Coded Caching with Asymmetric Cache Sizes and Link Qualities: The Two-User Case

Daming Cao, Deyao Zhang, Pengyao Chen, Nan Liu, Wei Kang, and Deniz Gündüz

***Abstract*—Centralized coded caching problem is studied for the two-user scenario, considering heterogeneous cache capacities at the users and private channels from the server to the users, in addition to a shared channel. Optimal caching and delivery strategies that minimize the worst-case delivery latency are presented for an arbitrary number of files. The converse proof follows from the sufficiency of file-index-symmetric caching and delivery codes, while the achievability is obtained through memory-sharing among a number of special memory–capacity pairs. The optimal scheme is shown to exploit the private link capacities by transmitting part of the corresponding user‘s request in an uncoded fashion. When there are no private links, the results presented here improve upon the two known results in the literature, namely, i) equal cache capacities and arbitrary number of files; and ii) unequal cache capacities and two files. The results are then extended to the caching problem with heterogeneous distortion requirements.**

I. Introduction

In their seminal paper [1], Maddah-Ali and Niesen propose a framework for coded caching and delivery to exploit the cache memories available at user devices to relieve the traffic burden at peak traffic periods. They consider a server holding N files of equal size, serving K users, each equipped with a local cache memory sufficient to store M files. Users' caches are proactively filled before they reveal their demands, called the *placement phase*, over a low-traffic period. In the ensuing *delivery phase*, each user requests a single file from the library, which are delivered simultaneously over an error-free shared link. The coded caching scheme proposed in [1] creates *multicasting* opportunities by jointly designing the content placement and delivery, resulting in a global caching gain. The optimal caching and delivery scheme for the general coded caching problem, in terms of the worst case delivery latency, remains open despite ongoing research efforts. While many schemes have been proposed in [2]–[8], and converse results are presented in [1], [9]–[12], the bounds obtained do not match in general except in some special cases, i.e., $N = K = 2$ [1], $K = 2$ and arbitrary N [12], $K = 3$ and $N = 2$ [12]. The optimal caching and delivery strategy is characterized in [9] when the cache placement is constrained to be *uncoded*.

Due to the difficulty of the problem, most of the literature follows the symmetric setting of [1], in which all the users are equipped with the same cache size, and the link between the server and the users is an error-free shared bit-pipe. However, in practice, owing to the heterogeneous nature of devices, the equal cache assumption is often not realistic. Furthermore, the delivery channel quality may be different for different users, while limiting the model to a single shared link is equivalent to targeting the user with the worst channel quality. Heterogeneous cache sizes with a shared link has been considered in [13]–[18], heterogeneous link qualities has been considered in [19]–[21], while a few works have studied heterogeneity in *both* the cache sizes and link qualities [22]–[26]. References [23]–[28] take a more general approach, and consider a broadcast channel from the server to the users during the delivery phase. These papers propose cache allocation among users with different channel qualities, where it is shown that a general rule of thumb is to assign more cache to users with weaker links. We note, however, that, the cache capacity, in practice, cannot be distributed across user devices dynamically, but rather given as a fixed parameter. For example, a mobile phone with a weak link to the server is unlikely to have a larger cache than a laptop with a stronger link. Hence, we assume that both the cache capacities and the link qualities are given, and we aim to find the best *centralized* caching and delivery strategy that minimizes the worst-case delivery latency. In centralized caching, we assume that the cache and link capacities of the users that participate in the delivery phase are known in advance during the placement phase, although their particular demands are not known. Therefore, their cache contents can be coordinated in a centralized manner.

To model the heterogeneous link qualities of K users we consider orthogonal common and private links from the server to the users. The multicast rate tuple is specified by $(R_{\mathcal{D}})_{D\subseteq\{1,2,\ldots,K\}}$, where $R_{\mathcal{D}}$ is the rate of the common message that can be reliably transmitted to the subset of users in $\mathcal{D}$. In practice, this might model a scenario with orthogonal error-free finite-capacity channels for each subset of users, either because an orthogonal frequency band is allocated for every subset of users, or because the underlying physical layer coding and modulation schemes that dictate these rates are fixed, and the coded caching scheme is implemented on a higher layer of the communication network stack. This setting

D. Cao and W. Kang are with the School of Information Science and Engineering and Shing-Tung Yau Center, Southeast University, Nanjing, China (email: {dmcao,wkang}@seu.edu.cn). Deyao Zhang, Pengyao Chen and Nan Liu are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China (email: {dyzhang, chenpy, nanliu}@seu.edu.cn). D. Gündüz is with the Imperial College London, London SW7 2AZ, U.K. (email: d.gunduz@imperial.ac.uk).

This work was partially supported by the National Natural Science Foundation of China under Grants 61571123 and 61571122, National Major Project (2017ZX03001002-004), the Research Fund of National Mobile Communications Research Laboratory, Southeast University (No. 2019A03) and Six talent peaks project in Jiangsu Province. The work of D. Gunduz was supported by the European Research Council (ERC) through Starting Grant BEACON (agreement 677854).

The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT) in Vail, Colorado in June 2018.

is also related to the multi-sender index coding problem [29]–[33], in which the transmitter does not have the freedom to design the placement phase.

Given the cache capacities $(M_1, M_2, \ldots, M_K)$, and the multicast rate tuple $(R_{\mathcal{D}})_{\mathcal{D}\subseteq\{1,2,\ldots,K\}}$ for the delivery phase, we are interested in finding the optimal centralized caching and delivery scheme that minimizes the delivery latency across all demand combinations. The optimal strategy will show us how to best utilize the heterogeneous caches at the users, and what to transmit over the shared and private links for the most efficient use of the communication resources.

In this paper, we focus on the special case of $K = 2$ users, while the number of files, N, is arbitrary. We reemphasize that the optimal solution has been open even in this limited setting. Moreover, the solution presented for this special case will provide insights into the more general problem. In particular, we characterize the optimal cache and delivery strategy for a generic scenario defined with five parameters $(M_1, M_2, R_c, R_{p1}, R_{p2})$, where R_c is the rate of the common message that can be transmitted to both users, while R_{pk} is the rate of the private message to User k, $k = 1, 2$. The main contributions of this paper can be summarized as:

1) We provide a converse result based on an observation by Tian [12] that it suffices to consider *file-index symmetric* caching schemes in this problem.
2) For $K = 2$ users with heterogeneous caches and only a shared common link, we identify the optimal cache and delivery strategy for an arbitrary number of $N \geq 3$ files. Previously, only the case of $M_1 = M_2, N \geq 2$ [12], and $M_1 \neq M_2$ and $N = 2$ [18] cases were solved.
3) For the general case with one common and two private links, we find the optimal caching and delivery strategy for $N \geq 2$ files. We show that: i) the private links are used to transmit part of the requested files in an uncoded fashion; ii) for the user with the smaller-capacity private link, part of the request will be transmitted over the shared common link in an uncoded fashion unless that part of all the files are cached in the said user's cache.
4) By identifying the parallels between the coded caching problem with one common and two private links studied here, and the coded caching problem with heterogeneous distortion requirements studied in [18] for the case of $K = 2$ users with heterogeneous caches, we prove the optimal caching and delivery strategy also for that problem for $N \geq 3$ files. In [18], the optimal cache and delivery strategy is characterized only for $N = 2$.

A. Notations

Throughout this paper, for $n \in \mathbb{Z}^+$, $[n]$ denotes the index set $\{1, 2, \ldots, n\}$. Entropy $H(X)$ and mutual information $I(X;Y)$ are defined in the standard way.

II. System Model

We consider a coded caching problem with one server connected to $K = 2$ users. The server has access to a database of N independent equal-size files, each consisting of F bits, denoted by $W_1, W_2, \ldots, W_N$. Both users are equipped with local caches, with capacities of M_1F and M_2F bits, respectively. The system operates in two phases. In the *placement phase*, the users are given access to the entire database and fill their caches in an error-free manner. The contents of the caches after the placement phase are denoted by Z_1 and Z_2, respectively. In the delivery phase, each user requests a single file from the server, where d_k denotes the index of the file requested by User k, $k = 1, 2$. After receiving the demand pair $D \triangleq (d_1, d_2)$, the server transmits messages over the available shared and private channels to the two users to satisfy their demands.

In [1] and most of the following literature, the delivery channel is modeled as an error-free shared link of limited capacity. However, in practice, the channels between the server and the users are typically of different quality. Thus, we model the delivery channel as consisting of two private error-free links with capacities $R_{p1}F$ and $R_{p2}F$ bits per unit time to User 1 and User 2, respectively, in addition to a shared link of capacity R_cF bits per unit time.

A *caching and delivery code* for this system consists of

1) two caching functions
$$\phi_k : [2^F]^N \to [2^{M_kF}], \quad k = 1, 2,$$
which map the database into cache contents of the users, denoted by $Z_k = \phi_k(W_1, W_2, \cdots, W_N)$, $k = 1, 2$.
2) N^2 encoding functions, one for each demand pair,
$$f^D : [2^F]^N \to [2^{r_c^DF}] \times [2^{r_{p1}^DF}] \times [2^{r_{p2}^DF}],$$
that map the files to the messages transmitted over the common and private links, denoted as X_c^D, X_{p1}^D and X_{p2}^D, respectively, i.e., $(X_c^D, X_{p1}^D, X_{p2}^D) \triangleq f^D(W_1, W_2, \cdots, W_N)$.
3) $2N^2$ decoding functions, one for each demand pair,
$$g_k^D : [2^{M_kF}] \times [2^{r_c^DF}] \times [2^{r_{pk}^DF}] \to [2^F], k = 1, 2,$$
which decodes the desired file W_{d_k} as $\hat{W}_{d_k}$ at User k from the cached content at User k, the messages transmitted over the shared link and the private link to User k, $k = 1, 2$.

The performance of a given caching and delivery code is measured by the worst-case delivery latency, which is defined as $T = \max_D T^D$, where $T^D \triangleq \max\{T_c^D, T_{p1}^D, T_{p2}^D\}$, and $T_c^D \triangleq \frac{r_c^D}{R_c}$, $T_{pk}^D \triangleq \frac{r_{pk}^D}{R_{pk}}$, $k = 1, 2$. In other words, T^D is the latency, under demand D, it takes for X_c^D to be received by both users while X_{pk}^D is received by User k, $k = 1, 2$.

Following the idea of symmetry in [12, Section 3] [34, Definitions 3 and 4], we will exploit the symmetry among the file indexes to simplify the proof of converse. Let $\pi(\cdot)$ be a permutation function on the file index set $\{1, 2, \cdots, N\}$, $\mathcal{Z}$ a subset of $\{Z_1, Z_2\}$, $\mathcal{W}$ a subset of $\{W_1, W_2, \cdots, W_N\}$, and $\mathcal{X}$ a subset of $\{X_c^D, X_{p1}^D, X_{p2}^D : D \in [N] \times [N]\}$. The mapping $\pi(\mathcal{W})$ is denoted by $\{W_{\pi(i)} : W_i \in \mathcal{W}\}$ and the mapping $\pi(\mathcal{X})$ is denoted by $\{X_{(\cdot)}^{(\pi(d_1),\pi(d_2))} : X_{(\cdot)}^{(d_1,d_2)} \in \mathcal{X}\}$. We define the *file-index-symmetric codes* as follows.

Definition 1: A caching and delivery code is called *file-index-symmetric* if for any permutation function $\pi(\cdot)$, any subset of caches $\mathcal{Z}$, any subset of files $\mathcal{W}$, and any subset of transmitted messages $\mathcal{X}$, the following relation holds:

$$H(\mathcal{W}, \mathcal{Z}, \mathcal{X}) = H(\pi(\mathcal{W}), \mathcal{Z}, \pi(\mathcal{X})). \tag{1}$$

Similarly to the argument on the existence of symmetric codes in [12, Proposition 1], we have the following lemma for the above problem.

Lemma 1: For any caching and delivery code, there exists a file-index-symmetric caching and delivery code with an equal or smaller worst-case delivery latency.

Proof: The proof follows similar steps to the one in [12, Proposition 1]. Intuitively, if we reorder the files and apply the same encoding function, the transmissions can also be changed accordingly to accommodate the requests, and it will lead to a new code that is equivalent to the original one. The proof can be completed by using a simple memory-sharing argument for these new codes. ■

File-index-symmetric caching and delivery codes have the following property: for any pair of distinct demands (d_1, d_2), i.e., $d_1 \neq d_2$, $(r_c^D, r_{p1}^D, r_{p2}^D)$ takes the same value, denoted by (r_c, r_{p1}, r_{p2}); similarly, for all the cases in which the two users demand the same file, i.e., $d_1 = d_2$, $(r_c^D, r_{p1}^D, r_{p2}^D)$ takes the same value, denoted by $(r_c^0, r_{p1}^0, r_{p2}^0)$. We are interested in the worst-case performance; hence, for the rest of the paper, we will assume $d_1 \neq d_2$. Hence, we have

$$T = \max\left\{\frac{r_c}{R_c}, \frac{r_{p1}}{R_{p1}}, \frac{r_{p2}}{R_{p2}}\right\}. \tag{2}$$

We will refer to the problem described above by $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$.

