Reflecting Intelligent Surfaces-Assisted Multiple-Antenna Coded Caching

Xiaofan Niu*, Minquan Cheng†, Kai Wan*, Robert Caiming Qiu*,
Giuseppe Caire§
*Huazhong University of Science and Technology, 430074 Wuhan, China,
{nxfan,kai_wan,caiming}@hust.edu.cn
† Guangxi Normal University, 541004 Guilin, China, chengqinshi@hotmail.com
§Technische Universität Berlin, 10587 Berlin, Germany, caire@tu-berlin.de

Abstract

Reconfigurable intelligent surface (RIS) has been treated as a core technique in improving wireless propagation environments for the next generation wireless communication systems. This paper proposes a new coded caching problem, referred to as Reconfigurable Intelligent Surface (RIS)-assisted multiple-antenna coded caching, which is composed of a server with multiple antennas and some single-antenna cache-aided users. Different from the existing multi-antenna coded caching problems, we introduce a passive RIS (with limited number of units) into the systems to further increase the multicast gain (i.e., degrees of freedom (DoF)) in the transmission, which is done by using RIS-assisted interference nulling. That is, by using RIS, we can 'erase' any path between one transmission antenna and one receive antenna. We first propose a new RIS-assisted interference nulling approach to search for the phase-shift coefficients of RIS for the sake of interference nulling, which converges faster than the state-of-the-art algorithm. After erasing some paths in each time slot, the delivery can be divided into several non-overlapping groups including transmission antennas and users, where in each group the transmission antennas serve the contained users without suffering interference from the transmissions by other groups. The division of groups for the sake of maximizing the DoF could be formulated into a combinatorial optimization problem. We propose a grouping algorithm which can find the optimal solution with low complexity, and the corresponding coded caching scheme achieving this DoF.

Index Terms

Coded caching, reconfigurable intelligent surface, zero-forcing

I. INTRODUCTION

Coded Caching was introduced to address content-related traffic congestion in networks. In addition to the local caching gain (i.e., when a user requires some content which has been already stored, the server does not need to transmit), the strategy of coded caching involves transmitting multicast messages and leveraging cached content to eliminate interference such that a single transmission can serve multiple users, resulting in a preferable multicast gain and showing a great potential to alleviate a significant part of the traffic. The original coded caching work was introduced by Maddah-Ali and Niesen (MN) in [1] for a Single Input Single Output (SISO) shared-link model, where a central server with access to a library containing N files connects to k cache-aided users via an error-free shared link. Each user is equipped with a cache of size M files. The coded caching process contains two phases: placement and delivery. During the placement phase, each user stores parts of content from each file without knowing future demands. During the delivery phase, each user requests a random file from the library. The server broadcasts coded messages based on users' caches and requests to satisfy all users' demands. The objective is to minimize the worst-case transmission load among all possible demands (or equivalently maximize the worst-cast coded caching gain). If each file is stored totally $t = KM/N$ times by users, the seminal MN coded caching scheme constructs multicast messages, each of which is useful to $t + 1$ users, i.e., the coded caching gain is $t + 1$ (the corresponding load is $(K - t)/(t + 1)$ where $K - t$ represents the local caching gain). It was shown in [2], [3] that under the constraint of uncoded cache placement (i.e., each user directly caches a subset of the library bits) and $\mathsf{N} \geq \mathsf{K}$, the achieved coded caching gain was proved to be optimal [2], [3].

Built upon the groundbreaking coded caching strategy, considerable research has been dedicated to wireless coded caching problems, with an emphasis on harnessing the benefits of coded caching gain. Multi-antenna coded caching problem was originally proposed in [4], where K_T transmitters with memory size M_T and K_R receivers with memory size M_R are connected through a wireless network. By smartly combing spatial multiplexing gain and coded caching gain, a multi-antenna coded caching scheme was proposed in [4] with a total degree-of-freedom (sum-DoF, i.e., overall multiplexing gain) $min\{\frac{K_T M_T + K_R M_R}{N}, K_R\}$, which was proved to be optimal under the constraint of uncoded cache placement and one-shot linear delivery. Following the work in [4], numerous works have been proposed for various multi-antenna coded caching problems to find out coded caching schemes to maximizing the multiplexing gain; just list a few works [4]–[10]. A new combinatorial structure on multi-antenna coded caching schemes with uncoded cache placement and one-shot zero-forcing delivery, referred to as Multiple-Antenna Placement Delivery Array (MAPDA), was proposed in [11], [12], by extending the placement delivery array (PDA) proposed in [13] for the original shared-link coded caching problem. Under this structure, designing a coded caching scheme could be transformed into designing an array satisfying some constraints, where the latter one is a combinatorial problem and could be solved by using combinatorial tools.

In this paper, we introduce Reconfigurable Intelligent Surfaces (RIS), seen as a promising technique for the next generation wireless systems to reconfigure wireless propagation environment [14]–[18], into coded caching systems in order to further increase the sum-DoF. As illustrated in Fig. 1, RIS serves as a passive configurable relay (by its phase-shift coefficients) in the channel, where the system channel transition matrix is equal to the product of the transition matrix from the transmitter to the RIS, the RIS phase-shift matrix, and the transition matrix from the RIS to the receiver. Several information theoretic works have been proposed on RIS-assisted interference management, in order to increase the sum-DoF. In [19] the authors considered RIS-assisted K-user interference channel. RIS-assisted interference alignment scheme was proposed for different numbers of RIS units. With the increasing of the RIS units, the sum-DoF can be increased from $K/2$ to K. Following the RIS-assisted K-user interference channel, the authors in [20] proposed a RIS-assisted interference nulling approach with zero-forcing transmission, which achieves the sum-DoF equal to K when the number of RIS units is over than a threshold approximately equal to $2K(K-1)$. This is done by using RIS to 'erase' all the interfering paths for the users. To find out the RIS phase-shift coefficients for interference nulling, an alternating projection algorithm was proposed in [20]. Then RIS-assisted zero-forcing and interference alignment were also extended to Multiple-Input-Multiple-Output (MIMO) networks in [21], [22], in order to increase the sum-DoF.

Main Contribution: We propose a RIS-assisted mutiple-input single-output (MISO) broadcast coded caching system, for fixed numbers of antennas at the transmitter and of RIS units. Our main contribution on constructing RIS-assisted multi-antenna coded caching schemes is as follows.

- We first propose a new RIS-assisted interference nulling approach to 'erase' some paths in the wireless channel, which outputs the needed RIS phase-shift coefficients with a faster convergence rate than the algorithm in [20].

- With the help of RIS-assisted interference nulling, we can divide the transmission in each time slot of the delivery phase into multiple groups including transmission antennas and users, where in each group the transmission antennas serve the contained users without suffering interference from the transmissions by other groups. By formulating the grouping optimization to maximize the sum-DoF as a combinatorial optimization problem, we propose an low-complexity algorithm to find out the optimal grouping method.
- After determining a new combinatorial structure, we design the two-phase coded caching scheme under the new struture, referred to as RMAPDA, by adding a new constraint on group-based delivery.

Notation Convention: Scalars are denoted by lowercase letters, vectors by bold lowercase letters, and matrices by bold uppercase letters. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^T$ represents the transpose of the matrix, $\boldsymbol{A}^*$ represents the conjugate. $|v|$ denotes the magnitude of a complex scalar. "Re$[\cdot]$" denotes taking the real part of a complex number. "$\cdot$" represents the element-wise multiplication of matrices. "$\odot$" indicates the Hadamard product. For a vector $\mathbf{v}$, $diag(\mathbf{v})$ represents the diagonal matrix whole diagonal elements are the elements in $\mathbf{v}$. Define that $[a] = \{1, 2, \ldots, a\}$.

