Coded Multi-User Information Retrieval with a Multi-Antenna Helper Node

Milad Abolpour*, MohammadJavad Salehi*, Soheil Mohajer†, Seyed Pooya Shariatpanahi‡, and Antti Tölli*
*Centre for Wireless Communications, University of Oulu, Finland, E-mail: {firstname.lastname}@oulu.fi
†Department of Electrical and Computer Engineering, University of Minnesota, USA, E-mail: soheil@umn.edu
‡School of Electrical and Computer Engineering, University of Tehran, Iran, E-mail: p.shariatpanahi@ut.ac.ir

***Abstract*—A novel coding design is proposed to enhance information retrieval in a wireless network of users with partial access to the data, in the sense of observation, measurement, computation, or storage. Information exchange in the network is assisted by a multi-antenna base station (BS), with no direct access to the data. Accordingly, the missing parts of data are exchanged among users through an uplink (UL) step followed by a downlink (DL) step. In this paper, new coding strategies, inspired by coded caching (CC) techniques, are devised to enhance both UL and DL steps. In the UL step, users transmit encoded and properly combined parts of their accessible data to the BS. Then, during the DL step, the BS carries out the required processing on its received signals and forwards a proper combination of the resulting signal terms back to the users, enabling each user to retrieve the desired information. Using the devised coded data retrieval strategy, the data exchange in both UL and DL steps requires the same communication delay, measured by normalized delivery time (NDT). Furthermore, the NDT of the UL/DL step is shown to coincide with the optimal NDT of the original DL multi-input single-output CC scheme, in which the BS is connected to a centralized data library.**

***Index Terms*—coded caching; multi-user information retrieval; coded distributed computing; multi-antenna communications**

I. Introduction

Multi-user information retrieval (MIR) is a generic field of research exploring mechanisms to enable each user in the network to recover specific pieces of information that are either aggregated at a central master node or distributed across the network [1]. An example use case is a distributed coded computing platform where the computation tasks are split among multiple servers, and the outputs are gathered and distributed by a master node (acting as the base station (BS)) such that each server has the result of a specific task [2]. Another example is a sensor network where the data is gathered by a BS from sensing nodes and then distributed to multiple actuators, each needing a specific type of data for their action [3]. Clearly, with the involvement of a BS, MIR consists of two consecutive steps: 1) an uplink (UL) step where the data is gathered by the BS, and 2) a downlink (DL) step where the gathered data is distributed to requesting nodes according to their needs. The goal of this paper is to introduce novel coding mechanisms to reduce the time needed to fulfill both UL and DL steps.

This research has been supported by the Academy of Finland, 6G Flagship program under Grants 346208, 343586 (CAMAIDE), and by the Finnish-American Research and Innovation Accelerator (FARIA).

The coding solutions devised in this paper draw inspiration from the coded caching (CC) technique, originally proposed to reduce the load at peak traffic times by employing the caches distributed in the network as a supplementary communication resource [4]. In a single-stream downlink network with K users, each with sufficient memory to store a fraction γ of the entire file library, CC boosts the achievable rate by the multiplicative factor of $K\gamma + 1$, which scales with the cumulative cache size in the entire network. This new gain is accomplished by multicasting carefully designed codewords to different subsets of users with size $K\gamma + 1$, and can also be aggregated with the spatial multiplexing gain to enable the speed-up factor of $K\gamma + L$ in a multi-input multi-output (MISO) setup with L antennas at the transmitter [5], [6]. Due to these exciting properties, CC has been extensively studied in the literature, to address its challenges, such as exponentially growing subpacketization [7], [8], complex beamformer design [9], [10], privacy [11], and applicability to dynamic setups [12], and to investigate its benefits in use cases such as large-scale video-on-demand (VoD) [13] and extended reality (XR) [14], [15]. Variations of the original CC models have also been studied, e.g., for data shuffling [16]–[18] and linear function retrieval [19], [20].

In the context of coded MIR, existing works in the literature have primarily considered the application of CC in distributed computing systems, which offer various advantages, such as enhanced scalability, reliability, and cost-effectiveness, over centralized computing solutions [21]. In such systems, the setup mainly consists of a network of K computing nodes, a library of files $\{d^n\}_n$, and a set of functions $\{\phi^n(\cdot)\}_n$. Each file d^n is split into non-overlapping portions $d^n_{\mathcal{P}}$, where $\mathcal{P}$ can be any subset of computing nodes with a predefined size. Each computing node k calculates the output of all the functions over all the file parts $\{d^n_{\mathcal{P}}\}_n$ for which $k \in \mathcal{P}$. Subsequently, the nodes exchange their calculated outputs, ensuring that each node k could ultimately reconstruct the output of its desired function $\phi^k(d^k)$ over its specified file d^k. Coding mechanisms resembling those of CC have been introduced to alleviate the communication load during the data exchange, where the excess computed elements at each node are used to remove undesired terms from the received signals, similar to cache-aided interference removal. The result is a balance between the excess computation power and the required communication load among servers [22]–[25].

