel is considered in [5], and an interference enhancement scheme is used to overcome the limitation of weak users. A cache-aided multicasting strategy over a Gaussian broadcast channel is presented in [6], with superposition coding and power allocation. The authors in [7] consider fading channels, and show that a linear increase in the sum delivery rate with the number of users can be achieved with user selection.

Another important line of research has focused on evaluating the performance of coded caching and delivery in the presence of multiple transmit antennas at the server. Multicast beamforming, where the multiple-antenna base station (BS) multicasts distinct data streams to multiple user groups, is an efficient physical layer technique [8]–[10]. In [11], the authors extend the results in [7] to multi-input single-output (MISO) fading channels, where the same linear increase in content delivery rate with respect to (w.r.t.) the number of users is achieved without channel state information at the transmitter (CSIT), and an improvement is obtained with spatial multiplexing when CSIT is available. In [12], coded delivery is employed along with zero-forcing to simultaneously exploit spatial multiplexing and caching gains. With multiple antennas at the BS, coded messages can be nulled at unintended user groups, which increases the number of users simultaneously served as compared to the single antenna setting. Particularly, this approach was found to achieve the near-optimal degrees-of-freedom (DoF) in [13]. In addition to the gain in content delivery rate, employing multiple transmit antennas also allows reducing the subpacketization level required in coded caching [14], [15].

By treating the transmission of coded subfiles as a coordinated beamforming problem, improved spectral efficiency

Part of this work was presented at the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Cannes, France, Jul. 2019 [1].

This work was partially supported by the European Research Council (ERC) through project BEACON (No. 677854), and by the European Union's Horizon 2020 Research and Innovation Programme through project SCAVENGE (No. 675891).

J. Zhao was with the Information Processing and Communications Lab, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK. He is now with the School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen, Guangdong 518172, P.R. China, and also with the University of Science and Technology of China, Anhui 230026, P.R. China (e-mail: j.zhao15@imperial.ac.uk)

M. Amiri was with the Information Processing and Communications Lab, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK. He is now with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: mamiri@princeton.edu)

D. Gündüz is with the Information Processing and Communications Lab, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK (e-mail: d.gunduz@imperial.ac.uk)

is achieved in [16] by optimizing the beamforming vectors, which is also shown to achieve the same DoF as in [12] in special cases. Memory-sharing is proposed in [17] to apply the content placement scheme of [2] for a fraction of the library, which exploits both the spatial multiplexing gain and the global caching gain by sending a common message together with user-dependent messages. The impact of imperfect CSIT on achievable DoF is considered for MISO broadcast channels in [18]. Similarly to [17], [19] adopts memory-sharing, and proposes a joint unicast and multicast beamforming approach.

In this paper, motivated by the results in [12] and [16], we consider a cache-aided MISO broadcast channel. Firstly, a general framework for cache-aided downlink beamforming is formulated, focusing on the minimum required transmit power for delivering the contents at a prescribed common rate. The resultant nonconvex optimization problem is tackled by successive convex approximation (SCA), which is guaranteed to converge to a stationary solution of the original nonconvex problem. As noted in [16], the beamforming design involves solving an optimization problem with exponentially increasing number of constraints with the number of coded messages each user decodes in each time slot. To limit the complexity, we propose a novel content delivery scheme, in which the coded subfiles, each targeted at a different subset of receivers, are delivered over multiple orthogonal time slots, while the number of coded messages each user decodes in each time slot can be flexibly adjusted. Unlike the scheme in [16], the scheme we propose does not limit the number of users served in each time slot, but directly limits the number of messages each user decodes, and hence, the complexity of the decoder. We propose a greedy algorithm that decides the multicast messages to be delivered at each time slot, and the number of time slots. A lower bound on the DoF achieved by the greedy scheme is also provided. We then consider a more general design of the beamforming vectors together with the content delivery scheme with a constraint on the maximum number of messages each user can decode in any time slot. We formulate this joint optimization as a power minimization problem with sparsity constraints, and solve it via SCA to obtain a stationary solution. Our numerical results show that the proposed greedy scheme has only a small performance gap to that of the optimization-based delivery scheme, and provide significant gains over the one proposed in [16] in terms of transmit power, particularly in the high rate/high signal-to-noise ratio (SNR) regime.

The remainder of the paper is organized as follows. Section II introduces the system model. In Section III, we present an achievable coded delivery scheme, and formulate the power minimization problem for a multi-antenna server. We introduce a low-complexity content delivery scheme in Section IV, and present its DoF analysis in Section V. In Section VI, we consider the joint design of beamforming and coded content delivery. Finally, we compare the proposed schemes with the state-of-the-art through numerical simulations in Section VII, and conclude the paper in Section VIII.

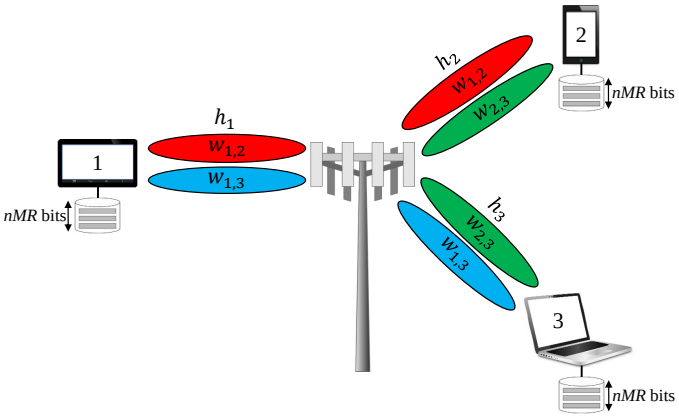



Figure 1: Illustration of a cache-aided MISO channel with $K = 3$ users. Multi-antenna BS employs multicast beamforming to deliver the missing parts of users' requests.

## II. System Model

We consider downlink transmission within a single cell, where a BS equipped with $N_T$ antennas serves $K$ single-antenna cache-equipped users, as illustrated in Fig. 1. We consider a library of $N$ files, denoted by $V \triangleq (V_1, \ldots, V_N)$, each distributed uniformly over the set $[2^{nR}]$[1], available at the BS, where $R$ and $n$ represent the rate of each file and the blocklength, respectively. Each user is equipped with a local cache that can store up to $M$ files, and the corresponding *caching factor*, $t$, is defined as the ratio of the total cache capacity across all the receivers to the library size, $t \triangleq MK/N$.

Contents are placed at users' caches during off-peak periods without any prior information on the user requests or the channel state information (CSI) to be experienced during the delivery phase. Caching function for user $k$ is denoted by $\phi_k^{(n)} : [2^{nR}]^N \to [2^{nMR}]$, which maps the library to the cache contents $Z_k$ at user $k$, i.e., $Z_k = \phi_k^{(n)}(V)$, $k \in [K]$. Once the users reveal their demands $d \triangleq (d_1, \ldots, d_K)$, where $d_k \in [N], \forall k \in [K]$, signal $X \in \mathbb{C}^{N_T \times n}$ is transmitted, where $X \triangleq [x_1 \ \ldots \ x_n]$, and $x_i \in \mathbb{C}^{N_T}$, $i = 1, \ldots, n$. An average power constraint $P$ is imposed on each channel input $X$. User $k$ receives

$$y_k^T = h_k^H X + n_k^T, \tag{1}$$

where $h_k \in \mathbb{C}^{N_T}$ is the channel vector from the BS to the $k$-th user, and $n_k \in \mathbb{C}^n$ is the additive white Gaussian noise at user $k$ with each entry independent and identically (i.i.d.) distributed according to $\mathcal{CN}(0, \sigma_k^2)$, $k \in [K]$. We assume that the CSI is perfectly known to the BS and the receivers in the delivery phase. Hence, the encoding function at the BS, $\psi^{(n)} : [2^{nR}]^N \times [N]^K \times \mathbb{C}^{N_T \times K} \to \mathbb{C}^{N_T \times n}$, maps the library, the demand vector, and the CSI to the channel input vector. We note here that, while the channel encoding function $\psi^{(n)}$ depends on the demand vector and the CSI, caching functions $\phi_k^{(n)}$ depend only on the library. After receiving $y_k$, user $k$ reconstructs $\hat{V}_{d_k}$ using its local cache

[1]For any positive real $X$, we define $[X]$ as the set of positive integers less than or equal to X.

content $Z_k$, channel vector $\mathbf{h}_k$, and demand vector $\mathbf{d}$ through function $\phi_k^{(n)} : \mathbb{C}^n \times \{1,\dots,2^{nMR}\} \times \mathbb{C}^{N_T \times 1} \times [N]^K \to \{1,\dots,2^{nR}\}$, i.e., $\hat{V}_k = \phi_k^{(n)}(Y_k, Z_k, \mathbf{h}_k, \mathbf{d})$, $k \in [K]$. The probability of error is defined as $P_e \triangleq \max_{\mathbf{d}} \max_{k\in[K]} \Pr\{V_{d_k} \neq \hat{V}_k\}$. An $(R, M, P)$ tuple is *achievable* if there exist a sequence of caching functions $\phi_1^{(n)}, \dots, \phi_K^{(n)}$, encoding function $\psi^{(n)}$, and decoding functions $\mu_1^{(n)}, \dots, \mu_K^{(n)}$, such that $P_e \to 0$ as $n \to \infty$. For file rate $R$ and cache size $M$, our goal is to characterize

$$P^*(R, M) = \inf\{P : (R, M, P) \text{ is achievable}\}; \quad (2)$$

which characterizes the minimum required transmit power that guarantees the reliable delivery of any demand vector.

## III. An Achievable Delivery Scheme

In this section, we present a multi-antenna transmission scheme with coded caching, where the cache placement and coded content generation follows [2], while beamforming is employed at the BS to multicast coded subfiles to the receivers.

### A. Placement and Delivery Schemes

For a caching factor $t \in \{1,\dots,K-1\}$, we represent $t$-element subsets of $[K]$ by $G_1^t, \dots, G_{\binom{K}{t}}^t$. File $V_i$, $i \in [N]$, is divided equally into $\binom{K}{t}$ disjoint subfiles $V_{i,G_1^t}, \dots, V_{i,G_{\binom{K}{t}}^t}$, each consisting of $n\frac{R}{\binom{K}{t}}$ bits. User $k$, $k \in [K]$, caches subfile $V_{i,G_j^t}$, if $k \in G_j^t$, $\forall j \in [1 : \binom{K}{t}]$. The cache content of user $k$ is then given by $\bigcup_{i\in[N]} \bigcup_{j\in[\binom{K}{t}]:k\in G_j^t} V_{i,G_j^t}$.