Definition 2: A tuple $(M_1, M_2, R_c, R_{p1}, R_{p2}, T)$ is said to be *achievable* if for large enough F, there exists a file-index-symmetric caching and delivery code with each user correctly decoding its requested file for any demand combination, i.e., $\hat{W}_{d_k} = W_{d_k}$, $k = 1, 2$ for all $(d_1, d_2) \in [N] \times [N]$. The minimum achievable worst-case delivery latency is defined as

$$\begin{aligned} &T^*(M_1, M_2, R_c, R_{p1}, R_{p2}) \\ &\quad = \inf\{T : (M_1, M_2, R_c, R_{p1}, R_{p2}, T) \text{ is achievable}\}. \end{aligned} \tag{3}$$

Remark 2.1: We adopt the zero-error decoding criterion in Definition 2, to simplify the converse proofs. We remark here that the diminishing-error decoding criterion (see [1]) is also applicable. More specifically, for the converse, the proofs and results still hold for the diminishing-error decoding criterion by using Fano's inequality (see a similar derivation in [35]); as for the achievablity, our schemes and the referred schemes are all zero-error achievability results, and thus, satisfies the diminishing-error decoding criterion.

The aim of the paper is to seek the minimum achievable worst-case delivery latency $T^*(M_1, M_2, R_c, R_{p1}, R_{p2})$ across all caching and delivery codes.

Remark 2.2: We have modeled the channel between the server and the two users as two private links and a shared common link of certain capacities. In practice, the channel between the server and the users may be a noisy wireless broadcast channel, which can be modeled as a broadcast erasure channel [22]–[25], a Gaussian broadcast channel [26]–[28], or a linear deterministic broadcast channel in [36]. The minimum achievable worst-case delivery latency T^* found in this paper would serve as an *achievable* worst-case delivery latency, where separate cache-channel coding is adopted. Joint cache-channel coding schemes that can provide a smaller latency can be studied for future work.

Note that for the problem of shared common link only, i.e., $\mathcal{Q}(M_1, M_2, R_c, 0, 0)$, the capacity R_c is of no significance as $r_c = TR_c$. Hence, minimizing T for a given R_c is equivalent to minimizing the data rate over the shared common link, i.e., r_c. As a result, we denote the problem $\mathcal{Q}(M_1, M_2, R_c, 0, 0)$ by $\mathcal{Q}^c(M_1, M_2)$, and the minimal achievable data rate over the shared common link is denoted by $r_c^*(M_1, M_2)$.

Since we are interested in the delivery latency, to simplify the notation in the rest of the paper, we drop the normalization measure F in the rest of the paper, where the value of $H(W_i)$ is normalized as 1, $\forall i$.

III. Shared Link Problem $\mathcal{Q}^c(M_1, M_2)$

We start by studying the case with heterogenous cache sizes and a shared common link only, i.e., the problem $\mathcal{Q}^c(M_1, M_2)$. For this problem, we would like to minimize the data rate over the shared common link, i.e., $r_c^*(M_1, M_2)$.

The case of $K = N = 2$ has been solved in [18], and the optimal rate is shown to be

$$\begin{aligned} r_c^*(M_1, M_2) = \max\Bigg\{&1 - \frac{M_1}{2}, 1 - \frac{M_2}{2}, \\ &2 - (M_1 + M_2), \frac{3}{2} - \frac{M_1 + M_2}{2}\Bigg\}. \end{aligned} \tag{4}$$

Note that [18] studied the case with heterogeneous cache sizes and distortion requirements. Thus, if we consider the special case of the problem studied in [18], in which the distortion requirements of the two users are the same, i.e., $D_1 = D_2$, or equivalently, $r_1 = r_2 = 1$, we obtain the problem $\mathcal{Q}^c(M_1, M_2)$, and [18, Corollary 1] provides the result in (4).

In the case of $K = 2$ and $N \geq 3$, we provide the following optimal data rate over the shared link, which was previously unknown.

Theorem 1: In the cache and delivery problem $\mathcal{Q}^c(M_1, M_2)$, when $N \geq 3$, we have

$$\begin{aligned} r_c^*(M_1, M_2) = \max\Bigg\{&1 - \frac{M_1}{N}, 1 - \frac{M_2}{N}, \\ &2 - \frac{3M_1}{N} - \frac{M_2 - M_1}{N - 1}, 2 - \frac{3M_2}{N} - \frac{M_1 - M_2}{N - 1}\Bigg\}. \end{aligned} \tag{5}$$

Remark 3.1: The special case of $M_1 = M_2 = M$ has been solved in [12], where the achievability follows from [1], while the converse proof utilizes the symmetry of optimal codes.

Remark 3.2: Compared to the uncoded placement result in [1], the optimal delivery rate depends on the number of files N, and not just the normalized cache size M_k/N. This is because the *coded* placement of Points G and F reduces the delivery rate.

Remark 3.3: Note that the optimal delivery rate takes different forms for $N = 2$ and $N \geq 3$. Intuitively this difference can be explained as follows:

- From the perspective of the converse: for $N = 2$, the worst-case demand is unique, in the sense that, there are two files, and each user requests one of these files. However, the worst-case demand for $N \geq 3$ is not unique. For example, in the case of $N = 3$ files in the server, the worst-case demand can be $(d_1, d_2) = (1, 2)$ or $(d_1, d_2) = (1, 3)$ or $(d_1, d_2) = (2, 3)$. The optimal caching scheme has to balance the need of all possible worst-case demands, and the converse proof steps need to reflect this, which is done in Lemma 2 of the following subsection. As a result, Lemma 2 holds only for $N \geq 3$.

- From the perspective of the achievability: comparing the two subfigures of Figure 1, we note that though the seven corner points $A - G$ are the same for $N = 2$ and $N \geq 3$, when performing memory-sharing, the linear combination of which three corner points leads to the lowest delivery rate is quite different. For example, for $(M_1, M_2) = (\frac{1}{2}, \frac{1}{3})$ and $N = 2$, the optimal (lowest) delivery rate is achieved by memory-sharing between points A, F and G, while for $N \geq 3$, by sharing between points A, F and B.

A. The converse proof of Theorem 1

The first two terms of (5) follow from the cut-set bound [1]. The third and fourth terms follow from the following lemma which will be useful throughout the paper.

Lemma 2: In problem $\mathcal{Q}^c(M_1, M_2)$ with $N \geq 3$, the common delivery rate r_c of any achievable scheme must satisfy

$$\begin{aligned} &NM_i + (2N-3)M_j + N(N-1)r_c \\ &\geq 2N(N-1), \quad \forall (i,j) \in \{(1,2),(2,1)\}. \end{aligned} \tag{6}$$

The details of the proof of Lemma 2 is given in Appendix A. In the following we comment on some of the proof ideas. The proof follows from the proof of Lemma 1 with the help of two major steps stated in the following two lemmas.

Lemma 3: In problem $\mathcal{Q}^c(M_1, M_2)$, for file-index-symmetric caching and delivery codes, we have:

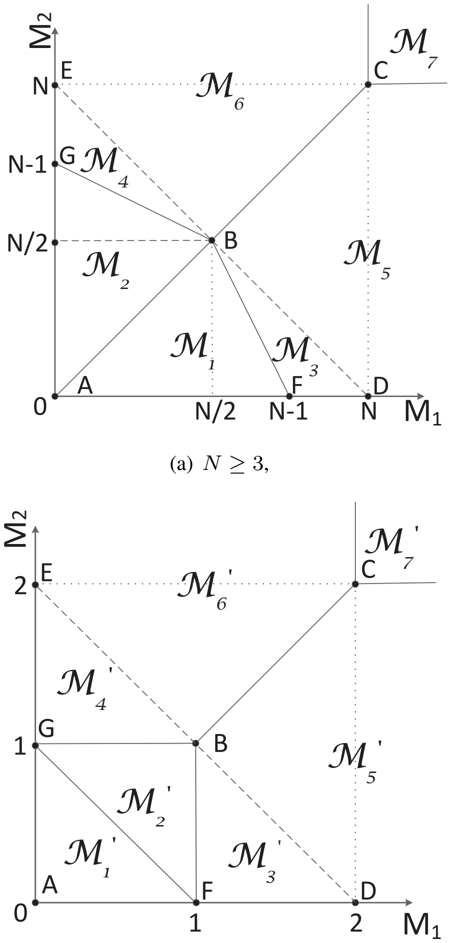
$$\begin{aligned} &H(X_c^{(1,2)}|Z_i, W_1) \\ &\geq 1 - \frac{1}{N-1}[H(Z_1|W_1) + H(Z_2|W_1)], \quad \forall i = 1, 2. \end{aligned} \tag{7}$$

Lemma 4: For file-index symmetric caching and delivery codes, we have

$$NH(Z_i|W_1) \geq (N-1)H(Z_i), \quad \forall i = 1, 2. \tag{8}$$

Please note that Lemma 4 holds for any file-index symmetric caching code, irrespective of the problem, i.e., it holds for the more general problem of $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$.

As it can be seen, Lemma 3 allow us to lower bound complicated terms, such as $H(X_c^{(1,2)}|Z_1, W_1)$, with simpler ones, such as $H(Z_1|W_1)$, while Lemma 4 further lower bounds terms, such as $H(Z_1|W_1)$, with even simpler ones, such as $H(Z_1)$, which is equal to the size of the cache of User 1, i.e., M_1. Hence, the main aim of the two lemmas is to provide a lower bound that depends only on the placement scheme, and is independent of the delivery scheme. The same idea appeared in [35, Lemma 1]. The proofs of Lemmas 3 and 4 are provided in Appendices B and C, respectively.



(a) $N \geq 3$,

(b) $N = 2$,

Fig. 1. The optimal tradeoff between $r_c^*(M_1, M_2)$ and (M_1, M_2).

The converse of Theorem 1 is completed with Lemma 2.

B. The achievability proof for Theorem 1

In Figure 1(a), we show the 2-dimensional plane of possible (M_1, M_2) pairs. For the following points on this figure, the minimum data rate on the shared common link, r_c^*, is known:

1) Point A: $(M_1, M_2, r_c^*) = (0, 0, 2)$. This is the case with no caches at the users.
2) Point B: $(M_1, M_2, r_c^*) = (\frac{N}{2}, \frac{N}{2}, \frac{1}{2})$. This is the symmetric cache capacity scenario with the achievability proposed in [1], and its converse proved in [12]. The corresponding caching-delivery scheme is the following: each file is split into two parts of equal size (W_i^1, W_i^2), $i = 1, 2, \cdots, N$. In the placement phase, User k caches $\{W_i^k : i = 1, 2, \cdots, N\}$, $k = 1, 2$. The delivery scheme upon receiving request (d_1, d_2) is to transmit $\{W_{d_1}^2 \oplus W_{d_2}^1\}$.
3) Point C: $(M_1, M_2, r_c^*) = (N, N, 0)$. This is the case in which the cache at each user is large enough to cache the entire library, and as such, nothing needs to be transmitted via the shared common link.

4) Point D: $(M_1, M_2, r_c^*) = (N, 0, 1)$. This is the case in which User 1 has a cache that is large enough to store the entire library, and User 2 has no cache. Thus, it is optimal to transmit only the requested file of User 2 via the shared common link.

We now add the achievability scheme for Point F, i.e., $(M_1, M_2, r_c^*) = (N-1, 0, 1)$. Note that the achievability for the points symmetric with respect to the AC line, i.e., points E and G, follow directly.

- **Placement phase**: User 1 fills its cache with the module sum of every two label-adjacent files, i.e. $Z_1 = \{W_1 \oplus W_2, W_2 \oplus W_3, \cdots, W_{N-1} \oplus W_N\}$.
- **Delivery phase**: The server transmits $X_c^{(d_1,d_2)} = \{W_{d_2}\}$. Therefore, User 2 can directly get W_{d_2}, while user 1 can decode W_{d_1} with the help of its own cache by successive cancellation. For example if $(d_1, d_2) = (1, 4)$, User 1 can firstly recover W_3 from $(W_3 \oplus W_4, X_c^{(1,4)} = W_4)$, it then goes on to obtain W_2 from $(W_3, W_2 \oplus W_3)$, and finally it decodes the requested file W_1 from $(W_2, W_1 \oplus W_2)$.