In this paper, we propose new coding schemes, inspired by CC, for wireless MIR scenarios where the information exchange among users is assisted by a multi-antenna BS as the helper node (similar to a two-way relay setup [26]). With the proposed solution, the aim is to generate the multicast signals in the DL step based on the signals received during the UL step, such that the DL step performs similarly to the downlink communication of the original MISO-CC scheme [5], [6]. For this purpose, a novel transmission design is required for the UL step, benefiting from the over-the-air addition of the signals transmitted by network users and the modest computation capability at the BS to create the codewords needed in the DL step over a small number of transmission slots. We show that the time required in the UL step can, in fact, be made equal to the time required in DL, which has already been shown to be information-theoretically optimal under simple conditions in downlink MISO-CC communications [27]. As multi-antenna connectivity is an integral part of all modern communication systems, including 5G and beyond cellular networks [28], the coding solutions devised in this paper for MIR are applicable to diverse scenarios, such as industrial IoT (exchanging measurements or observations) [29], distributed coded computing (exchanging computation results) [22], or distributed cache networks (exchanging cached contents) [4].

Notation: In this paper, bold lower-case and calligraphic letters show vectors and sets, respectively. Moreover, $\mathbf{v}[j]$ represents the j-entry of vector $\mathbf{v}$. We use $\mathbf{M}^{\mathrm{T}}$ and $\mathbf{M}^{\mathrm{H}}$ to demonstrate the transpose and conjugate-transpose (Hermitian) of matrix $\mathbf{M}$, respectively. For integers a and b, $[a:b]$ shows the set $\{a,\cdots,b\}$ and $[a]=\{1,\cdots,a\}$. $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$, and for $\mathcal{B}\subseteq\mathcal{A}$, $\mathcal{A}\backslash\mathcal{B}$ represents $\{x\in\mathcal{A}: x\notin\mathcal{B}\}$.

II. Problem Formulation

This paper aims to leverage the underlying coding mechanism of CC to improve multi-user information retrieval in application scenarios where each user has direct access only to a part of each content (i.e., data files), but requires all parts of one or more specific contents. We emphasize that in this context, the content may be generated online (e.g., in industrial IoT applications) [29] or be cached in advance [4]. In order to exchange data, as depicted in Fig. 1, K single-antenna users communicate with an L-antenna BS. Here, we consider the worst-case scenario where each user is required to recover a distinct content file. As the contents are only partially accessible by each user and the BS lacks direct access to the content library to satisfy users' demands, a relay-type two-way UL-DL model is used for data exchange. In the UL step, users transmit a portion of their accessible contents to the BS via a number of consecutive UL transmissions. Assuming the BS possesses sufficient computation capability to process and enough memory to store all received signals, it then appropriately processes and combines the received signals, and forwards them back to the users in the DL step.[1]

[1]In the context of distributed coded computing, the proposed model can be considered as a multi-antenna extension of the system models described in [2], [22].

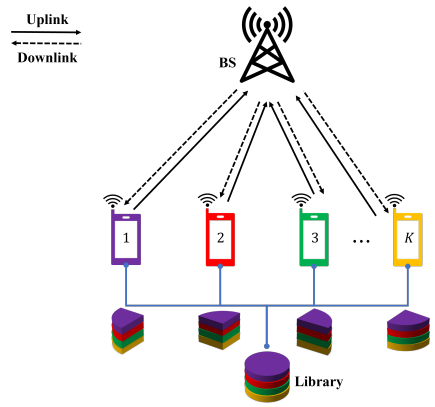

Fig. 1. System model for a coded data retrieval network with K single antenna users and a BS with L antennas: A γ portion of the entire library is either locally generated or stored beforehand in the cache memories of users.

The content library consists of N files W^n, $n\in[N]$, each with a size of F bits. Without loss of generality, we assume $N=K$. Let us use $W^n_{\mathcal{P}}$, $\mathcal{P}\subseteq[K]$, to denote the part of the content file W^n that can be accessed by every user in $\mathcal{P}$. We impose two simplifying assumptions: 1) each part of each file is accessible by the same number of users t, i.e., $W^n_{\mathcal{P}}$ has a size larger than zero bits if and only if $|\mathcal{P}|=t$, and 2) all the file parts $W^n_{\mathcal{P}}$ with a size larger than zero bits have the same size of $F/\binom{K}{t}$ bits. With these assumptions, the system is analogous to a distributed MISO cache network where each cache is sufficiently large to store a fraction γ of the entire library, and the CC gain is $t=K\gamma$. As a result, for the sake of simplicity, we may use the standard notation of MISO-CC systems: we use *packet* to denote a part of a content file and say a packet is *cached* by a user if the user has access to the respective file part. Respectively, the above-mentioned assumptions on the access of the users to content files could also be described by the so-called *content placement phase* of a MISO-CC system, where each content file W^n, $n\in[N]$, is split into $\binom{K}{t}$ equal-sized disjoint packets $W^n_{\mathcal{P}}$ as

$$W^n=\{W^n_{\mathcal{P}}:\mathcal{P}\subseteq[K],|\mathcal{P}|=t\},\tag{1}$$

and for each $\mathcal{P}\subseteq[K]$ with $|\mathcal{P}|=t$, each user $k\in\mathcal{P}$ stores $W^n_{\mathcal{P}}$ in its cache memory. Clearly, the total number of packets stored by each user $k\in[K]$ is $K\binom{K-1}{t-1}=t\binom{K}{t}=\gamma K\binom{K}{t}$, satisfying the cache size constraint.

Without loss of generality (up to a permutation of the indices of the users), let us assume that user $k\in[K]$ requires the content file W^k. To initiate data exchange, we first further split each packet $W^n_{\mathcal{P}}$ into $\binom{K-t-1}{L-1}$ equal-sized *subpackets* $W^n_{\mathcal{P},q}$ as $W^n_{\mathcal{P}}=\{W^n_{\mathcal{P},q}: q\in[\binom{K-t-1}{L-1}]\}$, where $W^n_{\mathcal{P},q}$ is comprised of f uniformly i.i.d. bits with $f=F/\binom{K}{t}\binom{K-t-1}{L-1}$.

