Beamforming Design for Wireless Coded Caching with Different Cache Sizes

Ayaka Urabe*, Koji Ishibashi*, MohammadJavad Salehi†, and Antti Tölli†
*Advanced Wireless & Communication Research Center (AWCC), The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
Email: urabe@awcc.uec.ac.jp, koji@ieee.org
†Centre for Wireless Communications (CWC), University of Oulu
Erkki Koiso-Kanttilan katu 3, 4.Floor, Linnanmaa, 90570, Oulu
Email: [mohammadjavad.salehi, antti.tolli]@oulu.fi

***Abstract*—This paper studies the performance of wireless coded caching over multiple-input and single-output (MISO) channels in a finite signal-to-noise power ratio (SNR) region when every user has a different cache memory size. We first propose multicast beamforming for the network with the conventional coded caching based on quadratic transform (QT) and then point out the non-optimality of the caching scheme when the spatial degree of freedom (DoF) is exploited. We hence formulate a new optimization problem to enhance the caching gain by minimizing the difference between the generated codewords. Numerical results confirm the non-optimality of the conventional coded caching in terms of the average transmission rate and the improvement of our proposed caching.**

***Index Terms*—coded caching, beamforming, cache size optimization, MISO channels, heterogeneous network**

I. Introduction

Recently, coded caching has been proposed as an enabler to mitigate peak-time traffic and pre-transmit it during the off-peak time, using local memory resources distributed in the network [1]. Similar to classic local caching scheme, coded caching operation is also composed of two phases. The first phase, named *placement phase*, is performed during the off-peak time, when the base station sends fractions of the data to users' cache memories. The second phase, named *delivery phase*, is performed during the peak time, when the base station sends the remainder of the data according to users' requests. However, unlike the classic local caching, coded caching jointly designs both placement and delivery phases to enable an additional *global caching gain* while also making the load on the shared-link independent of the user requests [1].

The seminal paper [1] focused only on the caching for a multi-user network over the error-free shared link with equal cache sizes. Coded caching for the error-free network with cache-size heterogeneity was later discussed in [2], [3], which minimized the sum of the codeword sizes in the delivery phase. The fundamental limits of the coded caching with cache-size heterogeneity and asymmetric link qualities for two user networks have been investigated in [4]. Also, multi-antenna coded caching for networks with two different cache sizes has been discussed in [5]. While these papers focused on the performances in the infinite signal-to-noise power ratio (SNR) region, beamforming design for the multi-user multiple-input single-output (MISO) channels with coded caching in the finite SNR region was studied in [6], where every user was assumed to have the same-sized cache memory, and multi-group multicast beamforming was proposed to maximize the minimum user rate while exploiting the combined benefit from the spatial multiplexing and global caching gains. However, the performance of multi-antenna coded caching with cache-size heterogeneity in the finite SNR region has not been discussed yet.

This paper studies the designs of beamforming and coded caching for multi-user MISO channels with cache-size heterogeneity in the finite SNR region. We apply the caching scheme proposed in [3] and consider the design of multi-group multicast beamforming for the heterogeneous network, using quadratic transform (QT) [7]. We further formulate a new optimization problem for the placement phase to minimize the difference of the codeword sizes in the delivery phase. Numerical results reveal that the conventional caching scheme is not optimal for MISO channels with multicast beamforming and confirm the improvement of our proposed caching. Moreover, the results show the existence of a better coded caching approach for the heterogeneous wireless networks.

Notation

In this paper, we use $\oplus$ to denote the bit-wise exclusive OR operator, and $\cup$ to represent the bit-wise concatenation operation. Similarly, $\bigcup$ denotes bit-wise concatenation of three or more binary sequences. For two binary sequences W^1 and W^2, $W^1 \backslash W^2$ represents removing the common elements of two sequences from W^1. For sets $\mathcal{A}$ and $\mathcal{B}$, $|\mathcal{A}|$ represents the cardinality of $\mathcal{A}$, $\mathcal{A} \subset \mathcal{B}$ means that $\mathcal{A}$ is one of the subsets of $\mathcal{B}$, and $\mathcal{A} \backslash \mathcal{B}$ means the set difference. The empty set is represented by $\emptyset$, and $\subsetneq_{\emptyset} \mathcal{A}$ means all subsets of $\mathcal{A}$ except for $\emptyset$. We denote the set of natural numbers from 1 to K, namely $\{1, 2, \cdots, K\}$, as $[K]$. $\mathbb{R}^+$ represents the set of positive real numbers. For a real number y, $\lfloor y \rfloor$ denotes the floor function of y and for a complex number x, $|x|$ represents the absolute value of x.