By performing memory-sharing [1], [18], [37] among the seven points, i.e., Point A to Point G, we can obtain the following achievable data rate on the shared common link:

$$r_c(M_1, M_2) = \begin{cases} 2 - \frac{3M_2}{N} - \frac{M_1 - M_2}{N-1} & (M_1, M_2) \in \mathcal{M}_1 \\ 2 - \frac{3M_1}{N} - \frac{M_2 - M_1}{N-1} & (M_1, M_2) \in \mathcal{M}_2 \\ 1 - \frac{M_2}{N} & (M_1, M_2) \in \mathcal{M}_3, \mathcal{M}_5 \\ 1 - \frac{M_1}{N} & (M_1, M_2) \in \mathcal{M}_4, \mathcal{M}_6 \end{cases}. \tag{9}$$

For completeness, we present the placement and delivery schemes for (M_1, M_2) pairs in the regions ABG and BEG to illuminate (9). When the cache size falls into the region of ABG, the cache and delivery scheme is as follows: each file is divided into four subfiles $W_{i,A}$, $W_{i,B1}$, $W_{i,B2}$ and $W_{i,G}$ with sizes $1 - 2M_1/N - (M_2 - M_1)/(N-1)$, M_1/N, M_1/N and $(M_2 - M_1)/(N-1)$, respectively.

- **Placement phase**: User 1 caches $\{W_{i,B1} : i \in [N]\}$; while User 2 caches $\{W_{i,B2} : i \in [N]\}$ and $\{W_{j,G} \oplus W_{j+1,G} : j \in [N-1]\}$.
- **Delivery phase**: The server transmits $X_c^{d_1,d_2} = \{W_{d_1,A}, W_{d_2,A}, W_{d1,B2} \oplus W_{d2,B1}, W_{d1,G}\}$.

It is easy to check that each user can recover its file of interest, and the scheme achieves a delivery rate of $2 - 3M_1/N - (M_2 - M_1)/(N-1)$. The placement and delivery schemes for (M_1, M_2) pairs in the regions BEG follow similarly.

The explicit caching schemes above show that the optimal schemes fully take advantage of *coded* cache placement schemes, i.e., Point G. More specifically, for $M_2 \geq M_1$, only when M_2 is sufficiently large, the optimal scheme will apply the weak *uncoded* cache placement scheme, i.e., Point E where weak means that, compared to Point G, Point E achieves the same delivery rate with a larger cache size.

The achievability part of Theorem 1 is complete. Note that without loss of generality, we may consider only the case $M_1 \leq M_2$, and the $M_1 \geq M_2$ case follows by symmetry. But, since we need to reuse the points A-G in the achievability proof of Theorem 2 in the next section, we presented the achievability proof of Theorem 1 for all (M_1, M_2) pairs.

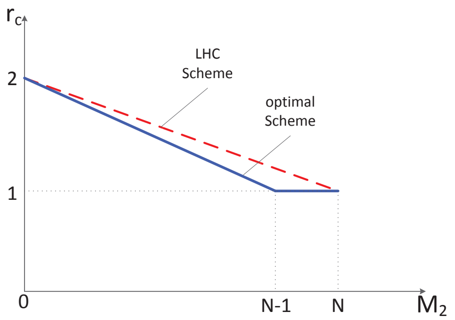


Fig. 2. The comparison between our scheme and the LHC scheme for the problem $\mathcal{Q}^c(0, M_2)$.

C. Comparison and analysis

As we mentioned before, the problem $\mathcal{Q}^c(M_1, M_2)$ with $N = 2$ has been solved in [18]. But for $N \geq 3$, the best known achievability schemes [18, Section III-C], [37], which will be denoted as the LHC scheme here, perform memory sharing between the five points of Fig. 1(a), i.e., Point A to Point E, and thus obtain an achievable data rate on the shared common link as

$$\bar{r}_c(M_1, M_2) = \begin{cases} 2 - \frac{2M_2}{N} - \frac{M_1}{N} & (M_1, M_2) \in \mathcal{M}_1, \mathcal{M}_3 \\ 2 - \frac{2M_1}{N} - \frac{M_2}{N} & (M_1, M_2) \in \mathcal{M}_2, \mathcal{M}_4 \\ 1 - \frac{M_2}{N} & (M_1, M_2) \in \mathcal{M}_5 \\ 1 - \frac{M_1}{N} & (M_1, M_2) \in \mathcal{M}_6 \end{cases}.$$

We see that the optimal delivery rate is lower than the rate achieved by the LHC scheme, in which the delivery phase is divided into layers of unicast and multicast. We improve the delivery rate from $(M_1, M_2, r_c) = (0, N, 1)$ to $(0, N-1, 1)$ with the help of *coded* placement. In particular, for the problem $\mathcal{Q}^c(0, M_2)$, i.e., $M_1 = 0$, the improvement of our scheme is plotted in Fig. 2. As for the converse, when $N \geq 3$, the best known converse to date is given by [18, Lemma 1], which is the minimum of the five terms

$$r_c(M_1, M_2) \geq \max\left\{1 - \frac{M_1}{N}, 1 - \frac{M_2}{N}, 2 - \frac{M_1 + M_2}{\lfloor N/2 \rfloor}, \right. \\ \left. \frac{3}{2} - \frac{M_1 + M_2}{2\lfloor N/2 \rfloor}, 2 - \frac{M_1 + M_2}{2\lfloor N/3 \rfloor}\right\}, \tag{10}$$

where the first two terms follow from the cut-set bound, the third and fourth terms follow from the straightforward generalization of the proof of the same problem for the case $N = 2$. In this proof, the step [18, Eqn. (40c)] may be loose because the content of two caches may not be independent even conditioned on the knowledge of some files. We transform terms like $H(X_{i,j}, Z_k|W_i)$ into $H(X_{i,j}|Z_k, W_i)$ and $H(Z_k|W_i)$, and then bound these two terms via Lemmas 3 and 4 to obtain a tighter converse. It has been argued in [18] that (10) is tight when N is an integer multiple of 3 and $M_1 = M_2$. Indeed, comparing (10) and (5), we see that when $N = 3$, the two bounds are the same, which means that the bound in (10) is tight for $N = 3$ and *arbitrary* (M_1, M_2). When $N = 4, 5$ and 6, we plot the two bounds

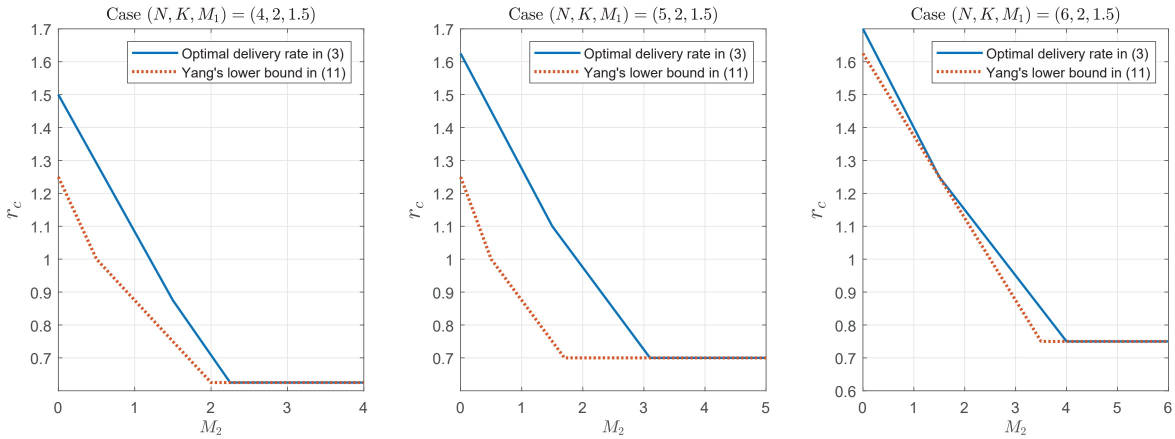


Fig. 3. The comparison between our lower bound and the one in [18] (see (10)) for the problem $\mathcal{Q}^c(1.5, M_2)$, for $N = 4, 5, 6$.

in Fig. 3 to illustrate that (5) improves upon the best known converse bound (10). Moreover, Theorem 1 proves that (5) is the minimum achievable data rate over the shared common link.

IV. General problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$

In this section, we study the general problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$, i.e., the problem with one shared common link and two private links, one for each user. We characterize the optimal delivery latency $T^*(M_1, M_2, R_c, R_{p1}, R_{p2})$ in the following theorem.

Theorem 2: For problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$ with $N = 2$, we have:

$$T^* = \max\left\{\frac{1-\frac{M_1}{2}}{R_c+R_{p1}}, \frac{1-\frac{M_2}{2}}{R_c+R_{p2}}, \frac{2-M_1-M_2}{R_c+R_{p1}+R_{p2}},\right.$$
$$\left.\frac{3-M_1-M_2}{2(R_c+R_{p2})+R_{p1}}, \frac{3-M_1-M_2}{2(R_c+R_{p1})+R_{p2}}\right\}, \quad (11)$$

while if $N \geq 3$, we have:

$$T^* = \max\left\{\frac{1-\frac{M_1}{N}}{R_c+R_{p1}}, \frac{1-\frac{M_2}{N}}{R_c+R_{p2}}, \frac{2-\frac{3M_2}{N}-\frac{M_1-M_2}{N-1}}{R_c+R_{p1}+R_{p2}},\right.$$
$$\frac{2-\frac{3M_1}{N}-\frac{M_2-M_1}{N-1}}{R_c+R_{p1}+R_{p2}}, \frac{N(2N-1)-2(N-1)M_1-NM_2}{N^2(R_c+R_{p2})+N(N-1)R_{p1}},$$
$$\left.\frac{N(2N-1)-2(N-1)M_2-NM_1}{N^2(R_c+R_{p1})+N(N-1)R_{p2}}\right\}. \quad (12)$$

Theorem 2 also takes on different forms for $N = 2$ and $N \geq 3$, which can be argued similarly to 3.3.

We note here that, while the proof ideas for both the converse and the achievability of Theorem 2 can be extended to the multiple users case, the results become highly complex with more than two users. Furthermore, they are not tight, which is unsurprising as the optimal performance of $N > 2$ users is open even for the original coded caching problem in [1]. Hence, due to the complexity and looseness of the achievability and converse results, we do not present the general results for multiple users.

A. Converse proof of Theorem 2

We define $\mathcal{S}$ as the set of all possible caching and delivering codes. Then, we have

$$\begin{aligned} T &= \min_{\mathcal{S}} \max\left\{\frac{r_c}{R_c}, \frac{r_{p1}}{R_{p1}}, \frac{r_{p2}}{R_{p2}}\right\} \\ &\geq \min_{\mathcal{S}} \frac{r_c + r_{p1}}{R_c + R_{p1}} \quad (13) \\ &\geq \frac{1 - M_1/N}{R_c + R_{p1}}, \quad (14) \end{aligned}$$

where (13) follows from the fact that for positive numbers a, b, c, d, α, we have $\max\left\{\frac{a}{b}, \frac{c}{d}\right\} \geq \frac{a+\alpha c}{b+\alpha d}$, and (14) is from the cut-set bound for User 1. Similarly, we also have

$$T \geq \frac{1 - M_2/N}{R_c + R_{p2}}. \quad (15)$$

Note that any achievable scheme for problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$ can be transformed to be achievable for problem $\mathcal{Q}^c(M_1, M_2)$, because we may transmit all three signals $X_c^{(d_1,d_2)}$ with rate r_c, $X_{p1}^{(d_1,d_2)}$ with rate r_{p1}, and $X_{p2}^{(d_1,d_2)}$ with rate r_{p2}, of $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$ over the shared common link of the problem $\mathcal{Q}^c(M_1, M_2)$, resulting in a common rate of $r_c + r_{p1} + r_{p2}$. Hence, $r_c + r_{p1} + r_{p2}$ must satisfy Lemma 2, i.e., when $N \geq 3$,

$$\begin{aligned} &NM_i + (2N-3)M_j + N(N-1)[r_c + r_{p1} + r_{p2}] \\ &\geq 2N(N-1), \quad \forall (i,j) \in \{(1,2),(2,1)\}. \quad (16) \end{aligned}$$

Therefore, we have

$$\begin{aligned} T &= \min_{\mathcal{S}} \max_{i=c,p1,p2} \{T_i\} \\ &\geq \min_{\mathcal{S}} \frac{r_c + r_{p1} + r_{p2}}{R_c + R_{p1} + R_{p2}} \quad (17) \\ &\geq \frac{\max\left\{2 - \frac{3M_2}{N} - \frac{M_1-M_2}{N-1}, 2 - \frac{3M_1}{N} - \frac{M_2-M_1}{N-1}\right\}}{R_c + R_{p1} + R_{p2}}, \quad (18) \end{aligned}$$

where (17) follows by applying twice the reasoning used for (13), and (18) follows from (16).

Note that any achievable scheme for problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$ can be transformed to be achievable for $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, 0)$, because we can transmit both signal $X_c^{(d_1,d_2)}$ with rate r_c and $X_{p2}^{(d_1,d_2)}$ with rate r_{p2} for problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$ over the shared common link in problem $\mathcal{Q}^c(M_1, M_2, R_c, R_{p1}, 0)$, resulting in a rate of $r_c + r_{p2}$, while the private rate r_{p1} to User 1 remaining the same. We can prove the following lemma for the problem of $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, 0)$, i.e., the problem with one shared common link and one private link to User 1.