Definition 1. For some $\mathcal{V}\subseteq[K]$ and an arbitrary $\theta^k_{\mathcal{P}_k,q_k}\in\mathbb{C}$, the superposition signal

$$\sum_{k\in\mathcal{V}}\theta^k_{\mathcal{P}_k,q_k}\mathsf{c}\big(W^k_{\mathcal{P}_k,q_k}\big),\tag{2}$$

is defined as a *codeword* of size $|\mathcal{V}|$, where $\mathcal{P}_k\subseteq[K]$ with $|\mathcal{P}_k|=t$, $q_k\in[\binom{K-t-1}{L-1}]$, and $\mathsf{c}\big(W^k_{\mathcal{P}_k,q_k}\big)$ is the encoded signal of $W^k_{\mathcal{P}_k,q_k}$ with $\mathsf{c}:\mathbb{F}_{2^f}\to\mathbb{C}$.

Assume that user $k \in [K]$ is connected to the BS through the channel $\mathbf{h}_k \in \mathbb{C}^{L\times 1}$. In this paper, we assume that the wireless links are semi-static and channel state information (CSI) is available at the BS and the users. The UL step comprises N_{UL} transmissions, where in transmission $j \in [N_{\mathrm{UL}}]$, each user in a subset $\mathcal{U}(j)$ of users transmits a fraction of its cache contents to the BS. Hence, after the transmission j, the BS receives the signal

$$\mathbf{y}_{\mathrm{BS}}(j) = \sum\nolimits_{k\in\mathcal{U}(j)} \mathbf{h}_k x^k(j) + \mathbf{n}_{\mathrm{BS}}, \tag{3}$$

where $\mathbf{n}_{\mathrm{BS}} \sim \mathcal{CN}(0, \mathbf{I}_L)$, and $x^k(j)$ is the transmitted signal of user k with power $P_k(j) = \mathbb{E}[|x^k(j)|^2]$. Here, it is assumed that the average transmit power of user $k \in [K]$ during the UL step is equal to P_{T}, i.e., $\frac{1}{N_{\mathrm{UL}}}\sum_{j=1}^{N_{\mathrm{UL}}} P_k(j) = P_{\mathrm{T}}$.

The DL step involves N_{DL} consecutive transmissions, such that in transmission $j \in [N_{\mathrm{DL}}]$, the BS transmits a superposition signal containing the required codewords for a specific set of users represented by $\mathcal{V}(j)$. Therefore, the received signal at user $k \in \mathcal{V}(j)$ in the j-th DL transmission can be written as

$$y_k(j) = \sum\nolimits_{i\in\mathcal{V}(j)} \mathbf{h}_k^{\mathrm{H}} \mathbf{s}_i(j) + n_k, \tag{4}$$

where $\mathbf{s}_i(j) \in \mathbb{C}^{L\times 1}$ is the precoded signal for user i during transmission j, $n_k \sim \mathcal{CN}(0,1)$, P_{BS} is the BS transmit power in each transmission, such that $P_{\mathrm{BS}} = \sum_{i\in\mathcal{V}(j)} \mathbb{E}[\|\mathbf{s}_i(j)\|^2]$.

In this paper, the aim is to design a UL-DL transmission strategy, minimizing the *normalized delivery time* (NDT) in the UL and DL steps at high-SNR regimes, i.e., when $P_{\mathrm{T}} \to \infty$ and $P_{\mathrm{BS}} \to \infty$. By following a similar approach presented in [24] and [25], the NDT for the UL step is defined as

$$T_{\mathrm{UL}} = \lim_{\mathrm{SNR}\to\infty} \sum_{i=1}^{N_{\mathrm{UL}}} \frac{D_i^{\mathrm{UL}} d}{F/\log(\mathrm{SNR})}, \tag{5}$$

where $D_i^{\mathrm{UL}} = \frac{FQ}{R_i}$ is the transmission time to deliver data in the i-th UL transmission with R_i expressing the achievable rate of the user with the worst channel condition, and $Q = 1/\binom{K}{t}\binom{K-t-1}{L-1}$. Moreover, d is the per-user degree-of-freedom (DoF) at high SNR [24]. Here, we note that $F/\log(\mathrm{SNR})$ is the transmission time of delivering a single file of F bits in a single-antenna point-to-point baseline system with Gaussian noise at the high-SNR regime. Considering the high-SNR condition, all optimal transmit and receive beamformers asymptotically behave as zero-forcing (ZF) precoders, and consequently, we have $R_i \approx d\log(\mathrm{SNR}) + \mathcal{O}(\log(\mathrm{SNR}))$. As a result, the NDT expression in (5) is simplified to

$$T_{\mathrm{UL}} = \lim_{\mathrm{SNR}\to\infty} \frac{N_{\mathrm{UL}} FQd\log(\mathrm{SNR})}{(d\log(\mathrm{SNR}) + \mathcal{O}(\log(\mathrm{SNR})))F} = N_{\mathrm{UL}} Q. \tag{6}$$

To give further insight into (6), one can say that FQd is the number of bits per uplink transmission. Now, dividing $FQd \times N_{\mathrm{UL}}$ by R_i (which includes the degrees of freedom per user), we get the absolute time taken for the uplink. Finally, normalizing the result by the time in a single-antenna point-to-point system, we get the NDT given in (6). Following the same process for the DL step, the NDT is expressed as

$$T_{\mathrm{DL}} = N_{\mathrm{DL}} Q. \tag{7}$$

III. Reference Strategies

In this section, we introduce two baseline UL-DL transmission strategies, that leverage either the spatial multiplexing gain L or the CC gain t during the UL step while adopting a transmission approach similar to the MISO-CC scheme of [5], [6] for the DL step.