II. System Model and Preliminary Results

A. System model

As illustrated in Fig. 1, we consider an (L, G, K, M, N) RIS-assisted MISO broadcast coded caching system. The RIS assists the communication, while the transmission follows the one-shot linear delivery scheme proposed in [4]. A server has L transmission antennas and can access to a library containing N files, denoted by $\mathcal{W} = \{\mathbf{W}_n \mid n \in [N]\}$. Each file $\mathbf{W}_n$ consists of F packets, defined as $\mathbf{W}_n \triangleq \{\mathbf{W}_{n,f} \mid f \in [F]\}$. Each user is equipped with one antenna and a cache capable of storing MF packets, where $0 \leq M \leq N$.

A coded caching scheme consists of two phases:

Placement Phase: During this phase, F data packets of each file are strategically placed into the users' caches without knowledge of the specific demands.

Delivery Phase: Each user $k \in [K]$ requests an arbitrary file $\boldsymbol{W}_{d_k}$ from the library, where $d_k \in [N]$. The required vector is denoted as $\boldsymbol{d} = (d_1, d_2, \ldots, d_K)$, we first encode each packet of desired files by Gaussian encoding with rate $\log P$ to obtain a coded packet, where P represents

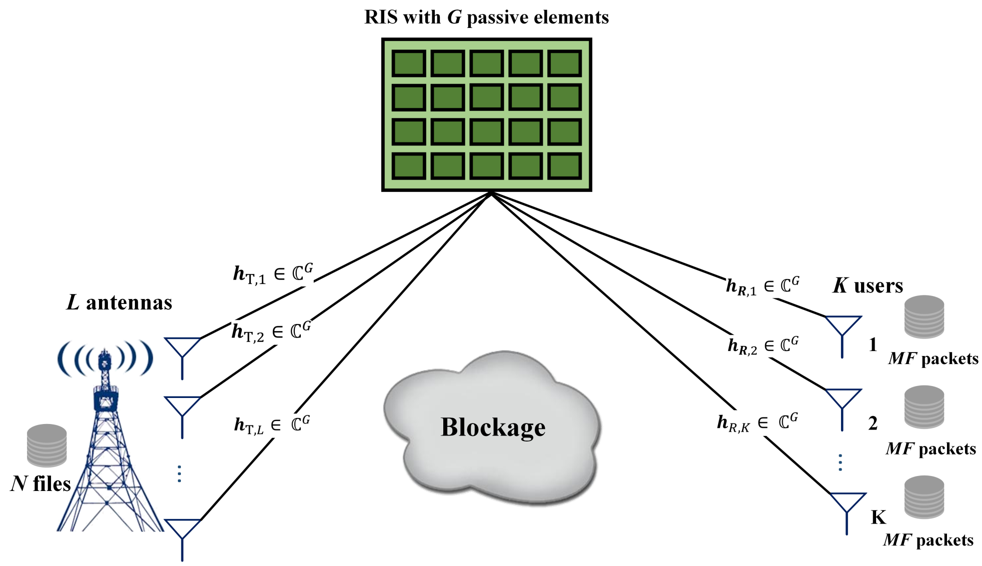

Fig. 1: RIS-assisted MISO broadcast system.

the Signal-to-Noise (SNR) ratio. If P is large enough, each coded packet carries one Degree-of Freedom (DoF). Then each packet is encoded as $\widetilde{W}_{n,f}$.

The communication process contains S time slots, each denoted by $s \in [S]$. In time slot s, the sever transmits a subset of requested packets, denoted by $\mathcal{D}_s = \{\widetilde{W}_{d_{k_1},f_1}, \widetilde{W}_{d_{k_2},f_2} \dots \widetilde{W}_{d_{k_{r_s}},f_{r_s}}\}$ desired by r_s users in $\mathcal{K}_s \subseteq [K]$. The signal transmitted by antenna i is denoted by $X_i(s)$, which is a linear combination of some coded packets given by

$$X_i(s) = \sum_{j \in [r_s]} m_{i,k_j}(s) \widetilde{W}_{d_{k_j},f_j} \tag{1}$$

where each $m_{i,k_j}(s)$ is a scalar complex coefficient in the precoding matrix to be designed.

In the wireless channel, there exists a passive RIS with G units. For the ease of description, we assume that the direct paths between the transmitters and receivers are blocked. As shown in [20], the RIS-assisted interference nulling method proposed in this case of no direct path could be directly extended to 'erase' the path between a transmission antenna and a user for the systems where direct paths exist. Define $\boldsymbol{h}_{T,j}$ as the channel transition vector with dimension G between the j-th transmission antenna (or equivalently called the j-th transmitter) and the RIS, and define $\boldsymbol{h}_{R,k}$ as the channel transition vector with dimension G between the RIS and the k-th receiver. It is assumed that the coefficients in these vectors are i.i.d over some continuous distributed and that channel state information (CSI) is fully known to the server and users. The reflection coefficients of the RIS are given by $\boldsymbol{v} = [e^{j\omega_1}, e^{j\omega_2}, \cdots, e^{j\omega_G}]^T \in \mathbb{C}^G$, where $\omega_i \in (0, 2\pi)$ indicates the phase-shift coefficient of the i-th unit of the RIS. Note that the phase-shift vector $\boldsymbol{v}$ could be re-configured during each time slot. Each receiver k receives a reflected signal from the RIS at time slot s, which can be expressed as

$$Y_k(s) = \sum_{j=1}^{L} \boldsymbol{h}_{R,k}^T diag(\boldsymbol{v}) \boldsymbol{h}_{T,j} X_j(s) + n_k(s) \tag{2a}$$

$$= \sum_{j=1}^{L} \boldsymbol{a}_{k,j}^T \boldsymbol{v} X_j(s) + n_k(s), \tag{2b}$$

where $\boldsymbol{a}_{k,j} \triangleq diag(\boldsymbol{h}_{T,j})\boldsymbol{h}_{R,k}$ and $n_k(s)$ represents the additive Gaussian white noise at receiver k. Assume that the delivery is one-shot: in time slot s, each user $k \in \mathcal{K}_s$ should recover one coded packet from $Y_k(s)$ and its cache content.

Objective: Consider the sum-DoF as the metric. Under the one-shot linear delivery, the sum-DoF indicates the average number of users served per time slot. Our objective is to maximize the worst-case sum-DoF among all possible demands.

B. MAPDA

Definition 1 (MAPDA, [11])**.** For any positive integers L, K, F, Z, and S, an $F \times K$ array $\mathbf{P} = (\mathbf{P}(j,k))_{j\in[F],k\in[K]}$ composed of "$*$" and $[S]$ is called an (L, K, F, Z, S) multiple-antenna placement delivery array (MAPDA) if it satisfies the following conditions:[1]

C1: The symbol $*$ appears Z times in each column;

C2: Each integer $s \in [S]$ occurs at least once in the array;

C3: Each integer s appears at most once in each column;

C4: For any integer $s \in [S]$, define $\mathbf{P}^{(s)}$ to be the subarray of $\mathbf{P}$ including the rows and columns containing s, and let $r_s' \times r_s$ denote the dimensions of $\mathbf{P}^{(s)}$. The number of integer entries in each row of $\mathbf{P}^{(s)}$ is less than or equal to $\min\{L, K\}$, i.e.,

$$\left|\{k_1 \in [r_s] \mid \mathbf{P}^{(s)}(f_1, k_1) \in [S]\}\right| \leq \min\{L, K\}, \quad \forall f_1 \in [r_s'].$$

If each integer appears g times in MAPDA $\mathbf{P}$, then $\mathbf{P}$ is a g-regular MAPDA, denoted by g-(L, K, F, Z, S) MAPDA.

Each MAPDA corresponds to a specific coded caching scheme where the cache ratio $\frac{M}{N}$ equals $\frac{Z}{F}$. This process is summarized in Algorithm 1.

