Decentralized Caching in Two-layer Networks: Algorithms and Limits

Lin Zhang, Student Member, IEEE, Zhao Wang, Ming Xiao, Senior Member, IEEE,
Gang Wu, Member, IEEE, and Shaoqian Li, Fellow, IEEE

Abstract—The decentralized caching is studied in two-layer networks, where users request contents through intermediate nodes (helpers) from a file server. By placing contents randomly and independently in each node and carefully designing the data delivery, the correlations of the pre-stored contents across layers can be utilized to reduce the transmission rates in the network. A hybrid caching scheme is developed by exploiting the cross-layer storage correlations, the single-layer multicast opportunities from the server (each helper) to helpers (the attached users), and the cross-layer multicast opportunities from the server to users. It is observed that, by the hybrid caching scheme, the achievable rate in the first layer is reduced without compromising the achievable rate in the second layer compared with the state of art. Furthermore, the achievable rate region is shown to be order-optimal and lies within constant margins to the information-theoretic optimum. In particular, the multiplicative and additive factors are carefully sharpened to be $1/n$ and $1$, respectively.

Index Terms—Decentralized coded caching, two-layer networks, hybrid caching scheme, feasible rate region, order-optimal caching

I. INTRODUCTION

Driven by the massive video clips in multimedia applications such as YouTube and Netflix, the Internet traffic caused by video-on-demand (VoD) is expected to increase by 1000× times in the next few years [1]. To accommodate the enormous traffic, a promising method is to utilize the available storage capacity across the whole network for content caching. The concept of content caching is created in the content distribution network (CDN) [2]–[6]. The principle of a CDN can be briefly interpreted as follows: In a communication networks, partial popular or preferred files are pre-stored in the storage memory close to users, e.g., web proxy, such that the local stored file can be accessed easily to reduce the total traffic from servers and offer low response latency. This heuristic manner of caching design is determined by the local storage capacity of each user. It is foreseen that this mechanism offers local caching gain, which is evident when the individual storage is large.

Recently, network coding is introduced in the content caching [7]. Different from the content caching in a traditional CDN, the pre-stored contents are viewed as the side-information, such that a coded delivery can be designed to create single-layer multicast opportunities (SMO) between the file server and users. By this means, the transmission rate from the file server can be further reduced based on a network coding method, revealing that the conventional way of caching in the CDN may be suboptimal. It is worth noting that the coding gain is shown to be proportional to the aggregated storage capacity of all users. Thus, content caching with network coding is demonstrated to leverage both local caching gain and global coding gain. In particular, Maddah-Ali & Niesen (MAU) develop a centralized coded caching scheme (we name it MAU-Centralized scheme hereafter) by a joint content placement and coded delivery protocol in [7].

Briefly speaking, files are first divided into segments based on the instantaneous network profile, e.g., the user number in the considered network and the storage capacity of each user. Then, file segments are placed to distributed storages in a deterministic and centralized manner, such that the SMO can be exploited based on network coding by viewing the pre-stored contents as side information during content delivery. Furthermore, the MAU-Centralized caching scheme is proved to be order-optimal by achieving the transmission rate within constant multiplicative and additive factors of the information-theoretic optimum.

However, there exists some important concerns on the robustness of the MAU-Centralized scheme. In particular, the centralized content placement together with the content delivery has to be re-designed once the instantaneous network profile is changed. To deal with this problem, a decentralized coded caching scheme (we name it MAU-Decentralized scheme hereafter) is proposed in [8]. The essential difference between the MAU-Centralized and the MAU-Decentralized schemes lies in the content placement. With the MAU-Decentralized scheme, contents are randomly and independently cached to distributed storages. Then, a greedy coded delivery algorithm is developed to complete all the request services. By adopting the decentralized placement approach, two different users have partial overlapped and distinct contents, which will stimulate SMO in the sequential content delivery. Meanwhile, this placement approach is independent from the instantaneous network profile and thus more robust than the centralized one. As a compromise, the MAU-Decentralized scheme needs a larger transmission rate in the delivery phase, in which contents are delivered from caching nodes to user terminals. However, it is remarkable that the MAU-Decentralized scheme is proved to be with the same order-optimality as the centralized one, i.e., the same constant multiplicative and additive factors.

After [7] and [8], the coded caching problems in various...
scenarios are studied. For instance, the coded caching with different storage sizes is studied in [9]. In [10] and [11], the randomness of user demands is considered and novel caching schemes are developed to reduce the average transmission rate. Heterogenous content popularity is investigated in [12]. Online coded caching schemes together with their performance limits are provided in [13]. In addition, the device-to-device (D2D) assisted content caching is considered in [14]. It is observed that single-layer networks are studied among these literature, in which users request files directly from the server. In practice, the network topology is usually tree-like and users have to obtain the requested files through intermediate nodes, i.e., helpers, from a file server. This scenario is studied in [15], where authors consider two-layer networks and intend to minimize the transmission rate in each layer. The important observation of [15] is that no tension exists between the rates of two layers, such that both layers can simultaneously operate at approximately the minimum rate. Yet, the problem still remains on whether the achievable rates and performance bounds can be improved further, which is both practically and theoretically important. In this paper, we focus on the two-layer caching network with decentralized algorithms similar to [15]. We observe that, the correlations between the pre-stored contents in two layers can be utilized by decentralized placement together with an optimized design of the delivery. This further reduces the transmission rate in the first layer without compromising the transmission rate in the second layer and leads to a cross-layer storage correlations (CSC) caching gain. By combining SMO, CMO, and CSC in a proper way, we propose a hybrid caching scheme for the two-layer network. It is worth noting that, the caching gains from SMO, CMO, and CSC can be accumulated in the hybrid caching scheme, achieving a better achievable rate region compared with [15]. To highlight our contributions, we summarize the main content as follows:

- By exploiting both SMO and CSC, an S&c caching scheme is developed. We note that if the MAU-Decentralized caching scheme is directly applied in individual layers (i.e., the caching scheme A in [15]), the whole requested file of a user will be resolved at its helper regardless of the pre-stored contents in the user’s storage during the data delivery. Then, only the SMO, specifically, from the server to helpers and from each helper to the attached users, is utilized in the delivery phase. We observe that, the requested contents available in the user’s storage can be locally accessed and do not need to be recovered at the helper. In other words, considering the correlation of the pre-stored contents of two layers in the placement phase, we aim to deliver the data selectively based on previous design, e.g., [15]. The goal is to make the data delivery phase more efficient. Motivated by this, an S&c caching scheme is proposed to make use of both SMO and CSC, and the transmission rates of two layers are also analytically derived in closed forms. It is observed that, with the S&c caching scheme, the achievable rate in the first layer can be reduced without compromising the achievable rate in the second layer compared with the caching scheme A in [15].
- By masking the caching capability in helpers and applying the MAU-Decentralized caching scheme between the server and users (i.e., the caching scheme B in [15]), the multicast opportunities between the server and users, i.e., CMO, can be elaborated. To exploit SMO, CMO, and CSC simultaneously, a hybrid algorithm is proposed by combining the CSC caching scheme and the caching scheme B in [15] in a memory-sharing manner. Specifically, the whole network is divided into two parallel sub-networks with one of the algorithms adopted in each sub-network. The memory sharing parameters \( \alpha, \beta \in [0, 1] \) are introduced into the hybrid algorithm to represent the network segmentation. We derive the achievable transmission rates of two layers with respect to \( \alpha, \beta \). Notably, for any \( \alpha, \beta \), with the hybrid caching scheme, the achievable transmission rate in the first layer can be reduced without compromising the achievable transmission rate in the second layer compared with the generalized caching algorithm in [15].

The order-optimality of the hybrid caching scheme is demonstrated as follows,

\[
R^H(M_1, M_2) \subset \mathcal{R}(M_1, M_2) \subset c_1 R^H(M_1, M_2) - c_2,
\]

where \( M_1 \) and \( M_2 \) are the storage sizes in two layers, \( R^H \) is the achievable rate region with the proposed hybrid caching scheme and \( \mathcal{R} \) is the feasible rate region. Both \( R^H \) and \( \mathcal{R} \) will be formally defined later on. In particular, the constant multiplicative and additive factors are carefully quantified to be \( c_1 = \frac{1}{38} \) and \( c_2 = 4 \), respectively. Comparing both factors with the generalized caching scheme in [15], where the constant multiplicative and additive factors are characterized to be \( c_1 = \frac{1}{16} \) and \( c_2 = 16 \) respectively, the margins has been reduced in our results. The reason behind the improvement contains two folds. The first one comes from the gain of exploiting the correlations of the pre-stored contents in two layers, which enables to achieve smaller transmission rates than the generalized caching scheme in [15]. The second one is the optimization of information-theoretic bounds proof.