1) Strategy A: As mentioned, a γ portion of the entire library is generated/stored by each user. Therefore, $K(1-\gamma)\binom{K}{t}\binom{K-t-1}{L-1}$ subpackets must be transmitted to the BS in the UL step. To this end, during each UL transmission in *Strategy A*, we select L users to simultaneously transmit L subpackets to the BS, which is then able to decode all of them as it is equipped with L antennas. Therefore, during the UL step, this strategy only benefits from the spatial multiplexing gain of L without incorporating the CC gain. Hence, employing (6), the NDT in the UL step via *Strategy A* is given by:

$$T_{\mathrm{UL}}^{A} = \frac{K(1-\gamma)\binom{K}{t}\binom{K-t-1}{L-1}}{L\binom{K}{t}\binom{K-t-1}{L-1}} = \frac{K-t}{L}. \tag{8}$$

Following the UL step, the BS has access to all the missing subpackets. As stated in [27], the optimal transmission strategy among all linear one-shot schemes with uncoded placement for the DL step is to follow a similar approach as the MISO-CC scheme of [5], [6]. With this scheme, using (7), the NDT in the DL step of *Strategy A* is $T_{\mathrm{DL}}^{A} = \frac{K-t}{t+L}$.

2) Strategy B: With this strategy, in each UL transmission, $t+1$ users are chosen to transmit $t+1$ subpackets to the BS simultaneously. The data transmitted by these users is then added over the air to form one of the codewords needed in the following DL step (more details are provided shortly after). Hence, during the UL step, this strategy only benefits from the CC gain without using the available spatial multiplexing gain. According to (6), the UL step via *Strategy B* achieves the NDT

$$T_{\mathrm{UL}}^{B} = \frac{K(1-\gamma)\binom{K}{t}\binom{K-t-1}{L-1}}{(t+1)\binom{K}{t}\binom{K-t-1}{L-1}} = \frac{K-t}{t+1}. \tag{9}$$

For the DL step, in order to optimize the transmission delay, the BS follows a similar scheme as the one proposed in [5], [6]. To this end, for each $\mathcal{T} \subseteq [K]$ with $|\mathcal{T}| = t+L$, the BS has received $\binom{t+L}{t+1}$ signals during the UL step. Therefore, it generates $\binom{t+L-1}{t}$ random linear combinations of these $\binom{t+L}{t+1}$ received signals and broadcasts them sequentially. For the decoding process, we adopt the signal-level decoding approach as expressed in [8], where the undesired terms are regenerated from the local memory and removed before the received signal is decoded by the users. The details of the DL step are presented in Section IV-C. Accordingly, utilizing (7), the NDT of the DL step via *Strategy B* is obtained as $T_{\mathrm{DL}}^{B} = \frac{K-t}{t+L}$.

As observed, both strategies achieve the same NDT during the DL step; however, in the UL step, neither *Strategy A* nor *Strategy B* achieves the NDT of $\frac{K-t}{t+L}$. In this work, we devise a UL transmission strategy that incorporates both spatial multiplexing and CC gains to achieve a UL NDT equal to that of the DL.

TABLE I
TRANSMITTED SIGNALS DURING THE UL STEP

$\mathcal{S}$ \ $x^k(\mathcal{S})$	$x^1(\mathcal{S})$	$x^2(\mathcal{S})$	$x^3(\mathcal{S})$
$\{12\}$	B_1	A_2	$-B_3-A_3$
$\{13\}$	C_1	$-C_2-A_2$	A_3

IV. THE NEW UL-DL TRANSMISSION STRATEGY

We first present an illustrative example to give further insight into the system performance and then design the generalized transmission strategies for the UL and DL steps.

A. An Illustrative Example

Consider a cache-aided MISO network with $K=3$ users, $\gamma=\frac{1}{3}$, $t=1$ and $L=2$. Users 1, 2 and 3 are required to recover the files W^1, W^2 and W^3, respectively. First, each file W^n is split into $\binom{K}{t}=3$ packets $W^n_{\mathcal{P}}$, where $\mathcal{P}\in[3]$. Accordingly, user $k\in[3]$ stores the packets W^n_k in its cache memory for all $n\in[3]$. For simplicity, let us use A, B, and C to denote the encoded signals of W^1, W^2, and W^3, respectively. Moreover, as $q=\binom{K-t-1}{L-1}=1$, we ignore the index q.

The UL step is comprised of $N_{\rm UL}=2$ transmissions, represented by $\{12\}$ and $\{13\}$. During the transmission $\mathcal{S}\in\{\{12\},\{13\}\}$, user $k\in[3]$ transmits the signal $x^k(\mathcal{S})$, as shown in Table I. Assuming noise-less channels,[2] the received signal of the BS during transmission $\mathcal{S}$, denoted by $\mathbf{y}_{\rm BS}(\mathcal{S})$ is given by:

$$\mathbf{y}_{\rm BS}(\{12\})=\mathbf{h}_1B_1+\mathbf{h}_2A_2+\mathbf{h}_3(-B_3-A_3),$$
$$\mathbf{y}_{\rm BS}(\{13\})=\mathbf{h}_1C_1+\mathbf{h}_2(-C_2-A_2)+\mathbf{h}_3A_3.$$