II. System Model

We assume a multi-user MISO downlink channel with an L-antenna transmitter (Tx) and K single-antenna receivers (Rxs),

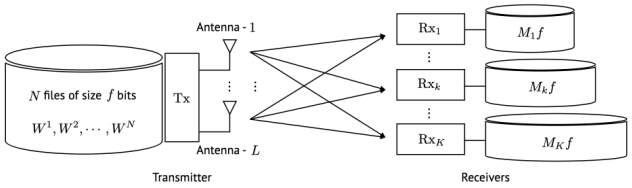

Fig. 1. Multi-user MISO downlink channel with differet cache sizes.

as shown in Figure 1. Tx has access to a library of N files, each with size f bits. The n-th file, $n \in [N]$, is labeled as $W^n \in \{0,1\}^f$, and the k-th receiver, $k \in [K]$, is labeled as Rx_k. The receiver Rx_k has a cache memory of size $M_k f$ bits, where $M_k \in [0, N]$. The cache size vector $\mathbf{M}$ is defined as $\mathbf{M} \triangleq [M_1 f, M_2 f, \cdots, M_K f]$.

In the placement phase, every file is split into 2^K subfiles of arbitrary sizes denoted by $W_{\mathcal{S}_s}$, where $s \in \left[2^K\right]$ denotes combination indexes of K Rxs, and $\mathcal{S}_s \subset [K]$ indicates the corresponding Rx indexes. More concretely, the subfile with label $\mathcal{S}_s$, namely $W_{\mathcal{S}_s}$, will be stored at Rx_k's cache memory where $k \in \mathcal{S}_s$. Let $a_{\mathcal{S}_s} \in [0,1]$ be allocation variables such that the size of the subfile $W_{\mathcal{S}_s}$ is $\lfloor a_{\mathcal{S}_s} f \rfloor$ bits. Moreover, let us define the allocation vector as $\mathbf{a} \in [0,1]^{2^K} = \left[a_{\mathcal{S}_1}, a_{\mathcal{S}_2}, \cdots, a_{\mathcal{S}_{2^K}}\right]$. Then, the relationship between the cache size vector $\mathbf{M}$ and the total amount of the subfiles in each receiver's cache memory can be formulated as

$$\mathfrak{U}(\mathbf{M}) = \left\{ \mathbf{a} \in [0,1]^{2^K} \left| \sum_{\mathcal{S}_s \subset [K]} a_{\mathcal{S}_s} = 1, \sum_{\substack{\mathcal{S}_s \subset [K]: \\ \mathcal{S}_s \ni k}} \lfloor a_{\mathcal{S}_s} f \rfloor \leq \frac{M_k f}{N} \right. \right\}.$$

The placement phase is performed during the off-peak time without any error.

In the subsequent delivery phase, every receiver Rx_k reveals its requested file W^{d_k}, where $d_k \in [N]$ represents the file index. Then, Tx generates and transmits a set of codewords according to the requested file indices and the contents cached in every Rx. The set of codewords consists of *unicast* and *multicast* ones. For efficient transmission, we first focus on multicast codewords intended to transmit data to two or more receivers. Each multicast codeword consists of multiple subfiles. However, some subfiles are further split into split-subfiles (i.e., smaller parts) as needed to avoid redundant transmissions. Therefore, the multicast codeword for $\mathrm{Rx}_k\,(k \in \mathcal{S}_s)$ can be obtained by

$$X_{\mathcal{S}_s} = \bigoplus_{k \in \mathcal{S}_s} W^{d_k} = \bigoplus_{k \in \mathcal{S}_s} \left(\bigcup_{\mathcal{S}' \in \mathcal{B}_k^{\mathcal{S}_s}} W_{\mathcal{S}'}^{d_k} \right), \quad \forall \mathcal{S}_s \in \{\mathcal{S}_s \subset [K] : 2 \leq |\mathcal{S}_s| \leq K\}, \tag{1}$$

where

$$\mathcal{B}_k^{\mathcal{S}_s} \triangleq \{\mathcal{S}' \subset [K] : \mathcal{S}_s \backslash \{k\} \subseteq \mathcal{S}', \mathcal{S}_s \cap \mathcal{S}' \neq \emptyset\}. \tag{2}$$