II. System Model

We consider a two-layer caching network consisting of a server, \( K_1 \) helpers, and \( K_1K_2 \) users as shown in Fig. 1. Specifically, the server is connected to \( K_1 (K_1 \geq 2) \) helpers \( H_i (i \in [1 : K_1]) \), each of which is connected to \( K_2 (K_2 \geq 2) \) users \( U_{i,j} (j \in [1 : K_2]) \) with orthogonal channels. In this network, the server hosts \( N \) files \( \mathcal{F} = \{ f_n : n \in [1 : N] \} \) each with size \( F \) bytes. Each helper and each user are equipped with the memory sizes of \( M_1F \) bits and \( M_2F \) bits \((M_1, M_2 \leq N)\), respectively. We denote the normalized memory sizes of each helper and each user as \( M_1 = \frac{M_1F}{P} \) and \( M_2 = \frac{M_2F}{P} \) files, respectively.

*For \( K_1 = 1 \) or \( K_2 = 1 \), the two-layer network can be reduced to a single-layer network. Thus, we study the two-layer network with \( K_1 \geq 2 \) and \( K_2 \geq 2 \) in this paper.
We consider a two-phase caching protocol which consists of a placement phase and a delivery phase in a proper sequence. In the placement phase, the server places contents into the memories of $K_1$ helpers by using $K_1$ placement functions $\mathbf{P}^H_i : \mathbb{F}^N \to \mathbb{F}^M (i \in [1 : K_1])$, and meanwhile feeds the caches of $K_1K_2$ users with normalized memory size by using $K_1K_2$ placement functions $\mathbf{P}^{U_{i,j}} : \mathbb{F}^N \to \mathbb{F}^M (i \in [1 : K_1], j \in [1 : K_2])$. In the delivery phase, each user $U_{i,j}$ requests for a file $f_{d_{i,j}}$ ($d_{i,j} \in [1 : N]$) from the server via the corresponding helper $H_i$. After receiving the requested set $D = \{f_{d_{i,j}} : i \in [1 : K_1], j \in [1 : K_2]\}$ from $K_1K_2$ users, the server transmits a message $Y_S(D, \mathbf{P}^H_i, \mathbf{P}^{U_{i,j}} : i \in [1 : K_1], j \in [1 : K_2])$ (referred as $Y_S$ hereafter) to $K_1$ helpers. Sequentially, combining the received message $Y_S$ from the server, the placement functions $\mathbf{P}^H_i$ and $\mathbf{P}^{U_{i,j}}$, and the requests set $D_i = \{f_{d_{i,j}} : j \in [1 : K_2]\}$, each helper $H_i$ transmits a message $Y_{H_i}(Y_S, D_i, \mathbf{P}^H_i, \mathbf{P}^{U_{i,j}} : j \in [1 : K_2])$ (referred as $Y_{H_i}$ hereafter) to the attached $K_2$ users. Then, each user $U_{i,j}$ recovers the requested file $f_{d_{i,j}}$ from the received message $Y_{H_i}$ and the placement function $\mathbf{P}^{U_{i,j}}$.

We denote the normalized (by the file size) delivery rates of the server and each helper as $R_1 = \frac{Y_S}{F}$ and $R_2 = \frac{\max\{|Y_{H_i}| : i \in [1 : K_1]\}}{F}$, respectively. For given $K_1, K_2$ (determined by a network topology), and an arbitrary given request set $D$, the tuple $(M_1, M_2, R_1, R_2)$ is achieved, if for a large enough file size $F$, each requested $f_{d_{i,j}} \in D$ can be recovered with error probability arbitrarily close to 0. Our major interests is the feasible rate region, which can be defined as follows:

**Definition 1 (Feasible rate region):** For memory sizes $M_1, M_2 \geq 0$, and the number of helpers $K_1$, the number of attached users $K_2$, the feasible rate region $\mathcal{R}$ is defined as the closure of rate pairs $(R_1, R_2)$, such that $(M_1, M_2, R_1, R_2)$ is achievable.

III. S&C CACHING SCHEME: EXPLOITING BOTH SMO AND CSC

In this section, we shall introduce an S&C caching scheme, which exploits the single-layer multicast opportunities from the server to helpers and from a helper to the attached users, as well as the correlations of the pre-stored contents in two layers to enhance the caching gain in a two-layer network. In what follows, we first present the principle and results of this scheme. The placement algorithm and delivery algorithm of this scheme are provided sequentially.

A. Principle and results of the S&C caching scheme

In the considered two-layer caching network, if we directly apply the MAU-Decentralized scheme in individual layers, all the requested files will be resolved at helpers from the server delivery without considering the cached contents in users. Thus, the server delivery simply neglects the existence of user storages. To deal with this problem, we develop the S&C scheme to consider the caches of both layers such that the unnecessary delivery data can be removed from directly applying the MAU-Decentralized scheme in individual layer. Therefore, the delivery rate in the two-layer network will be reduced.

**Lemma 1:** In the considered two-layer caching network, we assume the normalized memory sizes at each helper and each user as $M_1$ and $M_2$, respectively. Then, the transmission rates from the server and each helper with the S&C caching scheme are

$$R_1^{S&C}(M_1, M_2) = K_2 \left(1 - \frac{M_1}{N}\right) \left(1 - \frac{M_2}{N}\right) \frac{N}{M_1} \left(1 - \left(1 - \frac{M_1}{N}\right)^{K_1}\right) \tag{1}$$

and

$$R_2^{S&C}(M_2) = \left(1 - \frac{M_2}{N}\right) \frac{N}{M_2} \left(1 - \left(1 - \frac{M_2}{N}\right)^{K_2}\right) \tag{2}$$

respectively.

**Remark 1 (Performance Comparison with [15]):** If we directly apply the MAU-Decentralized scheme in individual layers (i.e., the decentralized caching scheme A in [15]), the transmission rates from the server and each helper are

$$R_1^A(M_1, M_2) = K_2 \left(1 - \frac{M_1}{N}\right) \frac{N}{M_1} \left(1 - \left(1 - \frac{M_1}{N}\right)^{K_1}\right) \tag{3}$$

and

$$R_2^A(M_2) = \left(1 - \frac{M_2}{N}\right) \frac{N}{M_2} \left(1 - \left(1 - \frac{M_2}{N}\right)^{K_2}\right) \tag{4}$$

respectively.

By comparing $R_1^{S&C}$ with $R_1^A$, the transmission rate from the server with the S&C caching scheme is scaled by a factor $1 - \frac{M_2}{N}$, which is smaller than 1. Therefore, the transmission rate from the server with the S&C caching scheme is smaller than that with the decentralized caching scheme A in [15].
Furthermore, we observe that $R_{S&C}^2$ is equal to $R_2^F$. Therefore, with the S&C caching scheme, the achievable rate in the first layer can be reduced without compromising the achievable rate in the second layer compared with the decentralized caching scheme A in [15].

### B. Placement and Delivery of the S&C scheme

In this part, we will present the detailed placement and delivery methods of the S&C scheme.

1) **Placement at helpers and users:** For the placement at helpers, each helper randomly and independently selects $\frac{M_F}{N}$ bits from each file $n \in [1 : K]$ and stores them into its storage. Similarly, for the placement at users, each user randomly and independently selects $\frac{M_F}{N}$ bits from each file $n \in [1 : K]$ and stores them into its storage.