However, in the upcoming DL step, the BS only needs $\alpha_1B_1+\alpha_2A_2$, $\alpha_3C_1+\alpha_4A_3$, and $\alpha_5C_2+\alpha_6B_3$, for some $\alpha_i\in\mathbb{C}$ with $i\in[6]$. To extract these combinations, it employs receive beamforming (row) vectors $\mathbf{v}_{\mathcal{R}}\in\mathbb{C}^{1\times2}$, such that $\mathcal{R}\subset[3]$ with $|\mathcal{R}|=t+1=2$, that satisfy

$$\begin{cases}\mathbf{v}_{\mathcal{R}}\mathbf{h}_k=0 & k\in[3]\text{ and }k\notin\mathcal{R}\\ \mathbf{v}_{\mathcal{R}}\mathbf{h}_k\neq0 & k\in[3]\text{ and }k\in\mathcal{R}\end{cases}.$$

First, using $\mathbf{v}_{\{12\}}$ and $\mathbf{v}_{\{13\}}$, the BS can compute

$$\mathbf{v}_{\{12\}}\mathbf{y}_{\rm BS}(\{12\})=\mathbf{v}_{\{12\}}\mathbf{h}_1B_1+\mathbf{v}_{\{12\}}\mathbf{h}_2A_2,$$
$$\mathbf{v}_{\{13\}}\mathbf{y}_{\rm BS}(\{13\})=\mathbf{v}_{\{13\}}\mathbf{h}_1C_1+\mathbf{v}_{\{13\}}\mathbf{h}_3A_3,$$

which yield $\alpha_1B_1+\alpha_2A_2$ and $\alpha_3C_1+\alpha_4A_3$ by setting $\alpha_1=\mathbf{v}_{\{12\}}\mathbf{h}_1$, $\alpha_3=\mathbf{v}_{\{13\}}\mathbf{h}_1$, and so on. Next, to calculate $\alpha_5C_2+\alpha_6B_3$, the BS simply uses $\mathbf{v}_{\{23\}}$ to get

$$\mathbf{v}_{\{23\}}(\mathbf{y}_{\rm BS}(\{12\})+\mathbf{y}_{\rm BS}(\{13\}))=-\mathbf{v}_{\{23\}}\mathbf{h}_2C_2-\mathbf{v}_{\{23\}}\mathbf{h}_3B_3.$$

Now, let us review the DL step in more detail. It consists of the following $N_{\rm DL}=2$ transmissions:

$$\mathbf{x}_{\rm BS}(j)=\beta_{1,j}(\alpha_1B_1+\alpha_2A_2)\mathbf{w}_{\{12\}}+\beta_{2,j}(\alpha_3C_1+\alpha_4A_3)\mathbf{w}_{\{13\}}+\beta_{3,j}(\alpha_5C_2+\alpha_6B_3)\mathbf{w}_{\{23\}},$$

where $\{\beta_{i,j}\}$, $i\in[3]$ and $j\in[2]$, is a set of scalars selected by the BS (more explanation is provided shortly), and for $\mathcal{Q}\subset[3]$ with $|\mathcal{Q}|=2$, $\mathbf{w}_{\mathcal{Q}}\in\mathbb{C}^{2\times1}$ is the beamforming vector that suppresses the interference at user $[3]\backslash\mathcal{Q}$. Let us consider the decoding process at user 1. Applying ZF precoders, in the transmission $j\in[2]$, it observes

$$y_1(j)=\beta_{1,j}(\alpha_1B_1+\alpha_2A_2)\mathbf{h}_1^{\rm H}\mathbf{w}_{\{12\}}+\beta_{2,j}(\alpha_3C_1+\alpha_4A_3)\mathbf{h}_1^{\rm H}\mathbf{w}_{\{13\}}.$$

User 1 is interested in A_2 and A_3. As it has B_1 and C_1 in its cache, it can regenerate and remove their interference terms from $y_1(j)$ to get $\hat{y}_1(j)$. Then, using

$$\begin{bmatrix}\alpha_2\mathbf{h}_1^{\rm H}\mathbf{w}_{\{12\}}A_2\\ \alpha_4\mathbf{h}_1^{\rm H}\mathbf{w}_{\{13\}}A_3\end{bmatrix}=\begin{bmatrix}\beta_{1,1}&\beta_{2,1}\\ \beta_{1,2}&\beta_{2,2}\end{bmatrix}^{-1}\begin{bmatrix}\hat{y}_1(1)\\ \hat{y}_1(2)\end{bmatrix},\tag{10}$$

and following a simple decoding step, it can recover A_2 and A_3 interference-free. Note that the scalars $\{\beta_{i,j}\}$ should be selected such that the matrix in (10) and similar matrices for other users are invertible (this can be done, e.g., using a predefined codebook). As observed, both UL and DL steps are comprised of two transmissions, and $Q=1/\binom{K}{t}\binom{K-t-1}{L-1}=\frac{1}{3}$. Therefore, by using (6), and (7), the NDT for the UL and DL steps is given by: $T_{\rm UL}=T_{\rm DL}=\frac{2}{3}$.