$\mathcal{S}'$ represents all the subsets of $\mathcal{S}_s$ that do not contain $k \in \mathcal{S}_s$, and $W_{\mathcal{S}'}^{d_k}$ denotes split-subfiles of the subfile of W^{d_k} that are available in the cache memory of Rx_j, where $j \in \mathcal{S}'$. The sizes of split-subfiles $W_{\mathcal{S}'}^{d_k}$, which constitute codeword $X_{\mathcal{S}_s}$, are defined as $u_{\mathcal{S}'}^{\mathcal{S}_s} f$ bits, where $u_{\mathcal{S}'}^{\mathcal{S}_s} \in [0, a_{\mathcal{S}_s}]$. Especially, the multicast codeword for all the Rxs is called the *broadcast codeword*. Next, we focus on unicast codewords intended to one specific receiver. The unicast codeword for Rx_k consists of subfiles that are not stored in Rx_k's cache memory and are not included in multicast codewords. Therefore, the unicast codeword for the receiver Rx_k can be obtained by

$$X_{\{k\}} = W^{d_k} \setminus \left(\bigcup_{\substack{\mathcal{S}_s \subset [K]: \\ \mathcal{S}_s \ni k}} W_{\mathcal{S}_s}^{d_k} \bigcup_{\substack{\mathcal{S}' \in \mathcal{B}_k^{\mathcal{S}'_s}: \mathcal{S}'_s \subset [K], \\ 2 \leq \left|\mathcal{S}'_s\right| \leq K}} W_{\mathcal{S}'}^{d_k} \right), \quad \forall k \in [K]. \tag{3}$$

If each Rx can reconstruct its requested file using only multicast codewords, its respective unicast codeword will not be generated. The multicast/unicast codeword $X_{\mathcal{S}_s}$ is composed of $v_{\mathcal{S}_s} f$-bits where $v_{\mathcal{S}_s} \in [0,1]$.

After generating the codewords, the Tx modulates each codeword $X_{\mathcal{S}_s}$ to a complex symbol $\tilde{X}_{\mathcal{S}_s} \in \mathbb{C}$. Without loss of generality, each symbol follows an independent and identically distributed (i.i.d) circularly symmetric complex Gaussian distribution. Then, multicast symbols (including the broadcast symbol) are transmitted simultaneously over MISO channels using appropriate beamforming vectors. At the time slot t, the signal received by $\mathrm{Rx}_k\ (k \in \mathcal{S}_s)$ can be written as

$$y_k(t) = \sum_{\mathcal{S}_s : 2 \leq |\mathcal{S}_s| \leq K} \mathbf{h}_k^H(t)\, \mathbf{w}_{\mathcal{S}_s}(t)\, \tilde{X}_{\mathcal{S}_s} + n_k(t), \tag{4}$$

where $\mathbf{h}_k^H(t) \in \mathbb{C}^{L \times 1}$ is the channel vector between the Tx and Rx_k at the time slot t. We assume that the channel is the quasi-statistic frequency non-selective Rayleigh fading channel, and each element of the channel vector $\mathbf{h}_k(t)$ follows the i.i.d complex Gaussian distribution. Also, $\mathbf{w}_{\mathcal{S}_s}(t) \in \mathbb{C}^{L \times 1}$ is the beamforming vector dedicated to $\mathrm{Rx}_k\,(k \in \mathcal{S}_s)$ at time slot t, and $n_k(t) \sim \mathcal{CN}\left(0, \sigma^2\right)$ represents the additive white Gaussian noise (AWGN).

After the transmission of multicast and broadcast codewords, each unicast codeword is sent to the Rx in subsequent time slots, following a time-division manner. In this paper, we define the transmission power in each time slot as P, and the SNR as $\mathrm{SNR} = P/\sigma^2$.

III. Conventional Caching Scheme

Based on the model described above, this section briefly describes conventional coded caching for the heterogeneous network as proposed in [3]. Note that this conventional approach is designed to minimize the total codeword size transmitted in the delivery phase, whereas multicast and broadcast symbols would be transmitted simultaneously via beamforming in our system model as described in the previous section.

There are three constraints on codewords' sizes $v_{\mathcal{S}_s}$ and split-subfile sizes $u_{\mathcal{S}'}^{\mathcal{S}_s}$ to guarantee the decodability. The first constraint can be written as

$$\sum_{\mathcal{S}_s \subsetneq \emptyset [K]: \mathcal{S}_s \ni k} v_{\mathcal{S}_s} \geq 1 - \sum_{\mathcal{S}_s \subset [K]: \mathcal{S}_s \ni k} a_{\mathcal{S}_s}, \forall k \in [K], \tag{5}$$

and represents that Tx must send all the subfiles that are not stored in the cache memory. The second constraint can be written as