2) **Delivery from the server to helpers:** To begin with, $V_{d_{i,j},S}$ denotes the contents of file $f_{d_{i,j}}$ and are exclusively cached at the nodes in set $S$. From the result in (7), the server needs to transmit $\{\oplus_{i \in S_1} V_{d_{i,j},S_{\mu_i}^H} : S_1^H \subset [1 : K_1], |S_1^H| = s_1, s_1 \in [1 : K_1], j \in [1, K_2]\}$, where $S_1^H$ denotes the set of helper indexes, such that each helper can recover the requested files of the attached users. Due to the randomness and independency of the contents placement at helpers and users, each bit in $V_{d_{i,j},S_{\mu_i}^H}$ is cached in user $U_{i,j}$ with probability $\frac{M_F}{N}$. As a result, we can divide each $V_{d_{i,j},S_{\mu_i}^H}$ in $\{\oplus_{i \in S_1} V_{d_{i,j},S_{\mu_i}^H} : S_1^H \subset [1 : K_1], |S_1^H| = s_1, s_1 \in [1 : K_1], j \in [1, K_2]\}$ into two parts, i.e.,

$$V_{d_{i,j},S_{\mu_i}^H} = \{V_{d_{i,j},S_{\mu_i}^H} : |V_{d_{i,j},S_{\mu_i}^H}| = \frac{M_2}{N} |V_{d_{i,j},S_{\mu_i}^H}| \text{ and } |V_{d_{i,j},S_{\mu_i}^H}| = 1 - \frac{M_2}{N} |V_{d_{i,j},S_{\mu_i}^H}| \}$$

Since $V_{d_{i,j},S_{\mu_i}^H}$ can be locally accessed by user $U_{i,j}$, helper $H_i$ does not need to recover $V_{d_{i,j},S_{\mu_i}^H}$. In other words, the server only needs to transmit $\{\oplus_{i \in S_1} V_{d_{i,j},S_{\mu_i}^H} : S_1^H \subset [1 : K_1], |S_1^H| = s_1, s_1 \in [1 : K_1], j \in [1, K_2]\}$, such that helper $H_i$ obtains all the requested subfiles in $f_{d_{i,j},U_{i,j}} = \{V_{d_{i,j},S_{\mu_i}^H} : S_1^H \subset [1 : K_1], i \in S_1, |S_1^H| = s_1, s_1 \in [1 : K_1], j \in [1, K_2]\}$ that are pre-stored in neither of helper $H_i$ and users. Thus, the transmission rate from the server can be obtained as (11).

3) **Delivery from helper $H_i$ to the attached users:** From the result in (7), helper $H_i$ needs to transmit $\{\oplus_{j \in S_2} V_{d_{i,j},S_{\mu_j}^U} : S_2^U \subset [1 : K_2], |S_2^U| = s_2, s_2 \in [1 : K_2]\}$, where $S_2^U$ denotes the set of user indexes attached to helper $H_i$, such that each user can recover the requested files of the attached users. In what follows, we shall prove that all the contents in $\{\oplus_{j \in S_2} V_{d_{i,j},S_{\mu_j}^U} : S_2^U \subset [1 : K_2], |S_2^U| = s_2, s_2 \in [1 : K_2]\}$ have been pre-stored or recovered by helper $H_i$.

According to whether the contents in $V_{d_{i,j},S_{\mu_j}^U}$ are pre-stored in helper $H_i$, $V_{d_{i,j},S_{\mu_j}^U}$ can be divided into two parts, i.e.,

$$V_{d_{i,j},S_{\mu_j}^U} = \{V_{d_{i,j},S_{\mu_j}^U} : |V_{d_{i,j},S_{\mu_j}^U}| = \frac{M_2}{N} |V_{d_{i,j},S_{\mu_j}^U}| \text{ and } |V_{d_{i,j},S_{\mu_j}^U}| = 1 - \frac{M_2}{N} |V_{d_{i,j},S_{\mu_j}^U}| \}$$

where $V_{d_{i,j},S_{\mu_j}^U}^{H_i}$ is the part pre-stored in helper $H_i$ and $V_{d_{i,j},S_{\mu_j}^U}^{H_i}$ is the part not pre-stored in helper $H_i$. It is clear that $V_{d_{i,j},S_{\mu_j}^U}^{H_i}$ is pre-stored in neither of helper $H_i$ nor user $U_{i,j}$.

Then, $V_{d_{i,j},S_{\mu_j}^U}^{H_i}$ must be in $f_{d_{i,j},U_{i,j}}$, which has been obtained by helper $H_i$. Thus, helper $H_i$ transmits $\{\oplus_{j \in S_2} V_{d_{i,j},S_{\mu_j}^U} : S_2^U \subset [1 : K_2], |S_2^U| = s_2, s_2 \in [1 : K_2]\}$, all the attached users are satisfied. This leads to the transmission rate of each helper as (7).

### IV. HYBRID CACHING IN TWO-LAYER NETWORKS

In this section, we shall present a hybrid caching scheme to obtain more caching gains. For illustrations, we briefly introduce the caching scheme B in [15], which exploits the CMO between the server and users. Then, to exploit SMO, CMO, and CSC simultaneously, we combine the S&C caching scheme with the caching scheme B of [15] in a memory-sharing manner and develop the hybrid caching scheme.

#### A. Caching scheme B in [15]

1) **Basic principles and results:** It is observed that the MAU-Decentralized caching scheme is able to benefit from the SMO in a single-layer network. To exploit the CMO, the MAU-Decentralized caching scheme is directly applied between the server and users. Then, to exploit SMO, CMO, and CSC simultaneously, we combine the S&C caching scheme with the caching scheme B of [15] in a memory-sharing manner and develop the hybrid caching scheme.
In the considered two-layer caching network, we assume the normalized memory sizes at each helper and each user as $M_1$ and $M_2$, respectively. Then, the achievable transmission rates of the server and each helper with the hybrid caching scheme are

$$R_1^H(\alpha, \beta) = \alpha K_2 \left( 1 - \frac{M_1}{\alpha N} \right) \left( 1 - \frac{\beta M_2}{\alpha N} \right) \frac{\alpha N}{M_1} \times \left( 1 - (1 - \frac{M_1}{\alpha N})^{K_1} \right) + (1 - \alpha) \left( 1 - \frac{1 - \beta M_2}{(1 - \alpha) N} \right) \frac{(1 - \alpha)N}{(1 - \beta)M_2} \times \left( 1 - (1 - \frac{1 - \beta M_2}{(1 - \alpha)N})^{K_1 K_2} \right)$$

and

$$R_2^H(\alpha, \beta) = \alpha \left( 1 - \frac{\beta M_2}{\alpha N} \right) \frac{\alpha N}{\beta M_2} \left( 1 - (1 - \frac{\beta M_2}{\alpha N})^{K_2} \right) + (1 - \alpha) \left( 1 - \frac{1 - \beta M_2}{(1 - \alpha) N} \right) \frac{(1 - \alpha)N}{(1 - \beta)M_2} \times \left( 1 - (1 - \frac{1 - \beta M_2}{(1 - \alpha)N})^{K_2} \right)$$

respectively.

**Remark 2 (Performance comparison):** Clearly, the hybrid caching scheme is composed of the S&G caching scheme and the caching scheme B of [15] in a memory-sharing manner with factor $\alpha$ and $\beta$. The proof is similar to that in [15] and will be omitted for space limitation.

**Lemma 3:** In the considered two-layer caching network, we assume the normalized memory sizes at each helper and each user as $M_1$ and $M_2$, respectively. Then, the achievable transmission rates of the server and each helper with the hybrid caching scheme are

$$R_1^H(\alpha, \beta) = \alpha K_2 \left( 1 - \frac{M_1}{\alpha N} \right) \left( 1 - \frac{\beta M_2}{\alpha N} \right) \frac{\alpha N}{M_1} \times \left( 1 - (1 - \frac{M_1}{\alpha N})^{K_1} \right) + (1 - \alpha) \left( 1 - \frac{1 - \beta M_2}{(1 - \alpha) N} \right) \frac{(1 - \alpha)N}{(1 - \beta)M_2} \times \left( 1 - (1 - \frac{1 - \beta M_2}{(1 - \alpha)N})^{K_1 K_2} \right)$$

and

$$R_2^H(\alpha, \beta) = \alpha \left( 1 - \frac{\beta M_2}{\alpha N} \right) \frac{\alpha N}{\beta M_2} \left( 1 - (1 - \frac{\beta M_2}{\alpha N})^{K_2} \right) + (1 - \alpha) \left( 1 - \frac{1 - \beta M_2}{(1 - \alpha) N} \right) \frac{(1 - \alpha)N}{(1 - \beta)M_2} \times \left( 1 - (1 - \frac{1 - \beta M_2}{(1 - \alpha)N})^{K_2} \right)$$

respectively.