B. UL Step

This step involves $N_{\rm S}=\binom{K}{t+L}$ stages, with each stage having $N_{\rm T}=\binom{t+L-1}{t}$ transmissions. During each transmission of stage $i\in[N_{\rm S}]$, a set of $t+L$ users denoted by $\mathcal{U}(i)$, encode their generated/stored subpackets and transmit them to the BS at the same time. Here, the key idea of the UL transmission strategy is to enable the BS to create the codewords of size $t+1$, required for the DL step. Without loss of generality (up to the permutation of users' indices), let us focus on the UL transmission strategy during stage 1, where users $\mathcal{U}(1)=[t+L]$ simultaneously send their encoded data to the BS. Now, for notational simplicity, for any subset $\mathcal{T}\subseteq[t+L]$ and $l\in\mathcal{T}$, we define the operator $\langle l\rangle_{\mathcal{T}}$ as follows

$$\langle l\rangle_{\mathcal{T}}=\begin{cases}\min\{i\in\mathcal{T}:i>l\} & l<\max\{i\in\mathcal{T}\}\\ \min\{i\in\mathcal{T}\} & l=\max\{i\in\mathcal{T}\}\end{cases}.\tag{11}$$

As per (11), $\langle l\rangle_{\mathcal{T}}$ returns the element in $\mathcal{T}$ that is next to l (according to a circular shift). For example, for $\mathcal{T}=\{1,2,3\}$, we have $\langle1\rangle_{\mathcal{T}}=2$, $\langle2\rangle_{\mathcal{T}}=3$ and $\langle3\rangle_{\mathcal{T}}=1$. In order to create the transmitted signals for users $k\in[t+L]$, first, we define

$$\mathcal{M}=\{\mathcal{S}:\mathcal{S}\subseteq[t+L],1\in\mathcal{S},|\mathcal{S}|=t+1\},\tag{12}$$

where $|\mathcal{M}|=N_{\rm T}$. As mentioned earlier, stage 1 is also comprised of $N_{\rm T}$ transmissions. Let us represent the transmitted signal of user $k\in[t+L]$ during stage 1 by $x^k(\mathcal{S})$, where $\mathcal{S}\in\mathcal{M}$. In this regard, for each $k\in[t+L]$ and $\mathcal{S}\in\mathcal{M}$, $x^k(\mathcal{S})$ is given by:

$$x^k(\mathcal{S})=\begin{cases}\mathsf{c}\big(W^{\langle k\rangle_{\mathcal{S}}}_{\mathcal{S}\backslash\{\langle k\rangle_{\mathcal{S}}\},q}\big) & k\in\mathcal{S}\\ -\sum\limits_{j\in\mathcal{S}}\mathsf{c}\Big(W^{\langle k\rangle_{\mathcal{S}\cup\{k\}\backslash\{j\}}}_{\mathcal{S}\cup\{k\}\backslash\{j,\langle k\rangle_{\mathcal{S}\cup\{k\}\backslash\{j\}}\},q}\Big) & k\notin\mathcal{S}\end{cases},\tag{13}$$

[2]The generalized UL and DL steps with noisy-channels are discussed in Section IV-B and Section IV-C.

where $q \in [\binom{K-t-1}{L-1}]$ increases sequentially after each transmission to ensure none of the subpackets is transmitted twice. As mentioned in Definition 1, $\mathsf{c}\big(W^n_{\mathcal{P},q}\big)$ is the encoded signal of $W^n_{\mathcal{P},q}$, where $\mathsf{c}: \mathbb{F}_{2^f} \to \mathbb{C}$. In addition, for all $\mathcal{Q} \subseteq [t+L]$ with $|\mathcal{Q}| = t+1$ and $k \in [t+L]$, it is assumed that:

$$\mathbb{E}\left[\left|\mathsf{c}\big(W^{\langle k\rangle_{\mathcal{Q}}}_{\mathcal{Q}\backslash\{\langle k\rangle_{\mathcal{Q}}\},q}\big)\right|^2\right] = \begin{cases} P_{\mathrm{UL}} & k=1 \\ \frac{P_{\mathrm{UL}}N_{\mathrm{T}}}{\binom{t+L-2}{t-1}+(t+1)\binom{t+L-2}{t}} & k\neq 1 \end{cases}, \tag{14}$$

where P_{UL} represents the transmit power of user 1 during each transmission of stage 1, i.e., $P_1(j) = P_{\mathrm{UL}}$ for all $j \in [N_{\mathrm{T}}]$. As per (13), in any stage $i \in [N_{\mathrm{S}}]$, among the $t+L$ users, there is one user that always transmits single encoded subpackets to the BS in all the transmissions. However, the other $t+L-1$ users involved in that stage will transmit a superposition of $t+1$ encoded files in at least one transmission. Accordingly, for any stage $i \in [N_{\mathrm{S}}]$, the transmit power of the user that transmits a single encoded subpacket to the BS in all transmissions is fixed at P_{UL}. In (14), the expectation operates on the uniformly i.i.d. random variable $\mathsf{c}\big(W^{\langle k\rangle_{\mathcal{Q}}}_{\mathcal{Q}\backslash\{\langle k\rangle_{\mathcal{Q}}\},q}\big)$, as it is assumed that $W^{\langle k\rangle_{\mathcal{Q}}}_{\mathcal{Q}\backslash\{\langle k\rangle_{\mathcal{Q}}\},q}$ is an i.i.d. random variable uniformly distributed on $\mathbb{F}_{2^f}$.

Lemma 1. Assuming E_k as the energy consumption of user $k \in [K]$ in the UL step, the proposed encoding scheme in (14) satisfies the users' energy constraints $E_k = P_{\mathrm{UL}}N_{\mathrm{T}}\binom{K-1}{t+L-1}$. Hence, although the transmit power of users may vary during each UL transmission, each user consumes an equal total amount of energy throughout all transmissions of the UL step.