During the *delivery phase*, for any demand combination $\mathbf{d}$, we aim to deliver the coded message

$$s_{G_j^{t+1}} = \bigoplus_{k\in G_j^{t+1}} V_{d_k, G_j^{t+1}\setminus\{k\}}, \quad (3)$$

to all the users in set $G_j^{t+1}$, for $j \in [\binom{K}{t+1}]$. Observe that, after receiving $s_{G_j^{t+1}}$, each user $k \in G_j^{t+1}$ can recover subfile $V_{d_k, G_j^{t+1}\setminus\{k\}}$ having access to $V_{d_l, G_j^{t+1}\setminus\{l\}}$, $\forall l \in G_j^{t+1}\setminus\{k\}$.

We define $\mathcal{S} \triangleq \{G_1^{t+1}, \dots, G_{\binom{K}{t+1}}^{t+1}\}$ as the set of all the multicast messages, with each message $T \in \mathcal{S}$ represented by the set of users it is targeting, and let $\mathcal{S}_k \subseteq \mathcal{S}$ denote the subset of messages targeting user $k$. We have $|\mathcal{S}| = \binom{K}{t+1}$ and $|\mathcal{S}_k| = \binom{K-1}{t}$.

The following examples will be used to better explain the proposed scheme:

***Example 1***: Let $N = 5$, $K = 5$, $M = 1$. We have $t = \frac{MK}{N} = 1$. Each file is split into $\binom{K}{t} = 5$ disjoint subfiles of the same size, where we represent file $i$, $i \in [N]$, as

$$V_i = \{V_{i,\{1\}}, V_{i,\{2\}}, V_{i,\{3\}} V_{i,\{4\}}, V_{i,\{5\}}\}. \quad (4)$$

The cache content of user $k$ is $Z_k = \bigcup_{i\in[N]} V_{i,\{k\}}$, $k \in [K]$, which satisfies the cache capacity constraint. All user demands can be fulfilled by delivering the following $\binom{K}{t+1} = 10$ subfiles:

$$\begin{aligned}
&s_{\{1,2\}} = V_{d_1,\{2\}} \oplus V_{d_2,\{1\}}, \; s_{\{1,3\}} = V_{d_1,\{3\}} \oplus V_{d_3,\{1\}},\\
&s_{\{1,4\}} = V_{d_1,\{4\}} \oplus V_{d_4,\{1\}}, \; s_{\{1,5\}} = V_{d_1,\{5\}} \oplus V_{d_5,\{1\}},\\
&s_{\{2,3\}} = V_{d_2,\{3\}} \oplus V_{d_3,\{2\}}, \; s_{\{2,4\}} = V_{d_2,\{4\}} \oplus V_{d_4,\{2\}},\\
&s_{\{2,5\}} = V_{d_2,\{5\}} \oplus V_{d_5,\{2\}}, \; s_{\{3,4\}} = V_{d_3,\{4\}} \oplus V_{d_4,\{3\}},\\
&s_{\{3,5\}} = V_{d_3,\{5\}} \oplus V_{d_5,\{3\}}, \; s_{\{4,5\}} = V_{d_4,\{5\}} \oplus V_{d_5,\{4\}}.
\end{aligned}$$

***Example 2***: Let $N = 4$, $K = 4$, $M = 1$. We have $t = \frac{MK}{N} = 1$. Each file is split into $\binom{K}{t} = 4$ disjoint subfiles of the same size. All user demands can be fulfilled by delivering the following $\binom{K}{t+1} = 6$ subfiles:

$$s_{\{1,2\}}, \; s_{\{1,3\}}, \; s_{\{1,4\}}, \; s_{\{2,3\}}, \; s_{\{2,4\}}, \; s_{\{3,4\}}. \quad (5)$$

Note that the message $s_T$ is intended for users in set $T$, but interferes with users in set $[K]\setminus T$. Moreover, for any demand combination $\mathbf{d}$, all the users are required to decode the same number of messages, which is $\binom{K-1}{t}$.

### B. Multi-Antenna Transmission Scheme

The delivery of the coded messages in set $\mathcal{S}$ to their respective receivers is a multi-antenna multi-message multicasting problem. Before introducing our low-complexity scheme in the next section, we present here a general transmission strategy based on message-splitting and time-division transmission. The messages in $\mathcal{S}$ can be transmitted over $B$ orthogonal time slots, the $i$-th of which is of blocklength $n_i$, $i \in [B]$, where $\sum_{i=1}^{B} n_i = n$. The transmitted signal $X(i) \triangleq [x_{\sum_{j=1}^{i-1} n_j + 1} \; \dots \; x_{\sum_{j=1}^{i} n_j}]$ at time slot $i \in [B]$ is given by

$$X(i) = \sum_{T\in\mathcal{S}} \mathbf{w}_T(i) s_T^T(i), \quad (6)$$

where $s_T(i) \in \mathbb{C}^{n_i \times 1}$ is a unit-power complex Gaussian signal of blocklength $n_i$, modulated from the corresponding message $s_T$ in (3), intended for the users in set $T$, transmitted in time slot $i$, encoded by the beamforming vector $\mathbf{w}_T(i) \in \mathbb{C}^{N_T \times 1}$.

The received signal at user $k$ in time slot $i$ is

$$Y_k^T(i) = \underbrace{\mathbf{h}_k^H \sum_{T\in\mathcal{S}_k} \mathbf{w}_T(i) s_T^T(i)}_{\text{desired messages}} + \underbrace{\mathbf{h}_k^H \sum_{I\in\mathcal{S}_k^C} \mathbf{w}_I(i) s_I^T(i)}_{\text{interference}} + n_k^T(i), \quad (7)$$

where $\mathcal{S}_k^C$ is the complement of set $\mathcal{S}_k$ in $\mathcal{S}$. Let $\mathcal{T}_{\mathcal{S}_k}$ denote the collection of all non-empty subsets of $\mathcal{S}_k$, with each element of $\mathcal{T}_{\mathcal{S}_k}$ denoted by $\mathcal{S}_k^j$, $j \in [2^{\binom{K-1}{t}} - 1]$. We denote $\mathcal{S}(i) \subseteq \mathcal{S}$ as the subset of messages transmitted in time slot $i$, i.e., $T \in \mathcal{S}(i)$ if $\mathbf{w}_T(i) \neq 0$.

Note that each user may receive more than one message in each transmission slot. From the capacity region of the associated Gaussian multiple access channel, following conditions must be satisfied for successful decoding of all the intended messages at user $k$, $k \in [K]$, at time slot $i$:

$$\sum_{T\in\mathcal{S}_k^j} R^T(i) \le \frac{n_i}{n}\log_2\left(1 + \sum_{T\in\mathcal{S}_k^j} \gamma_k^T(i)\right), \; \forall \mathcal{S}_k^j \in \mathcal{T}_{\mathcal{S}_k}, \quad (8)$$

where $R^T(i)$ is the rate of message $s_T(i)$, and $\gamma_k^T(i)$ is the received signal-to-interference-plus-noise ratio (SINR) of message $s_T(i)$ at user $k$ at time slot $i$, given by

$$\gamma_k^T(i) = \frac{|\mathbf{h}_k^H \mathbf{w}_T(i)|^2}{\sum_{l \notin \mathcal{S}_k^c} |\mathbf{h}_k^H \mathbf{w}_l(i)|^2 + \sigma_k^2}; \tag{9}$$

for any $T \ni k$, or equivalently, any $T \in \mathcal{S}_k$. The rate of message $T$ is the sum of the rate of submessages $s_T(i)$, and must satisfy

$$\sum_{i=1}^{B} R^T(i) \le \frac{R}{K}, \ \forall T. \tag{10}$$

Note that this scheme is quite flexible; each multicast message can be split into $B$ messages and transmitted over $B$ time slots. It can be specialized to different content delivery schemes by specifying the subset of transmitted subfiles in each time slot and the blocklength of each time slot, i.e., $\{\mathcal{S}(i)\}_{i=1}^B$ and $\{n_i\}$. Let

$$v_T(i) = \begin{cases} 1 & \text{if } T \in \mathcal{S}(i) \\ 0 & \text{if } T \notin \mathcal{S}(i) \end{cases} \tag{11}$$

be the indicator function specifying whether message $T$ is transmitted at time slot $i$ or not. Note that $\|\mathbf{v}_T\|_1 \ge t+1$ is required to fulfill users' demands, where $\mathbf{v}_T \triangleq [v_T(1) \cdots v_T(B)]$. It is readily seen that $v_T(i)$ can be inferred by the corresponding beamforming vector $\mathbf{w}_T(i)$, or equivalently, by the message rate $R_T(i)$.

### C. Transmit Power Minimization

For any given delivery scheme specified by $v_T(i)$ and $n_i$, $\forall i \in [B], \forall T \subseteq \mathcal{S}$, the associated minimum required transmit power problem is obtained as follows:

$$P: \min_{\{\mathbf{w}_T(i)\},\{R^T(i)\}} \sum_{T \subseteq \mathcal{S}} \sum_{i=1}^{B} \frac{n_i}{n} \|\mathbf{w}_T(i)\|^2 \tag{12a}$$

$$\text{s.t.} \sum_{T \supseteq \mathcal{S}_k^j} R^T(i) \le \frac{n_i}{n} \log_2\left(1 + \sum_{T \supseteq \mathcal{S}_k^j} \gamma_k^T(i)\right), \quad \forall \mathcal{S}_k^j \subseteq \mathcal{S}_k, \forall i \in [B], \forall k, \tag{12b}$$

$$\sum_{i=1}^{B} R^T(i) \le \frac{R}{K}, \ \forall T, \tag{12c}$$

$$R^T(i) \ge 0, \ \forall i, T, \tag{12d}$$

$$R^T(i) = 0, \text{ if } v_T(i) = 0, \ \forall i, \tag{12e}$$

where $\gamma_k^T(i)$ is defined in (9). Here, constraints in (12b) guarantee that the rates of the messages targeting each user in each time slot are within the capacity region, constraints in (12c) ensure that sufficient information is delivered for each coded subfile over $B$ time slots, while (12d)-(12e) represent the specific content delivery scheme.