Lemma 5: In problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, 0)$ with $N \geq 2$, the data rate on the shared common link r_c and the only private link r_{p1}, must satisfy:

$$\begin{aligned} &N^2 r_c + N(N-1) r_{p1} \\ &\geq N(2N-1) - 2(N-1)M_1 - NM_2. \end{aligned} \tag{19}$$

The details of the proof of Lemma 5, which follows similarly to Lemma 2, are relegated to Appendix D. In the proof, the following lemma, whose proof is provided in Appendix E, replaces the role of Lemma 3.

Lemma 6: In problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, 0)$, for file-index-symmetric caching and delivery codes, we have

$$\begin{aligned} &H(X_c^{(1,2)}, X_{p1}^{(1,2)} | Z_1, W_1) \\ &\geq 1 - \frac{1}{N-1}[H(Z_1|W_1) + H(Z_2|W_1)], \end{aligned} \tag{20}$$

$$\begin{aligned} &H(X_c^{(2,1)} | Z_2, W_1) + r_c + r_{p1} + M_1 \\ &\geq 2 + \frac{N-2}{N-1} H(Z_1|W_1) - \frac{1}{N-1} H(Z_2|W_1). \end{aligned} \tag{21}$$

Again, Lemma 6 provides a way to lower bound terms, such as $H(X_c^{(i,j)}, X_p^{(i,j)} | Z_1, W_1)$, with simpler ones, such as $H(Z_1|W_1)$, and then, we again use Lemma 4 to lower bound terms, such as $H(Z_1|W_1)$, with simpler ones, such as $H(Z_1)$, to obtain Lemma 5. Thus, for problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$ with $N \geq 2$, we have

$$\begin{aligned} &N^2 [r_c + r_{p2}] + N(N-1) r_{p1} \\ &\geq N(2N-1) - 2(N-1)M_1 - NM_2. \end{aligned} \tag{22}$$

We can obtain

$$\begin{aligned} T &= \min_{\mathcal{S}} \max_{i=c,p1,p2} \{T_i\} \\ &\geq \min_{\mathcal{S}} \frac{N^2(r_c + r_{p2}) + N(N-1) r_{p1}}{N^2(R_c + R_{p2}) + N(N-1) R_{p1}} \qquad (23) \\ &\geq \frac{N(2N-1) - 2(N-1)M_1 - NM_2}{N^2(R_c + R_{p2}) + N(N-1) R_{p1}}, \qquad (24) \end{aligned}$$

where (23) follow similarly to (13); and (24) from (22). By exploring the symmetry between Users 1 and 2, similarly to (24), we also have

$$T \geq \frac{N(2N-1) - 2(N-1)M_2 - NM_1}{N^2(R_c + R_{p1}) + N(N-1) R_{p2}}. \tag{25}$$

Hence, from (14), (15), (18), (24), (25), the proof of (12) is completed. Note that the above upper bounds (14), (15), (24), (25) hold for any $N \geq 2$.

Finally, for the case $N = 2$, we only need to prove the third term, i.e.,

$$T^* \leq \frac{2 - M_1 - M_2}{R_c + R_{p1} + R_{p2}},$$

which follows from the cut-set bound

$$M_1 + M_2 + r_c + r_{p1} + r_{p2} \geq H(W_1, W_2) = 2,$$

and (17). Hence, the proof of (11) is also complete.

B. Achievability proof of Theorem 2 for $N \geq 3$

The proof of achievability consists of three parts. In the first part, we find achievable schemes for a set of special points. More specifically, the achievable scheme we propose for each special point is a generalization of the achievable scheme proposed for the special point (M_1, M_2) of problem $\mathcal{Q}^c(M_1, M_2)$, studied in Section III-B. In the second part, we perform memory-sharing and time-sharing among the special points obtained in the first part to construct a set of achievable schemes for the current problem. In the third part, we show that there exists an achievable point $(M_1, M_2, r_c, r_{p1}, r_{p2})$ within the set of achievable points, whose peak delivery latency meets the converse bound.

Without loss of generality, we assume $R_{p1} \geq R_{p2}$. Based on the achievable scheme for problem $\mathcal{Q}^c(M_1, M_2)$, we consider the rate of the message transmitted over the shared common link, r_c, for a given $(M_1, M_2, r_{p1}, r_{p2})$ tuple.

The seven points considered in Section III-B for the achievability of problem $\mathcal{Q}^c(M_1, M_2)$, i.e., points A to G, correspond to the following seven points in the format $(M_1, M_2, r_{p1}, r_{p2}, r_c)$: $P_A = (0, 0, 0, 0, 2)$, $P_B = (\frac{N}{2}, \frac{N}{2}, 0, 0, \frac{1}{2})$, $P_C = (N, N, 0, 0, 0)$, $P_D = (N, 0, 0, 0, 1)$, $P_E = (0, N, 0, 0, 1)$, $P_F = (N-1, 0, 0, 0, 1)$ and $P_G = (0, N-1, 0, 0, 1)$. We add five new points:

1) Point $P_H = (0, 0, 1, 1, 0)$. This is the case with no caches at the users. The server transmits W_{d1} to User 1 and W_{d_2} to User 2 via the corresponding private links, respectively.
2) Point $P_I = (0, 0, 1, 0, 1)$. In this case the server transmits W_{d1} to User 1 via its private link and W_{d_2} to User 2 via the shared common link.
3) Point $P_J = (0, 0, 0, 1, 1)$. This case is symmetric to Point P_I.
4) Point $P_K = (0, N, 1, 0, 0)$. This is the case in which User 2 can cache the entire library, while User 1 has no cache. The server transmits W_{d_1} to User 1 via its private link.
5) Point $P_L = (N, 0, 0, 1, 0)$. This case is symmetric to Point P_K.

These twelve points are achievable for problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$.

By using memory-sharing for the cache capacity values and time-sharing for the transmitted rates (r_{p1}, r_{p2}), the convex hull of these twelve points and the corresponding r_c value, i.e., $(M_1, M_2, r_{p1}, r_{p2})$ as the independent variables and r_c as the dependent variable, are also achievable.

Therefore, we obtain a set of achievable tuples for problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, R_{p2})$, denoted by Δ.

For a $(M_1, M_2, r_{p1}, r_{p2})$ tuple, let $f(M_1, M_2, r_{p1}, r_{p2})$ be the smallest rate $\bar{r}_c$ in Δ, i.e.,

$$\begin{aligned}\bar{r}_c &= f(M_1, M_2, r_{p1}, r_{p2}) \\ &= \min\{r_c : (M_1, M_2, r_{p1}, r_{p2}, r_c) \in \Delta\}.\end{aligned}$$

To obtain $f(M_1, M_2, r_{p1}, r_{p2})$ in closed form, we consider its projection for fixed values of (r_{p1}, r_{p2}), and derive $f_{(r_{p1},r_{p2})}(M_1, M_2)$ in closed form. Before we delve into the details, we provide some insights on the achievable scheme corresponding to $f_{(r_{p1},r_{p2})}(M_1, M_2)$. Suppose that rates $0 \leq r_{pk} \leq 1$, $k = 1, 2$, will be transmitted over the private link.

How to use the private links: The private links will be used to transmit part of the desired messages in an uncoded fashion. Then the delivery strategy is designed for file sizes reduced by the rates transmitted over the private links. For example, for $r_{p1} \geq r_{p2}$, we split each file into three parts W_i^c, W_i^{p1} and $W_i^{p12}, i = 1, \dots, N$, with sizes $l_1, l_2 - l_1, 1 - l_2$, respectively, where $l_1 \triangleq 1 - r_{p1}$ and $l_2 \triangleq 1 - r_{p2}$. In the delivery phase, the server transmits $\{W_{d_1}^{p1}, W_{d_1}^{p12}\}$ and $W_{d_2}^{p12}$ to Users 1 and 2, respectively, via their private links. Thus, we only need to deliver $(W_{d_1}^c, W_{d_2}^c)$ among sub-files $\{W_1^c, W_2^c, \cdots, W_N^c\}$ to Users 1 and 2, and $W_{d_2}^{p1}$ among sub-files $\{W_1^{p1}, W_2^{p1}, \cdots, W_N^{p1}\}$ to User 2 over the shared links.

How to deal with the sub-files from $\{W_1^{p1}, W_2^{p1}, \cdots, W_N^{p1}\}$ requested by one user only: Memory-sharing is performed among certain special achievable points. In each point, the achievable scheme is to either transmit $W_{d_2}^{p1}$ uncoded through the shared common link, or cache all files $\{W_1^{p1}, W_2^{p1}, \cdots, W_N^{p1}\}$ (of file size $l_2 - l_1$) in the cache of User 2. The caching and delivery strategy over the common shared link for files $\{W_1^c, W_2^c, \cdots, W_N^c\}$ (of file size l_1) is the same as those proposed for problem $\mathcal{Q}^c(M_1, M_2)$.

We obtain the following lemma for the closed-form expression of $f_{(r_{p1},r_{p2})}(M_1, M_2)$.

Lemma 7: For a given (r_{p1}, r_{p2}) pair with $r_{p1} \geq r_{p2}$, by memory-sharing among the nine points illustrated in Fig. 4(a), the smallest achievable rate over the shared common link, $\bar{r}_c = f_{(r_{p1},r_{p2})}(M_1, M_2)$, is given as

$$\bar{r}_c = \begin{cases} 2 - r_{p1} - r_{p2} - \frac{3M_2}{N} - \frac{M_1 - M_2}{N-1} \\ \qquad\qquad (M_1, M_2) \in \mathcal{M}_1(r_{p1}, r_{p2}) \\ 2 - r_{p1} - r_{p2} - \frac{3M_1}{N} - \frac{M_2 - M_1}{N-1} \\ \qquad\qquad (M_1, M_2) \in \mathcal{M}_2(r_{p1}, r_{p2}) \\ \frac{2N-1}{N} - \frac{N-1}{N} r_{p1} - r_{p2} - \frac{2(N-1)M_1}{N^2} - \frac{M_2}{N} \\ \qquad\qquad (M_1, M_2) \in \mathcal{M}_3(r_{p1}, r_{p2}) \\ 1 - r_{p2} - \frac{M_2}{N} \qquad (M_1, M_2) \in \mathcal{M}_4(r_{p1}, r_{p2}) \\ 1 - r_{p1} - \frac{M_1}{N} \qquad (M_1, M_2) \in \mathcal{M}_5(r_{p1}, r_{p2}) \end{cases}, \tag{26}$$

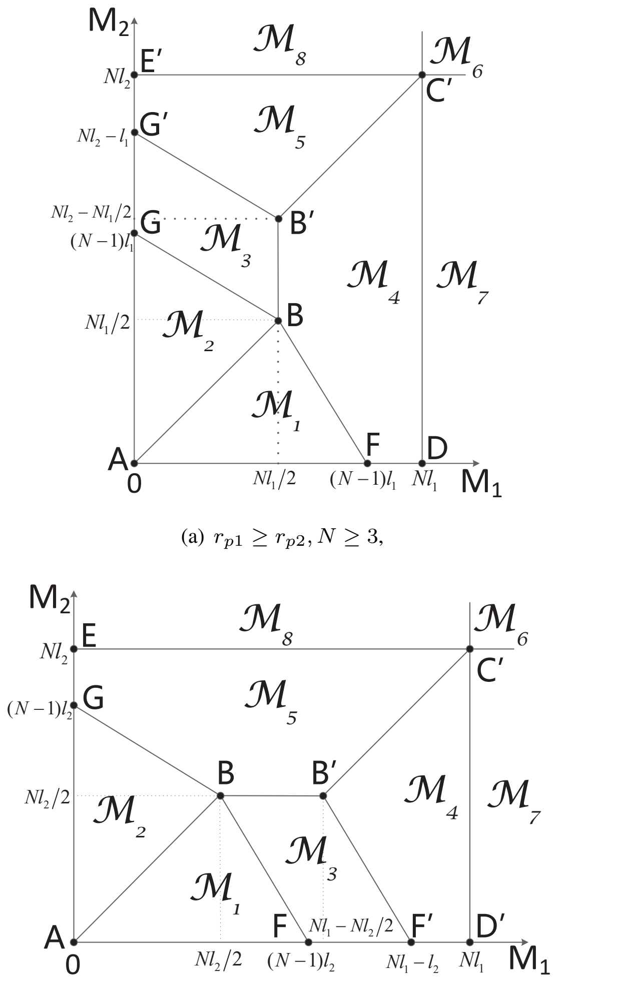
where the regions $\mathcal{M}_1(r_{p1}, r_{p2})$ to $\mathcal{M}_5(r_{p1}, r_{p2})$ are shown in Fig 4(a).