**Proof:** $R_1^H(\alpha, \beta)$ and $R_2^H(\alpha, \beta)$ can be obtained by applying the S&G caching scheme and the caching scheme B of [15] in a memory-sharing manner with factor $\alpha$ and $\beta$. The proof is similar to that in [15] and will be omitted for space limitation. □

**Remark 2 (Performance comparison):** Clearly, the hybrid caching scheme is composed of the S&G caching scheme and the caching scheme B of [15] in a memory-sharing manner, which is similar to the generalized caching scheme in [15], where the generalized caching scheme is obtained by combining the caching scheme A and B in a memory-sharing manner. From Remark 1, with the S&G caching scheme, the achievable transmission rate in the first layer can be reduced without compromising the achievable transmission rate in the second layer compared with the caching scheme A in [15]. Thus, for any memory-sharing factors $\alpha$ and $\beta$, with the hybrid caching scheme, the achievable transmission rate in the first layer is smaller than that with the generalized caching scheme in [15]. Meanwhile, the achievable transmission rates in the second layer with both schemes are identical. In other words, the achievable rate region with the proposed hybrid caching scheme is better than that with the generalized caching scheme in [15].

For illustration, we compare the achievable transmission rates of the hybrid caching scheme and the generalized caching scheme in Fig. 3, where $N = 50$, $M_1 = 10$, $M_2 = 20$, $K_1 = 10$, and $K_2 = 2$. Since the proposed hybrid scheme is able to reduce the transmission rate in the first layer and achieves the same transmission rate in the second layer, we only compare the transmission rates in the first layer with two schemes. In Fig. 3(a), we set $\beta = 0.5$ and increase $\alpha$ from 0.2 to 0.9. In Fig. 3(b), we set $\alpha = 0.5$ and increase $\beta$ from 0.2 to 0.9. From this figure, given $\alpha$ ($\beta$), the transmission rates of the server can be significantly reduced by the proposed hybrid.
caching scheme, especially when $\beta$ ($\alpha$) is large. The result shows the advantages of the proposed hybrid caching scheme over the generalized caching scheme in [15].

Lemma 3 indicates that the transmission rates in (11) and (12) can be achieved simultaneously for given ($\alpha$, $\beta$). The performance limits of $R_1^H$ and $R_2^H$ are investigated in the following.

2) Performance limit: In what follows, we adopt the achievable rate region to measure the performance of the hybrid caching scheme, and characterize the gap between the achievable rate region of the hybrid caching scheme and its information-theoretic lower bound.

**Definition 2:** For normalized memory size $M_1, M_2 \geq 0$, we define

\[ \mathcal{R}^H(M_1, M_2) = \{(R_1^H(\alpha, \beta), R_2^H(\alpha, \beta)) : \alpha, \beta \in [0, 1]\} \]

as the achievable rate region of the hybrid caching scheme, where $\mathbb{R}^2_+$ denotes the positive quadrant and the addition corresponds to the Minkowski sum between sets.

Besides, let $\mathcal{R}(M_1, M_2)$ represent the feasible rate region as in Definition 1. The achievable rate region of the proposed hybrid caching scheme can be shown to be order optimal as shown in the following theorem.

**Theorem 1:** In the considered two-layer caching network, we assume the normalized memory sizes at each helper and each user as $M_1$ and $M_2$, respectively. Then, we have

\[ \mathcal{R}^H(M_1, M_2) \subset \mathcal{R}(M_1, M_2) \subset \frac{1}{48} \mathcal{R}^H(M_1, M_2) - 4. \]

**Sketches of proof:** Since it is straightforward to obtain $\mathcal{R}^H(M_1, M_2) \subset \mathcal{R}(M_1, M_2)$, we only prove $\mathcal{R}(M_1, M_2) \subset \frac{1}{48} \mathcal{R}^H(M_1, M_2) - 4$.

Firstly, we let $R_1^{lb}(M_1, M_2)$ and $R_2^{lb}(M_2)$ represent the information-theoretic lower bounds of achievable rates $R_1$ and $R_2$, respectively. Recalling that by cut-set theorem [10], they can be written as [15]:

\[ R_1^{lb}(M_1, M_2) = \max_{s_1 \in \{1, \ldots, K_1\}} s_1 s_2 (N - s_1 M_1 - s_1 s_2 M_2) \]

and

\[ R_2^{lb}(M_2) = \max_{t \in \{1, \ldots, K_2\}} t(N - t M_2) \]

Secondly, we carefully choose ($\alpha^*, \beta^*$) and obtain the upper bounds $R_1^{ub}(M_1, M_2)$ and $R_2^{ub}(M_2)$ of the achievable transmission rates $R_1^H(\alpha^*, \beta^*)$ and $R_2^H(\alpha^*, \beta^*)$ in Appendix A. In particular, we divide the feasible region of $(M_1, M_2)$ into Region I and Region II as shown in Fig. 4. Region I denotes small memory aggregation when the total memory in the sub-network shown in Fig. 1 is smaller than the file number, i.e., $M_1 + K_2 M_2 < N$, and Region II denotes large memory aggregation, i.e., $M_1 + K_2 M_2 \geq N$. Then, we select several pairs of ($\alpha$, $\beta$) with typical values in each region and evaluate the corresponding achievable rates. We note that the constant multiplicative and additive factors are highly related to the gap between the upper bounds $R_1^{ub}(M_1, M_2)$ and information-theoretic lower bound $R_1^{lb}(M_1, M_2)$. Thus, we choose the ($\alpha^*, \beta^*$) with the minimum $R_1^H(\alpha, \beta)$ in each region. In this way, we obtain a unique pair of ($\alpha^*, \beta^*$), and the corresponding upper bounds $R_1^{ub}(M_1, M_2)$ and $R_2^{ub}(M_2)$ in each region.

Thirdly, we prove that $R_1^{ub}(M_1, M_2) \geq \frac{1}{48} R_1^{lb}(M_1, M_2) - 4$ and $R_2^{ub}(M_2) \geq \frac{1}{48} R_2^{lb}(M_2) - 4$ can be satisfied simultaneously in Appendix B and Appendix C, respectively. More specifically, we divide Region I and Region II into two sub-regions based on the value of $M_1$, i.e., whether $M_1$ is larger or smaller than $N/2$. Therefore, we have four sub-regions in total to investigate, i.e., Sub-region I-(1): $M_1 + K_2 M_2 < N$ and $M_1 < N/2$, Sub-region I-(2): $M_1 + K_2 M_2 < N$ and $M_1 \geq N/2$, Sub-region II-(1): $M_1 + K_2 M_2 \geq N$ and $M_1 < N/2$, and Sub-region II-(2): $M_1 + K_2 M_2 \geq N$ and $M_1 \geq N/2$. After characterizing the gap between $R_1^{lb}(M_1, M_2)$ and $R_1^{ub}(M_1, M_2)$, and the gap between $R_2^{lb}(M_2)$ and $R_2^{ub}(M_2)$ in the four regions, we summarize that $R_1^{lb}(M_1, M_2) \geq \frac{1}{48} R_1^{lb}(M_1, M_2) - 4.$
Remark 3 (Performance comparison): Similar to the definition in [14], we define \( R^G(M_1, M_2) \) as the achievable rate region of the generalized caching scheme in [15]. Thus, we have

\[
R^G(M_1, M_2) \subset \mathcal{R}(M_1, M_2) \subset \frac{1}{66} R^G(M_1, M_2) - 16 \quad (17)
\]

from the results in [15]. By comparing the multiplicative and additive factors in (14) and (17), the gap of achievable rates and their lower bounds has been decreased in the hybrid caching scheme. The improvement reason of the multiplicative and additive factors lies in two folds. The first one comes from the gain of exploiting the correlations of the pre-stored contents in two layers, which is able to reduce the transmission rate in the first layer without compromising the transmission rate in the second layer compared with the generalized caching scheme in [15]. The second one is the optimization of information-theoretic bounds compared with those in [15].