Note that the problem in (12) is a generalization of various well-known NP-hard problems depending on the specific content delivery scheme. For $B = |\mathcal{S}|$ with $|\mathcal{S}(i)| = 1, \forall i$, the problem boils down to a series of standard multicast beamforming problems, where a common message is broadcast to a different subset of $t+1$ users in each time slot [8]. When $|\mathcal{S}(i)| > 1$, $T \cap T' = \emptyset$ if $T \neq T' \in \mathcal{S}(i), \forall i$, and $\mathcal{S}(i) \cap \mathcal{S}(j) = \emptyset$ if $i \neq j$, we need to solve the conventional multigroup multicast beamforming problem at each time slot [9]. It can be seen from (12) that the content delivery scheme specified by $v_T(i)$ and $n_i$ affects the minimum required power. A straightforward solution (as done in [17], [20]) would transmit a single coded message in each time slot. However, this does not exploit the spatial multiplexing gain provided by multiple antennas, and results in a poor DoF performance in the high SNR regime. Another approach studied in [21] is to deliver the coded messages targeting non-overlapping user groups in parallel. Obviously, the content delivery scheme is an important factor on the system performance and needs to be carefully designed.

We remark here that, even when the delivery scheme is specified, the problem in (12) is computationally intractable due to the non-convex constraints in (12b). However, we show in the Appendix that SCA methods [22] can be employed to obtain a stationary point of the problem, which serves as an upper bound on the optimal solution. Starting with a feasible initial point, the SCA algorithm solves a sequence of subproblems in an iterative manner, where the subproblem in the $n$-th iteration is derived by convexifying the original problem at the solution point of the $(n-1)$-th subproblem. More detailed discussions on the SCA algorithm are provided in the Appendix.

## IV. A Low-Complexity Design

In this section, we propose a low-complexity content delivery scheme with the flexibility to adjust the number of coded messages intended for each user at each time slot. Observing that if a set $\mathcal{S}(i) = \{T | v_T(i) = 1\}$ of messages are transmitted in time slot $i$, $c_k(i) \triangleq |\mathcal{S}(i) \cap \mathcal{S}_k|$ messages are transmitted to user $k$, which results in $2^{c_k(i)} - 1$ constraints only for user $k$ in time slot $i$ in problem (12). Computational complexity of problem (12) increases drastically with the number of constraints, rendering the numerical optimization problem practically infeasible. More importantly, a multi-user detection scheme needs to be employed at the users, whose complexity also increases with $c_k(i)$.

A low complexity scheme is proposed in [16] by limiting the number of users to be served in each time slot, thereby indirectly reducing the number of coded messages to be decoded by each user. Specifically, an integer parameter $t \in [\min\{N_T, K-1\}]$ is leveraged in [16] to control the number of active users in each time slot, which is set to $t+1$, and leads to a content delivery scheme with $B = \binom{K}{t+1}$ time slots. In each time slot, a fraction of the desired coded messages for all the active users are transmitted. In addition to $t$, another integer parameter $\alpha$ determines the possible set partitions of the user subset in each time slot. When $t+1$ is divisible by $\alpha(t+1)$, the user subset can be partitioned into $\frac{t+1}{\alpha(t+1)}$ non-overlapping subsets, and a fraction of the desired coded messages for each partition can be transmitted simultaneously. It is shown in [16] that the system performance can be improved if multiple groups of messages can be transmitted in parallel, i.e., $\frac{t+1}{\alpha(t+1)} \ge 2$, as compared to the case $\alpha = 1$.

Moreover, the number of messages for each user to decode in each time slot is $\binom{t+}{t}^{1}$, which is an exponential function of . Therefore, by adjusting the value of , the number of coded messages for each user in each time slot is indirectly adjusted.

Instead of limiting the subsets of users to be served in each time slot, we propose to directly adjust the number of coded messages targeted to each user. We will show that this results in a more efficient delivery scheme than the one in [16]. In Example 1, if we transmit all the messages in one time slot, i.e., $B = 1$, a total of $jSj = \binom{K}{t+1} = 10$ coded subfiles are transmitted simultaneously, with each user decoding $\binom{K}{t}^{1} = 4$ messages. Accordingly, in the optimization problem in (12) we will have $K \quad (2^{jS_kj} \quad 1) = 75$ constraints. To alleviate the computational complexity, the low complexity scheme in [16] splits each subfile into 3 minifiles, and the coded messages are grouped to serve a subset of $t + \quad = 3$ users in each of the $B = \binom{K}{t+} = 10$ time slots. Within each time slot, each user needs to decode 2 messages. Note that the power minimization problem for each time slot can be solved independently; therefore, we would need to solve 10 smaller optimization problems, each with $3 \quad 3 = 9$ constraints.

In contrast, we propose to serve as many users as needed at each time slot while keeping $c_k(i)$ under a given threshold $s$ for each user $k$. In our Example 1, we can satisfy all the user requests in only 2 time slots, by setting nonzero rate targets for the messages in

$$S(1) = ff1;2g;f2;3g;f3;4g;f4;5g;f1;5gg; \text{ and}$$
$$S(2) = ff1;3g;f2;4g;f3;5g;f1;4g;f2;5gg$$

in time slots 1 and 2, respectively. Note that each user $k$ decodes only $c_k(i) = s = 2$ messages in each time slot, the same as the delivery scheme in [16], requiring the same implementation complexity at each user; however, 5 users are served in each time slot, which results in a significantly smaller number of time slots. Thus, we need to solve only two optimization problems at the BS, each with $5 \quad 3 = 15$ constraints.

In general, the number of constraints in the optimization problem in (12) increases exponentially with $s$, which results in exponentially increasing number of constraints in the problem in each SCA iteration. Thus the computational complexity of the delivery scheme can be largely alleviated by choosing a small $s$ value, which also simplifies the multi-user detection algorithm.

The key idea of our proposed low-complexity scheme is to divide set $S$ into disjoint subsets $S(1); \quad ;S(B)$, with $c_k(i) \quad s, 8k;i$, while keeping $B$ as small as possible. Since the total number of subfiles to transmit is fixed, choosing a small value of $B$, i.e., completing the delivery phase within a small number of time slots, requires multiplexing more messages in each time slot, without increasing the complexity of the receivers. To obtain this low-complexity scheme, the following optimization problem can be formulated:

$$\min_{fv_T g;B} \quad B \tag{13a}$$
$$\text{s.t.} \quad \sum_{T \ 3k} v_T(i) \quad s;8i \ 2 \ [B]; \ k \ 2 \ [K]; \tag{13b}$$
$$\sum_{i=1}^{B} v_T(i) = 1;8T; \tag{13c}$$
$$v_T(i) \ 2 \ f0;1g;8T;i \ 2 \ [B]; \tag{13d}$$

where constraint (13b) imposes that each user decodes no more than $s$ messages in each time slot, while (13c) requires that each message will be transmitted in only one time slot. However, since the problem itself varies with variable $B$, the problem is not in a tractable form. By introducing $L \quad B$ as a prescribed parameter that determines the dimension of the problem, and an auxiliary variable $q \ 2 \ f0;1g^L$, problem (13) can be equivalently written as

$$B = \min_{fv_T g;q} \quad 1^T q \tag{14a}$$
$$\text{s.t.} \quad \sum_{T \ 3k} v_T(i) \quad s;8i \ 2 \ [L]; \ k \ 2 \ [K]; \tag{14b}$$
$$\sum_{i=1}^{L} v_T(i) = 1;8T; \tag{14c}$$
$$\sum_{T} v_T(i) \quad \binom{K}{t+1} q_i;8i \ 2 \ [L]; \tag{14d}$$
$$v_T(i) \ 2 \ f0;1g;8T;i \ 2 \ [L]; \tag{14e}$$
$$q \ 2 \ f0;1g^L; \tag{14f}$$

where $1$ denotes a column vector of all ones. Since $\binom{K}{t+1}$ is a bound on $\sum_T v_T(i)$, the optimal $q_i$ is 1 if $\sum_T v_T(i)$ is nonzero, and 0 otherwise, in order to minimize $\sum_{i=1}^{L} q_i$ in the objective. Note that problem (14) can be considered as minimizing the number of time slots employed out of a maximum $L$ available time slots. We can set $L = \binom{K}{t+1}$ which guarantees the existence of a solution; however, choosing a smaller $L$ will reduce the complexity of the problem. The problem in (14) is a $0 \quad 1$ integer programming problem, which is generally NP-hard [23].

**Algorithm 1** Low-complexity greedy delivery scheme

**Input:** $N, K, M, s, R$
**Output:** $B$, $\{S(i)\}_{i=1}^{B}$, $\{n_i\}_{i=1}^{B}$, $T$

Set $t = \frac{MK}{N}$, $i = 1$, and $E = S$
**while** $E \neq \emptyset$ **do**
  Set $c(i) = [c_1(i), \ldots, c_K(i)] = 0$, $S(i) = \emptyset$, $C = E$
  **while** $c_k(i) \geq s, \forall k \in [K]$ and $C \neq \emptyset$ **do**
    $K' = \{k \mid k \in \arg\min_k c(i)\}$
    Find $f = \arg\max_{T \in C} |K' \cap T|$
    $C = C \setminus \{f\}$
    **if** $c_k(i) + 1 \leq s, \forall k \in f$ **then**
      $c_k(i) = c_k(i) + 1, \forall k \in f$
      $S(i) = S(i) \cup f$, $E = E \setminus \{f\}$
    **else**
      **break**
    **end if**
  **end while**
  $i = i + 1$
**end while**
Set $B = i - 1$
**for** $i = 1 : B$ **do**
  $n_i = \frac{|S(i)|}{\binom{K}{t+1}} n$
  $R^T(i) = \begin{cases} \frac{R}{\binom{K}{t}}, & \forall T \supseteq S(i), |T| = t+1 \\ 0, & \text{otherwise} \end{cases}$
**end for**

In Algorithm 1, we propose a greedy solution that constructs disjoint $S(i)$ sets for any $s$ value. Specifically, $S(i)$'s are generated in a sequential manner: to construct $S(i)$, we initialize $c(i) = [c_1(i), \ldots, c_K(i)] = 0$, $S(i) = \emptyset$, and the set $E = S \setminus \bigcup_{j=1}^{i-1} S(j)$ of remaining messages for assignment, we first identify the user(s) that have decoded the least number of messages so far, i.e., user(s) in set $K' = \{k \mid k \in \arg\min_k c(i)\}$, and check whether there exists a message $f \in E$ such that the condition $c_k(i) + 1 \leq s$ for $\forall k \in f$ holds. If no such $f$ can be found, the process of constructing $S(i)$ is completed, and we start constructing $S(i+1)$ in the same manner. The whole procedure is completed when $E = \emptyset$, i.e., all the messages have been assigned to a subset. Note that our proposed greedy scheme covers the case of $B = 1$, where all the messages are sent simultaneously, if $s \geq \binom{K}{t+1}$.