Proof: The proof is relegated to [30, Appendix A]. ■

Definition 2. For the subset $\mathcal{Q} \subseteq [t+L]$ with $|\mathcal{Q}| = t+1$, define the beamforming (row) vector $\mathbf{v}_{\mathcal{Q}} \in \mathbb{C}^{1\times L}$ as follows

$$\begin{cases} \mathbf{v}_{\mathcal{Q}}\mathbf{h}_k = 0 & k \in [t+L] \text{ and } k \notin \mathcal{Q} \\ \mathbf{v}_{\mathcal{Q}}\mathbf{h}_k \neq 0 & k \in \mathcal{Q} \end{cases}. \tag{15}$$

Theorem 1. Using the proposed UL transmission strategy, during stage 1, the BS is able to create all codewords of size $t+1$ in the set $\mathcal{A} \cup \mathcal{B}$, where

$$\mathcal{A} = \Big\{ \sum_{k\in\mathcal{S}} \mathbf{v}_{\mathcal{S}}\mathbf{h}_k \mathsf{c}\big(W^{\langle k\rangle_{\mathcal{S}}}_{\mathcal{S}\backslash\{\langle k\rangle_{\mathcal{S}}\},q}\big) + \mathbf{v}_{\mathcal{S}}\mathbf{n}_{\mathrm{BS}} : \mathcal{S} \in \mathcal{M} \Big\}, \tag{16}$$

$$\mathcal{B} = \Big\{ \sum_{k\in\mathcal{R}} -\mathbf{v}_{\mathcal{R}}\mathbf{h}_k \mathsf{c}\big(W^{\langle k\rangle_{\mathcal{R}}}_{\mathcal{R}\backslash\{\langle k\rangle_{\mathcal{R}}\},q}\big) + \mathbf{v}_{\mathcal{R}}\mathbf{n}_{\mathrm{BS}} : \mathcal{R} \subseteq [2:t+L], |\mathcal{R}| = t+1 \Big\} \tag{17}$$

include noisy versions of the codewords of size $t+1$ (that are needed in the DL step). Moreover, the NDT for UL is

$$T_{\mathrm{UL}} = \frac{K-t}{t+L}. \tag{18}$$

Proof: The proof is provided in [30, Appendix B]. ■

Using Theorem 1 and changing the users' indices in the sets A and B, for any stage $i \in [N_{\mathrm{S}}]$, one can simply show that the BS is able to create all codewords of size $t+1$, required for the DL step. Fig. 2 compares the achievable UL NDT of the proposed scheme with the reference strategies in Section III.

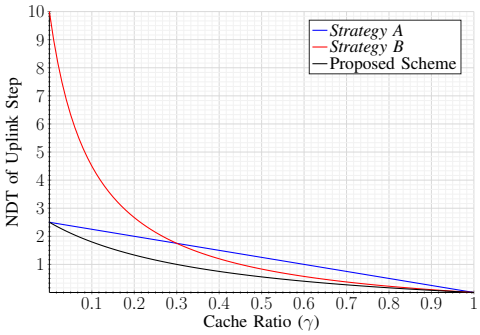

Fig. 2. Comparison between the achievable UL normalized delivery time of the proposed scheme with strategies *A* and *B* in a MISO network with $K = 10$ and $L = 4$.

C. DL Step

The DL step, inspired by [5], [6], is comprised of N_{S} stages, each with N_{T} DL transmissions. Similar to Section IV-B, we focus on stage 1, where the BS serves users $\mathcal{U}(1) = [t+L]$. During stage 1, for each $\mathcal{Q} \subseteq [t+L]$ with $|\mathcal{Q}| = t+1$, the BS generates a random vector $\mathbf{a}_{\mathcal{Q}} \in \mathbb{R}^{1\times N_{\mathrm{T}}}$. Then, using (16) and (17), in transmission $j \in [N_{\mathrm{T}}]$, it broadcasts a superposition of the codewords in the sets $\mathcal{A}$ and $\mathcal{B}$ as follows

$$\mathbf{x}_{\mathrm{BS}}(j) = \sum_{\mathcal{Q}\subseteq[t+L],|\mathcal{Q}|=t+1} \mathbf{a}_{\mathcal{Q}}[j] f_{\mathcal{Q}} \mathbf{w}_{\mathcal{Q}}, \tag{19}$$

where $f_{\mathcal{Q}} = \sum_{l\in\mathcal{Q}} \theta_{\mathcal{Q}}\mathbf{v}_{\mathcal{Q}}\mathbf{h}_l \mathsf{c}\big(W^{\langle l\rangle_{\mathcal{Q}}}_{\mathcal{Q}\backslash\{\langle l\rangle_{\mathcal{Q}}\},q}\big) + \mathbf{v}_{\mathcal{Q}}\mathbf{n}_{\mathrm{BS}}$, such that $\theta_{\mathcal{Q}} = 1$ if $1 \in \mathcal{Q}$ and $\theta_{\mathcal{Q}} = -1$ otherwise. Moreover, $\mathbf{w}_{\mathcal{Q}} \in \mathbb{C}^{L\times 1}$ is the precoder that suppresses the interference at the set $[t+L]\backslash\mathcal{Q}$.

Theorem 2. By adopting the introduced DL transmission strategy, each user is able to retrieve its requested file, and the achievable NDT for the DL step takes the form of:

$$T_{\mathrm{DL}} = \frac{K-t}{t+L}. \tag{20}$$

Proof: The proof is relegated to [30, Appendix C]. ■

In [30, Appendix D], we present an example to give further insight into the system performance for the proposed scheme.