**Remark 4 (Optimization of information-theoretic bounds):** Comparing with the proof of the order-optimality in [15], there are two main aspects of optimization, both of which are based on an observation that the constant multiplicative and additive factors, i.e., the gap between the achievable rate region and the feasible rate region, is dominated by the gap between the upper bound \( R^{ub}(M_1, M_2) \) and the lower bound \( R^{lb}(M_1, M_2) \). In particular, the first optimization is the choice of \((\alpha, \beta)\). In [15], a pair of \((\alpha, \beta)\) is firmly chosen in each region to balance the transmission rates in two layers and minimize the gap between the achievable rate region \( R^G \) of the generalized caching scheme and the feasible rate region \( \mathcal{R} \). In our proof, we prepare several pairs of \((\alpha, \beta)\) and choose the one corresponding to the minimum \( R^{ub}(M_1, M_2) \). In this way, we reduce the gap between \( R^{ub}(M_1, M_2) \) and \( R^{lb}(M_1, M_2) \). The second optimization is the division of \((M_1, M_2)\) subregions. We observe that the gap between the upper bound \( R^{ub}(M_1, M_2) \) and the lower bound \( R^{lb}(M_1, M_2) \) is dominated by the characterization of the multiplicative and additive factors in Region I, i.e., \( M_1 + K_2 M_2 < N \). By further carefully dividing Region I into two subregions, we quantify the gap between the upper bound \( R^{ub}(M_1, M_2) \) and the lower bound \( R^{lb}(M_1, M_2) \) with better multiplicative and additive factors in Region I compared with results in [15]. Thus, we have four subregions instead of three subregions in [15]. Specifically, Region I and Region II are each divided into two subregions as shown in Fig. 4.

V. CONCLUSIONS

We studied the decentralized caching in two-layer networks, where users request contents through intermediate nodes (helpers) from a file server. In particular, we first exploited the correlations of the pre-stored contents in two layers, namely, cross-layer storage correlations (CSC), as well as the single-layer multicast opportunities (SMO) from the file server (each helper) to helpers (the attached users), and developed an S&C caching scheme to reduce the transmission rates in the two-layer network. Then, by combining CSC, SMO, and cross-layer multicast opportunities (CMO) from the server to users in a proper manner, we proposed a hybrid caching scheme. With the hybrid caching scheme, the transmission rate in the first layer is reduced without compromising the transmission rate in the second layer compared with the generalized caching scheme in [15]. Therefore, the achievable rate region of the proposed hybrid caching scheme is better than the generalized caching scheme in [15]. Furthermore, we demonstrate that the achievable rate region of the proposed hybrid caching scheme is order-optimal with constant multiplicative and additive factors \( \frac{1}{16} \) and \( 4 \), respectively.

**APPENDIX A**

**Upper Bounds \( R_{1}^{ub}(M_1, M_2) \) and \( R_{2}^{ub}(M_2) \)**

Since the achievable rates \( (R_1^{H}(\alpha, \beta), R_2^{H}(\alpha, \beta)) \) are highly related to the values of \((\alpha, \beta)\), the upper bounds \( R_{1}^{ub}(M_1, M_2) \) and \( R_{2}^{ub}(M_2) \) are also determined by \((\alpha, \beta)\). Meanwhile, achievable rates differ in variable \((M_1, M_2)\) regimes. Thus, we first consider two regimes, i.e., Regime I) \( M_1 + K_2 M_2 \leq N \) and Regime II) \( M_1 + K_2 M_2 > N \). In what follows, we will discuss the upper bounds \( R_{1}^{ub}(M_1, M_2) \) and \( R_{2}^{ub}(M_2) \) in the two regimes, respectively.

**A. Upper Bounds \( R_{1}^{ub}(M_1, M_2) \) and \( R_{2}^{ub}(M_2) \) in Regime I**

In this regime, we consider the tuples of \((\alpha, \beta)\) as follows,

\[
(\alpha, \beta) = \begin{cases} 
\left( \frac{M_1}{N}, \frac{M_1}{N} \right), & \text{Tuple I,} \\
\left( \frac{M_1}{M_1 + K_2 M_2}, 0 \right), & \text{Tuple II,} \\
(1, 1), & \text{Tuple III,}
\end{cases}
\]

(18)

Firstly, we substitute \((\alpha, \beta) = (\frac{M_1}{N}, \frac{M_1}{N})\) into (11) and (12). Then, we have

\[
R_1^{H}(\alpha, \beta) = \left( 1 - \frac{M_1}{N} \right) K_1 K_2 \left( 1 - \frac{M_2}{N} \right) \frac{N}{K_1 K_2 M_2} \times \left( 1 - (1 - \frac{M_2}{N}) K_1 K_2 \right) \leq \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right) \leq \min \{ K_1 K_2, \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right) \} 
\]

(19)

and

\[
R_2^{H}(\alpha, \beta) = K_2 \left( 1 - \frac{M_2}{N} \right) \frac{N}{K_2 M_2} \left( 1 - (1 - \frac{M_2}{N}) K_2 \right) \leq \min \{ K_2, \frac{N}{M_2} \} 
\]

(20)

Substitute \((\alpha, \beta) = (\frac{M_1}{M_1 + K_2 M_2}, 0)\) into (11) and (12), we
have
\[
R_1^H(\alpha, \beta) \leq \frac{M_1K_2}{M_1 + K_2M_2} \min\{K_1, \frac{N}{M_1 + M_2K_2}\} \\
+ \frac{K_2M_2}{M_1 + K_2M_2} \min\{K_1, \frac{N}{M_1 + M_2K_2}\} \\
= \frac{M_1 + K_2M_2}{M_1 + K_2M_2} \min\{K_1, \frac{N}{M_1 + M_2K_2}\} \\
+ \frac{M_1 + K_2M_2}{M_1 + K_2M_2} \min\{K_1, \frac{N}{M_1 + M_2K_2}\} \\
\leq \min\{K_1, \frac{N}{M_1 + M_2K_2}\}
\]
(21)

and
\[
R_2^H(\alpha, \beta) \leq K_2 \left( \frac{N}{M_2} \right)
\]
(22)

where \((a)\) follows from \(K_2M_2 < N\) in regime I.

Substitute \((\alpha, \beta) = (1, 1)\) into (11) and (12). Then, we have
\[
R_1^H(\alpha, \beta) = K_1K_2 \left( 1 - \frac{M_1}{N} \right) \left( 1 - \frac{M_2}{N} \right) \frac{N}{K_1M_1} \\
\times \left( 1 - (1 - \frac{M_1}{N})K_1 \right) \\
= \frac{K_2N}{M_1} \left( 1 - \frac{M_1}{N} \right)
\]
(23)

and
\[
R_2^H(\alpha, \beta) \leq K_2 \left( \frac{N}{M_2} \right)
\]
(24)

where \((a)\) follows from \(K_2M_2 < N\) in regime I.

Thus, if we choose \((\alpha^*, \beta^*)\) in (18) corresponding to the minimum \(R_1^H(\alpha, \beta)\), we can achieve the upper bounds \(R_1^{ub}(M_1, M_2)\) and \(R_2^{ub}(M_2)\) in Regime I as
\[
R_1^{ub}(M_1, M_2) \leq \min\left\{K_1K_2 \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right) \frac{NK_2}{M_1 + M_2K_2} \frac{K_2N}{M_1} \left( 1 - \frac{M_1}{N} \right) \right\}.
\]
(25)

and
\[
R_2^{ub}(M_2) \leq \min\left\{K_2 \frac{N}{M_2} \right\}.
\]
(26)

**B. Upper Bounds \(R_1^{ub}(M_1, M_2)\) and \(R_2^{ub}(M_2)\) in Regime II**

In this regime, we consider the tuples of \((\alpha, \beta)\) as follows,
\[
(\alpha, \beta) = \left\{ \left( \frac{M_1}{N}, \frac{M_2}{N} \right), \left( \frac{M_1}{N}, \frac{M_2}{2} \right) \right\}, \text{Tuple I,}
\]
(27)

and
\[
(\alpha, \beta) = \left\{ \left( \frac{M_1}{N}, \frac{M_2}{2} \right) \right\}, \text{Tuple II.}
\]

Since we have obtained the upper bounds with Tuple I in (19) and (20), we will only calculate the upper bounds with Tuple II in the following.