Next we elaborate the proposed greedy content delivery algorithm in Examples 1 and 2.

***Example 1*** *(continued)*: $N = 5$, $K = 5$, $M = 1$. $t = \frac{MK}{N} = 1$. Suppose $s = 2$. As illustrated in Fig. 2, the algorithm starts by constructing $S(1)$, i.e., identifying the coded messages to be delivered in the first time slot, initialized as $c(1) = [c_1(1), \ldots, c_K(1)] = 0$, $S(1) = \emptyset$. Firstly, it is obvious that $K' = [K]$ since $c_k(1) = 0, \forall k$. Hence, one may choose any of the available messages in $E = S$. Suppose message $s_{\{1,2\}}$ is chosen. We update $c_1(1) = c_2(1) = 1$, $E = E \setminus \{1,2\}$. The algorithm then identifies the updated $K' = [3,4,5]$, according to which one may choose from $s_{\{3,4\}}, s_{\{3,5\}}, s_{\{4,5\}}$ without violating the constraint $c_k(1) + 1 \leq s, \forall k$. Suppose $s_{\{3,4\}}$ is chosen, and we have $c_1(1) = c_2(1) = c_3(1) = c_4(1) = 1$, $E = \{\{3,4\}\}$, and $K' = [5]$; and accordingly one may choose from $s_{\{1,5\}}, s_{\{2,5\}}, s_{\{3,5\}}, s_{\{4,5\}}$, while still keeping the constraint $c_k(1) + 1 \leq s, \forall k$. Similarly, messages $s_{\{2,3\}}$ and $s_{\{4,5\}}$ can be chosen, and the algorithm for $S(1)$ is completed since $c_k(1) = 2, \forall k$, and adding any of the remaining messages will violate the constraint. The algorithm then turns to construct $S(2)$ similarly, until all the messages have been chosen, i.e., $E = \emptyset$.

*Remark 1*: As it can be seen above, the content delivery scheme obtained via Algorithm 1 is not unique. For instance, another feasible content delivery scheme with $s = 2$ for Example 1 is:

$$S(1) = \{\{1,2\}, \{3,4\}, \{1,5\}, \{2,4\}, \{4,5\}\}, \text{ and}$$
$$S(2) = \{\{1,3\}, \{2,3\}, \{3,5\}, \{1,4\}, \{2,5\}\}.$$

***Example 2***: $N = 4$, $K = 4$, $M = 1$. $t = \frac{MN}{K} = 1$.

We present the content delivery schemes obtained from Algorithm 1 for different values of $s$:

**Case 1:** $s = 1$. Algorithm 1 leads to $B = 3$ time slots, each with $\frac{1}{3}n$ channel uses, and

$$S(1) = \{\{1,2\}, \{3,4\}\},$$
$$S(2) = \{\{1,3\}, \{2,4\}\},$$
$$S(3) = \{\{1,4\}, \{2,3\}\}.$$

It is noted that the scheme is the same as the one in [16] obtained for $= 3$ and $= 1$, in the sense that the same sets of messages are transmitted over the same number of time slots.

**Case 2:** $s = 2$. Algorithm 1 leads to $B = 2$ time slots, and

$$S(1) = \{\{1,2\}, \{3,4\}, \{1,3\}, \{2,4\}\},$$
$$S(2) = \{\{1,4\}, \{2,3\}\}.$$

It is noted that there is no value to induce a scheme in [16] in this scenario, since can only take the value of 2 to have $s = 2$, making $t +$ not divisible by $t +$ .

*Remark 2*: The proposed greedy content delivery scheme can be easily extended by limiting the number of active users as in [16]. Specifically, instead of serving as many users as possible, which is up to $K$, Algorithm 1 can be applied for a user subset of size $t +$ to obtain a content delivery scheme under the constraints on the number of messages to decode. While $\frac{t+}{t+}$ must be an integer to induce a content delivery scheme in [16], Algorithm 1 always provides a delivery scheme for any $s$ value. Therefore, our proposed greedy scheme can be considered as a generalization of the one in [16].

*Remark 3*: It is noted that Algorithm 1 may lead to unequal number of messages transmitted in different time slots, which can be highly sub-optimal. An intuitive way to enhance the performance is to allocate more channel uses to the time slot with more messages to deliver. In general, once the non-overlapping partition of $S$, i.e., $\{S(i)\}_{i=1}^{B}$, is obtained, we can set the blocklength for the transmission of $S(i)$ proportionally to the number of messages $|S(i)|$. For instance, in the case of $s = 2$ in Example 2, we can allocate $\frac{2n}{3}$ channel uses for $S(1)$ and $\frac{n}{3}$ channel uses for $S(2)$.

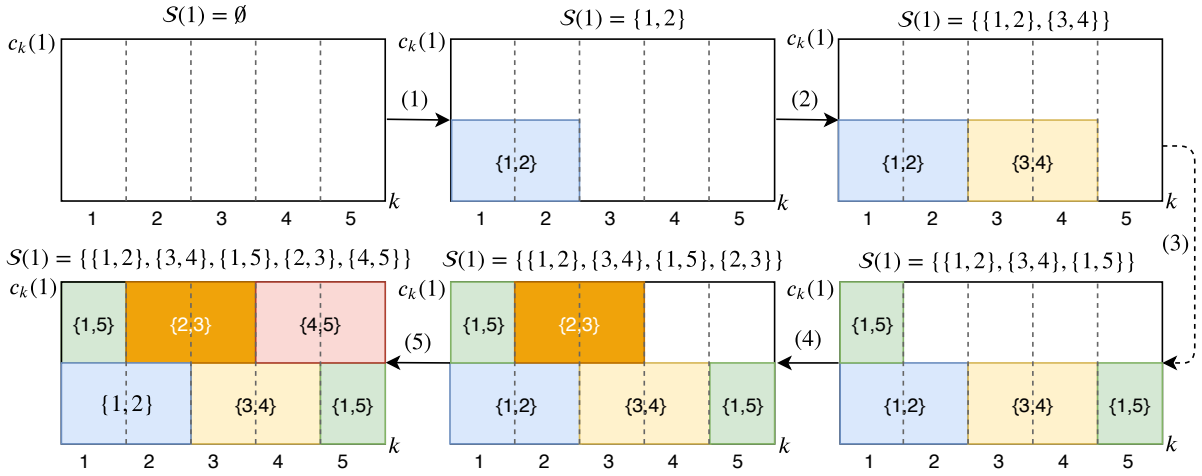



Figure 2: Illustration of the proposed low-complexity greedy scheme for the network with $N = K = 5$, and $M = 1$.

*Remark 4*: The proposed low-complexity scheme focuses on limiting $s$, which determines the complexity of multiuser detection at the receivers. On the other hand, the beamformer design resorts to solving the optimization problem in (12), whose computational complexity depends on the content delivery scheme. Given the proposed low-complexity scheme in Algorithm 1, the optimization problem in (12) can be solved as detailed in the Appendix, which is firstly decomposed into at most $B_u$ problems in the form of (32). For each problem, we have at most $K(2^s \quad 1)$ constraints to characterize the achievable rate region. For the same $s$ value, the optimization problem for the scheme in [16] can be decomposed into $B_l$ , $\binom{K}{t+}\frac{(t+)!}{!(t+)!}$ subproblems, each with $(t+)(2^s \quad 1)$ constraints for the achievable rate region characterization, where $= \frac{t+}{t+}$. Therefore, each optimization problem in the proposed scheme involves $\frac{K}{t+}$ times more constraints compared to [16], but may require solving much fewer problems.

Table I compares the proposed scheme in Algorithm 1 and the scheme in [16] in terms of the number of time slots, or equivalently, the number of optimization problems to be solved, denoted by $B_u$ and $B_l$, respectively, and the ratio of the number of constraints in each of these problems in the former scheme to the latter. We observe that, by simultaneously serving as many users as possible, the optimization problem in each time slot of the proposed low-complexity scheme has comparable number of constraints, but much fewer such optimization problems are needed, as compared to the scheme in [16].

Numerical results for the minimum required power for the proposed greedy transmission scheme, and the comparison with the one proposed in [16] will be presented in Section VII.

## V. Degrees-of-Freedom (DoF) Analysis

In this section we analyze the performance of the scheme proposed in Algorithm 1 in terms of the DoF it achieves in the high SNR regime. To this end, we develop a content delivery scheme which upper bounds the number of time slots $B$ of the scheme presented in Algorithm 1.

For caching factor $t = MK/N$, Algorithm 1 constructs, at each time slot, a set of distinct subsets of users of size $t + 1$, such that no user index appears more than $s$ times in the set. This procedure is repeated until all the $\binom{K}{t+1}$ distinct subsets of users are selected exactly once. While Algorithm 1 aims at minimizing the number of time slots required to deliver all the multicast messages, it is not possible to know in advance how many time slots will be needed. To overcome this uncertainty, we develop a more relaxed scheme, which utilizes a higher number of time slots than the one presented in Algorithm 1.

We assume that the BS is equipped with $N_T \quad K \quad t$ antennas, to simultaneously transmit each coded packet of rate $R = \binom{K}{t}$ to the $t + 1$ users that are interested in this message, while zero-forcing it at the remaining $K \quad t \quad 1$ users [13]. At each time slot all the $K$ users are targeted, and each user receives no more than $s$ coded packets, and the coded delivery is performed for a total of $B$ time slots. Assuming equal blocklength for different time slots in the high SNR regime, i.e., $n_i = n = B$, $8i \ 2 \ [B]$, since $N_T \quad K \quad t$, we can lower bound the per-user DoF as follows [24], [25]:

$$\text{DoF} \quad \frac{\binom{K}{t}}{sB}; \tag{16}$$

where $B$ is the number of time slots obtained from Algorithm 1.