Note that when $L = 1$ the MAPDA reduces to the PDA in [13]. Given a g-MAPDA, we can obtain a multi-antenna coded caching scheme for the system containing a server with L antenas

[1]An array is composed of $*$ and integers. Each column represents a user and each row represents a packet of files. A $*$ appears in the position (i, j) of the array, meaning that user j stores the i-th packet of each file. Each integer corresponds to a multicast message transmitted in one time slot.

Algorithm 1 Caching Scheme Based on MAPDA in [11]

procedure PLACEMENT(P, $\mathcal{W}$)
    Split each file $W_n \in W$ into F packets, i.e., $W_n = \{W_{n,f} | f = 1, 2, \ldots, F\}$.
    for $k \in [K]$ **do**
        $Z_k \leftarrow \{W_{n,f} | \mathrm{P}(f,k) = *, n \in [N], f \in [F]\}$
    end for
end procedure
procedure DELIVERY(P, $\mathcal{W}, d$)
    for $s = 1, 2, \ldots, S$ **do**
        Server uses L antennas to send $\mathbf{W}_{d_k,j}$ where $\mathrm{P}(j,k) = s$ to the users.
    end for
end procedure

and K single-antenna users with $M/N = Z/F$, which has sum-DoF g. We review some PDA and MAPDA that will be used later.

Lemma 1 (Maximum DoF [11])**.** *Under the MAPDA structure, the maximum achievable sum-DoF is* $\min\{K, L + KM/N\}$*, when* KM/N *is an integer.*

Lemma 2 (MN PDA [1])**.** *For any positive integers* K *and* t *with* $t < K$*, there exists a* $(K, \binom{[K]}{t}, \binom{K-1}{t-1}, \binom{K}{t+1})$ *PDA with the maximum DoF* $t+1$

Construction 1 (MN PDA [1])**.** For any integer t within the set $[K]$, we can construct a $(K, \binom{K}{t}, \binom{K-1}{t-1}, \binom{K}{t+1})$ PDA called MN PDA $\mathbf{P} = (\mathbf{P}(\mathcal{T}, k)_{\mathcal{T} \subseteq \binom{[K]}{t}, k \in [K]})$ with the sum-DoF $t+1$. For each $\mathcal{T} \subseteq \binom{[K]}{t}$ and $k \in [K]$, the entry of P is defined as follows

$$\mathbf{P}(\mathcal{T}, k) = \begin{cases} * & \text{if } k \in \mathcal{T}, \\ \mathcal{T} \cup \{k\} & \text{otherwise.} \end{cases}$$

Example 1. For $K = 7, t = 1$, we have a $(7, 7, 1, 21)$ MN PDA in Table I.

	1	2	3	4	5	6	7
1	$*$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$	$\{1,6\}$	$\{1,7\}$
2	$\{1,2\}$	$*$	$\{2,3\}$	$\{2,4\}$	$\{2,5\}$	$\{2,6\}$	$\{2,7\}$
3	$\{1,3\}$	$\{2,3\}$	$*$	$\{3,4\}$	$\{3,5\}$	$\{3,6\}$	$\{3,7\}$
4	$\{1,4\}$	$\{2,4\}$	$\{3,4\}$	$*$	$\{4,5\}$	$\{4,6\}$	$\{4,7\}$
5	$\{1,5\}$	$\{2,5\}$	$\{3,5\}$	$\{4,5\}$	$*$	$\{5,6\}$	$\{5,7\}$
6	$\{1,6\}$	$\{2,6\}$	$\{3,6\}$	$\{4,6\}$	$\{5,6\}$	$*$	$\{6,7\}$
7	$\{1,7\}$	$\{2,7\}$	$\{3,7\}$	$\{4,7\}$	$\{5,7\}$	$\{6,7\}$	$*$

TABLE I: A $(7, 7, 1, 21)$ MN PDA

	1	2	3	4	5	6	7
1	*	{1,2,3}	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
2	{1,2,3}	*	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
3	{1,2,3}	{1,2,3}	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}
4	{1,2,4}	{1,2,4}	{1,3,4}	*	{1,4,5}	{1,4,6}	{1,4,7}
5	{1,2,5}	{1,2,5}	{1,3,5}	{1,4,5}	*	{1,5,6}	{1,5,7}
6	{1,2,6}	{1,2,6}	{1,3,6}	{1,4,6}	{1,5,6}	*	{1,6,7}
7	{1,2,7}	{1,2,7}	{1,3,7}	{1,4,7}	{1,5,7}	{1,6,7}	*
1	*	{1,2,4}	{1,3,4}	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}
2	{1,2,4}	*	{2,3,4}	{2,3,4}	{2,3,5}	{2,3,6}	{2,3,7}
3	{1,3,4}	{2,3,4}	*	{2,3,4}	{2,3,5}	{2,3,6}	{2,3,7}
4	{1,3,4}	{2,3,4}	{2,3,4}	*	{2,4,5}	{2,4,6}	{2,4,7}
5	{1,3,5}	{2,3,5}	{2,3,5}	{2,4,5}	*	{2,5,6}	{2,5,7}
6	{1,3,6}	{2,3,6}	{2,3,6}	{2,4,6}	{2,5,6}	*	{2,6,7}
7	{1,3,7}	{2,3,7}	{2,3,7}	{2,4,7}	{2,5,7}	{2,6,7}	*
1	*	{1,2,5}	{1,3,5}	{1,4,5}	{1,4,5}	{1,4,6}	{1,4,7}
2	{1,2,5}	*	{2,3,5}	{2,4,5}	{2,4,5}	{2,4,6}	{2,4,7}
3	{1,3,5}	{2,3,5}	*	{3,4,5}	{3,4,5}	{3,4,6}	{3,4,7}
4	{1,4,5}	{2,4,5}	{3,4,5}	*	{3,4,5}	{3,4,6}	{3,4,7}
5	{1,4,5}	{2,4,5}	{3,4,5}	{3,4,5}	*	{3,5,6}	{3,5,7}
6	{1,4,6}	{2,4,6}	{3,4,6}	{3,4,6}	{3,5,6}	*	{3,6,7}
7	{1,4,7}	{2,4,7}	{3,4,7}	{3,4,7}	{3,5,7}	{3,6,7}	*
1	*	{1,2,6}	{1,3,6}	{1,4,6}	{1,5,6}	{1,5,6}	{1,5,7}
2	{1,2,6}	*	{2,3,6}	{2,4,6}	{2,5,6}	{2,5,6}	{2,5,7}
3	{1,3,6}	{2,3,6}	*	{3,4,6}	{3,5,6}	{3,5,6}	{3,5,7}
4	{1,4,6}	{2,4,6}	{3,4,6}	*	{4,5,6}	{4,5,6}	{4,5,7}
5	{1,5,6}	{2,5,6}	{3,5,6}	{4,5,6}	*	{4,5,6}	{4,5,7}
6	{1,5,6}	{2,5,6}	{3,5,6}	{4,5,6}	{4,5,6}	*	{4,6,7}
7	{1,5,7}	{2,5,7}	{3,5,7}	{4,5,7}	{4,5,7}	{4,6,7}	*
1	*	{1,2,7}	{1,3,7}	{1,4,7}	{1,5,7}	{1,6,7}	{1,6,7}
2	{1,2,7}	*	{2,3,7}	{2,4,7}	{2,5,7}	{2,6,7}	{2,6,7}
3	{1,3,7}	{2,3,7}	*	{3,4,7}	{3,5,7}	{3,6,7}	{3,6,7}
4	{1,4,7}	{2,4,7}	{3,4,7}	*	{4,5,7}	{4,6,7}	{4,6,7}
5	{1,5,7}	{2,5,7}	{3,5,7}	{4,5,7}	*	{5,6,7}	{5,6,7}
6	{1,6,7}	{2,6,7}	{3,6,7}	{4,6,7}	{5,6,7}	*	{5,6,7}
7	{1,6,7}	{2,6,7}	{3,6,7}	{4,6,7}	{5,6,7}	{5,6,7}	*