Substitute \((\alpha, \beta) = \left( \frac{M_1}{N}, \frac{M_2}{2} \right)\) into (11) and (12), we have
\[
R_1^H(\alpha, \beta) = \left( 1 - \frac{M_1}{N} \right) \left( 1 - \frac{M_2}{2(N - M_1)} \right) \\
\times \frac{2(N - M_1)}{M_2} \left( 1 - \left( 1 - \frac{M_2}{2(N - M_1)} \right)^{K_2} \right) \\
\leq \left( 1 - \frac{M_1}{N} \right) \min\{K_2, \frac{2(N - M_1)}{M_2}\}
\]
(28)

and
\[
R_2^H(\alpha, \beta) = \frac{M_1}{N} \left( 1 - \frac{M_2}{2M_1} \right) \frac{2M_1}{M_2} \left( 1 - \left( 1 - \frac{M_2}{2M_1} \right)^{K_2} \right) \\
+ \frac{1 - \frac{M_1}{N}}{1 - \frac{M_1}{N}} \frac{M_2}{2(N - M_1)} \frac{2(N - M_1)}{M_2} \\
\times \left( 1 - \left( 1 - \frac{M_2}{2(N - M_1)} \right)^{K_2} \right) \\
\leq \frac{M_1}{N} \min\{K_2, \frac{2M_1}{M_2}\} \\
+ \left( 1 - \frac{M_1}{N} \right) \min\left\{K_2, \frac{2(N - M_1)}{M_2}\right\}
\]
(29)

Thus, if we choose \((\alpha^*, \beta^*)\) in (27) corresponding to the minimum \(R_1^H(\alpha, \beta)\), we can achieve the upper bounds \(R_1^{ub}(M_1, M_2)\) and \(R_2^{ub}(M_2)\) in Regime II as
\[
R_1^{ub}(M_1, M_2) \leq \min\left\{K_2 \frac{N}{M_2} \right\}.
\]
(30)

and
\[
R_2^{ub}(M_2) \leq 2 \min\left\{K_2 \frac{N}{M_2} \right\},
\]
(31)

respectively.

**APPENDIX B**

**GAP BETWEEN THE UPPER BOUND \(R_1^{ub}(M_1, M_2)\) AND LOWER BOUND \(R_1^{lb}(M_1, M_2)\)**

In this section, we will characterize the gap between the upper bound \(R_1^{ub}(M_1, M_2)\) and lower bound \(R_1^{lb}(M_1, M_2)\). Recall that we consider \(K_1 \geq 2, K_2 \geq 2, \) and \(N \geq K_1K_2\).

**A. Gap between \(R_1^{ub}(M_1, M_2)\) and \(R_1^{lb}(M_1, M_2)\) in Regime I**

In regime I, we consider two subregimes, i.e., Subregime I) \(0 \leq M_1 \leq \frac{N}{2}\) and Subregime II) \(\frac{N}{2} \leq M_1 \leq N\). Then, we will discuss the gap in the two subregimes respectively.
1) Gap in Subregime I: In this subregime, we have $M_1 + K_2 M_2 \leq N$ and $0 \leq M_1 \leq \frac{N}{2}$. Then, we have $0 \leq M_2 \leq \frac{N - M_1}{K_2}$ using $K_2 \geq 2$. Thus, we consider

A) $0 \leq M_1 < \frac{N}{2}$ and $0 \leq M_2 \leq \frac{N - M_1}{K_2}$;

B) $0 \leq M_1 < \frac{N}{2}$ and $\frac{N - M_1}{K_2} \leq M_2 \leq \frac{N}{2}$;

C) $0 \leq M_1 < \frac{N}{2}$ and $\frac{N - M_1}{K_2} \leq M_2 \leq \frac{N}{2}$;

D) $\frac{N}{2} < M_1 < \frac{N}{K_2}$ and $0 \leq M_2 \leq \frac{N}{K_2}$;

E) $\frac{N}{2} < M_1 < \frac{N}{K_2}$ and $\frac{N}{2} < M_2 \leq \frac{N}{K_2}$;

F) $\frac{N}{K_2} < M_1 < \frac{N}{2}$ and $0 \leq M_2 \leq \frac{N}{2}$;

G) $\frac{N}{K_2} < M_1 < \frac{N}{2}$ and $\frac{N}{2} < M_2 \leq \frac{N}{2}$.

• A) $0 \leq M_1 < \frac{N}{2K_1}$ and $0 \leq M_2 \leq \frac{N}{K_2}$; we choose $s_1 = \frac{K_1 M_1}{K_2}$ and $s_2 = K_2$ in the lower bound \[15\]. This is a valid choice since $K_1 \geq 2$, and thus $s_1 = \frac{K_1 M_1}{K_2} \geq 1$. Then, we have

$$R^b_1(M_1, M_2) \geq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$

• B) $0 \leq M_1 < \frac{N}{2K_1}$ and $\frac{N}{K_2} < M_2 \leq \frac{N}{2}$; we choose $s_1 = \frac{K_1 M_1}{K_2}$ and $s_2 = K_2$ in the lower bound \[15\]. Note that this is a valid choice since

$$1 - \left( \frac{N}{2M_2 K_2} \right) \leq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$

B) $0 \leq M_1 < \frac{N}{2K_1}$ and $\frac{N}{K_2} < M_2 \leq \frac{N}{2}$; we choose $s_1 = \frac{K_1 M_1}{K_2}$ and $s_2 = K_2$ in the lower bound \[15\]. Note that this is a valid choice since

$$1 - \left( \frac{N}{2M_2 K_2} \right) \leq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$

where $(a)$ follows from $x \geq \left\lfloor \frac{x}{2} \right\rfloor \geq \frac{x}{2}$ for any $x \geq 1$ and $(b)$ follows from $\frac{K_1 M_1}{2K_2} \leq \frac{N}{2}$ by using $M_1 \leq \frac{N}{K_1 K_2 M_2} \leq \frac{N}{2}$ by using $N \geq K_1 K_2$. Combining with \[25\], we have

$$R^b_1(M_1, M_2) \geq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$

where $(a)$ follows from $x \geq \left\lfloor \frac{x}{2} \right\rfloor \geq \frac{x}{2}$ for any $x \geq 1$ and $(b)$ follows from $\frac{K_1 M_1}{2K_2} \leq \frac{N}{2}$ by using $M_1 \leq \frac{N}{K_1 K_2 M_2} \leq \frac{N}{2}$ by using $N \geq K_1 K_2$. Combining with \[25\], we have

$$R^b_1(M_1, M_2) \geq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$

where $(a)$ follows from $x \geq \left\lfloor \frac{x}{2} \right\rfloor \geq \frac{x}{2}$ for any $x \geq 1$ and $(b)$ follows from $\frac{K_1 M_1}{2K_2} \leq \frac{N}{2}$ by using $M_1 \leq \frac{N}{K_1 K_2 M_2} \leq \frac{N}{2}$ by using $N \geq K_1 K_2$. Combining with \[25\], we have

$$R^b_1(M_1, M_2) \geq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$

where $(a)$ follows from $x \geq \left\lfloor \frac{x}{2} \right\rfloor \geq \frac{x}{2}$ for any $x \geq 1$ and $(b)$ follows from $\frac{K_1 M_1}{2K_2} \leq \frac{N}{2}$ by using $M_1 \leq \frac{N}{K_1 K_2 M_2} \leq \frac{N}{2}$ by using $N \geq K_1 K_2$. Combining with \[25\], we have

$$R^b_1(M_1, M_2) \geq \frac{K_1 K_2}{24} \left( N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor \right) \frac{K_2 (N - \left\lfloor \frac{K_1 K_2}{2} \right\rfloor)}{M_1 + M_2 K_2}.$$
\[ \geq \frac{1}{24} R_1^b(M_1, M_2). \]  

(39)

- D) \( \frac{N}{2K_1} \leq M_1 < \frac{N}{4} \) and \( 0 \leq M_2 \leq \frac{N}{4K_2} \); we choose \( s_1 = \frac{N}{2(M_1 + M_2K_2)} \) and \( s_2 = K_2 \) in the lower bound (15).

This is a valid choice since

\[
1 = \left\lfloor \frac{N}{2(N/4 + N/4)} \right\rfloor \leq \left\lfloor \frac{N}{2(M_1 + M_2K_2)} \right\rfloor \leq \frac{N}{2M_1} \leq K_1, \quad (40)
\]

where (a) follows from \( M_1 \leq \frac{N}{4} \) and \( M_2 \leq \frac{N}{4K_2} \), (b) follows from \( x \geq |x| \) for any \( x \geq 1 \), and (c) follows from \( \frac{N}{2K_1} \leq M_1 \).