Next, we present an upper bound on $B$. We further divide each time slot to $s$ sub-time-slots for the new content delivery approach. At each sub-time-slot, the goal is to create a set of $(t + 1)$-element subsets of users, such that no user index appears in more than one subset. Due to the symmetry, it is easy to verify that the maximum number of such subsets at each sub-time-slot is $\left\lfloor \frac{K}{t+1} \right\rfloor$. Therefore, we can generate a set of distinct $(t + 1)$-element subsets of users, with no user appearing more than $s$ times by repeating this procedure for $s$ sub-time-slots. Accordingly, we can generate a set of $s\left\lfloor \frac{K}{t+1} \right\rfloor$ distinct subsets of users, each of size $t + 1$, such that each user index appears no more than $s$ times. This provides us with the set of coded packets for delivery at each time slot. For example, considering Example 1, where $N = 5$; $K =$

Table I: Comparison between Algorithm 1 and [16] in terms of the number of constraints they need to consider for each optimization problem, and the number of time slots, or equivalently, the number of optimization problems to be solved.

| $(K;t;s)=(10;1;1)$ | | | | | $(K;t;s)=(10;2;1)$ | | | | | $(K;t;s)=(10;3;1)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $(\ ;\ )$ | $\frac{K}{t+s}$ | $B_u$ | $B_l$ | $\frac{B_u}{B_l}$ | $(\ ;\ )$ | $\frac{K}{t+s}$ | $B_u$ | $B_l$ | $\frac{B_u}{B_l}$ | $(\ ;\ )$ | $\frac{K}{t+s}$ | $B_u$ | $B_l$ | $\frac{B_u}{B_l}$ |
| $(1;1)$ | 5 | 9 | 45 | 0.2 | $(1;1)$ | $\frac{10}{3}$ | 40 | 120 | 0:3333 | $(1;1)$ | $\frac{5}{2}$ | 105 | 210 | 0.5 |
| $(1;3)$ | $\frac{5}{2}$ | 9 | 630 | 0:0143 | $(1;4)$ | 2 | 40 | 2100 | 0:0190 | $(1;5)$ | $\frac{5}{4}$ | 105 | 1575 | 0:0667 |
| $(1;5)$ | $\frac{5}{3}$ | 9 | 3150 | 0:0029 | $(1;7)$ | $\frac{10}{7}$ | 40 | 2800 | 0:0143 | | | | | |
| $(1;7)$ | $\frac{5}{4}$ | 9 | 4725 | 0:0019 | | | | | | | | | | |
| $(1;9)$ | 1 | 9 | 945 | 0:0095 | | | | | | | | | | |

$5;\ t=1$, and $s=2$, we can generate the following sets of subsets of users, each of size 2, in the two sub-time-slots of the first time slot:

$$\mathcal{S}(1;1) = \{\{1;2\};\{3;4\}\}, \text{ and}$$
$$\mathcal{S}(1;2) = \{\{1;3\};\{4;5\}\}, \tag{17}$$

where $\mathcal{S}(1) = \{\mathcal{S}(1;1);\mathcal{S}(1;2)\}$, and this partitioning is not unique. We exploit the symmetry in the subsets of users of size $t+1$, to which the coded packets are targeted, to obtain the total number of time slots required for transmission. By creating $s\left\lfloor\frac{K}{t+1}\right\rfloor$ $(t+1)$-element distinct subsets of users at each time slot, we need no more than $\frac{\binom{K}{t+1}}{s\left\lfloor\frac{K}{t+1}\right\rfloor}$ time slots to deliver all the coded messages.

We highlight that, this approach resembles the set generation process in Algorithm 1, but is stricter as it requires each user index to appear no more than once in each sub-time-slot, compared to Line 8 in Algorithm 1. Hence, it results in fewer selected messages at each time slot, and more time slots. Moreover, it is guaranteed in Algorithm 1 that each of the $K$ users decodes at least $s-1$ messages in each time slot, except for the final time slot. Therefore, the number of time slots is also bounded by $\frac{\binom{K-1}{t}}{s-1}+1$. Consequently, we have

$$B_u \geq \min\left\{\frac{\left\lfloor\frac{K}{t+1}\right\rfloor}{s\left\lfloor\frac{K}{t+1}\right\rfloor};\ \frac{K-t}{s-1}+1\right\}; \tag{18}$$

and the DoF of the proposed low-complexity scheme satisfies

$$\text{DoF} \geq \frac{\frac{K}{t}}{sB_u}. \tag{19}$$

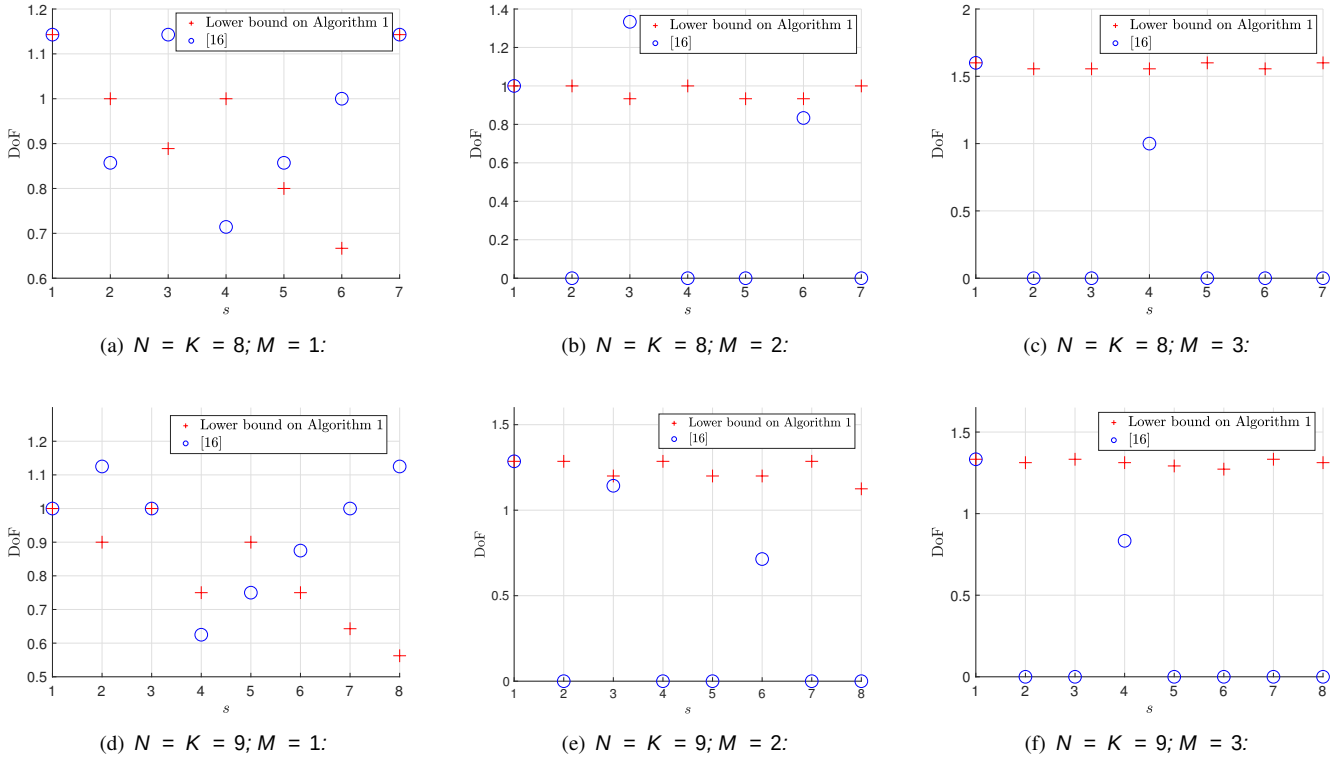
The lower bound on the DoF of the proposed greedy scheme is depicted in Fig. 3 for different network parameters, in comparison with the scheme in [16]. As the greedy scheme in Algorithm 1, a low-complexity scheme that achieves the DoF lower bound can always be obtained for any $s$, while the scheme in [16] may not exist for some values of $s$, for which the DoF is set to 0 in Fig. 3. Note that the achievable DoF in [16] monotonically increases with . For any $s$ value, if an integer is found such that $\binom{t+}{t}^{1} = s$, then we find the maximum possible such that $t+$ is divided by $t+$, yielding the highest DoF of the scheme in [16] as depicted in Fig. 3. It can be seen from Fig. 3 that, the proposed greedy scheme can outperform [16] for certain values of $s$, especially for small $s$ that is of particular interest in practice. We remark that, the derived lower bound on DoF can be loose due to the floor and ceiling operations in (18), and the proposed greedy scheme may achieve a DoF strictly higher than the lower bound illustrated here. For instance, in the scenario of $N=K=9$ and $M=1$ as considered in Fig. 3(d), the total number of messages for each user to decode is 8. The proposed low-complexity scheme for $s=8$ delivers all the coded messages in only one time slot, achieving the same DoF of $\frac{9}{8}$ as the scheme in [16]. Hence, the lower bound on DoF at $s=8$ is very loose as can be seen in Fig. 3(d). We finally remark that, while the DoF lower bound is in general loose, it is tight at certain values of $s$ regardless of the values of $N;K$, and $M$. Specifically, when $s=1$, or when $s$ is sufficiently large such that $B_u=1$, the scheme obtained via the relaxed algorithm, which derives the DoF lower bound in (19), is identical to the one obtained via Algorithm 1 with equal blocklength allocation over all the time slots; and hence, the right-hand-side term in (19) is the exact DoF achieved by the scheme in Algorithm 1 when simply allocating equal blocklength across all the time slots.