$$\sum_{\mathcal{S}'\in\mathcal{B}_k^{\mathcal{S}_s}} u_{\mathcal{S}'}^{\mathcal{S}_s} = v_{\mathcal{S}_s}, \forall \mathcal{S}_s, \forall k \in \mathcal{S}_s, \tag{6}$$

and denotes that the total size of split-subfiles $W_{\mathcal{S}'}^{d_k}$ that are components of the codeword $X_{\mathcal{S}_s}$ must be equal to the size of $X_{\mathcal{S}_s}$. Finally, the third constraint can be written as

$$\begin{aligned} &\sum_{\substack{\mathcal{S}_s\subsetneq_\emptyset[K]:\mathcal{S}_s\ni k,\\ \mathcal{S}_s\cap\mathcal{S}'\neq\emptyset,\mathcal{S}_s\setminus\{k\}\subset\mathcal{S}'}} u_{\mathcal{S}'}^{\mathcal{S}_s} \le a_{\mathcal{S}_s}, \forall k \notin \mathcal{S}_s, \\ &\forall \mathcal{S}_s \in \left\{\tilde{\mathcal{S}} \subset [K] : 2 \le \left|\tilde{\mathcal{S}}\right| \le K-1\right\}, \end{aligned} \tag{7}$$

and means avoiding redundant transmissions. For a given allocation vector $\mathbf{a}$, the function of the delivery phase can be written as

$$\begin{aligned} \mathfrak{D}(\mathbf{a}) = \Bigg\{ (\mathbf{v},\mathbf{u}) \Bigg| & \sum_{\substack{\mathcal{S}_s\subsetneq_\emptyset[K]:\\ \mathcal{S}_s\ni k}} v_{\mathcal{S}_s} \ge 1 - \sum_{\substack{\mathcal{S}_s\subset[K]:\\ \mathcal{S}_s\ni k}} a_{\mathcal{S}_s}, \forall k \in [K], \\ & \sum_{\mathcal{S}'\in\mathcal{B}_k^{\mathcal{S}_s}} u_{\mathcal{S}'}^{\mathcal{S}_s} = v_{\mathcal{S}_s}, \forall \mathcal{S}_s \subsetneq_\emptyset [K], \forall k \in \mathcal{S}_s, \\ & \sum_{\substack{\mathcal{S}_s\subsetneq_\emptyset[K]:\mathcal{S}_s\ni k,\\ \mathcal{S}_s\cap\mathcal{S}'\neq\emptyset,\mathcal{S}_s\setminus\{k\}\subset\mathcal{S}'}} u_{\mathcal{S}'}^{\mathcal{S}_s} \le a_{\mathcal{S}_s}, \forall k \notin \mathcal{S}_s, \\ & \forall \mathcal{S}_s \in \left\{\tilde{\mathcal{S}} \subset [K] : 2 \le \left|\tilde{\mathcal{S}}\right| \le K-1\right\}, \\ & 0 \le u_{\mathcal{S}'}^{\mathcal{S}_s} \le a_{\mathcal{S}_s}, \forall \mathcal{S}_s \subsetneq_\emptyset [K], \forall \mathcal{S}' \in \bigcup_{k\in\mathcal{S}_s} \mathcal{B}_k^{\mathcal{S}_s} \Bigg\}, \end{aligned} \tag{8}$$

where the vector $\mathbf{v}$ includes codeword size values $v_{\mathcal{S}_s}$, and the vector $\mathbf{u}$ contains split-subfile size elements $u_{\mathcal{S}'}^{\mathcal{S}_s}$. In [3], the optimization problem for maximizing the caching gain is formulated as

$$\begin{aligned} \underset{\mathbf{a},\mathbf{u},\mathbf{v}}{\text{minimize}} \quad & \sum_{\mathcal{S}_s} v_{\mathcal{S}_s}, \\ \text{subject to} \quad & \mathbf{a} \in \mathfrak{U}(\mathbf{M}) \text{ and } (\mathbf{u},\mathbf{v}) \in \mathfrak{D}(\mathbf{a}). \end{aligned} \tag{9}$$

The optimized allocation variables and codewords are then obtained via a linear programming problem.

IV. Beamforming for the MISO Setup

In this section, we propose our multicast beamformer design for MISO transmissions, assuming that the conventional coded caching scheme for the heterogeneous setup in [3] is applied. As clear from the optimization problem in (9), the sizes of the generated codewords could be different, whereas the transmission for all the Rxs has to be done simultaneously. This leads to the symmetric rate design (i.e., maximizing the minimum rate) for the beamformers. Note that conventional works assuming identical cache sizes are also based on the symmetric rate design [6], [8]. However, in our approach, we have to consider the difference of the generated codeword sizes besides the effect of the channel differences among Rxs.