By symmetry, for a given (r_{p1}, r_{p2}), where $r_{p1} \leq r_{p2}$, the smallest achievable rate on the shared common link, $\bar{r}_c$, is given by

$$\bar{r}_c = \begin{cases} 2 - r_{p1} - r_{p2} - \frac{3M_2}{N} - \frac{M_1 - M_2}{N-1} \\ \qquad\qquad (M_1, M_2) \in \mathcal{M}_1(r_{p1}, r_{p2}) \\ 2 - r_{p1} - r_{p2} - \frac{3M_1}{N} - \frac{M_2 - M_1}{N-1} \\ \qquad\qquad (M_1, M_2) \in \mathcal{M}_2(r_{p1}, r_{p2}) \\ \frac{2N-1}{N} - \frac{N-1}{N} r_{p2} - r_{p1} - \frac{2(N-1)M_2}{N^2} - \frac{M_1}{N} \\ \qquad\qquad (M_1, M_2) \in \mathcal{M}_3(r_{p1}, r_{p2}) \\ 1 - r_{p2} - \frac{M_2}{N} \qquad (M_1, M_2) \in \mathcal{M}_4(r_{p1}, r_{p2}) \\ 1 - r_{p1} - \frac{M_1}{N} \qquad (M_1, M_2) \in \mathcal{M}_5(r_{p1}, r_{p2}) \end{cases}, \tag{27}$$

where the regions $\mathcal{M}_1(r_{p1}, r_{p2})$ to $\mathcal{M}_5(r_{p1}, r_{p2})$ are shown in Fig. 4(b).

The proof of Lemma 7 is provided in Appendix F. Note that (26) and (27) achieve the lower bound of (16), (22) and the cut-set bound. For an arbitrary (M_1, M_2) pair, $0 \leq M_1 \leq N, 0 \leq M_2 \leq N$, the set Δ, i.e., the three-dimensional achievable region of (r_{p1}, r_{p2}, r_c), is characterized by (26) and (27). The remaining task is to find the $(M_1, M_2, r_{p1}, r_{p2}, r_c)$



(a) $r_{p1} \geq r_{p2}, N \geq 3$,

(b) $r_{p1} \leq r_{p2}, N \geq 3$,

Fig. 4. The illustration of possible (M_1, M_2) pairs for arbitrary r_{p1}, r_{p2} when $N \geq 3$.

tuple within the achievable region Δ that minimizes $T = \max\{\frac{r_{p1}}{R_{p1}}, \frac{r_{p2}}{R_{p2}}, \frac{r_c}{R_c}\}$.

Lemma 8: For any $(M_1, M_2, R_c, R_{p1}, R_{p2})$, there exists an achievable scheme $(M_1, M_2, r_{p1}, r_{p2}, r_c)$ in Δ with a delivery latency equal to one of the six terms in (12).

The proof of Lemma 8 is provided in Appendix G.

This completes the achievability part of Theorem 2 for $N \geq 3$ and $R_{p1} \geq R_{p2}$. Before we proceed to the achievability for $N = 2$, we make the following connection between the achievability scheme proposed here and the one in [18].

Remark: In [18] the authors study the caching problem in which the users request different quality descriptions of the files, due to, for example, different processing or display capabilities. For given distortion targets (D_1, D_2), assuming $D_1 \geq D_2$ without loss of generality, the authors suggest using scalable coding [38] of the files in the library at rates (r_1, r_2), such that the *base layer* of rate r_1 allows the first receiver to obtain an average reconstruction distortion of D_1, while the base layer together with the *refinement layer* of rate r_2 allows an average reconstruction distortion of D_2 at the second receiver. This successive coding scheme is known to be rate-distortion optimal for Gaussian sources under squared error distortion.

Once we specify how the private links are used, the (l_1, l_2) parameters in our problem correspond to (r_1, r_2) in the achievable scheme of [18], where r_1 corresponds to the number of bits transmitted over the common link, while $r_2 - r_1$ to the number of bits transmitted over the private link to the user that request a higher quality description. As such, we may make a comparison of the achievable scheme proposed here and the one in [18] for $K = 2$ users with $N \geq 3$ files. The scheme in [18] is a suboptimal memory-sharing scheme between points A, B, B', C', D, E', ignoring the three points G, G' and F. We can show that memory-sharing among all the nine points is optimal for the coded caching with heterogeneous distortion requirements problem for $K = 2, N \geq 3$, and a converse is provided in Appendix H.

Theorem 3: For the coded caching problem with heterogeneous distortion requirements, defining $l_k = \frac{1}{2}\log\frac{\sigma^2}{D_k}$, $k = 1, 2$, the optimal cache capacity-delivery trade-off is given by

$$R^*(M_1, M_2) = \max\Bigg\{l_1 + l_2 - \frac{3M_2}{N} - \frac{M_1 - M_2}{N-1}, \\ l_1 + l_2 - \frac{3M_1}{N} - \frac{M_2 - M_1}{N-1}, l_2 - \frac{M_2}{N}, l_1 - \frac{M_1}{N}, \\ \frac{N-1}{N}l_1 + l_2 - \frac{2(N-1)M_1}{N^2} - \frac{M_2}{N}, \\ \frac{N-1}{N}l_2 + l_1 - \frac{2(N-1)M_2}{N^2} - \frac{M_1}{N}\Bigg\}.$$

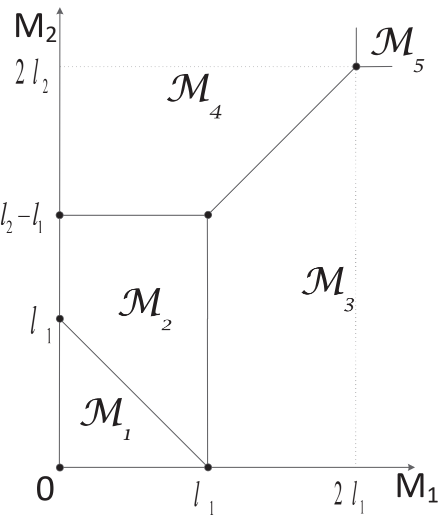


Fig. 5. The illustration of possible r_{p1}, r_{p2} values that satisfy $r_{p1} \geq r_{p2}$ for $N = 2$.

C. The achievability of Theorem 2 for $N = 2$

Based on the above discussion of the similarity between the studied problem and that of [18], we can use the optimal achievability found in [18, Section III.B] and obtain the smallest achievable rate on the shared common link, r_c, as follows:

$$\bar{r}_c = \begin{cases} l_1 + l_2 - M_1 - M_2 = 2 - r_{p1} - r_{p2} - M_1 - M_2 \\ \qquad\qquad\qquad (M_1, M_2) \in \mathcal{M}_1(r_{p1}, r_{p2}) \\ \frac{l_1}{2} + l_2 - \frac{M_1}{2} - \frac{M_2}{2} = \frac{3 - r_{p1} - 2r_{p2} - M_1 - M_2}{2} \\ \qquad\qquad\qquad (M_1, M_2) \in \mathcal{M}_2(r_{p1}, r_{p2}) \\ l_2 - \frac{M_2}{2} = 1 - r_{p2} - \frac{M_2}{2} \quad (M_1, M_2) \in \mathcal{M}_3(r_{p1}, r_{p2}) \\ l_1 - \frac{M_1}{2} = 1 - r_{p1} - \frac{M_1}{2} \quad (M_1, M_2) \in \mathcal{M}_4(r_{p1}, r_{p2}) \end{cases},$$

where $\mathcal{M}_1(r_{p1}, r_{p2})$ to $\mathcal{M}_4(r_{p1}, r_{p2})$ are shown in Fig 5.

Similarly to the discussion on the $N \geq 3$ case, we find the achievable $T = \max\left\{\frac{r_{p1}}{R_{p1}}, \frac{r_{p2}}{R_{p2}}, \frac{r_c}{R_c}\right\}$ to coincide with (11). Thus, the achievability proof of Theorem 2 is complete.

V. Conclusions

We have studied the problem of centralized coded caching for two users with different cache capacities, where, in addition to the shared common link, each user also has a private link from the server. We have characterized the optimal caching and delivery strategies for any number of files in the library. In the case of a shared common link only, we have improved upon the known results in the literature by proposing a new achievable scheme for a special (M_1, M_2) pair, and performing memory-sharing among a total of nine special memory pairs. In the case of two private links in addition to the shared common link, we have shown that it is optimal to use all the capacity available over the private links to transmit the file requested by the corresponding user in an uncoded fashion. A connection between the problem of coded caching with a private link to each user considered here and that of coded caching with heterogeneous distortion requirements studied in [18] has also been established, which allowed us extending the proposed results to improve the state of the art in the latter problem as well.

APPENDIX

A. Proof of Lemma 2

We will provide the proof for the case $(i,j)=(1,2)$ of (6), and the other case where $(i,j)=(2,1)$ follows by symmetry. For any caching-delivery scheme, we have

$$\begin{aligned}
& r_c + M_1 \\
&\geq H(X_c^{1,2}) + H(Z_1) && (28)\\
&\geq H(Z_1, X_c^{(1,2)}) \\
&= H(Z_1, X_c^{(1,2)}, W_1) && (29)\\
&= H(W_1) + H(Z_1|W_1) + H(X_c^{(1,2)}|Z_1, W_1) && (30)\\
&\geq 1 + H(Z_1|W_1) \\
&\qquad + (1 - \frac{1}{N-1}[H(Z_1|W_1) + H(Z_2|W_1)]) && (31)\\
&\geq 2 + \frac{N-2}{N-1} H(Z_1|W_1) - \frac{1}{N-1} H(Z_2|W_1), && (32)
\end{aligned}$$

where (28) follows from the problem definition in Section II, (29) follows from the fact that User 1 can decode W_1 from $(Z_1, X_c^{(1,2)})$, (31) is from Lemma 3.

Similarly, by exchanging the indices of 1 and 2, we have

$$r_c + M_2 \geq 2 + \frac{N-2}{N-1} H(Z_2|W_1) - \frac{1}{N-1} H(Z_1|W_1). \quad (33)$$

By cancelling the term $H(Z_1|W_1)$ in (32) and (33), we obtain

$$\begin{aligned}
& M_1 + r_c + (N-2)[r_c + M_2] \\
&\geq 2(N-1) + (N-3) H(Z_2|W_1) \\
&\geq 2(N-1) + \frac{(N-3)(N-1)}{N} H(Z_2), && (34)
\end{aligned}$$

where (34) follows from Lemma 4. Hence, following from (34), we have

$$N M_1 + (2N-3) M_2 + N(N-1) r_c \geq 2N(N-1),$$

which completes the proof of Lemma 2.

B. Proof of Lemma 3

The proof of Lemma 3 is given here for completeness, but it follows the proof of [12, Lemma 1] very closely. By setting $n=1$ in [12, Lemma 1] and not using symmetry, i.e., [12, Eqn. (13)], to replace Z_2 with Z_1, we would obtain Lemma 3. For completeness, the proof of Lemma 3 is as follows:

In the problem $\mathcal{Q}(M_1, M_2)$, we have

$$\begin{aligned}
& (N-1) H(X_c^{(1,2)}|Z_1, W_1) \\
&= \sum_{i=2}^{N} H(X_c^{(1,i)}|Z_1, W_1) && (35)\\
&\geq H(X_c^{(1,[2:N])}|Z_1, W_1) \\
&\geq H(X_c^{(1,[2:N])}, Z_2|W_1) - H(Z_1|W_1) - H(Z_2|Z_1, W_1) \\
&= H(X_c^{(1,[2:N])}, Z_2, W_{[2:N]}|W_1) \\
&\qquad - H(Z_2|W_1) - H(Z_1|Z_2, W_1) && (36)\\
&\geq (N-1) - [H(Z_2|W_1) + H(Z_1|W_1)], && (37)
\end{aligned}$$

where (35) is from Lemma 1, (36) follows because given $(X_c^{(1,[2:N])}, Z_2)$, User 2 can recover $W_{[2:N]}$, and (37) is from $H(X_c^{(1,[2:N])}, Z_2|W_{[1:N]}) = 0$. Thus, we have proved (7), and the rest case follows by symmetry.

C. Proof of Lemma 4

For any $i \in \{1 : N-1\}$, we have

$$\begin{aligned}
& H(W_{[1:i]}, Z_1) - H(W_{[1:i-1]}, Z_1) \\
&= H(W_i|W_{[1:i-1]}, Z_1) \\
&= H(W_{i+1}|W_{[2:i]}, Z_1) && (38)\\
&\geq H(W_{i+1}|W_{[1:i]}, Z_1) \\
&= H(W_{[1:i+1]}, Z_1) - H(W_{[1:i]}, Z_1),
\end{aligned}$$

where (38) is from Lemma 1.