## VI. Joint Optimization of Beamforming and Coded Content Delivery

In this section, we formulate a sparsity constrained power minimization problem to jointly optimize the beamformers and the content delivery scheme. The sparsity induced problem directly limits the number of messages to be decoded by each user at any time slot, and the indicator function $v_T(i)$ is identified by setting $v_T(i) = \|R^T(i)\|_0;\ \forall i;T$, where $\|\ \|_0$ denotes the $\ell_0$-norm and is equal to the number of non-zero elements of a vector. Therefore, we impose an $\ell_0$-norm constraint on the rates of messages at any time slot $i$ as follows:

$$\sum_{T\supseteq\mathcal{S}_k} \|R^T(i)\|_0 \leq s;\ \forall k;i. \tag{20}$$

In this section, we assume equal blocklength allocation over all the $B$ time slots for simplicity. Then, the minimum required power problem with the constrains on the number of messages to be decoded by any user at any time slot can be formulated




(a) $N = K = 8; M = 1$: (b) $N = K = 8; M = 2$: (c) $N = K = 8; M = 3$:

(d) $N = K = 9; M = 1$: (e) $N = K = 9; M = 2$: (f) $N = K = 9; M = 3$:

Figure 3: DoF lower bound for the proposed greedy scheme in Algorithm 1 compared to the DoF of the scheme in [16] as a function of $s$ for various network parameters.

as follows:

$$\min_{\{W_T(i)\},\{R^T(i)\}} \frac{1}{B}\sum_{i=1}^{B}\sum_{T\subseteq\mathcal{S}} \|W_T(i)\|^2 \tag{21a}$$

$$\text{s.t.} \sum_{T\supseteq\{j_{S_k}\}} R^T(i) \le \frac{1}{B}\log_2\Big(1+\sum_{T\supseteq\{j_{S_k}\}} \gamma_k^T(i)\Big), \quad \forall\, j_{S_k}\subseteq S_k,\ \forall k, i, \tag{21b}$$

$$\sum_{i=1}^{B} R^T(i) \ge \frac{R}{K}, \quad \forall T, \tag{21c}$$

$$\sum_{T\supseteq S_k} |R^T(i)|_0 \le s, \quad \forall k, i, \tag{21d}$$

where $\gamma_k^T(i)$ is defined in (9). In this problem, the objective is to minimize the average transmission power over all the time slots; constraints (21b)-(21c) guarantee successful decoding of all the required messages at each user in each time slot; and constraint (21d) limits the number of messages decoded by each user in any time slot. Since (21d) limits only the number of messages decoded by each user in any time slot, without assuming any specific content delivery scheme, the problem in (21) includes the content delivery schemes in [16] as a special case. This formulation also generalizes the one presented in Section III when the time slots are of equal duration. However, note that the number of time slots $B$ is a free variable for the greedy scheme, while it is assumed to be given in (21).

To deal with the discontinuous $\ell_0$-norm constraint in (21d), we approximate it with a differentiable continuous function [26]

$$f(R^T(i);t) \triangleq \frac{2}{\pi}\arctan\frac{R^T(i)}{t}, \tag{22}$$

where $t > 0$ is a prescribed constant that determines the approximation accuracy. The function in (22) is concave w.r.t. $R^T(i)$, therefore the approximate constraint for (21d)

$$\sum_{T\supseteq S_k} f(R^T(i);t) \le s, \quad \forall k, i, \tag{23}$$

is concave and can be treated as a difference of convex function. Overall, the original problem in (21) can be approximated by the following problem:

$$\min_{\{W_T(i)\},\{R^T(i)\}} \sum_{i=1}^{B}\sum_{T\subseteq\mathcal{S}} \frac{1}{B}\|W_T(i)\|^2 \tag{24a}$$

$$\text{s.t.} \ (21b), (21c) \text{ and } (23).$$

Similarly to (12), the problem in (21) can be solved via the SCA method. Specifically, we introduce $\gamma_{j_{S_k}}(i) \triangleq \sum_{T\supseteq\{j_{S_k}\}} \gamma_k^T(i)$, then the constraint in (21b) can be rewritten

similarly to (30b) and (30c), given by

$$\sum_{T \supseteq \mathcal{S}_k^j} R^T(i) \quad \frac{1}{B}\log_2(1+\quad {}_{\mathcal{S}_k^j}(i)),\ \forall\ \mathcal{S}_k^j \supseteq \quad \mathcal{S}_k, \forall k, i, \tag{25}$$

$$\sum_{I \subseteq \mathcal{S}_k^C} |\boldsymbol{h}_k^H \boldsymbol{w}_I(i)|^2 \quad \frac{\sum_{T \supseteq \mathcal{S}_k^j} |\boldsymbol{h}_k^H \boldsymbol{w}_T(i)|^2}{{}_{\mathcal{S}_k^j}(i)} + \quad {}_k^2 \quad 0,\ \forall\ \mathcal{S}_k^j \supseteq \quad \mathcal{S}_k, \forall k, i, \tag{26}$$

where the constraints in (25) are convex, and the constraints in (26) are in the form of difference of convex functions. In the $(\quad +1)$-th iteration of the SCA algorithm, the constraints in (26) can be linearized with the first order Taylor expansion at $\{\boldsymbol{w}_T(i)\}$ and $\{ {}_{\mathcal{S}_k^j}(i)\}$, leading to stricter constraints given by

$$\sum_{I \subseteq \mathcal{S}_k^C} |\boldsymbol{h}_k^H \boldsymbol{w}_I(i)|^2 + \frac{\sum_{T \supseteq \mathcal{S}_k^j} |\boldsymbol{h}_k^H \boldsymbol{w}_T(i)|^2}{{}^2_{\mathcal{S}_k^j}(i)} \quad {}_{\mathcal{S}_k^j}(i) \quad \frac{2\sum_{T \supseteq \mathcal{S}_k^j} \boldsymbol{w}_T{}^H(i)\boldsymbol{h}_k\boldsymbol{h}_k^H\boldsymbol{w}_T(i)}{{}_{\mathcal{S}_k^j}(i)} + \quad {}_k^2 \quad 0,\ \forall\ \mathcal{S}_k^j \supseteq \quad \mathcal{S}_k, \forall k, i, \tag{27}$$

where $\{\boldsymbol{w}_T(i)\}$ and $\{ {}_{\mathcal{S}_k^j}(i)\}$ are the solutions to the subproblem in the -th iteration. The same linearization technique can also be performed for the constraints in (23), yielding stricter constrains given by

$$\sum_{T \supseteq \mathcal{S}_k} \arctan\left(\frac{R^T(i)}{t}\right) + \frac{t}{t^2 + R^T{}^2(i)}\left(R^T(i) \quad R^T(i)\right) \quad \frac{s}{2},\ \forall k, i, \tag{28}$$

where $\{R^T(i)\}$ are the solutions to the subproblem in the -th iteration. Overall, the convex subproblem to be solved in the $(\quad +1)$-th iteration is

$$\min_{\{\boldsymbol{w}_T(i)\},\{R^T(i)\},\{ {}_{\mathcal{S}_k^j}(i)\}} \quad \frac{1}{B}\sum_{i=1}^{\quad}\sum_{T \supseteq \mathcal{S}} \|\boldsymbol{w}_T(i)\|^2 \tag{29a}$$

$$\text{s.t.} \qquad (21c),\ (25),\ (27),\ \text{and}\ (28):$$

The initialization of the SCA algorithm for problem (29) for a given value of $s$ requires a content delivery scheme with less or equal complexity, which can be obtained via Algorithm 1 in Section IV. The associated beamforming design can be readily obtained similarly to (31). From problem (21), we can also conclude that the minimum required power of a content delivery scheme is a non-decreasing function of $s$, as the problem becomes more relaxed as $s$ increases.

## VII. Simulation Results

We consider a single-cell with radius 500m, and users uniformly randomly distributed in the cell. Channel vectors $\boldsymbol{h}_k$ are written as $\boldsymbol{h}_k = (10^{-\mathrm{PL}/10})^{1/2}\tilde{\boldsymbol{h}}_k$, $\forall k$, where $\tilde{\boldsymbol{h}}_k$ denotes an i.i.d. vector accounting for Rayleigh fading of unit power, and the path loss exponent is modeled as $\mathrm{PL} = 148.1 + 37.6\log_{10}(v_k)$, with $v_k$ denoting the distance between the BS and the user (in kilometers). The noise variance is set to $ {}_k^2 = \quad {}^2 = \quad 134$ dBW for all the users. Throughout this section, we assume that the number of transmit antennas is $N_T \quad K \quad t$ as required for achieving a satisfactory DoF performance. All simulation results are averaged over 300 independent trials computed with CVX [27].

The scheme with $B = 1$ time slot will be referred to as the full superposition (FS) scheme. FS has the best performance in terms of transmit power given enough spatial DoF, and serves as a baseline, but it also has the highest complexity. To compare our results with those in [16], same number of coded messages are transmitted to each user in each time slot for both schemes. We note here that with the use of parameter, the scheme in [16] can be improved by serving disjoint subsets of users simultaneously without increasing the complexity, but the improvement is only applicable when the size of user subset can be partitioned equally and exactly. Therefore, the scheme in [16] cannot handle certain settings such as the case of $s = 2$ in Example 1.

We first present the average transmit power as a function of the target rate $R$ in Fig. 4 for Example 1, assuming that the BS is equipped with $N_T = 6$ antennas. The scheme in [16] that satisfies $s = 2$ is adopted for fair comparison, where $t + \quad = 3$ users are served in each time slot. We observe that the proposed greedy scheme provides significant savings in the transmit power compared to [16] at all rates. The power savings increase with rate $R$ as a result of the increased superposition coding gain. Furthermore, the gap between the proposed greedy scheme and FS is quite small, and remains almost constant with rate. At $R = 8$ bps/Hz, the power loss of the scheme in [16] and ours compared to FS are about 8.5 dB and 0.5 dB, respectively. Hence, we can conclude that the proposed greedy scheme provides significant reduction in the computational complexity without sacrificing the performance much.

The average transmit power as a function of file rate $R$ is further investigated for the setting with $N = 6$ files, $K = 6$ users, $M = 1$, and $N_T = 6$ antennas. Similarly to Fig. 4, it is observed in Fig. 5 that our proposed low-complexity greedy scheme substantially outperforms the scheme in [16] with the same value of $s$ in the high SNR regime. For example, the power savings of the greedy delivery scheme compared to [16] are 8dB and 2dB, for $s = 3$ and $s = 4$, respectively, at $R = 10$ bps/Hz. The power gain is again observed to be larger as the rate increases, while in the low SNR/rate regime, all the schemes achieve comparable performance regardless of $s$. Also, for the proposed greedy scheme, a larger $s$ allows achieving the same rate with lower transmit power in the high SNR regime, at the expense of increased complexity at the receivers. It is noted that the proposed greedy scheme yields the same content delivery scheme in terms of the transmitted coded messages in each time slot as the one in [16] when $s = 1$ and $s = 2$, which correspond to $\quad = 5;\ \quad = 1$, and $\quad = 5;\ \quad = 2$ in [16], respectively.