Then, we have

\[
R_1^b(M_1, M_2) \geq \frac{N}{2(M_1 + M_2K_2)} K_2 \left( N - \frac{N}{2(M_1 + M_2K_2)} (M_1 + M_2K_2) \right) \left( \frac{N}{2M_1} \right) \leq \frac{NK_2}{16(M_1 + M_2K_2)},
\]

(41)

where (a) follows from \( x \geq |x| \geq \frac{x}{2} \) for any \( x \geq 1 \), (b) follows from \( \frac{N}{2(M_1 + M_2K_2)} \leq \frac{N}{2K_1} \), and (c) follows from \( \frac{N}{2K_1} \leq K_2 \leq N \) by using \( \frac{N}{2K_1} \leq M_1 \). Combining with (25), we have

\[
R_1^b(M_1, M_2) \geq \frac{NK_2}{16(M_1 + M_2K_2)} \geq \frac{1}{16} \min \left\{ K_1K_2, \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right), \frac{NK_2}{M_1 + M_2K_2}, \frac{K_2N}{M_1} \left( 1 - \frac{M_1}{N} \right) \right\} \geq \frac{1}{16} R_1^b(M_1, M_2). \quad (42)
\]

- E) \( \frac{N}{2K_1} \leq M_1 < \frac{N}{4} \) and \( \frac{N}{4K_2} \leq M_2 \leq \frac{N}{2K_1} \). Let

\[
(s_1, s_2) = \begin{cases} \left( \frac{N}{4M_1}, \frac{M_1}{M_2} \right), & \text{if } M_1 \geq M_2, \\ \left( \frac{N}{4M_2}, 1 \right), & \text{otherwise}. \end{cases} \quad (43)
\]

in the lower bound (15). This is a valid choice since for \( M_1 \geq M_2 \), we have

\[
1 = \left\lfloor \frac{N}{4 \cdot N/4} \right\rfloor \leq \left\lfloor \frac{N}{4M_1} \right\rfloor \leq \frac{N}{4M_1} \leq \frac{K_1}{2}, \quad (44)
\]

and

\[
1 = \left\lfloor \frac{M_1}{M_2} \right\rfloor \leq \frac{M_1}{M_2} \leq \frac{N/4}{N/(4K_2)} = K_2, \quad (45)
\]

where (a) follows from \( M_1 < \frac{N}{4} \), (b) follows from \( x \geq |x| \geq \frac{x}{2} \) for any \( x \geq 1 \), (c) follows from \( \frac{N}{2K_1} \leq M_1 \), and (d) follows from \( \frac{N}{4M_1} \leq M_2 \) and \( M_2 \geq \frac{N}{4K_2} \).

For \( M_1 < M_2 \), we have

\[
1 = \left\lfloor \frac{N}{4 \cdot N/4} \right\rfloor \leq \left\lfloor \frac{N}{4M_2} \right\rfloor \leq \frac{N}{4M_2} \leq \frac{N}{2K_1} \leq K_1. \quad (46)
\]

Note that \( s_1 \leq \frac{N}{16M_2} \) and \( s_2 \leq \frac{N}{4M_2} \) due to \( x \geq \frac{x}{2} \). Then, we have

\[
R_1^b(M_1, M_2) \geq \frac{N}{48M_2} \left( 1 - \frac{M_2}{N} \right), \quad (47)
\]

where (a) follows from \( \frac{N}{4M_2} \leq K_2 \leq \frac{N}{16M_2} \), using \( \frac{N}{16M_2} \leq M_2 \), \( N > K_1K_2 \), and \( K_1 \geq 2 \). Combined with (25), we have

\[
R_1^b(M_1, M_2) \geq \frac{N}{48M_2} \left( 1 - \frac{M_2}{N} \right) \geq \frac{1}{48} \min \left\{ K_1K_2, \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right), \frac{NK_2}{M_1 + M_2K_2}, \frac{K_2N}{M_1} \left( 1 - \frac{M_1}{N} \right) \right\} \geq \frac{1}{48} R_1^b(M_1, M_2). \quad (48)
\]

- F) \( \frac{N}{2K_1} \leq M_1 < \frac{N}{4} \) and \( 0 \leq M_2 \leq \frac{N}{4K_2} \); we choose \( s_1 = 1 \) and \( s_2 = K_2 \) in the lower bound (15). Then, we have

\[
R_1^b(M_1, M_2) \geq \frac{K_2(N - M_1 - M_2K_2)}{N + K_2} \geq \frac{1}{3N} \min \left\{ K_1K_2, \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right), \frac{NK_2}{M_1 + M_2K_2}, \frac{K_2N}{M_1} \left( 1 - \frac{M_1}{N} \right) \right\} \geq \frac{1}{12} R_1^b(M_1, M_2). \quad (49)
\]

where (a) follows from \( M_1 < \frac{N}{4} \), (b) follows from \( x \geq |x| \geq \frac{x}{2} \) for any \( x \geq 1 \) and \( K_2 \leq \frac{N}{16M_2} \), using \( N \geq K_1K_2 \) and \( K_1 \geq 2 \), (b) follows from \( \frac{N}{2K_1} \leq M_1 \) and \( M_2 \leq \frac{N}{4K_2} \). Combining with (25), we have

\[
R_1^b(M_1, M_2) \geq \frac{K_2N}{12M_1} \left( 1 - \frac{M_1}{N} \right) \geq \frac{1}{12} \min \left\{ K_1K_2, \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right), \frac{NK_2}{M_1 + M_2K_2}, \frac{K_2N}{M_1} \left( 1 - \frac{M_1}{N} \right) \right\} \geq \frac{1}{12} R_1^b(M_1, M_2). \quad (50)
\]
\[ R^b_1(M_1, M_2) \geq \frac{1}{12} R^b_1(M_1, M_2). \]  

\( \bullet \) G) \( \frac{N}{2} \leq M_1 < \frac{N}{2} \) and \( \frac{N-M_1}{2M_2} \leq M_2 \leq \frac{N-M_1}{K_2} \): we choose \( s_1 = 1 \) and \( s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \) in the lower bound (15). This is a valid choice since

\[ 1 \leq \frac{N - M_1}{K_2 M_2} \leq \frac{1}{2M_2} \leq \frac{N - M_1}{2M_2} \leq K_2, \quad (51) \]

where (a) follows from \( M_1 + K_2 M_2 \leq N \), (b) follows from \( K_2 \geq 2 \), (c) follows from \( x \geq \lfloor x \rfloor \) for any \( x \geq 1 \), and (d) follows from \( \frac{N-M_1}{2M_2} \leq M_2 \).

Then, we have

\[ R^b_1(M_1, M_2) \geq \frac{N - M_1}{N} + \frac{\left\lfloor \frac{N-M_1}{2M_2} \right\rfloor M_2}{N} \geq (N - M_1)^2 \frac{N - M_1}{12N M_2} = K_2(N - M_1) \frac{N - M_1}{12M_1}, \quad M_1 \leq M_2 K_2, \quad M_1 \leq N \]

\[ \text{Substituting } s_1 = 1 \text{ and } s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \text{ in the lower bound (15). This is a valid choice since} \]

\[ 1 \leq \frac{N - M_1}{K_2 M_2} \leq \frac{1}{2M_2} \leq \frac{N - M_1}{2M_2} \leq K_2, \quad (56) \]

\[ \text{Substituting } s_1 = 1 \text{ and } s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \text{ in the lower bound (15), we obtain} \]

\[ R^b_1(M_1, M_2) \geq \frac{K_2 N}{48 M_1} \left( 1 - \frac{M_1}{N} \right) \]

\[ \geq \frac{1}{48} \frac{12 M_2 N}{K_2 M_2} \left( K_2 N \frac{N - M_1}{M_1} \left( 1 - \frac{M_1}{N} \right) \right) \]

\[ \geq \frac{1}{48} R^b_1(M_1, M_2). \]  

\[ 2) \text{Gap between } R^b_1(M_1, M_2) \text{ and } R^b_1(M_1, M_2) \text{ in Subregime II: In this subregime, i.e., } M_1 + K_2 M_2 \leq N \text{ and } \frac{N}{2} \leq M_1 \leq N, \text{ we have } M_2 \leq \frac{N - M_1}{K_2}. \]