TABLE II: A $(2,7,35,70)$ MS MAPDA

Lemma 3 (MS MAPDA [6]). *For any positive integers K and t, there exists a $(L, K, \binom{K}{t}\binom{K-t-1}{L-1}, \binom{K-1}{t-1}\binom{K-t-1}{L-1}, \binom{K}{t+L}\binom{t+L-1}{t})$ MAPDA with the maximum DoF $t+L$.*

Construction 2 (MS MAPDA [6]). For any integers t, L and K with $t+L \in [K]$, we can construct a $(L, K, \binom{K}{t}\binom{K-t-1}{L-1}, \binom{K-1}{t-1}\binom{K-t-1}{L-1}, \binom{K}{t+L}\binom{t+L-1}{t})$ MAPDA with the sum-DoF $t+L$ called MS MAPDA $\mathbf{P} = (\mathbf{P}(\mathcal{T}, \mathcal{L}), k))$ where $\mathcal{T} \in \binom{[K]}{t}, \mathcal{L} \in \binom{[K-t-1]}{L-1}, k \in [K]$, whose dimension is $\binom{K}{t}\binom{K-t-1}{L-1} \times K$:

$$\mathbf{P}(\mathcal{T}, \mathcal{L}), k) = \begin{cases} * & \text{if } k \in \mathcal{T}, \\ (\mathcal{S}(\mathcal{T}, \mathcal{L}, k), \text{order}(\mathcal{S}(\mathcal{T}, \mathcal{L}, k)) & \text{otherwise,} \end{cases}$$

where $\mathcal{S}(\mathcal{T}, \mathcal{L}, k) = \mathcal{T} \cup ([K] \setminus (\mathcal{T} \cup \{k\})[\mathcal{L}] \cup \{k\}$ and order$(\mathcal{S})$ as the order of appearance of the set $\mathcal{S}$ in each column.

Example 2. For $K = 7, t = 1, L = 2$,we have a $(2, 7, 35, 70)$ MS MAPDA in Table II.

III. Improved RIS-assisted Interference Nulling Algorithm

In this section, we will propose a new RIS-assisted interference nulling algorithm to find out the phase-shift vector to eliminate some paths in the channel. When we want to eliminate the path from the j-th transmission antenna to user k, we need to have $\boldsymbol{a}_{k,j}^T \boldsymbol{v} = 0$ in (2b). It was shown in [20] that to eliminate p paths, the number of RIS units needs to be slightly larger than $2p$. In this paper, for ease of presentation, we assume that the elimination p paths requires $2p$ RIS units.

Assume that for user k, we want to eliminate the paths between itself and antennas $j_{k,1}, j_{k,2}, \ldots,$ j_{k,q_k} where $j_q \in [L]$, and q_k denotes the total number of paths that user k needs to eliminate. Considering all users, the total number of paths needed to eliminate is $\sum_{k\in[K]} q_k$, which requires slighetly more than $G = 2\sum_{k\in[K]} q_k$. Subsequently, we will assume that this is approximately equal to this value.

Define that

$$\boldsymbol{A} = [\boldsymbol{a}_{1,j_{1,1}}, \ldots, \boldsymbol{a}_{1,j_{1,q_1}}, \boldsymbol{a}_{2,j_{2,1}}, \ldots, \boldsymbol{a}_{K,j_{K,q_K}}]. \tag{3}$$

Then our aim can be rewritten as

$$\mathcal{S}_1 = \{\boldsymbol{A}^T \boldsymbol{v} = 0\}. \tag{4}$$

Since the RIS is passive, it is essential to also ensure that

$$\mathcal{S}_2 = \{|v_i| = 1\}. \tag{5}$$

The algorithm presented in [20] operates by performing alternating projections between these two sets, i.e.

$$\Pi_{\boldsymbol{\mathcal{S}}_1}(\boldsymbol{v}) = \boldsymbol{v} - \boldsymbol{A}^*(\boldsymbol{A}^T\boldsymbol{A}^*)^{-1}\boldsymbol{A}^T\boldsymbol{v} \tag{6}$$

$$\Pi_{\boldsymbol{\mathcal{S}}_2}(\boldsymbol{v}) = \frac{\boldsymbol{v}}{|\boldsymbol{v}|} \tag{7}$$

thereby gradually converging to the intersection of these sets.

Different from the alternative projection algorithm in [20], we adopt the approach of projecting onto the tangent space to accelerate convergence rate which increases the step size of each projection to speeds up the convergence. The pseudo code of the improved algorithm is given in Algorithm 2.

Algorithm 2 Improved Alternating Projection Algorithm

1: Initialize reflection coefficients $\boldsymbol{v}$, channel matrix $\boldsymbol{A}_k$, and number of iterations m
2: Start with initial value $\boldsymbol{v} = \boldsymbol{v}^0 \in \boldsymbol{S}_2$
3: **for** $t = 0, 1, 2, \ldots, m$ **and** interference not nullified **do**
   $\boldsymbol{y}^t = \boldsymbol{v}^t - \Pi_{\boldsymbol{S}_1}(\boldsymbol{v}^t)$
   $\boldsymbol{m}^t = \boldsymbol{y}^t - \Pi_{\boldsymbol{v}^t}(\boldsymbol{y}^t)$
   $\tilde{\boldsymbol{v}}^t = \boldsymbol{v}^t - 2\boldsymbol{m}^t$
   $\boldsymbol{v}^{t+1} = \Pi_{\boldsymbol{S}_2}(\tilde{\boldsymbol{v}}^t)$
4: **end for**
5: **if** iteration limit reached or interference nullified **then**
   Stop iteration
6: **end if**
7: Output the resulting reflection coefficients $\boldsymbol{v}^{t+1}$

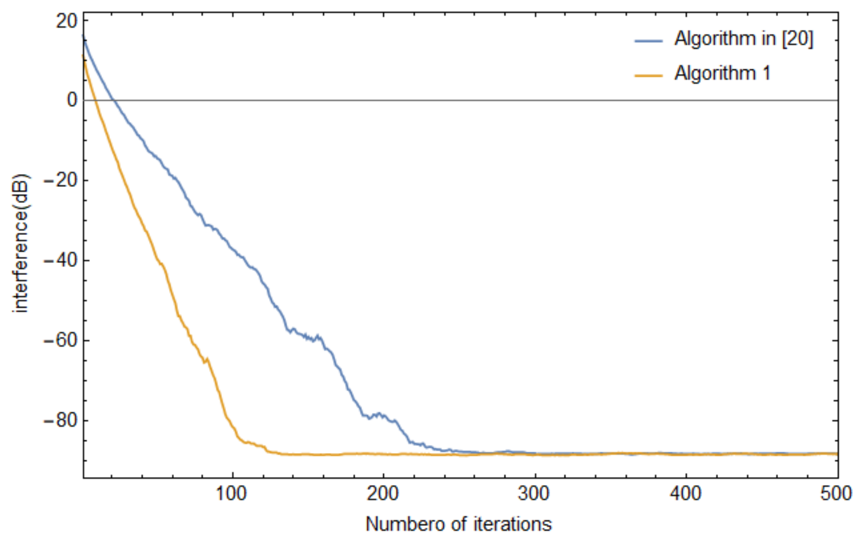

Fig. 2: Comparison of two algorithms in interference

In Algorithm 2, $\boldsymbol{y} = \boldsymbol{A}^*(\boldsymbol{A}^T\boldsymbol{A}^*)^{-1}\boldsymbol{A}^T\boldsymbol{v}$ represents the projection of $\boldsymbol{v}$ onto the orthogonal complement space of set $\boldsymbol{\mathcal{S}}_1$. Then we project vector $\boldsymbol{y}$ onto $\boldsymbol{v}$. i.e. $\boldsymbol{m} = \boldsymbol{y} - \mathrm{Re}(\boldsymbol{v}^* \odot \boldsymbol{y}) \cdot \boldsymbol{v}$. Subtract $2\boldsymbol{m}$ from $\boldsymbol{v}$, where the coefficient 2 is to achieve a larger step size. In essence, the operation still entails subtracting the direction that is orthogonal to the set $\boldsymbol{\mathcal{S}}_1$.[2] After that, $\boldsymbol{v}$ is projected onto $\boldsymbol{\mathcal{S}}_2$, resulting in a larger step size compared to the projection onto $\boldsymbol{\mathcal{S}}_2$ as described in (7).