Fig. 6 and Fig. 7 show the average transmit power versus rate $R$ for Example 2, with $N = 4$ files, $K = 4$ users, $M = 1$, and $N_T = 3$ antennas. In Fig. 6, we compare our greedy content delivery scheme with the one obtained

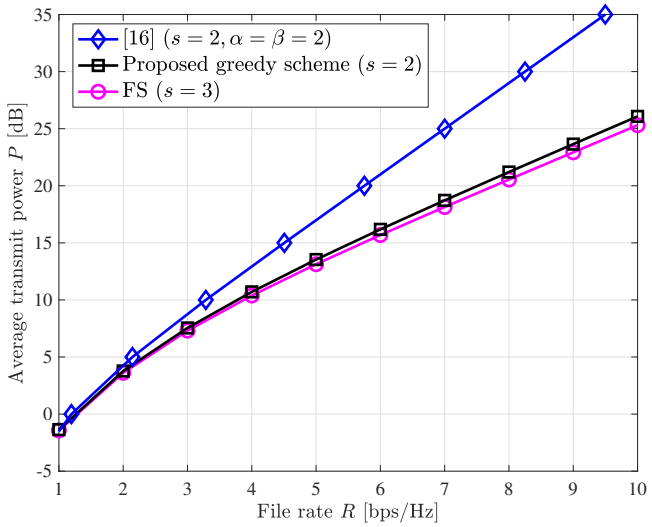



Figure 4: Average transmit power $P$ as a function of rate $R$ for $N = K = 5$, $M = 1$, and $N_T = 6$.

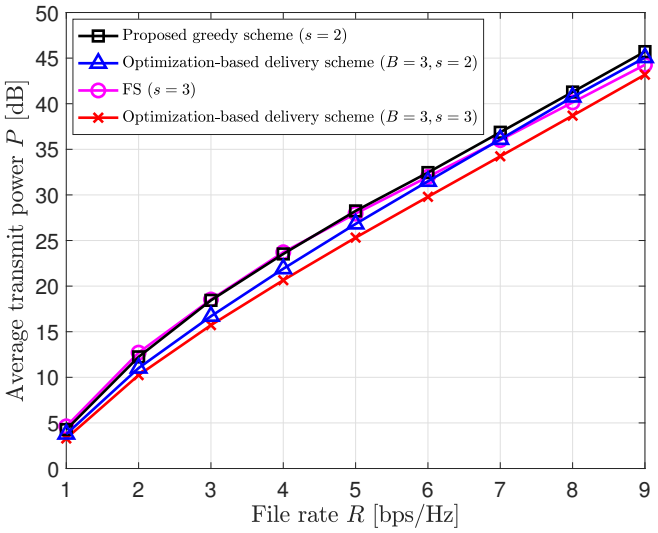



Figure 6: Average transmit power $P$ as a function of rate $R$ for $N = K = 4$, $M = 1$, and $N_T = 3$.

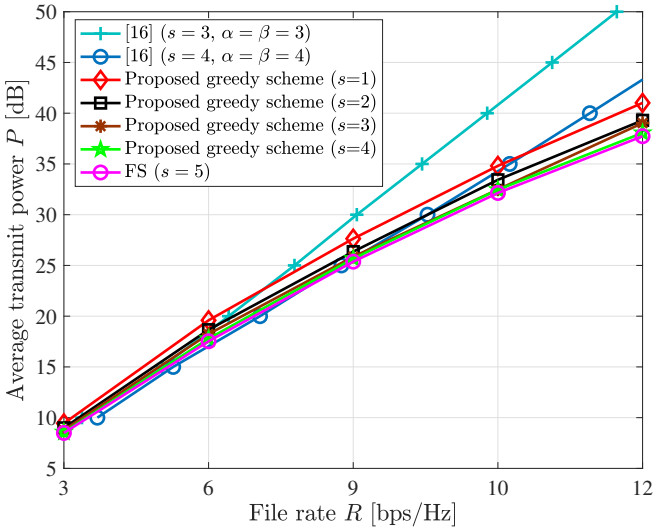



Figure 5: Average transmit power $P$ as a function of rate $R$ for $N = K = 6$, $M = 1$, and $N_T = 6$.

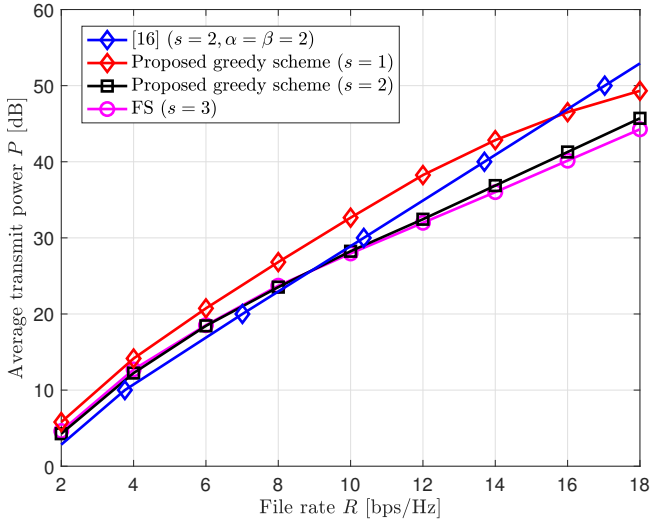



Figure 7: Average transmit power $P$ as a function of rate $R$ for $N = K = 4$, $M = 1$, and $N_T = 3$.

by solving the problem in (21) for $s = 2$ and $s = 3$, by setting $B = 3$. It is seen that the greedy scheme can achieve comparable performance, and the performance gap is small especially for high rates. The optimization-based content delivery scheme with $s = 3$ is found to outperform the one with $s = 2$ as expected, and the improvement is larger as the rate increases. In Fig. 7, it is interesting to see that when the rate is low, the scheme in [16] slightly outperforms both the FS and the proposed schemes. A similar observation has been made in [16] showing that a higher rate can be achieved when transmitting a smaller number of coded messages at low SNR, reducing the interference from coded packets that are not decoded at each user. Due to insufficient spatial degrees of freedom, both the FS and the proposed schemes fail to manage the interference between data streams. We conclude that this effect occurs only for low rates, as the benefit of superposition coding becomes more dominant at higher rates. We note that the greedy scheme coincides with the one in [16] for $s = 1$ in terms of the transmitted coded messages in each time slot, but this does not always happen. For instance, when $s = 2$, the only option in [16] to keep the same level of complexity is to serve 3 users in each time slot.

We plot in Fig. 8 the power loss of the proposed scheme in Algorithm 1 compared to FS as a function of $s$, that is, how much more the required transmit power is compared to the transmit power required by FS. This results in a more clear plot. Assuming $N = 6$, $K = 6$, $M = 1$, we let $s$ take values from $\{1, 2, 3, 4, 5\}$, where $s = 5$ corresponds to the

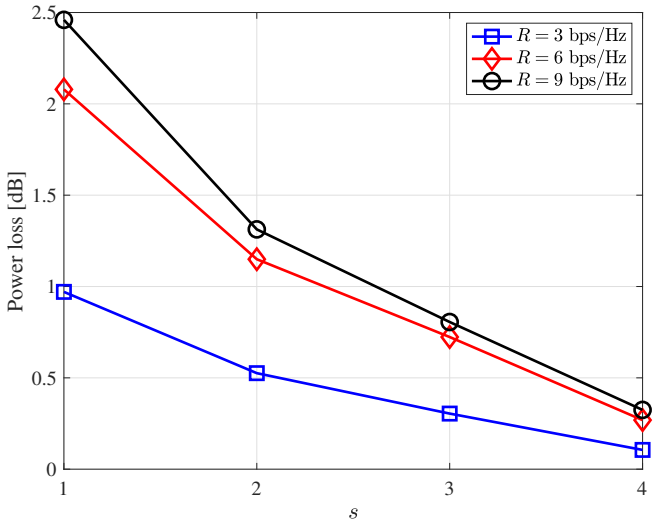



Figure 8: Power loss w.r.t. FS as a function of $s$ for $N = K = 6$, $M = 1$, and $N_T = 6$.

FS scheme in which all the $\binom{K}{t+1} = 15$ coded messages are transmitted simultaneously. FS requires less transmit power at the expense of a high decoding complexity. When $s = 1$, the model boils down to the single-cell multigroup multicasting problem, which has the lowest computation and implementation complexity. In general, Fig. 8 can be considered as the trade-off curve between the performance and complexity for each rate value, both of them increasing with $s$.

## VIII. Conclusions

In this paper, we have studied cache-aided content delivery from a multi-antenna BS in the finite SNR regime. We have formulated a general beamforming scheme that multicasts coded files over multiple orthogonal time slots. We have then specialized this general formulation to a low-complexity greedy scheme by limiting the number of coded messages targeted at each user at each time slot. This scheme provides the flexibility to adjust the computational complexity of the optimization problem and the receiver complexity. We have then formulated the constraint on the number of coded messages targeted at each user at each time slot as a sparsity constraint, and solved the resulting mixed-integer non-convex optimization problem using the SCA method. Compared with FS, where all the coded messages are transmitted simultaneously, and the scheme obtained via the sparsity-constrained optimization framework, the greedy scheme achieves comparable performance, and outperforms the one proposed in [16] for all values of SNR and rate with sufficient spatial degrees of freedom, while the improvement is limited to high data rate values when the BS does not have sufficiently many transmit antennas. Furthermore, the gap between the greedy delivery scheme and the optimization-based delivery scheme decreases as the SNR/power increases. When considering practical implementations, one must choose a suitable value of $s$ that yields an acceptable performance while keeping the complexity feasible. The satisfactory DoF performance of the proposed low-complexity scheme is guaranteed with at least $K - t$ antennas, while the analysis of overloaded systems with $K$ users served by less than $K - t$ antennas is left as future work.