Then we have

$$\begin{aligned}
& \sum_{i=1}^{N-1} (N-i)[H(W_{[1:i]}, Z_1) - H(W_{[1:i-1]}, Z_1)] \\
&\qquad \geq \sum_{i=1}^{N-1} (N-i)[H(W_{[1:i+1]}, Z_1) - H(W_{[1:i]}, Z_1)] \\
&\Leftrightarrow [\sum_{i=1}^{N-1} H(W_{[1:i]}, Z_1)] - (N-1) H(Z_1) \\
&\qquad \geq [\sum_{i=1}^{N-1} H(W_{[1:i+1]}, Z_1)] - (N-1) H(W_1, Z_1) \\
&\Leftrightarrow (N-1) H(W_1, Z_1) - (N-1) H(Z_1) \\
&\qquad \geq H(W_{[1:N]}, Z_1) - H(W_1, Z_1) \\
&\Leftrightarrow N H(W_1, Z_1) - H(W_{[1:N]}) \geq (N-1) H(Z_1) && (39)\\
&\Leftrightarrow N H(Z_1|W_1) \geq (N-1) H(Z_1),
\end{aligned}$$

where (39) is from $H(Z_1|W_{\{1:N\}}) = 0$. Thus, we have proved (8), and the rest case follows from symmetry.

D. Proof of Lemma 5

For User 2, we have

$$\begin{aligned}
& M_2 + r_c \\
&\geq H(Z_2, X_c^{(2,1)}) \\
&= H(W_1) + H(Z_2|W_1) + H(X_c^{(2,1)}|Z_2, W_1) && (40)\\
&\geq 3 + \frac{N-2}{N-1} [H(Z_2|W_1) + H(Z_1|W_1)] \\
&\qquad - r_c - r_{p1} - M_1, && (41)
\end{aligned}$$

where (40) follows from the same steps as (30), and (41) is from (21) in Lemma 6.

And similarly to (41), we have

$$\begin{aligned}
& M_1 + r_c + r_{p1} \\
&\geq H(Z_1, X_c^{(1,2)}, X_{p1}^{(1,2)}) \\
&= H(W_1) + H(Z_1|W_1) + H(X_c^{(1,2)}, X_{p1}^{(1,2)}|Z_1, W_1) \\
&\geq 2 + \frac{N-2}{N-1} H(Z_1|W_1) - \frac{1}{N-1} H(Z_2|W_1), && (42)
\end{aligned}$$

where (42) follows from (20) in Lemma 6.

Therefore, by cancelling the term $H(Z_2|W_1)$ in (41) and (42), we obtain (19), which completes the proof.

E. The proof of Lemma 6

In problem $\mathcal{Q}(M_1, M_2, R_c, R_{p1}, 0)$, substituting $X_c^{(i,j)}$ in the proof of (37) with $(X_c^{(i,j)}, X_{p1}^{(i,j)})$, we get (20). Similarly, for (21), we have

$$\begin{aligned}
&(N-1)H(X_c^{(2,1)}|Z_2, W_1)\\
&= \sum_{i=2}^{N} H(X_c^{(i,1)}|Z_2, W_1) \qquad (43)\\
&\geq H(X_c^{([2:N],1)}|Z_2, W_1)\\
&\geq H(X_c^{([2:N],1)}, X_{p1}^{([2:N],1)}, Z_1|W_1) - H(Z_2|W_1)\\
&\quad - H(X_{p1}^{([2:N],1)}, Z_1|Z_2, W_1)\\
&= H(X_c^{([2:N],1)}, X_{p1}^{([2:N],1)}, Z_1, W_{[2:N]}|W_1) - H(Z_2|W_1)\\
&\quad - H(Z_1|Z_2, W_1) - H(X_{p1}^{([2:N],1)}|Z_1, Z_2, W_1)\\
&\geq (N-1) - [H(Z_2|W_1) + H(Z_1|W_1)]\\
&\quad - (N-1)H(X_{p1}^{(2,1)}|Z_1, Z_2, W_1), \qquad (44)
\end{aligned}$$

where (43) follows from Lemma 1, and (44) from $H(X_c^{([2:N],1)}, X_{p1}^{([2:N],1)}, Z_1|W_{[1:N]}) = 0$ and Lemma 1.

Finally, we upper bound $H(X_{p1}^{(2,1)}|Z_1, Z_2, W_1)$ as follows:

$$\begin{aligned}
&H(X_{p1}^{(2,1)}|W_1, Z_1, Z_2) \leq H(X_{p1}^{(2,1)}, X_c^{(2,1)}|W_1, Z_1, Z_2)\\
&= H(X_{p1}^{(2,1)}, X_c^{(2,1)}, Z_1, Z_2, W_1) - H(W_1, Z_1, Z_2)\\
&= H(X_{p1}^{(2,1)}, X_c^{(2,1)}, Z_1, Z_2, W_2) - H(W_1, Z_1, Z_2)\\
&= H(X_{p1}^{(2,1)}|W_2, X_c^{(2,1)}, Z_1, Z_2)\\
&\quad + H(W_2, X_c^{(2,1)}, Z_1, Z_2) - H(W_2, Z_1, Z_2) \qquad (45)\\
&= H(X_{p1}^{(2,1)}|W_2, X_c^{(2,1)}, Z_1, Z_2) + H(X_c^{(2,1)}|W_2, Z_1, Z_2)\\
&\leq H(X_{p1}^{(2,1)}|W_2, X_c^{(2,1)}, Z_1) + H(X_c^{(2,1)}|W_2, Z_1)\\
&= H(X_{p1}^{(2,1)}, X_c^{(2,1)}|W_2, Z_1)\\
&= H(X_{p1}^{(2,1)}, X_c^{(2,1)}|Z_1) - H(W_2) - H(Z_1|W_2) + H(Z_1)\\
&\leq r_c + r_{p1} + M_1 - 1 - H(Z_1|W_1), \qquad (46)
\end{aligned}$$

where (45) and (46) follow from Lemma 1. From (44) and (46), we obtain (21), which completes the proof.

F. Proof of Lemma 7

We will characterize $f_{(r_{p1}, r_{p2})}(M_1, M_2)$ for a given (r_{p1}, r_{p2}) pair. To do so, we consider the (M_1, M_2) plane for a fixed (r_{p1}, r_{p2}) pair, as illustrated in Fig. 4(a). The achievability follows from performing memory-sharing among the nine points specified below. These correspond to points A to G in Fig. 1(a), plus either transmitting $W_{d_2}^{p1}$ uncoded through the shared common link, or caching all files $\{W_1^{p1}, W_2^{p1}, \cdots, W_N^{p1}\}$ at User 2, which is also reflected in the notation used to refer to these points. Recall that all these points can be achieved from the twelve points P_A to P_L described in Section IV-B via memory-sharing. The points used in memory-sharing and the corresponding fractions for these nine points are given as follows.

1) Point A: it can be achieved by memory-sharing among Points P_A, P_H and P_I with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
2) Point B: it can be achieved by memory-sharing among Points P_B, P_H and P_I with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
3) Point B': it can be achieved by memory-sharing among Points P_B, P_H and P_K with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
4) Point C': it can be achieved by memory-sharing among Points P_C, P_H and P_K with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
5) Point D: it can be achieved by memory-sharing among Points P_D, P_H and P_I with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
6) Point E': it can be achieved by memory-sharing among Points P_E, P_H and P_K with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
7) Point F: it can be achieved by memory-sharing among Points P_F, P_H and P_I with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
8) Point G: it can be achieved by memory-sharing among Points P_G, P_H and P_I with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.
9) Point G': it can be achieved by memory-sharing among Points P_G, P_H and P_K with fractions $l_1, 1-l_2$ and l_2-l_1, respectively.

Next, we present the coding scheme for Points B and B' to illustrate our observation that the schemes either transmit $W_{d_2}^{p1}$ uncoded over the shared common link, or cache all the files $\{W_1^{p1}, W_2^{p1}, \cdots, W_N^{p1}\}$ at User 2. Similarly for the other points.

For point B with $(M_1, M_2, r_c) = (\frac{N}{2}l_1, \frac{N}{2}l_1, l_2 - \frac{l_1}{2})$, we use the scheme for Point B of Fig. 1(a) for subfiles $\{W_i^c, i \in [N]\}$, and transmit $W_{d_2}^{p1}$ through the common link. In other words, each subfile W_i^c is split into two parts of equal size (W_i^{c1}, W_i^{c2}), $i \in [N]$. User k caches $\{W_i^{ck}, i \in [N]\}$, $k = 1, 2$. In the delivery phase, $\{W_{d_1}^{c2} \oplus W_{d_2}^{c1}, W_{d_2}^{p1}\}$ is transmitted over the shared link.

For point B' with $(M_1, M_2, r_c) = (\frac{N}{2}l_1, Nl_2 - \frac{N}{2}l_1, \frac{l_1}{2})$, we also use the scheme for Point B of Fig. 1(a) for subfiles $\{W_i^c, i \in [N]\}$, i.e., each subfile W_i^c is split into two parts of equal size (W_i^{c1}, W_i^{c2}), $i \in [N]$. Compared with point B, instead of transmitting $W_{d_2}^{p1}$ through the common link, we cache $\{W_i^{p1}, i \in [N]\}$ at User 2. In other word, User k caches $\{W_i^{ck}, i \in [N]\}$, $k = 1, 2$, and furthermore, User 2 caches $\{W_1^{p1}, W_2^{p1}, \cdots, W_N^{p1}\}$.In the delivery phase, $\{W_{d_1}^{c2} \oplus W_{d_2}^{c1}\}$ is transmitted over the shared link.

In Fig. 4 (a), for $(M_1, M_2) \in \mathcal{M}_1$, we perform memory-sharing among Points A, B, F; for $(M_1, M_2) \in \mathcal{M}_2$, among Points A, B and G; for $(M_1, M_2) \in \mathcal{M}_3$, among B, B', G, G'; for $(M_1, M_2) \in \mathcal{M}_4$, among B, B', F, D, C'; for $(M_1, M_2) \in \mathcal{M}_5$, among Points C', B', G', E'. When $(M_1, M_2) \in \mathcal{M}_6$, the caches at both users are large enough, so we do not need to transmit any data over the shared link. When $(M_1, M_2) \in \mathcal{M}_7$, we waste the extra cache at User 1 and achieve the same performance as point $(Nl_1, M_2) \in \mathcal{M}_4$. Similarly, when $(M_1, M_2) \in \mathcal{M}_8$, we waste the extra cache at User 2 and achieve the same performance as point $(M_1, Nl_2) \in \mathcal{M}_5$. Hence, we focus on the non-trivial cases of

$\mathcal{M}_1 \bigcup \mathcal{M}_2 \bigcup \cdots \bigcup \mathcal{M}_5$, and the memory-sharing expressions are given by (26). By symmetry, we can also obtain (27).

G. Proof of Lemma 8

In this proof, we consider another projection of $f(M_1, M_2, r_{p1}, r_{p2})$ where we fix the pair (M_1, M_2) and focus on the function $f_{(M_1,M_2)}(r_{p1}, r_{p2})$ for the remaining parameters (r_{p1}, r_{p2}).

Note that $\bar{r}_c = f_{(M_1,M_2)}(r_{p1}, r_{p2})$ can be found explicitly from (26) or (27), albeit the expressions may be tedious to write explicitly. However, we do not need the explicit expression of $f_{(M_1,M_2)}(\cdot)$, only its following properties: i) Since $f(M_1, M_2, r_{p1}, r_{p2})$ is continuous and the closed-form expression of $f_{(r_{p1},r_{p2})}(M_1, M_2)$ in (26) and (27) is monotonically decreasing in (r_{p1}, r_{p2}), $f_{(M_1,M_2)}(r_{p1}, r_{p2})$ is a continuous and monotonically decreasing function of (r_{p1}, r_{p2}), where the monotonicity is defined as $f_{(M_1,M_2)}(r_{p1}, r_{p2}) \geq f_{(M_1,M_2)}(r'_{p1}, r'_{p2})$ if $r_{p1} \leq r'_{p1}, r_{p2} \leq r'_{p2}$; ii) The value of $f_{(M_1,M_2)}(r_{p1}, r_{p2})$ can take only one of the five values in (26) or (27).

For a given and fixed (M_1, M_2) pair, we pick an achievable (r_{p1}, r_{p2}, r_c) tuple as follows:

Note that, since none of the points with coded cache, i.e., P_F and P_G, lie on the boundary in this projection, it is sufficient to only consider the rectangle $0 \leq r_{pi} \leq 1 - \frac{M_i}{N}$, $i = 1, 2$, since the rate $r_{pi} = 1 - \frac{M_i}{N}$, $i = 1, 2$, is enough for User i, $i = 1, 2$, to recover the file, respectively.

We have the following cases as shown in Figure 6, which shows the projection to the space with parameters (r_{p1}, r_{p2}):

- **Case 1:** $\{M_1 > M_2, 0 \leq \frac{R_{p2}}{R_{p1}} \leq 1\}$ or $\{M_1 < M_2, \frac{R_{p2}}{R_{p1}} \leq \frac{N-M_2}{N-M_1}\}$, i.e., Fig. 6(a) and (b). For this case, we further have the following two sub-cases:
  - $0 \leq \frac{R_{p1}}{R_c+R_{p2}} \leq \frac{N-M_1}{N-M_2}$: The achievable (r_{p1}, r_{p2}) we pick is inside the rectangle, and also lies on line $r_{p2} = \frac{R_{p2}}{R_{p1}} r_{p1}$, i.e., it is the line segment of OP in Fig. 6(a) or Fig. 6(b).