Simulation: Next we compare Algorithm 2 with the RIS-assisted interference nulling algorithm in [20], in the K-transmitter and K-user interference channel. In this channel, we need

[2]This step is the primary difference from the algorithm presented in [20], namely, projecting onto the tangent plane of the current $\boldsymbol{v}$ prior to projecting onto the set $\boldsymbol{\mathcal{S}}_2$.

to use RIS eliminate $K-1$ inteference paths for each user; thus there are totally $K(K-1)$ paths to eliminate. Assume that $K=10$ and the number of RIS units G is 300. The channel coefficients are i.i.d. over a standard normal distribution. The number of algorithm's iterations is set to be 500. For each iteration, we compute the interference power at users in dB, defined as $\sum_{j\neq k}|\boldsymbol{a}_{k,j}\boldsymbol{v}|^2$.

It can be seen from Fig. 2 that Algorithm 2 performs significantly better than the algorithm in [20] in terms of interference power with the same number of iterations.

IV. Optimal Grouping Approach

By RIS-assisted interference nulling method proposed in the last section, we can divide the transmission in each time slot of the delivery phase into multiple groups; inside each group the contained antennas serve the contained users without suffering interference from the transmissions by other groups. In this section, we aim to find out the optimal grouping method leading to the maximum sum-DoF.

Given the total number of antennas L, memory ration M/N and K users, assume that the number of active antennas is $L_0 \leq L$, i.e., we can keep some antenna 'silent'. L_0 antennas are divided into r groups each of which contains L_i antennas, $\sum_{i=1}^{r} L_i = L_0$. Consider an integer $t = KM/N$. By Lemma 1 and the grouping method, we have the following sum-DoF:[3]

$$g = \sum_{i=1}^{r}(L_i + t) = L_0 + tr. \tag{8}$$

Since the number of RIS elements can be expressed as twice the number of interference paths and the maximum of users served by each group is $L_i + t$, the number of RIS units G needed is at least

$$G = 2\sum_{i=1}^{r}(L_i + t)(L_0 - L_i). \tag{9}$$

We consider an alternative optimization task, which is more feasible: given a task sum DoF g, we want to minimize the needed number of RIS units; i.e.,

$$\begin{aligned} \underset{L_i \in \mathbb{Z}^+}{\text{minimize}} \quad & G \\ \text{subject to} \quad & L_0 \leq L, r \leq L, g = L_0 + tr. \end{aligned} \tag{10}$$

[3] For the ease of description, we assume that g in (8) is no larger than K.

Algorithm 3 Optimal Grouping Scheme

Input: Initialize total antennas L, t, sum-DoF g
Compute initial $L_{\text{min}} = \lceil \frac{g}{t+1} \rceil$ and $L_{\text{max}} = \min(L, g-t)$
while $g \leq L(t+1)$ **do**
    valid_found = false
    for $L_0 = L_{\text{max}}$ **down to** L_{min} **do**
        Compute $r = \frac{g-L_0}{t}$
        if r is an integer and $r > 0$ **then**
            valid_found = true
            break
        end if
    end for
    if valid_found then
        break
    else
        $g \leftarrow g + 1$
    end if
end while
Calculate $\lambda_0 =$ solve for $f(\lambda) = 0$ and $\lambda_0 \neq 0$
Compute maximum $\lambda_{\text{max}} = \left\lfloor \frac{L_{\text{max}} - r}{t+1} \right\rfloor$
if $\lambda_{\text{max}} \geq \lambda_0$ **then**
    $L_{\text{opt}} = L_{\text{max}} - \lambda_{\text{max}} \cdot t$
    $r_{\text{opt}} = r + \lambda_{\text{max}}$
    $G_{\text{opt}} = G + f(\lambda_{\text{max}})$
else
    $L_{\text{opt}} = L_{\text{max}}$, $r_{\text{opt}} = r$, $G_{\text{opt}} = G$
end if
Output: $L_{\text{opt}}, r_{\text{opt}}, G_{\text{opt}}$

Theorem 1. *Given the sum-DoF g, the number of groups r and the number of active antennas L_0, the grouping scheme with minimum number of RIS elements is that $L_1 = L_0 - (r-1)$ and $L_2 = \cdots = L_r = 1$. The minimum of RIS elements is*

$$G = 2((r-1)[(t+2)L_0 - r). \tag{11}$$

The proof of Theorem 1 (i.e., the proof of the closed-form solution for the optimization problem in (10)) could be found in Appendix A. By Theorem 1, we only need to determine r and L_0 that minimize (11).

Algorithm 3 can find the optimal grouping scheme with complexity up to $O(L^2)$. or the case that there exists a scheme where the sum-DoF exactly equal to the given g, Algorithm 3 can find the optimal grouping scheme with complexity up to $O(L)$. Detailed analysis and design process of it can be find in Appendix B.

V. RIS-Assisted Grouping Coded Caching Scheme

After finding the solution for the grouping optimization problem, we then propose how to construct a coded caching scheme based on the L_{opt} and r_{opt}. Recall that by grouping, the transmission for each time slot could be divided into multiple separate transmissions, each of which is for one group.

Inspired by MAPDA, our scheme can also be represented in the form of an array. Similar to MAPDA, the symbol "$*$" is used to denote cached content at the users. For transmission, our objective is to enable simultaneous broadcasting across r groups.

From our previous analyses, the antenna groups are limited to two configurations: one consisting of $L_1 = L_{opt} - r_{opt} + 1$ antennas with $t + L_1$ users, and the other consisting of $L_2 = L_3 = \cdots = L_{r_{opt}=1}$ antenna with $t+1$ users. So our objective is to serve $g = (t+1)(r_{opt}-1) + (t + L_{opt} - r_{opt} + 1)$ users per time slot. Besides meeting requirements C1, C2, C3 in definition 1, it must also satisfy the following condition.

C4: For any integer $s \in [S]$, define $\mathbf{P}_1^{(s)}, \mathbf{P}_2^{(s)}, \ldots, \mathbf{P}_r^{(s)}$ to be the subarrays of $\mathbf{P}$ including the rows and columns containing s. These subarrays are mutually disjoint in columns. One of the subarrays has integer entries in each row less than or equal to $L - r + 1$, while the remaining $r - 1$ subarrays each have integer entries in each row no more than 1.

The new structure proposed for the RIS-assisted MISO coded caching problem is called RMAPDA.

Definition 2 (RMAPDA)**.** For any positive integers L, K, F, Z, r and S, an $F \times K$ array P composed of "$*$" and $[S]$ is called an (L, K, F, Z, r, S) multiple-antenna placement delivery array (MAPDA) if it satisfies the following conditions:

C1: The symbol "$*$" appears Z times in each column;

C2: Each integer from 1 to S occurs at least once in the array;

C3: Each integer s appears at most once in each column;

C4: For any integer $s \in [S]$, define $\mathbf{P}_1^{(s)}, \mathbf{P}_2^{(s)}, \ldots, \mathbf{P}_r^{(s)}$ to be the subarrays of $\mathbf{P}$ including the rows and columns containing s. These subarrays are mutually disjoint in columns. One of the subarrays has integer entries in each row less than or equal to $L - r + 1$, while the remaining $r - 1$ subarrays each have integer entries in each row no more than 1.

We are now equipped to begin the construction of a RMAPDA that fulfills the specified condition based on the MN PDA in lemma 2 and MS MAPDA in lemma 3. For simplicity, let

L' represent the number of antennas in a multi-antenna group, i.e., L_1 and denote the optimal number of active antennas L_{opt} as L.