## Appendix A
## An upper bound on the solution of Problem (12)

It is noted that the constrains are in the form of difference of convex functions, which can be approximated by linearizing the concave functions, resulting in a convex problem that can be solved via SCA techniques. To see this, we first rewrite the problem in (12) as

$$\min_{\{\boldsymbol{w}_T(i)\},\{R^T(i)\},\{\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)\}} \sum_{i=1}^{\binom{K}{t+1}} \frac{n_i}{n} \|\boldsymbol{w}_T(i)\|^2 \tag{30a}$$

$$\text{s.t.} \sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} R^T(i) \le \frac{n_i}{n}\log_2(1+\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)), \quad \forall\, \mathcal{J}_{\mathcal{S}_k}\supseteq\mathcal{S}_k, \forall k, i, \tag{30b}$$

$$\sum_{l\in\mathcal{S}_k^C} |\boldsymbol{h}_k^H\boldsymbol{w}_l(i)|^2 \le \frac{\sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} |\boldsymbol{h}_k^H\boldsymbol{w}_T(i)|^2}{\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)} + \sigma_k^2, \quad \forall\, \mathcal{J}_{\mathcal{S}_k}\supseteq\mathcal{S}_k, \forall k, i, \tag{30c}$$

(12c), (12d) and (12e);

where $\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i) \triangleq \sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} \gamma_k^{\mathcal{T}}(i)$. The constraint in (30c) is the difference of convex function, since $\sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} |\boldsymbol{h}_k^H\boldsymbol{w}_T(i)|^2 = \gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)$ is the sum of quadratic-over-linear functions of $\boldsymbol{w}_T(i)$ and $\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)$. Therefore, a sequence of convex subproblems can be solved iteratively to approximately tackle this convex-concave problem [28], with the subproblem in the $(\ell+1)$-th iteration given by

$$\min_{\{\boldsymbol{w}_T(i)\},\{R^T(i)\},\{\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)\}} \sum_{i=1}^{\binom{K}{t+1}} \frac{n_i}{n} \|\boldsymbol{w}_T(i)\|^2 \tag{31a}$$

$$\text{s.t.} \sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} R^T(i) \le \frac{n_i}{n}\log_2(1+\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)), \ \forall\, \mathcal{J}_{\mathcal{S}_k}\supseteq\mathcal{S}_k; \forall k, i; \tag{31b}$$

$$\sum_{l\in\mathcal{S}_k^C} |\boldsymbol{h}_k^H\boldsymbol{w}_l(i)|^2 + \frac{\sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} |\boldsymbol{h}_k^H\boldsymbol{w}_T(i)|^2}{\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)^2}\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i) - \frac{2\sum_{\mathcal{T}\supseteq\mathcal{J}_{\mathcal{S}_k}} \boldsymbol{w}_T^H(i)\boldsymbol{h}_k\boldsymbol{h}_k^H\boldsymbol{w}_T(i)}{\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)} \le \sigma_k^2, \quad \forall\, \mathcal{J}_{\mathcal{S}_k}\supseteq\mathcal{S}_k; \forall k, i; \tag{31c}$$

(12c), (12d) and (12e);

given the solution of $\boldsymbol{w}_T(i)$, $R^T(i)$, and $\gamma_{\mathcal{S}_k}^{\mathcal{J}}(i)$ obtained in the $\ell$-th SCA iteration. Each of the convex subproblems can be efficiently solved with standard interior-point algorithms or off-the-shelf solvers, and the SCA approach is guaranteed to converge to a stationary solution of the original problem in (12) [29]. Details of the SCA algorithm are outlined in Table. II.

An initial point in the feasible set of problem (12) is required to initialize the SCA algorithm. We first observe that for any feasible target rates $\{R_T^{[i]} : \forall T \supseteq \mathcal{S}_k^{[i]}\}_{i=1}^{B}$ that satisfy the constraints in (12c) and (12e), the problem in (12) can be decoupled and decomposed into $B$ parallel subproblems, each for a distinct time slot $i \in [B]$, given by

$$\{\mathbf{w}_T[i]\} = \arg\min_{\{\mathbf{w}_T[i]\}} \sum_{T \supseteq \mathcal{S}[i]} \|\mathbf{w}_T[i]\|^2 \tag{32a}$$

$$\text{s.t.} \sum_{T \supseteq \mathcal{S}_k^{[i]}} R_T^{[i]} \le \frac{n_i}{n} \log_2\Big(1 + \frac{1}{\gamma_k^{[i]}}\Big), \quad \forall \mathcal{S}_k^{[i]} \subseteq \mathcal{S}_k; \ \forall k, i; \tag{32b}$$

$$\gamma_k^{[i]} = \frac{|\mathbf{h}_k^H \mathbf{w}_T[i]|^2}{\sum_{l \in \mathcal{S}_k^c} |\mathbf{h}_k^H \mathbf{w}_l[i]|^2 + \sigma_k^2}, \ \forall k, i; \tag{32c}$$

$$\|\mathbf{w}_T[i]\|^2 \ge 0 \text{ for } \forall v_T[i] \ne 0, \tag{32d}$$

$$\|\mathbf{w}_T[i]\|^2 = 0 \text{ for } \forall v_T[i] = 0, \tag{32e}$$

which is nonconvex. Nevertheless, it can be transformed into a semidefinite programming problem by introducing $\mathbf{W}_T[i] \succeq \mathbf{w}_T[i]\mathbf{w}_T^H[i]$ and dropping the rank-1 constraints on $\mathbf{W}_T[i]$, which is given by

$$\{\mathbf{W}_T[i]\} = \arg\min_{\{\mathbf{W}_T[i]\}} \sum_{T \supseteq \mathcal{S}[i]} \mathrm{Tr}\{\mathbf{W}_T[i]\} \tag{33a}$$

$$\text{s.t.} \quad 2^{\frac{n}{n_i}\sum_{T \supseteq \mathcal{S}_k^{[i]}} R_T^{[i]}} - 1 \ge \sum_{l \in \mathcal{S}_k^c} \mathrm{Tr}\{\mathbf{H}_k \mathbf{W}_l[i]\} + \sigma_k^2 - \sum_{T \supseteq \mathcal{S}_k^{[i]}} \mathrm{Tr}\{\mathbf{H}_k \mathbf{W}_T[i]\} \ge 0; \ \forall \mathcal{S}_k^{[i]} \subseteq \mathcal{S}_k; \forall k; \tag{33b}$$

$$\mathbf{W}_T[i] \succeq 0; \forall v_T[i] \ne 0, \tag{33c}$$

$$\mathbf{W}_T[i] = 0; \forall v_T[i] = 0, \tag{33d}$$

and can be efficiently solved with standard interior-point algorithms. However, the solution obtained with semidefinite relaxation is not necessarily rank-1. If the obtained $\mathbf{W}_T[i]$'s are all rank-1, then the optimal solution of (32) can be readily recovered from $\mathbf{W}_T[i]$. Otherwise, Gaussian randomization can be adopted to obtain a feasible approximation to the optimal solution of (32). Note that the solution given by (32) is an upper bound on the minimum required power in (12) as the rates $\{R_T^{[i]} : \forall T \supseteq \mathcal{S}_k^{[i]}\}_{i=1}^{B}$ are not optimized, which hence can serve as an initial point in the successive convex approximation algorithm to obtain a tighter upper bound on the problem in (12).

Table II: SCA Algorithm for the Multicast Beamforming Problem with a Given Coded Delivery Scheme

STEP 0: Set $t = 1$. Set a step size $\epsilon$.
Initialize $\mathbf{w}_T[i]$, $R_T[i]$, and $\gamma_{\mathcal{S}_k^{[i]}}[i]$ with feasible values
STEP 1: If a stopping criterion is satisfied, then STOP
STEP 2: Solve the optimization problem in (31)
STEP 3: Update $\mathbf{w}_T^{t+1}[i] = \mathbf{w}_T[i] + \epsilon(\mathbf{w}_T[i] - \mathbf{w}_T[i])$,
$R_T^{t+1}[i] = R_T[i] + \epsilon(R_T[i] - R_T[i])$,
$\gamma_{\mathcal{S}_k^{[i]}}^{t+1}[i] = \gamma_{\mathcal{S}_k^{[i]}}[i] + \epsilon(\gamma_{\mathcal{S}_k^{[i]}}[i] - \gamma_{\mathcal{S}_k^{[i]}}[i])$,
STEP 4: Set $t = t + 1$, and go to STEP 1

## REFERENCES


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

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

[3] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6996–7016, Nov. 2018.

[4] M. Mohammadi Amiri and D. Gündüz, "Cache-aided content delivery over erasure broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 370–381, Jan. 2018.

[5] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," *CoRR*, vol. abs/1606.08253, 2016. [Online]. Available: http://arxiv.org/abs/1606.08253

[6] M. Mohammadi Amiri and D. Gündüz, "Caching and coded delivery over Gaussian broadcast channels for energy efficiency," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1706–1720, Aug. 2018.

[7] A. Ghorbel, K. Ngo, R. Combes, M. Kobayashi, and S. Yang, "Opportunistic content delivery in fading broadcast channels," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec. 2017, pp. 1–6.

[8] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[9] E. Karipidis, N. D. Sidiropoulos, and Z. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[10] Z. Xiang, M. Tao, and X. Wang, "Coordinated multicast beamforming in multicell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 12–21, Jan. 2013.

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

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

[13] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[14] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[15] M. Salehi, A. Tölli, and S. P. Shariatpanahi, "A multi-antenna coded caching scheme with linear subpacketization," *CoRR*, vol. abs/1910.10384, 2019. [Online]. Available: http://arxiv.org/abs/1910.10384

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

[17] E. Piovano, H. Joudeh, and B. Clerckx, "On coded caching in the overloaded MISO broadcast channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 2795–2799.

[18] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May. 2017.

[19] S. Zhong and X. Wang, "Joint multicast and unicast beamforming for coded caching," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3354–3367, Aug 2018.

[20] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827–2839, Apr. 2018.

[21] T. X. Vu, L. Lei, S. Chatzinotas, B. Ottersten, and T. A. Vu, "Energy efficient design for coded caching delivery phase," in *2019 3rd International Conference on Recent Advances in Signal Processing, Telecommunications Computing (SigTelCom)*, Mar. 2019, pp. 165–169.

[22] B. R. Marks and G. P. Wright, "A general inner approximation algorithm