V. Conclusion

This paper introduced a novel communication strategy to enhance distributed information retrieval in a network comprising multiple users. To design an inclusive setup supporting various user access scenarios, we modified the operational framework of the conventional coded caching (CC) model. In this modified CC model, all types of user access were considered as data partially stored in the cached content of users, while the multi-antenna base station (BS) is not directly connected to the library. The missing data portions were exchanged among users through an uplink (UL) step followed by a downlink (DL) step. During the UL step, users transmitted a smart combination of their generated/cached contents to the BS. In the DL step, an appropriate combination of these signals was forwarded back to the users, enabling each user to retrieve its desired data. It was shown that the UL step achieved an UL delivery time equal to that of the DL step. For future work, we aim to prove the optimality of the proposed scheme.

REFERENCES

[1] J. S. Ng, W. Y. B. Lim, N. C. Luong, Z. Xiong, A. Asheralieva, D. Niyato, C. Leung, and C. Miao, "A Comprehensive Survey on Coded Distributed Computing: Fundamentals, Challenges, and Networking Applications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1800–1837, 2021.

[2] S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, "A Fundamental Tradeoff Between Computation and Communication in Distributed Computing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 109–128, 2018.

[3] Y. Xu and H. Qi, "Distributed Computing Paradigms for Collaborative Signal and Information Processing in Sensor Networks," *Journal of Parallel and Distributed Computing*, vol. 64, no. 8, pp. 945–959, 2004.

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

[5] 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.

[6] 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.

[7] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tölli, "Low-Complexity High-Performance Cyclic Caching for Large MISO Systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3263 – 3278, May 2022.

[8] 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.

[9] 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.

[10] M. Salehi, A. Tolli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-Rate Trade-Off in Multi-Antenna Coded Caching," in *IEEE Global Communications Conference (GLOBECOM)*, December 2019, pp. 1–6.

[11] K. Wan and G. Caire, "On Coded Caching With Private Demands," *IEEE Trans. Inf. Theory*, vol. 67, no. 1, pp. 358–372, 2021.

[12] M. Abolpour, M. Salehi, and A. Tölli, "Cache-Aided Communications in MISO Networks with Dynamic User Behavior: A Universal Solution," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2023, pp. 132–137.

[13] M. Bayat, K. Wan, and G. Caire, "Coded Caching Over Multicast Routing Networks," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3614–3627, 2021.

[14] M. Salehi, K. Hooli, J. Hulkkonen, and A. Tölli, "Enhancing Next-Generation Extended Reality Applications with Coded Caching," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1371–1382, Jun. 2023.

[15] H. B. Mahmoodi, M. Salehi, and A. Tölli, "Multi-Antenna Coded Caching for Location-Dependent Content Delivery," *IEEE Trans. Wireless Commun.*, 2023, "Early Access".

[16] K. Wan, D. Tuninetti, M. Ji, G. Caire, and P. Piantanida, "Fundamental Limits of Decentralized Data Shuffling," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3616–3637, 2020.

[17] A. Elmahdy and S. Mohajer, "On the Fundamental Limits of Coded Data Shuffling for Distributed Machine Learning," *IEEE Trans. Information Theory*, vol. 66, no. 5, pp. 3098–3131, 2020.

[18] M. Adel Attia and R. Tandon, "Near Optimal Coded Data Shuffling for Distributed Learning," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7325–7349, 2019.

[19] K. Wan, H. Sun, M. Ji, D. Tuninetti, and G. Caire, "On the Optimal Load-Memory Tradeoff of Cache-Aided Scalar Linear Function Retrieval," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 4001–4018, 2021.

[20] Q. Yan and D. Tuninetti, "Robust, Private and Secure Cache-Aided Scalar Linear Function Retrieval From Coded Servers," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 968–981, 2022.

[21] V. Cristea, C. Dobre, C. Stratan, and F. Pop, "Large-Scale Distributed Computing and Applications: Models and Trends," in *Information Science Reference - Imprint of: IGI Publishing*, 2010.

[22] S. Li, Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "A Scalable Framework for Wireless Distributed Computing," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2643–2654, 2017.

[23] F. Li, J. Chen, and Z. Wang, "Wireless MapReduce Distributed Computing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6101–6114, 2019.

[24] F. Xu, M. Tao, and K. Liu, "Fundamental Tradeoff Between Storage and Latency in Cache-Aided Wireless Interference Networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, 2017.

[25] K. Yuan and Y. Wu, "Coded Wireless Distributed Computing via Interference Alignment," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2022, pp. 1016–1021.

[26] I. Hammerstrom, M. Kuhn, C. Esli, J. Zhao, A. Wittneben, and G. Bauch, "MIMO Two-Way Relaying with Transmit CSI at the Relay," in *Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2007, pp. 1–5.

[27] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the Feedback Bottleneck of Multi-Antenna Coded Caching," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2331–2348, 2022.

[28] N. Rajatheva *et. al.*, "White Paper on Broadband Connectivity in 6G," *arXiv preprint arXiv:2004.14247*, 2020.

[29] S. H. Shah and I. Yaqoob, "A survey: Internet of Things (IOT) Technologies, Applications and Challenges," in *IEEE Smart Energy Grid Engineering (SEGE)*, 2016, pp. 381–385.

[30] M. Abolpour, M. Salehi, S. Mohajer, S. P. Shariatpanahi, and A. Tölli, "Coded Multi-User Information Retrieval with a Multi-Antenna Helper Node," *arXiv preprint: 2402.00465*, 2024.