Given that under the one-shot transmission strategy, the DoF $g \leq K$, it is sufficient to study cases where $g \leq K$.

In each MN PDA, groups of $t+1$ users require only one time slot each for data handling. In addition, in the MAPDA, each group of $t+L'$ users necessitates $\binom{t+L'-1}{t}$ time slots. To ensure that the number of time slots across $r-1$ MN PDAs and a MAPDA matches, we replicate them vertically until the total number of slots aligns.

Let $L' = L - r + 1$ and the number of antenna in lemma 3 is L'.

First, assume each MS MAPDA is replicated vertically x times, and the MN PDA is replicated vertically y times. To confirm that the time slots of a $t+L'$ users group is the same of the $r-1$ groups with $t+1$ users, x and y should satisfied:

$$x\binom{t+L'-1}{t} = y'\binom{K-(t+L')}{(r-1)(t+1)} \times \frac{[(r-1)(t+1)]!}{((t+1)!)^{(r-1)}(r-1)!} \tag{12}$$

A PDA copied y' times only satisfies the matching of a single multi-user group and the corresponding single-user group. To ensure that each multi-user group can be matched, the copied PDA should be copied n more times, where

$$n = \binom{K-t-1}{t+L'}\binom{K-(t+1)-(t+L')}{(r-2)(t+1)} \times \frac{[(r-2)(t+1)]!}{((t+1)!)^{(r-2)}(r-2)!} \tag{13}$$

Therefore, the total number y of times the PDA is ultimately copied is

$$y = ny' \tag{14}$$

The subpacketization F is

$$F = F_1 + F_2 = x\binom{K}{t}\binom{K-t-1}{L-1} + y\binom{K}{t} \tag{15}$$

The construction process is summarized as Algorithm 4. Then we will propose an example to illustrate the construction process in Algorithm 4 in detail.

Example 3. Consider a system with $K = 7$, $L = 4$, $M = 1$, $t = KM/N = 1$, dividing antennas into $r = 3$ groups, one group with $L' = 2$ antennas and the other two groups both have only one antenna. Assume each MS MAPDA is replicated vertically x times, and the MN PDA is

	1	2	3	4	5	6	7	
1	*	{1,2,3}	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}	The first three rows of Q_0
2	{1,2,3}	*	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}	
3	{1,2,3}	{1,2,3}	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}	
	1	2	3	4	5	6	7	
1	*	{1,2,3}	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}	The first three rows of Q_0
2	{1,2,3}	*	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}	
3	{1,2,3}	{1,2,3}	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}	
	1	2	3	4	5	6	7	
1	*	{1,2,3}	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}	The first three rows of Q_0
2	{1,2,3}	*	{1,2,3}	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}	
3	{1,2,3}	{1,2,3}	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}	

	1	2	3	4	5	6	7	
1	*	{1,2}	{1,3}	{1,4}	{1,5}	{1,6}	{1,7}	The first P_1
2	{1,2}	*	{2,3}	{2,4}	{2,5}	{2,6}	{2,7}	
3	{1,3}	{2,3}	*	{3,4}	{3,5}	{3,6}	{3,7}	
4	{1,4}	{2,4}	{3,4}	*	{4,5}	{4,6}	{4,7}	
5	{1,5}	{2,5}	{3,5}	{4,5}	*	{5,6}	{5,7}	
6	{1,6}	{2,6}	{3,6}	{4,6}	{5,6}	*	{6,7}	
7	{1,7}	{2,7}	{3,7}	{4,7}	{5,7}	{6,7}	*	
	1	2	3	4	5	6	7	
1	*	{1,2}	{1,3}	{1,4}	{1,5}	{1,6}	{1,7}	
2	{1,2}	*	{2,3}	{2,4}	{2,5}	{2,6}	{2,7}	
3	{1,3}	{2,3}	*	{3,4}	{3,5}	{3,6}	{3,7}	
4	{1,4}	{2,4}	{3,4}	*	{4,5}	{4,6}	{4,7}	
5	{1,5}	{2,5}	{3,5}	{4,5}	*	{5,6}	{5,7}	
6	{1,6}	{2,6}	{3,6}	{4,6}	{5,6}	*	{6,7}	
7	{1,7}	{2,7}	{3,7}	{4,7}	{5,7}	{6,7}	*	

Fig. 3: Example of $t+L'$-usres $\{1,2,3\}$

replicated vertically y times.

- **Step 1.** Construction of MS MAPDA $\mathbf{Q}_0$ and MN PDA $\mathbf{P}_0$.
  - **Step 1.1.** We first get a 7×7 table $\mathbf{P}_0$ from construction 1. In this example, the table I is the $\mathbf{P}_0$ we want.
  - **Step 1.2.** Then we get a 35×7 table $\mathbf{Q}_0$ which can be constructed by construction 2. In fact, tabel II is the $\mathbf{Q}_0$ that we need.
- **Step 2.** Construction of $\mathbf{Q}$ and $\mathbf{P}$.
  - **Step 2.1.** The $\mathbf{Q}$ represent the table after replicating $\mathbf{Q}_0$ x times, so it's necessary to decide x. The time slots required to transmit for each $(t+L')$-user group in $\mathbf{Q}$ is $S_1=\binom{t+L'-1}{t}=\binom{2}{1}=2$, while each $(t+1)$-user group only need $S_2=1$ time slot. To ensure time slot matching, we can consider selecting a group of $(t+L')$ users first, and then traversing all possible combinations of the remaining users.In this example, after excluding $t+L'=3$ users, there are $\binom{K-(t+L')}{(r-1)(t+1)}\frac{[(r-1)(t+1)]!}{((t+1)!)^{(r-1)}(r-1)!}=\binom{4}{4}\frac{4!}{(2!)^2 2!}=3$ cases in

Algorithm 4 General Construction of RMAPDA

procedure CONSTRUCTRMAPDA(K, L, M, N, r)

$t \leftarrow \frac{KM}{N}$

$L' \leftarrow L - r + 1$ ▷ Divide antennas into r groups, one with L' antennas, others with one antenna

Step 1: Construct MS MAPDA $\mathbf{Q}_0$ and MN PDA $\mathbf{P}_0$

$\mathbf{P}_0 \leftarrow$ CONSTRUCTION OF MN_PDA(K, t)

$\mathbf{Q}_0 \leftarrow$ CONSTRUCTION OF MS_MAPDA(K, t, L')

Step 2: Construct $\mathbf{Q}$ and $\mathbf{P}$

Determine x and y' such that

$$x\binom{t+L'-1}{t} = y'\binom{K-(t+L')}{(r-1)(t+1)}\frac{[(r-1)(t+1)]!}{((t+1)!)^{r-1}(r-1)!}$$

$$\mathbf{Q} \leftarrow \left.\begin{bmatrix} \mathbf{Q}_0 \\ \hline \mathbf{Q}_0 \\ \hline \vdots \\ \hline \mathbf{Q}_0 \end{bmatrix}\right\} x \quad times$$

$$\mathbf{P}_1 \leftarrow \left.\begin{bmatrix} \mathbf{P}_0 \\ \hline \mathbf{P}_0 \\ \hline \vdots \\ \hline \mathbf{P}_0 \end{bmatrix}\right\} y' \quad times$$

Compute n as:

$$n \leftarrow \binom{K-t-1}{t+L'}\binom{K-(t+1)-(t+L')}{(r-2)(t+1)}\left(\frac{[(r-2)(t+1)]!}{((t+1)!)^{r-2}(r-2)!}\right)$$

$$\mathbf{P} \leftarrow \left.\begin{bmatrix} \mathbf{P}_1 \\ \hline \mathbf{P}_1 \\ \hline \vdots \\ \hline \mathbf{P}_1 \end{bmatrix}\right\} n \quad times$$

Step 3: Combine $\mathbf{P}$ and $\mathbf{Q}$

$$\mathbf{T} \leftarrow \begin{bmatrix} P \\ \hline \mathbf{Q} \end{bmatrix}$$

return $\mathbf{T}$

end procedure

which $(r-1)(t+1)$ users are selected from the remaining users and divided into $(r-1)$ groups of $(t+1)$ users each. Therefore, select x and y' to satisfy the equation 12, i.e.

$$2x = 3y' \tag{16}$$

Let $x = 3$, $y' = 2$, which ensures that the time slots of a multi-user group can match those of the single-user groups during transmission.