For simplicity of explanation and without loss of generality, we consider a specific network model as *Scenario 1*: $(L,N,K)=(2,3,3)$ and $\mathbf{M}=[1.2f, 1.5f, 2.1f]$. In this scenario, the required file indices are assumed to be $[d_1,d_2,d_3]=[1,2,3]$. Also, we define three subsets of Rxs' indicies as $\mathcal{S}_1 \triangleq \{1,2\}$, $\mathcal{S}_2 \triangleq \{1,3\}$, and $\mathcal{S}_3 \triangleq \{2,3\}$.

Solving (9) results in the following three codewords:

$$X_{\mathcal{S}_1} = \left(W_{\{2\}}^1 \cup W_{\{2,3\}}^{1'}\right) \oplus \left(W_{\{1\}}^2 \cup W_{\{1,3\}}^2\right), \tag{10a}$$

$$X_{\mathcal{S}_2} = W_{\{1\}}^3 \oplus \left(W_{\{3\}}^1 \cup W_{\{2,3\}}^{1''}\right), \tag{10b}$$

$$X_{\mathcal{S}_3} = W_{\{2\}}^3 \oplus W_{\{3\}}^1, \tag{10c}$$

where $W_{\{2,3\}}^{1'}$ and $W_{\{2,3\}}^{1''}$ are split-subfiles of the subfile $W_{\{2,3\}}^1$. The sizes of $X_{\mathcal{S}_1}$, $X_{\mathcal{S}_2}$, and $X_{\mathcal{S}_3}$ are $2f/5$, $f/5$, and $f/10$ bits, respectively. Also, the total transmission size is $7f/10$ bits.

Let us use $\tilde{X}_{\mathcal{S}_1}$, $\tilde{X}_{\mathcal{S}_2}$, and $\tilde{X}_{\mathcal{S}_3}$ to denote the transmission symbols corresponding to $X_{\mathcal{S}_1}$, $X_{\mathcal{S}_2}$, and $X_{\mathcal{S}_3}$, respectively. These symbols are transmitted via multicast beamforming exploiting spatial degree-of-freedom (DoF). In the following, we omit the time index t for the brevity of notation. Moreover, we focus on the received signal at Rx_1. From (10), the symbol $\tilde{X}_{\mathcal{S}_3}$ does not include Rx_1's required subfiles, and thus appears as an interference term. Hence, the received signal-to-interference-plus-noise power ratio (SINR) values at Rx_1 corresponding to the symbols $X_{\mathcal{S}_1}$ and $X_{\mathcal{S}_2}$ can be written as:

$$\gamma_{\mathcal{S}_1}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3}) = \left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_1}\right|^2 \Big/ \left(\left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_3}\right|^2 + \sigma^2\right), \tag{11a}$$

$$\gamma_{\mathcal{S}_2}^{(1)}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) = \left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_2}\right|^2 \Big/ \left(\left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_3}\right|^2 + \sigma^2\right), \tag{11b}$$

respectively. From (4), the channel to each Rx is a Gaussian multiple access channel, and hence, the corresponding rates to these SINR values would be

$$R_{\mathcal{S}_1}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3}) = 2\log_2\left(1+\gamma_{\mathcal{S}_1}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})\right), \tag{12a}$$

$$R_{\mathcal{S}_2}^{(1)}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) = 2\log_2\left(1+\gamma_{\mathcal{S}_2}^{(1)}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})\right), \tag{12b}$$

$$\begin{aligned} R_{\mathrm{sum}}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) = \log_2\Big(1+\gamma_{\mathcal{S}_1}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3}) \\ + \gamma_{\mathcal{S}_2}^{(1)}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})\Big). \end{aligned} \tag{12c}$$

Now, as $X_{\mathcal{S}_1}$ and $X_{\mathcal{S}_2}$ should be decoded simultaneously, the achievable rate at Rx_1 would be the minimum value of the rates in (12a) – (12c). In other words

$$\begin{aligned} R^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) \triangleq \min\Big\{ & R_{\mathcal{S}_1}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3}), \\ & R_{\mathcal{S}_2}^{(1)}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}), R_{\mathrm{sum}}^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})\Big\}. \end{aligned} \tag{13}$$

In order to calculate the transmission time, as the total size of the codewords for Rx_1 is $0.6f$ bits, its required transmission time, denoted by $T^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})$, is given by (cf. [6]):

$$T^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) = 0.6f \,/ R^1(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}), \tag{14}$$

and the transmission times for Rx_2 and Rx_3, denoted by $T^{(2)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})$ and $T^{(3)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})$, could be calculated similarly. Since the transmissions for all the Rxs should be done simultaneously, the resulting downlink transmission time to decode all the required files, denoted by $T(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})$, is given by

$$T(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) = \max\{T^{(1)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}),$$

$$T^{(2)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}),T^{(3)}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})\}. \quad (15)$$

Finally, as the total required file size for every Rx is f bits, the total symmetric rate can be calculated as

$$R(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}) \triangleq f/T(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}). \quad (16)$$

As clear from (12) to (16), this symmetric rate depends not only on the channel coefficients given in the delivery phase but also the codeword sizes pre-defined in the placement phase.