    Consider the following function of r_{p1}:

    $$g_1(r_{p1}) \triangleq \frac{r_{p1}}{f_{(M_1,M_2)}(r_{p1}, \frac{R_{p2}}{R_{p1}} r_{p1}) + \frac{R_{p2}}{R_{p1}} r_{p1}}.$$

    Since $f_{(M_1,M_2)}(r_{p1}, r_{p2})$ is continuous and monotonically decreasing, $g_1(r_{p1})$ is continuous and montonically increasing. At the point O in Fig. 6(a) and 6(b), i.e.,
    $(r_{p1}, r_{p2}) = (0, 0)$, $g_1(0) = 0$. At the point P in Fig. 6(a) and 6(b), i.e., $(r_{p1}, r_{p2}) = \left(1 - \frac{M_1}{N}, \frac{R_{p2}}{R_{p1}}\left(1 - \frac{M_1}{N}\right)\right)$, we have $M_1 = N(1 - r_{p1}) = Nl_1$, which in region $\mathcal{M}_4(r_{p1}, r_{p2})$ in (26). This gives us $f_{(M_1,M_2)}(r_{p1}, \frac{R_{p2}}{R_{p1}} r_{p1}) + \frac{R_{p2}}{R_{p1}} r_{p1} = 1 - \frac{M_2}{N}$, and as a result, $g_1(1 - \frac{M_1}{N}) = \frac{N-M_1}{N-M_2}$. Since we are considering the case $0 \leq \frac{R_{p1}}{R_c+R_{p2}} \leq \frac{N-M_1}{N-M_2}$, we may find a $\tilde{r}_{p1}$, where $\left(\tilde{r}_{p1}, \frac{R_{p2}}{R_{p1}} \tilde{r}_{p1}\right)$ lies on the line segment OP in Fig. 6(a) and 6(b), that satisfies

    $$g_1(\tilde{r}_{p1}) = \frac{R_{p1}}{R_c + R_{p2}},$$

    and the (r_{p1}, r_{p2}, r_c) point we pick to calculate $T = \max\{\frac{r_{p1}}{R_{p1}}, \frac{r_{p2}}{R_{p2}}, \frac{r_c}{R_c}\}$ is $(\hat{r}_{p1}, \hat{r}_{p2}, \hat{r}_c) = \left(\tilde{r}_{p1}, \frac{R_{p2}}{R_{p1}} \tilde{r}_{p1}, f_{(M_1,M_2)}(\tilde{r}_{p1}, \frac{R_{p2}}{R_{p1}} \tilde{r}_{p1})\right)$. Note that this point satisfies

    $$\frac{\hat{r}_{p1}}{R_{p1}} = \frac{\hat{r}_{p2}}{R_{p2}} = \frac{\hat{r}_c}{R_c}. \qquad (47)$$

    Since $\left(\tilde{r}_{p1}, \frac{R_{p2}}{R_{p1}} \tilde{r}_{p1}\right)$ can take all values on the line segment OP for some (R_{p1}, R_{p2}, R_c), then the pair (M_1, M_2) can appear in these five regions $\mathcal{M}_1, \cdots, \mathcal{M}_5$ in (26) for some (R_{p1}, R_{p2}, R_c). Therefore, since the value of $f_{M_1,M_2}(\hat{r}_{p1}, \hat{r}_{p2})$ can take only one of the five corresponding values in (26), combining with (47), we see that $T = \max\left\{\frac{\hat{r}_{p1}}{R_{p1}}, \frac{\hat{r}_{p2}}{R_{p2}}, \frac{\hat{r}_c}{R_c}\right\}$ can only take one of the following values

    $$\left\{\frac{2 - \frac{3M_2}{N} - \frac{M_1-M_2}{N-1}}{R_c + R_{p1} + R_{p2}}, \frac{2 - \frac{3M_1}{N} - \frac{M_2-M_1}{N-1}}{R_c + R_{p1} + R_{p2}}, \right. \\ \frac{N(2N-1) - 2(N-1)M_1 - NM_2}{N^2(R_c + R_{p2}) + N(N-1)R_{p1}}, \\ \left. \frac{1 - \frac{M_1}{N}}{R_c + R_{p1}}, \frac{1 - \frac{M_2}{N}}{R_c + R_{p2}}\right\}. \qquad (48)$$

    Note that, in this sub-case, it is easy to check that the optimal latency T^* showed in (12) is equal to the maximum value of (48). Therefore, we have shown that $T = T^*$ in this sub-case due to the fact that T^* is the lower bound of T.
  - $\frac{R_{p1}}{R_c+R_{p2}} > \frac{N-M_1}{N-M_2}$: The achievable (r_{p1}, r_{p2}) lies on the line segment QR in Fig. 6(a) or Fig. 6(b), i.e., $r_{p1} = 1 - \frac{M_1}{N}$. Now, we pick r_c within the three-dimensional achievable region, and this will determine to which point on line segment QR it corresponds.

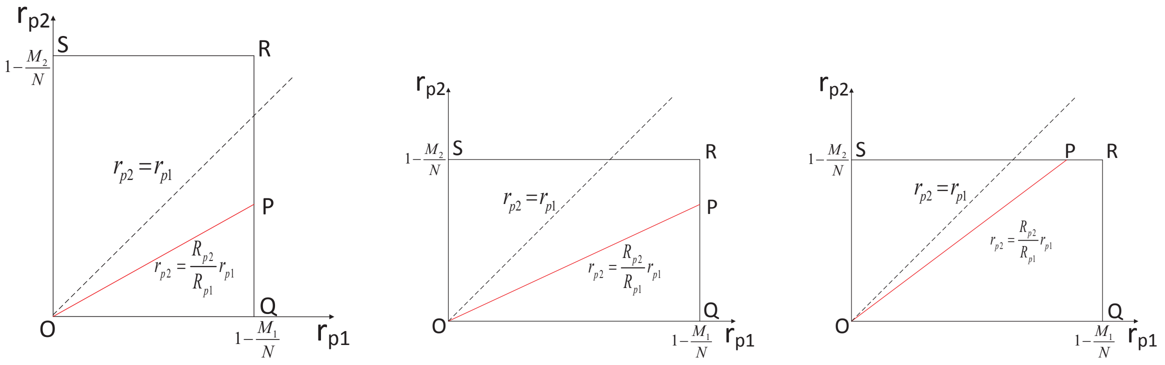
    Consider the following function of r_{p2}:

    $$g_2(r_{p2}) \triangleq \frac{f_{(M_1,M_2)}(1 - \frac{M_1}{N}, r_{p2})}{r_{p2}}.$$

    Since $f_{(M_1,M_2)}(r_{p1}, r_{p2})$ is continuous and monotonically decreasing, so is $g_2(r_{p2})$. At point Q in Fig. 6(a) and 6(b), i.e., $(r_{p1}, r_{p2}) = (1 - \frac{M_1}{N}, 0)$, $g_2(0) = \infty$, we have $M_1 = N(1 - r_{p1}) = Nl_1$. At the point R in Fig. 6(a) and 6(b), i.e., $(r_{p1}, r_{p2}) = \left(1 - \frac{M_1}{N}, 1 - \frac{M_2}{N}\right)$, we have $M_1 = N(1 - r_{p1}), M_2 = N(1 - r_{p2})$ which is the point C' in Fig. 4(a) or 4((b). This gives us $f_{(M_1,M_2)}\left(1 - \frac{M_1}{N}, 1 - \frac{M_2}{N}\right) = 0$, and as a result, $g_2(1 - \frac{M_2}{N}) = 0$. Hence, we may find a point $(1 - \frac{M_1}{N}, \tilde{r}_{p2})$ on line segment QR that satisfies

    $$g_2(\tilde{r}_{p2}) = \frac{R_c}{R_{p2}},$$

    and the (r_{p1}, r_{p2}, r_c) point we pick to calculate $T = \max\{\frac{r_{p1}}{R_{p1}}, \frac{r_{p2}}{R_{p2}}, \frac{r_c}{R_c}\}$ is $(\hat{r}_{p1}, \hat{r}_{p2}, \hat{r}_c) =$



(a) $M_1 \geq M_2$ and $0 \leq \frac{R_{p2}}{R_{p1}} \leq 1$ (b) $M_1 \leq M_2$ and $0 \leq \frac{R_{p2}}{R_{p1}} \leq \frac{N-M_2}{N-M_1}$ (c) $M_1 \leq M_2$ and $\frac{N-M_2}{N-M_1} \leq \frac{R_{p2}}{R_{p1}} \leq 1$

Fig. 6. For a fixed (M_1, M_2) pair, the achievable (r_{p1}, r_{p2}) region.

$\left(1-\frac{M_1}{N}, \tilde{r}_{p2}, f_{(M_1,M_2)}(1-\frac{M_1}{N}, \tilde{r}_{p2})\right)$. Note that this point satisfies

$$\frac{\hat{r}_{p2}}{R_{p2}} = \frac{\hat{r}_c}{R_c} \geq \frac{\hat{r}_{p1}}{R_{p1}}. \tag{49}$$

where the last $\geq$ follows from $\tilde{r}_{p2} + f_{(M_1,M_2)}(1-\frac{M_1}{N}, \tilde{r}_{p2}) = 1 - \frac{M_2}{N}$ and $\frac{R_{p1}}{R_c+R_{p2}} > \frac{N-M_1}{N-M_2}$. In this sub-case, (M_1, M_2) is always in the line segment of $C'D$ in Fig. 4 (a) or $C'D'$ in Fig. 4 (b), i.e., $M_1 = N(1-r_{p1}) = Nl_1$. Therefore, the value of $f_{(M_1,M_2)}(1-\frac{M_1}{N}, \tilde{r}_{p2})$ is $1-\tilde{r}_{p2}-\frac{M_2}{N}$. Combining with (49), we see that $T = \max\left\{\frac{\hat{r}_{p1}}{R_{p1}}, \frac{\hat{r}_{p2}}{R_{p2}}, \frac{\hat{r}_c}{R_c}\right\}$ can only take the following value

$$T = \frac{1-\frac{M_2}{N}}{R_c + R_{p2}}.$$

Note that, in this sub-case, it is easy to check that the optimal latency T^* showed in (12) is equal to T.

- **Case 2:** the remaining case in Fig. 6 (c). For this case, we again consider two sub-cases: 1)$0 \leq \frac{R_{p2}}{R_c+R_{p1}} \leq \frac{N-M_2}{N-M_1}$ and 2) $\frac{R_{p2}}{R_c+R_{p1}} \geq \frac{N-M_2}{N-M_1}$. The proof can be completed by using a similar argument in Case 1. Due to the space limit, we omit the details.

H. Converse proof of Theorem 3

Firstly, we denote S_i as the i-th source and $\hat{S}_i^k$ as the i-th source recovered by the k-th user, in which $i = 1, \cdots, N$ and $k = 1, 2$. Due to the independence of the sources and the constraints of users' decoding, the Lemmas 1 and 4 apply to this model, i.e., there must be an optimal source-index-symmetric caching and delivery code, for which we have:

$$NH(Z_i|S_1) \geq (N-1)H(Z_i), \quad \forall i = 1, 2 \tag{50}$$

Then, similarly to Lemma 3, we have

$$\begin{aligned}
&(N-1)H(X_c^{(1,2)}|Z_1, S_1) \\
&= \sum_{i=2}^{N} H(X_c^{(1,i)}|Z_1, S_1) && (51)\\
&\geq H(X_c^{(1,[2:N])}|Z_1, S_1) \\
&\geq H(X_c^{(1,[2:N])}, Z_2|S_1) - H(Z_1|S_1) - H(Z_2|Z_1, S_1) \\
&= H(X_c^{(1,[2:N])}, Z_2, \hat{S}^2_{[2:N]}|S_1) \\
&\qquad - H(Z_2|S_1) - H(Z_1|Z_2, S_1) && (52)\\
&= H(\hat{S}^2_{[2:N]}|S_1) + H(X_c^{(1,[2:N])}, Z_2|\hat{S}^2_{[2:N]}, S_1) \\
&\qquad - H(Z_2|S_1) - H(Z_1|Z_2, S_1) \\
&\geq H(\hat{S}^2_{[2:N]}|S_1) + H(X_c^{(1,[2:N])}, Z_2|S_{[1:N]}) \\
&\qquad - H(Z_2|S_1) - H(Z_1|S_1) \\
&\geq \sum_{i=2}^{N} H(\hat{S}^2_i) + H(X_c^{(1,[2:N])}, Z_2|S_{[1:N]}) \\
&\qquad - H(Z_2|S_1) - H(Z_1|S_1) && (53)\\
&\geq (N-1)l_2 - [H(Z_2|S_1) + H(Z_1|S_1)], && (54)
\end{aligned}$$

where (51) follows since we consider source-index-symmetric codes; (52) from the recovery of requests from the transmitted messages and cache contents; (53) from the independence of sources; and (54) from the definition of the rate distortion function. Similarly,

$$\begin{aligned}
&(N-1)H(X_c^{(2,1)}|Z_2, S_1) \\
&\geq (N-1)l_1 - [H(Z_1|S_1) + H(Z_2|S_1)].
\end{aligned}$$

Then, similarly to Lemma 3, we have

$$\begin{aligned}
&r_c + M_1 \\
&\geq H(X_c^{(1,2)}) + H(Z_1) \\
&\geq H(Z_1, X_c^{(1,2)}) \\
&= H(Z_1, X_c^{(1,2)}, \hat{S}_1^1) \\
&= H(\hat{S}_1^1) + H(Z_1|\hat{S}_1^1) + H(X_c^{(1,2)}|Z_1, \hat{S}_1^1) \\
&\geq H(\hat{S}_1^1) + H(Z_1|S_1) + H(X_c^{(1,2)}|Z_1, S_1) \\
&\geq l_1 + H(Z_1|S_1)
\end{aligned}$$

$$+ (l_2 - \frac{1}{N-1}[H(Z_1|S_1) + H(Z_2|S_1)]) \tag{55}$$

$$\geq l_1 + l_2 + \frac{N-2}{N-1}H(Z_1|S_1) - \frac{1}{N-1}H(Z_2|S_1), \tag{56}$$

where (55) follows from (54) and the definition of the rate distortion function.