- **Step 2.2.** Then we replicate $\mathbf{Q}_0$ x times vertically to get a new table $\mathbf{Q}$ and replicate $\mathbf{P}_0$ y' times vertically to get $\mathbf{P}_1$.
- **Step 2.3.** The $\mathbf{P}_0$ replicated vertically y' times obtain $\mathbf{P}_1$ can only meet the time slot requirements of one $(t+L')$ group. For example, for the $(t+L')$ group composed of users

	1	2	3	4	5	6	7
1	*	1	1	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
2	1	*	2	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
3	2	2	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}
	1	2	3	4	5	6	7
1	*	3	3	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
2	3	*	4	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
3	4	4	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}
	1	2	3	4	5	6	7
1	*	5	5	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
2	5	*	6	{1,2,4}	{1,2,5}	{1,2,6}	{1,2,7}
3	6	6	*	{1,3,4}	{1,3,5}	{1,3,6}	{1,3,7}
	1	2	3	4	5	6	7
1	*	{1,2}	{1,3}	{1,4}	{1,5}	{1,6}	{1,7}
2	{1,2}	*	{2,3}	{2,4}	{2,5}	{2,6}	{2,7}
3	{1,3}	{2,3}	*	{3,4}	{3,5}	{3,6}	{3,7}
4	{1,4}	{2,4}	{3,4}	*	1	2	3
5	{1,5}	{2,5}	{3,5}	1	*	3	2
6	{1,6}	{2,6}	{3,6}	2	3	*	1
7	{1,7}	{2,7}	{3,7}	3	2	1	*
	1	2	3	4	5	6	7
1	*	{1,2}	{1,3}	{1,4}	{1,5}	{1,6}	{1,7}
2	{1,2}	*	{2,3}	{2,4}	{2,5}	{2,6}	{2,7}
3	{1,3}	{2,3}	*	{3,4}	{3,5}	{3,6}	{3,7}
4	{1,4}	{2,4}	{3,4}	*	4	5	5
5	{1,5}	{2,5}	{3,5}	4	*	6	6
6	{1,6}	{2,6}	{3,6}	5	6	*	4
7	{1,7}	{2,7}	{3,7}	6	5	4	*

Fig. 4: Example of time slots 1-6

$\{1, 2, 3\}$, after replication, it should occupy a total of 6 time slots. There are three ways to select two two-user groups from the remaining users. Each grouping occupies two time slots after being replicated y' times. Take the user group $\{4, 5\}$ as an example. Assuming that the three-user group becomes $\{1, 2, 6\}$, in its corresponding two-user group, $\{4, 5\}$ still needs to occupy two time slots. This means that replicating y' times can only meet the time slot requirements of one three-user group for the two-user groups, and one $(t+1)$ group may be needed n times in equation 13. In this example, n is $\binom{5}{3} = 10$ Therefore, $\mathbf{P}_1$ needs to be replicated 10 times vertically to get $\mathbf{P}$.

- **Step 3.** Combine $\mathbf{P}$ and $\mathbf{Q}$
  - **Step 3.1** Combine the generated $\mathbf{P}$ and $\mathbf{Q}$ vertically to form a new table $\mathbf{T}$, which corresponds to the array of the entire cache placement scheme.

– **Step 3.2** Rewrite the sets in the table as numbers, which determine the time slot in which each file is sent. First, select a $(t+L')$-users group from $\mathbf{Q}$, for example, choose the set $\{1,2,3\}$. This set occupies two time slots in $\mathbf{Q}_0$. Each $\mathbf{Q}$ contains 3 $\mathbf{Q}_0$, thus occupying a total of 6 time slots. One time slot's $\{1,2,3\}$ corresponds to two $(t+1)$-users group, i.e., two sets of two users. The remaining four users are divided into two groups of two users with three possible combinations: $(\{4,5\},\{6,7\}),(\{4,6\},\{5,7\}),(\{4,7\},\{5,6\})$. In each time slot of $\{1,2,3\}$, select one of these three combinations in $\mathbf{P}$ to correspond to it, choosing each combination twice, which exactly matches the time slots. Similarly, other $(t+L')$-users in $\mathbf{Q}$ also select the corresponding combinations in $\mathbf{P}$ to match them, writing the same number for the same time slot.

  For example, when the $(t+L')$-users group is selected as $\{1,2,3\}$, we extract the first three rows of each $\mathbf{Q}_0$ containing $\{1,2,3\}$ and the first $\mathbf{P}_1$. In Fig. 3, each color represents a time slot

  So the final time slots corresponding to $\{1,2,3\}$ can be filled as shown in Fig. 4. The filling of time slots occupied by other $(t+L')$-users groups is similar. When finding the corresponding $(t+1)$-users group in $\mathbf{P}$, you can follow the principle of left to right and top to bottom.

  If the remaining users after excluding the $(t+L')$ group exceed $(r-1)(t+1)$, then each possible combination should be considered. Therefore, first select $(r-1)(t+1)$ users from the remaining users, resulting in $\binom{K-(t+L')}{(r-1)(t+1)}$ combinations. For each combination, assign the user groups for each time slot using the method described in the previous example. After completing one combination, move to the next $(r-1)(t+1)$ combination and repeat the above steps. Once all combinations are exhausted, move to the next $(t+L')$ group and repeat the allocation process.

.

References

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

[2] K. Wan, D. Tuninetti, and P. Piantanida, "An index coding approach to caching with uncoded cache placement," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1318–1332, Mar. 2020.

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

[4] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3092–3107, 2017.

[5] 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.

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

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

[8] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2792–2807, 2018.

[9] S. Mohajer and I. Bergel, "Miso cache-aided communication with reduced subpacketization," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[10] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized degrees of freedom of the symmetric cache-aided miso broadcast channel with partial csit," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5799–5815, 2019.

[11] T. Yang, K. Wan, M. Cheng, R. C. Qiu, and G. Caire, "Multiple-antenna placement delivery array for cache-aided miso systems," *IEEE Transactions on Information Theory*, 2023.

[12] K. K. Namboodiri, E. Peter, and B. S. Rajan, "Extended placement delivery arrays for multi-antenna coded caching scheme," *IEEE Transactions on Communications*, 2023.

[13] 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, 2017.

[14] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light: science & applications*, vol. 3, no. 10, pp. e218–e218, 2014.

[15] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[16] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. De Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE journal on selected areas in communications*, vol. 38, no. 11, pp. 2450–2525, 2020.

[17] C. Pan, H. Ren, K. Wang, J. F. Kolb, M. Elkashlan, M. Chen, M. Di Renzo, Y. Hao, J. Wang, A. L. Swindlehurst, X. You, and L. Hanzo, "Reconfigurable intelligent surfaces for 6g systems: Principles, applications, and research directions," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 14–20, 2021.

[18] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE communications surveys & tutorials*, vol. 23, no. 3, pp. 1546–1577, 2021.

[19] A. H. A. Bafghi, V. Jamali, M. Nasiri-Kenari, and R. Schober, "Degrees of freedom of the k-user interference channel in the presence of intelligent reflecting surfaces," *arXiv preprint arXiv:2012.13787*, 2020.

[20] T. Jiang and W. Yu, “Interference nulling using reconfigurable intelligent surface,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 5, pp. 1392–1406, 2022.