As discussed in [6], an appropriate approach for finding optimized beamformers is to first write the optimization problem in the epigraph form. Using an auxiliary variable τ as the inverse of the transmission time, this can be done as

$$\underset{\tau,\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}}{\text{maximize}} \quad \tau \quad (17a)$$

$$\text{subject to} \quad \|\mathbf{w}_{\mathcal{S}_1}\|^2+\|\mathbf{w}_{\mathcal{S}_2}\|^2+\|\mathbf{w}_{\mathcal{S}_3}\|^2 \le P, \quad (17b)$$

$$\tau \le \frac{R^{(1)}_{\mathcal{S}_1}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})}{0.6f}, \tau \le \frac{R^{(1)}_{\mathcal{S}_2}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.6f},$$
$$\tau \le \frac{R^{(1)}_{\text{sum}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.6f}, \quad (17c)$$

$$\tau \le \frac{R^{(2)}_{\mathcal{S}_1}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2})}{0.5f}, \tau \le \frac{R^{(2)}_{\mathcal{S}_3}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.5f},$$
$$\tau \le \frac{R^{(2)}_{\text{sum}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.5f}, \quad (17d)$$

$$\tau \le \frac{R^{(3)}_{\mathcal{S}_2}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2})}{0.3f}, \tau \le \frac{R^{(3)}_{\mathcal{S}_3}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})}{0.3f},$$
$$\tau \le \frac{R^{(3)}_{\text{sum}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.3f}. \quad (17e)$$

Obviously, the SINR functions are non-convex, and hence, it is difficult to solve this optimization problem efficiently. We thus apply the QT [7], [9] to reformulate the non-convex optimization problem as a convex one.

Focusing on $R^{(1)}_{\mathcal{S}_1}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})$, the SINR after applying QT can be written as below with auxiliary variable $y^{(1)}_{\mathcal{S}_1}$:

$$G^{(1)}_{\text{QT}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3}) \triangleq \left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_1}\right|^2 \Big/ \left\{\left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_3}\right|^2+\sigma^2\right\}$$
$$= 2y^{(1)}_{\mathcal{S}_1}\sqrt{\left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_1}\right|^2} - \left(y^{(1)}_{\mathcal{S}_1}\right)^2\left(\left|\mathbf{h}_1^H\mathbf{w}_{\mathcal{S}_3}\right|^2+\sigma^2\right), \quad (18)$$

where the auxiliary variable $y^{(1)}_{\mathcal{S}_1}$ is built using approximate fixed beamforming vectors $\bar{\mathbf{w}}_{\mathcal{S}_1}$ and $\bar{\mathbf{w}}_{\mathcal{S}_3}$, as follows

$$y^{(1)}_{\mathcal{S}_1} = \sqrt{\left|\mathbf{h}_1^H\bar{\mathbf{w}}_{\mathcal{S}_1}\right|^2} \Big/ \left(\left|\mathbf{h}_1^H\bar{\mathbf{w}}_{\mathcal{S}_3}\right|^2+\sigma^2\right). \quad (19)$$

Applying QT, the resulting rate is given by

$$R^{(1),\text{QT}}_{\mathcal{S}_1}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3}) = 2\log_2\left(1+G^{(1)}_{\text{QT}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})\right). \quad (20)$$