Similarly, by exchanging the indices of 1 and 2, we have

$$\begin{aligned} & r_c + M_2 \\ & \geq l_1 + l_2 + \frac{N-2}{N-1}H(Z_2|S_1) - \frac{1}{N-1}H(Z_1|S_1). \end{aligned} \tag{57}$$

By cancelling the term $H(Z_1|S_1)$ in (56) and (57), we obtain for $N \geq 3$

$$\begin{aligned} & M_1 + r_c + (N-2)[r_c + M_2] \\ & \geq (N-1)(l_1 + l_2) + (N-3)H(Z_2|S_1) \\ & \geq (N-1)(l_1 + l_2) + \frac{(N-3)(N-1)}{N}H(Z_2), \end{aligned} \tag{58}$$

where (58) is from (50). Hence, following from (58), we have

$$NM_1 + (2N-3)M_2 + N(N-1)r_c \geq N(N-1)(l_1 + l_2). \tag{59}$$

Symmetrically,

$$NM_2 + (2N-3)M_1 + N(N-1)r_c \geq N(N-1)(l_1 + l_2). \tag{60}$$

Then

$$\begin{aligned} & M_1 + M_2 + 2r_c \\ & \geq H(Z_1, X_c^{(1,2)}) + H(Z_2, X_c^{(2,1)}) \\ & = H(Z_1, X_c^{(1,2)}, \hat{S}_1^1) + H(Z_2, X_c^{(2,1)}, \hat{S}_1^2) \\ & = H(\hat{S}_1^1) + H(Z_1|\hat{S}_1^1) + H(X_c^{(1,2)}|Z_1, \hat{S}_1^1) \\ & \quad + H(\hat{S}_1^2) + H(Z_2|\hat{S}_1^2) + H(X_c^{(2,1)}|Z_2, \hat{S}_1^2) \\ & \geq H(\hat{S}_1^1) + H(Z_1|S_1) + H(X_c^{(1,2)}|Z_1, S_1) \\ & \quad + H(\hat{S}_1^2) + H(Z_2|S_1) \\ & \geq l_1 + 2l_2 + \frac{N-2}{N-1}\left[H(Z_2|S_1) + H(Z_1|S_1)\right], \end{aligned} \tag{61}$$

where (61) follows from (54).

Recall that

$$\begin{aligned} & r_c + M_1 \\ & \geq l_1 + l_2 + \frac{N-2}{N-1}H(Z_1|S_1) - \frac{1}{N-1}H(Z_2|S_1). \end{aligned} \tag{62}$$

Therefore, by cancelling the term $H(Z_2|S_1)$ in (61) and (62), we obtain

$$\begin{aligned} & M_1 + M_2 + 2r_c + (N-2)(r_c + M_1) \\ & \geq (N-1)l_1 + Nl_2 + (N-2)H(Z_1|S_1) \\ & \geq (N-1)l_1 + Nl_2 + \frac{(N-2)(N-1)}{N}H(Z_1), \end{aligned} \tag{63}$$

where (63) follows from (50). Hence, we have

$$2(N-1)M_1 + NM_2 + N^2 r_c \geq N(N-1)l_1 + N^2 l_2. \tag{64}$$

Similarly, we have

$$2(N-1)M_2 + NM_1 + N^2 r_c \geq N(N-1)l_2 + N^2 l_1. \tag{65}$$

Finally, from (59), (60), (64), (65) and the cut-set bound proved in [18], the converse proof is completed.

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] Z. Chen, P. Fan, and K. B. Letaief, “Fundamental limits of caching: Improved bounds for users with small buffers,” *IET Communications*, vol. 10, no. 17, pp. 2315–2318, 2016.

[3] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, “Caching and coded multicasting: Multiple groupcast index coding,” in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 881–885.

[4] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, “Finite-length analysis of caching-aided coded multicasting,” *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5524–5537, 2016.

[5] J. Gómez-Vilardebó, “Fundamental limits of caching: Improved bounds with coded prefetching,” *arXiv preprint arXiv:1612.09071*, 2016.

[6] C. Tian and J. Chen, “Caching and delivery via interference elimination,” in *IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 830–834.

[7] M. Mohammadi Amiri, Q. Yang, and D. Gunduz, “Coded caching for a large number of users,” in *IEEE Information Theory Workshop (ITW)*, 2016, pp. 171–175.

[8] M. Mohammadi Amiri and D. Gunduz, “Fundamental limits of coded caching: improved delivery rate-cache capacity trade-off,” *IEEE Transactions on Communications*, vol. 65, no. 2, pp. 806–815, 2017.

[9] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “The exact rate-memory tradeoff for caching with uncoded prefetching,” in *IEEE International Symposium on Information Theory (ISIT)*, vol. 63, no. 7. IEEE, 2017, pp. 1613–1617.

[10] A. Sengupta, R. Tandon, and T. C. Clancy, “Improved approximation of storage-rate tradeoff for caching via new outer bounds,” in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1691–1695.

[11] H. Ghasemi and A. Ramamoorthy, “Improved lower bounds for coded caching,” *IEEE Transactions on Information Theory*, pp. 4388–4413, 2017.

[12] C. Tian, “Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching,” *Entropy*, vol. 20, no. 8, p. 603, 2018.

[13] S. Wang, W. Li, X. Tian, and H. Liu, “Coded caching with heterogenous cache sizes,” *arXiv preprint arXiv:1504.01123*, 2015.

[14] M. Mohammadi Amiri, Q. Yang, and D. Gunduz, “Decentralized caching and coded delivery with distinct cache capacities,” *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4657 – 4669, 2017.

[15] A. M. Ibrahim, A. A. Zewail, and A. Yener, “Centralized coded caching with heterogeneous cache sizes,” in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[16] ——, “Benefits of coded placement for networks with heterogeneous cache sizes,” *arXiv preprint arXiv:1811.04067*, 2018.

[17] ——, “Coded caching for heterogeneous systems: An optimization perspective,” *arXiv preprint arXiv:1810.08187*, 2018.

[18] Q. Yang and D. Gündüz, “Coded caching and content delivery with heterogeneous distortion requirements,” *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4347–4364, 2018.

[19] J. Zhang and P. Elia, “Wireless coded caching: A topological perspective,” in *IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 401–405.

[20] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “On the optimality of separation between caching and delivery in general cache networks,” *arXiv preprint arXiv:1701.05881*, 2017.

[21] A. M. Ibrahim, A. A. Zewail, and A. Yener, “Optimization of heterogeneous caching systems with rate limited links,” in *IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[22] A. Ghorbel, M. Kobayashi, and S. Yang, “Content delivery in erasure broadcast channels with cache and feedback,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6407–6422, 2016.

[23] S. S. Bidokhti, M. Wigger, and R. Timo, “Erasure broadcast networks with receiver caching,” in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 1819–1823.

[24] M. Mohammadi Amiri and D. Gunduz, “Cache-aided content delivery over erasure broadcast channels,” *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 370–381, 2018.

[25] S. S. Bidokhti, M. Wigger, and R. Timo, “Noisy broadcast networks with receiver caching,” *arXiv preprint arXiv:1605.02317*, 2016.

[26] S. S. Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," *arXiv preprint arXiv:1702.08044*, 2017.

[27] M. Mohammadi Amiri and D. Gunduz, "Caching and coded delivery over gaussian broadcast channels for energy efficiency," *to appear, IEEE Journal on Selected Areas in Communications*, vol. 36, no. 8, pp. 1706–1720, Aug. 2018.

[28] ——, "On the capacity region of a cache-aided gaussian broadcast channel with multi-layer messages," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, June 2018, pp. 1909–1913.

[29] L. Ong, C. K. Ho, and F. Lim, "The single-uniprior index-coding problem: The single-sender case and the multi-sender extension," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3165–3182, June 2016.

[30] P. Sadeghi, F. Arbabjolfaei, and Y. Kim, "Distributed index coding," in *IEEE Information Theory Workshop (ITW)*, Sep. 2016, pp. 330–334.

[31] C. Thapa, L. Ong, and S. J. Johnson, "Graph-theoretic approaches to two-sender index coding," in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2016, pp. 1–6.

[32] M. Li, L. Ong, and S. J. Johnson, "Improved bounds for multi-sender index coding," in *IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 3060–3064.

[33] ——, "Cooperative multi-sender index coding," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1725–1739, March 2019.

[34] C. Tian, "Characterizing the rate region of the (4,3,3) exact-repair regenerating codes," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 5, pp. 967–975, May 2014.

[35] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 647–663, Jan 2019.

[36] H. S. Kang, M. G. Kang, and W. Choi, "The k-user linear deterministic broadcast channel with receiver memory," in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2017, pp. 1–6.

[37] A. Sengupta, R. Tandon, and T. C. Clancy, "Layered caching for heterogeneous storage," in *50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 719–723.

[38] T. M. Cover and W. H. Equitz, "Successive refinement of information," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 269–275, 1991.

Daming Cao received the B.Eng. degree in information engineering from Southeast University, Nanjing, China, in 2013. He is currently working toward the Ph.D. degree from the school of information science and engineering, Southeast University. His research interests mainly include information theory, network coding and security.

Deyao Zhang received the B.Eng. degree in telecommunications engineering from Xidian University, Xian, China, in 2016. He is currently working toward the M.Eng. degree at Southeast University. His research interest include information theory and network coding.

Pengyao Chen received the B.Eng. degree in communication engineering from Southwest Jiaotong University, Chengdu, P. R. China in 2015, and the M.Eng. degree in communication engineering from Southeast University, Nanjing, P. R. China in 2018. Her research interests are in network information theory for wireless networks.

Nan Liu received the B.Eng. degree in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, P. R. China in 2001, and the Ph.D. degree in electrical and computer engineering from University of Maryland, College Park, MD in 2007.

From 2007-2008, she was a postdoctoral scholar in the Wireless Systems Lab, Department of Electrical Engineering, Stanford University. In 2009, she became a professor in the National Mobile Communications Research Laboratory, School of Information Science and Engineering in Southeast University, Nanjing, China. Her research interests are in network information theory for wireless networks, SON algorithms for next generation cellular networks and energy-efficient communications. She is an associate editor for China Communications.

Wei Kang [S05-M08] received the B.Eng. degree in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2001, the M.Eng. degree in electrical engineering from McGill University, Montreal, Canada, in 2003, and the Ph.D. degree in electrical engineering from University of Maryland, College Park, in 2008. He joined the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2009, where he is currently an Associate Professor. His research mainly focuses on network information theory.

Deniz Gündüz [S03-M08-SM13] received the B.S. degree in electrical and electronics engineering from METU, Turkey in 2002, and the M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively. After his PhD, he served as a postdoctoral research associate at Princeton University, and as a consulting assistant professor at Stanford University. He was a research associate at CTTC in Barcelona, Spain until September 2012, when he joined the Electrical and Electronic Engineering Department of Imperial College London, UK, where he is currently a Reader (Associate Professor) in information theory and communications, and leads the Information Processing and Communications Laboratory (IPC-Lab). His research interests lie in the areas of communications and information theory, machine learning, and privacy. He is the recipient of the IEEE Communications Society - Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, a Starting Grant of the European Research Council (ERC) in 2016, IEEE Communications Society Best Young Researcher Award for the Europe, Middle East, and Africa Region in 2014, Best Paper Award at the 2016 IEEE Wireless Communications and Networking Conference (WCNC), and the Best Student Paper Awards at the 2018 IEEE Wireless Communications and Networking Conference (WCNC) and the 2007 IEEE International Symposium on Information Theory (ISIT).