[21] M. Fu, Y. Zhou, and Y. Shi, “Reconfigurable intelligent surface for interference alignment in mimo device-to-device networks,” in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

[22] S. Zheng, B. Lv, T. Zhang, Y. Xu, G. Chen, R. Wang, and P. Ching, “On dof of active ris-assisted mimo interference channel with arbitrary antenna configurations: When will ris help?” *IEEE transactions on vehicular technology*, 2023.

APPENDIX A
PROOF OF THEOREM 1

Given the number of active antennas L_0, the total number G of RIS elements used must satisfy the following inequality.

$$\begin{aligned}
\frac{G}{2} &\geq \sum_{i=1}^{r}(L_i+t)(L_0-L_i) \\
&= \sum_{i=1}^{r}(L_0L_i+tL_0-L_i^2-tL_i) \\
&= (L_0-t)\sum_{i=1}^{r}L_i+trL_0-\sum_{i=1}^{r}L_i^2 \\
&= L_0(L_0-t)+trL_0-\sum_{i=1}^{r}L_i^2
\end{aligned} \tag{17}$$

We prefer to design a scheme such that the value G is as small as possible. This implies that given the parameters L, r and t, the value of $\sum_{i=1}^{r} L_i^2$ is as large as possible under the the condition $\sum_{i=1}^{r} L_i = L_0$. Without loss of generality, we assume that $L_1 \geq L_2 \geq \cdots \geq L_r \geq 1$. When we move a_2, a_3, $\ldots$, a_r antennas from the second, third, $\ldots$, rth group respectively to the first group, then the first group has $L_1' = L_1 + a_1 + a_2 + \cdots + a_{k-1}$ and the ith group has $L_i' = L_i - a_i$ antennas for each $2 \leq i \leq r$ where $a_i < L_i$. Now let us consider the value of $\sum_{i=1}^{r} {L_i'}^2$ and obtain the following equality

$$\begin{aligned}
&{L_1'}^2+{L_2'}^2+\cdots+{L_r'}^2 \\
=&(L_1+a_2+a_3+\cdots+a_r)^2+(L_2-a_2)^2+(L_3-a_3)^2 \\
&+\cdots+(L_r-a_r)^2 \\
=&L_1^2+L_2^2+L_3^2+\cdots+L_r^2+(a_2+a_3+\cdots+a_r)^2 \\
&+2a_2(L_1-L_2)+2a_3(L_1-L_3)+\cdots+2a_r(L_1-L_r).
\end{aligned}$$

Clearly L_1-L_2, L_1-L_3, $\ldots$, L_1-L_r are all non-negative. Therefore, we have $\sum_{i=1}^{r} {L_i'}^2 \geq \sum_{i=1}^{r} {L_i}^2$. This implies that the more antennas $a_1, a_2, \ldots, a_r$ are moved to the group with the highest number of antennas, the greater increase of the sum of squares $\sum_{i=1}^{r} {L_i}^2$. Clearly the extreme cases is that the first group has exactly L_0-r+1 antennas and each of the last $r-1$ groups has exactly one antenna.

APPENDIX B

Detailed Analysis and Design Process of the Algorithm 3

Based on the analysis of scenarios where the number of used antennas is known, it can be deduced that even if some antennas is allowed to be silenced, dividing antennas into r groups, with $r-1$ groups containing one antenna each, and the remaining group consisting of $L_0-(r-1)$ antennas is still is considered optimal only under the conditions specified for grouping.

Assuming two schemes have the same sum-DoF, denoted as g, and the number of antennas for these schemes are L and L' respectively. Let us assume that $L > L'$ and the L antennas and L' antennas are divided into r and r' groups respectively. By our hypothesis we have $g = L + tr = L' + tr'$ for each t, i.e., $L - L' = t(r' - r) > 0$. So we have $r < r'$ and let $\lambda = r' - r$.

From (17), the required number of RIS elements G for L antennas is

$$\begin{aligned}\frac{G}{2} &= (t+1)(L-1) + (L-(r-1)+t)(r-1) \\ &= (r-1)[(t+2)L - r]\end{aligned} \tag{18}$$

and the number of RIS elements G' for L' antennas is

$$\begin{aligned}\frac{G'}{2} &= (t+1)(L'-1) + (L'-(r'-1)+t)(r'-1) \\ =&(\lambda+r-1)[(t+2)(L-\lambda t)-(\lambda+r)] \\ =&\lambda[(t+2)(L-\lambda t)-(\lambda+r)] + (r-1)((t+2)L-r)] \\ &+ (r-1)[-(t+2)\lambda t - \lambda]\end{aligned} \tag{19}$$

Then we have

$$\begin{aligned}f(\lambda) =&G' - G \\ =&2(\lambda((t+2)(L-\lambda t)-(\lambda+r))+ \\ &(r-1)(-(t+2)\lambda t-\lambda)) \\ =&2(\lambda((-t^2-2t-1)\lambda + (t+2)L- \\ &2r + (1-r)t^2 - 2tr + 2t + 1))\end{aligned} \tag{20}$$

The value of (20) represents the difference in the number of RIS elements between the original scheme and the scheme with fewer antennas.

For a given sum-DoF g, assume L represents the maximum number of antennas used to achieve this sum-DoF. Consider this scheme as the original scheme referred to in (20).

Starting from this maximum value, consider other schemes with fewer antennas(i.e. consider the value of λ). It is not difficult to check that $f(\lambda)$ in (20) is a downward-opening parabola that passes through the origin,with λ as the independent variable. This implies that if $f(\lambda) \geq 0$, the reducing the number of antennas $L - L'$ leads to an increase or no change in RIS elements; if $f(\lambda) < 0$, the scheme with L' antennas utilizes fewer RIS elements than the original scheme, making it more optimal. Additionally, the larger the value of λ, the greater the reduction in the number of RIS elements.

Since λ is positive, it is only necessary to determine whether λ can exceed another zero point. If it can, then by reducing the number of antennas, a more optimal scheme can be found. Otherwise, the scheme using the maximum number of antennas is the best solution.

- To find the scheme that uses the maximum number of antennas under a given sum-DoF g, it is necessary first to determine the range of L.

  First, when the used antennas is L, the maximum achievable sum-DoF is $L(t+1)$. Therefore, $g \leq L(t+1)$, which implies $L \geq \frac{g}{t+1}$. Since L must be a positive integer, we have $L \geq \lceil \frac{g}{t+1} \rceil$. From (8), we have the number of groups $r = \frac{(g-L)}{t} \geq 1$, i.e., $L \leq g - t$. So we have the range for L as follows,

$$\lceil \frac{g}{t+1} \rceil \leq L \leq \min\{L_0, g - t\} \tag{21}$$

  where L_0 represents the total number of existing antennas.

  In addition, the value of L must also ensure that $r = \frac{(g-L)}{t}$ is a positive integer.

  After determining the range and constraints for L, it's possible to verify from the upper bound of L downwards, checking each value whether it meets the condition that r is a positive integer. This process will help identify the scheme that uses the maximum number of antennas, along with the corresponding number of groups and RIS elements. Then substitute these values into the function $f(\lambda)$ and compute the non-zero roots λ_0 of $f(\lambda)$.

$$\lambda_0 = \frac{(t+2)L - 2r + (1-r)t^2 - 2tr + 2t + 1}{(t+1)^2} \tag{22}$$

- After determining λ_0, it is important to consider the range of possible values for λ.

  Given that reducing the number of antennas by t results in an increase of one group, the reduced number of antennas L' must be greater than or equal to the new number of groups r', that is, $L' \geq r'$. Substituting in the expressions for L' and r', we get $L - \lambda t \geq \lambda + r$. Considering also that λ must be a positive integer, the range for λ can be determined as follows,

$$0 \leq \lambda \leq \lfloor \frac{L-r}{t+1} \rfloor \tag{23}$$

Therefore, we only need to compare the maximum value of λ with λ_0.