Using the same procedure, the resulting convex optimization problem is

$$\underset{\tau,\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3}}{\text{maximize}} \quad \tau \quad (21a)$$

$$\text{subject to} \quad \|\mathbf{w}_{\mathcal{S}_1}\|^2+\|\mathbf{w}_{\mathcal{S}_2}\|^2+\|\mathbf{w}_{\mathcal{S}_3}\|^2 \le P, \quad (21b)$$

$$\tau \le \frac{R^{(1),\text{QT}}_{\mathcal{S}_1}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})}{0.6f}, \tau \le \frac{R^{(1),\text{QT}}_{\mathcal{S}_2}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.6f},$$
$$\tau \le \frac{R^{(1),\text{QT}}_{\text{sum}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.6f}, \quad (21c)$$

$$\tau \le \frac{R^{(2),\text{QT}}_{\mathcal{S}_1}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2})}{0.5f}, \tau \le \frac{R^{(2),\text{QT}}_{\mathcal{S}_3}(\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.5f},$$
$$\tau \le \frac{R^{(2),\text{QT}}_{\text{sum}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.5f}, \quad (21d)$$

$$\tau \le \frac{R^{(3),\text{QT}}_{\mathcal{S}_2}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2})}{0.3f}, \tau \le \frac{R^{(3),\text{QT}}_{\mathcal{S}_3}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_3})}{0.3f},$$
$$\tau \le \frac{R^{(3),\text{QT}}_{\text{sum}}(\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3})}{0.3f}. \quad (21e)$$

By repeating the update of the auxiliary variables and solving the optimization problem in (21) iteratively, beamforming vectors are optimized. This process is summarized in Algorithm IV.

Algorithm 1 Proposed Beamforming Design

Input: $\bar{\mathbf{w}}_{\mathcal{S}_1},\bar{\mathbf{w}}_{\mathcal{S}_2},\bar{\mathbf{w}}_{\mathcal{S}_3}$: Initial values of beamforming vectors.
τ : The objective value in (17).
$\text{itr}=1$: The index of the number of iteration.
$\text{itr}_{\max}\in\mathbb{N}$:The number of maximum iteration.

1: **for** $\text{itr}=1$ to $\text{itr}_{\max}$ **do**
2: Update the auxiliary variables given by (19).
3: Update the beamforming vectors given by (21).
4: **end for**
5: **return** $\mathbf{w}_{\mathcal{S}_1},\mathbf{w}_{\mathcal{S}_2},\mathbf{w}_{\mathcal{S}_3},\tau$

V. Proposed caching scheme

In the previous section, we discussed beamformer design for minimizing the total transmission time, considering the difference of channel coefficients and codeword sizes in the delivery phase among Rxs. However, the placement was done using (9), which is not necessarily optimal for MISO setups where it is possible to transmit multiple multicast codewords using the spatial DoF.

In order to consider the possibility of multi-group multicast transmissions, we use a new optimization problem given by

$$\underset{\mathbf{a},\mathbf{u},\mathbf{v}}{\text{minimize}} \quad \sum_{\mathcal{S}_s} v_{\mathcal{S}_s} - \lambda_1 v_{[K]} + \lambda_2 \sum_{\mathcal{S}_s:|\mathcal{S}_s|=1} v_{\mathcal{S}_s} \quad (22a)$$

$$\text{subject to} \quad \mathbf{a}\in\mathfrak{U}(\mathbf{m}) \text{ and } (\mathbf{u},\mathbf{v})\in\mathfrak{D}(\mathbf{a}), \quad (22b)$$

where $\lambda_1 \in \mathbb{R}^+$ is a hyperparameter for the broadcast codeword size, and $\lambda_2 \in \mathbb{R}^+$ is another hyperparameter for the total size of unicast codewords. Using these two hyperparameters, the second and the third terms in (22) force generating a broadcast codeword and avoiding unicast codewords, respectively. Generating the broadcast codeword is preferred as it mitigates the size difference in unicast codewords, which require additional time slots and hence deteriorate the performance. The optimization problem (22) is solved via linear programming problem as in (9). In this

paper, we optimize these hyperparameters numerically using a brute-force search process.

VI. Numerical Results

This section evaluates our proposed beamforming design and caching. In the following, we assume two scenarios with different total cache sizes: Scenario 1 as described in Section IV, and *Scenario 2* with $(L, N, K) = (2, 3, 3)$ and $\mathbf{M} = [1.2f, 1.5f, 1.8f]$. Note that the total cache size of users in Scenario 1 is larger than that of Scenario 2. Also, the codewords generated by the conventional caching scheme for Scenario 2 are three multicast codewords of size $10f/30$ bits, $7f/30$ bits, and $4f/30$ bits, and a broadcast codeword of size $f/30$ bits. Moreover hyperparamters λ_1 and λ_2 were searched over 0 to 50, and both are set as 1 in the rest of the paper, which achieved the highest average rate.

Table I shows codewords obtained by the conventional caching scheme [3] and our proposed caching in Scenario 1, where $\mathcal{S}_4 \triangleq [3]$. Note that, in case of Scenario 2, the codewords obtained by the proposed caching are identical to ones resulting from the conventional caching scheme.

Figure 2 shows the average rates for both scenarios. As references, the performances of time division multiple access (TDMA) are also presented, where the beamformer is designed to maximize the worst SNR among Rxs. Furthermore, the performances of TDMA with the original coded caching for equal cache sizes [1] and multi-antenna coded caching for equal cache sizes [6], [8] are presented, where all the cache sizes are set to the minimum one among Rxs. From the figure, when TDMA is used, Scenario 1 with the larger total cache size (blue dashed curve with square markers) shows the higher average rate than Scenario 2 (red dashed curve with circle markers). However, when our proposed multicast beamforming is applied, Scenario 2 (blue dotted curve with square markers) performs better than Scenario 1 (red dotted curve with circle markers) even though Scenario 2 has a smaller total cache size in the network than Scenario 1. Moreover, Scenario 1 with our proposed caching (red solid curve with star markers) is superior to that with the conventional caching. These facts clearly reveal that the conventional caching is not optimal over MISO channels when spatial multiplexing is used, and prove the existence of better caching schemes for MISO channels with heterogeneous cache memories. Note that the performance of Scenario 1 with the proposed caching scheme is still inferior to that of Scenario 2. Hence, one judicious option is to use the caching designed for Scenario 2 even in case of Scenario 1. In general, the optimum caching scheme design for wireless heterogeneous network is still an open problem.

VII. Conclusion and Future Works

This paper studied the design of coded caching schemes for multi-user MISO channels with different cache memory sizes. We discussed multicast beamforming design for the considered network setup and proposed a new optimization problem for the placement phase. Numerical results revealed that the conventional caching scheme is not optimal for MISO channels with multicast beamforming and proved the existence of better coded caching approaches for heterogeneous wireless networks. Theoretical design and DoF analysis of this network remain as future work.

TABLE I
Generated codewords in *Scenario 1*.

codewords	Conventional Caching Scheme [3]	Proposed Caching Scheme
$X_{\mathcal{S}_1}$	$12f/30$	$8f/30$
$X_{\mathcal{S}_2}$	$6f/30$	$8f/30$
$X_{\mathcal{S}_3}$	$3f/30$	$5f/30$
$X_{\mathcal{S}_4}$	-	$2f/30$

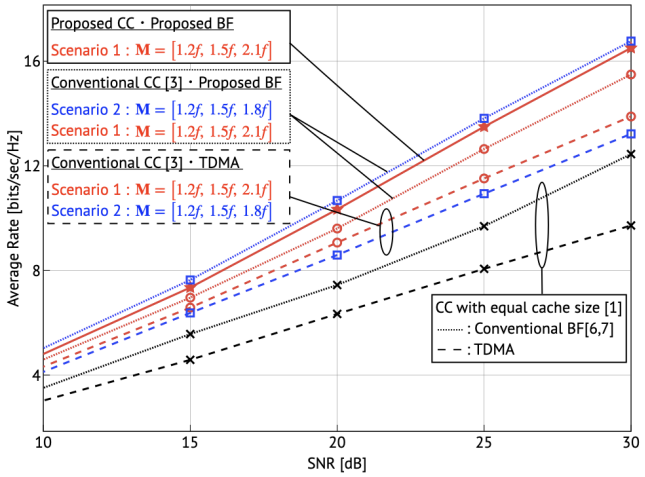

Fig. 2. Average rates of conventional and proposed wireless coded caching (CC) approaches with proposed multicast beamforming (BF) or TDMA.

Acknowledgement

This research was supported by the Ministry of Internal Affairs and Communications in Japan (JPJ000254).

References

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

[2] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Device-to-device coded caching with heterogeneous cache sizes," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[3] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Coded caching for heterogeneous systems: An optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5321–5335, May 2019.

[4] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gündüz, "Coded Caching With Asymmetric Cache Sizes and Link Qualities: The Two-User Case," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6112–6126, 2019.

[5] E. Lampiris and P. Elia, "Full Coded Caching Gains for Cache-Less Users," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7635–7651, 2020.

[6] 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, Jan. 2020.

[7] K. Shen and W. Yu, "Fractional programming for communication systems - Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, Mar. 2018.

[8] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multicast beamformer design for coded caching," in *Proc. 2018 IEEE Int. Symp. Inf. Theory (ISIT)*, June 2018, pp. 1914–1918.

[9] Z. Wang, L. Vandendorpe, M. Ashraf, Y. Mou, and N. Janatian, "Minimization of sum inverse energy efficiency for multiple base station systems," in *2020 IEEE Wireless Commun. Netw. Conf.(WCNC)*, May 2020, pp. 1–7.