\( \bullet \) A) \( \frac{N}{2} \leq M_1 < N \) and \( 0 \leq M_2 \leq \frac{N - M_1}{K_2} \)

\( \bullet \) B) \( \frac{N}{2} \leq M_1 < N \) and \( \frac{N - M_1}{2M_2} \leq M_2 \leq \frac{N - M_1}{K_2} \)

• A) \( \frac{N}{2} \leq M_1 < N \) and \( 0 \leq M_2 \leq \frac{N - M_1}{K_2} \): we choose \( s_1 = 1 \) and \( s_2 = K_2 \) in the lower bound (15). Then, we have

\[ R^b_1(M_1, M_2) \geq \frac{K_2(N - M_1 - K_2 M_2)}{N + K_2} \geq \frac{K_2(N - M_1 - \frac{N-M_1}{K_2})}{N + \frac{N}{2}} = \frac{K_2(N - M_1)}{3N} \]

\[ = \frac{K_2 N}{3N} \left( 1 - \frac{M_1}{N} \right) \]

\[ \geq \frac{1}{24} R^b_1(M_1, M_2). \]  

where (a) follows from \( M_2 \leq \frac{N - M_1}{K_2} \) and \( K_2 = \frac{K_1 K_2}{K_1} \leq \frac{K_1 K_2}{K_1} \leq \frac{K_1 K_2}{K_1} \) using \( N > K_1 K_2 \) and \( K_1 \geq 2 \), and (b) follows from \( \frac{N-M_1}{2M_2} \leq M_2 \). Combining with (25), we have

\[ R^b_1(M_1, M_2) \geq \frac{K_2 N}{6M_1} \left( 1 - \frac{M_1}{N} \right) \]

\[ \geq \frac{1}{6} R^b_1(M_1, M_2). \]  

\[ \bullet \) B) \( \frac{N}{2} \leq M_1 < N \) and \( \frac{N - M_1}{2M_2} \leq M_2 \leq \frac{N - M_1}{K_2} \): we choose \( s_1 = 1 \) and \( s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \) in the lower bound (15). This is a valid choice since

\[ 1 \leq \frac{N - M_1}{K_2 M_2} \leq \frac{1}{2M_2} \leq \frac{N - M_1}{2M_2} \leq K_2, \quad (56) \]

where (a) follows from \( M_2 \leq \frac{N - M_1}{2M_2} \), (b) follows from \( K_2 \geq 2 \), (c) follows from \( x \geq \lfloor x \rfloor \) for any \( x \geq 1 \), and (d) follows from \( \frac{N-M_1}{2M_2} \leq M_2 \). Substituting \( s_1 = 1 \) and \( s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \) in the lower bound (15), we obtain

\[ R^b_1(M_1, M_2) \geq \frac{K_2(N - M_1 - \frac{N-M_1}{2M_2})}{N + \frac{N-M_1}{2M_2}} \geq \frac{K_2 N}{48 M_1} \left( 1 - \frac{M_1}{N} \right) \]

\[ \geq \frac{1}{48} \frac{12 M_2 N}{K_2 M_2} \left( K_2 N \frac{N - M_1}{M_1} \left( 1 - \frac{M_1}{N} \right) \right) \]

\[ \geq \frac{1}{48} R^b_1(M_1, M_2). \]  

where (a) follows from \( x \geq \lfloor x \rfloor \) for any \( x \geq 1 \), and \( \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \leq \frac{N-M_1}{2M_2} \leq K_2 \leq \frac{K_1 K_2}{K_1} \leq \frac{K_1 K_2}{K_1} \) using \( \frac{N-M_1}{2M_2} \leq M_2 \), \( N \geq K_1 K_2 \), and \( K_1 \geq 2 \), and (b) follows from \( \frac{N-M_1}{2M_2} \geq 1 \) and \( M_2 \geq \frac{N}{2} \). Combining with (25), we have

\[ R^b_1(M_1, M_2) \geq \frac{K_2 N}{24 M_1} \left( 1 - \frac{M_1}{N} \right) \]

\[ \geq \frac{1}{24} \frac{12 M_2 N}{K_2 M_2} \left( K_2 N \frac{N - M_1}{M_1} \left( 1 - \frac{M_1}{N} \right) \right) \]

\[ \geq \frac{1}{24} R^b_1(M_1, M_2). \]  

(58)
B. Gap between $R_1^{lb}(M_1, M_2)$ and $R_1^{ub}(M_1, M_2)$ in Regime II

In regime II, we consider two subregimes, i.e., Subregime I) $0 \leq M_1 \leq \frac{N}{2}$ and Subregime II) $\frac{N}{2} < M_1 \leq N$. Then, we will discuss the gap in the two subregimes, respectively.

1) Gap in Subregime I: In this subregime, i.e., $M_1 + K_2 M_2 \geq N$ and $0 \leq M_1 \leq \frac{N}{2}$, we have $M_2 \geq \frac{N - M_1}{K_2} \geq \frac{N}{2K_2} \geq \frac{N}{4K_2}$.

Then, we consider

A) $0 \leq M_1 < \frac{N}{2}$ and $\frac{N}{2} \leq M_2 \leq \frac{N}{2}$.

B) $0 \leq M_1 < \frac{N}{4}$ and $\frac{N}{4} \leq M_2 \leq N$.

C) $\frac{N}{4} \leq M_1 < \frac{N}{2}$ and $\frac{N}{2} \leq M_2 \leq \frac{N}{2}$.

D) $\frac{N}{2} \leq M_1 < \frac{N}{4}$ and $\frac{N}{4} \leq M_2 \leq \frac{N}{2}$.

E) $\frac{N}{4} \leq M_1 < \frac{N}{2}$ and $\frac{N}{2} \leq M_2 \leq N$.

F) $\frac{N}{2} \leq M_1 < \frac{N}{4}$ and $\frac{N}{4} \leq M_2 \leq N$.

• A) $0 \leq M_1 < \frac{N}{4}$ and $\frac{N}{4} \leq M_2 \leq \frac{N}{2}$: we choose $s_1 = 1$ and $s_2 = \left\lfloor \frac{N}{2M_2} \right\rfloor$ in the lower bound (15). This is a valid choice since

$$1 \leq \left\lfloor \frac{N}{2 \cdot N/2} \right\rfloor \leq \frac{N}{2} \leq \frac{N}{2M_2} \leq K_2,$$

where (a) follows from $M_2 \leq \frac{N}{2}$, (b) follows from $x \geq \lceil x \rceil$ for any $x \geq 1$, and (c) follows from $\frac{N}{4} \leq M_2 \leq \frac{N}{2}$.

Substitute $s_1$ and $s_2$ into the lower bound (15), we have

$$R_1^{lb}(M_1, M_2) \geq \frac{N}{2M_2} \left( N - M_1 - \left\lfloor \frac{N}{2M_2} \right\rfloor M_2 \right)$$

$$\geq \frac{N}{2M_2} \left( N - M_1 - \frac{N}{2M_2} \right)$$

$$= \frac{N}{2M_2} \left( 1 - \frac{M_2}{N} \right),$$

where (a) follows from $x \geq \lceil x \rceil$ for any $x \geq 1$, (b) follows from $\frac{N}{4} \leq M_2 \leq \frac{N}{2}$ and $K_1 \geq 2$, and (c) follows from $\frac{N}{4} \leq M_2 \leq \frac{N}{2}$ and $K_1 \geq 2$. Combining with (25), we have

$$R_1^{lb}(M_1, M_2) \geq \frac{N}{24M_2} \left( 1 - \frac{M_2}{N} \right).$$

Combining with (30), we have

$$R_1^{ub}(M_1, M_2) \geq \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right) - 1$$

$$\geq \min \left\{ K_1 K_2, \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right), \frac{2(N - M_1)^2}{NM_2} \right\} - 1$$

$$\geq R_1^{lb}(M_1, M_2) - 1.$$  \hspace{1cm} (63)

• B) $0 \leq M_1 < \frac{N}{2K_1}$ and $\frac{N}{2} \leq M_2 \leq N$: We have

$$R_1^{lb} \geq 0 \geq \frac{N}{2M_2} - 2 = \frac{N}{M_2} \left( 1 - \frac{M_2}{N} \right) - 1.$$
Combined with \((30)\), we have
\[
R_1^{ib}(M_1, M_2) \\
\geq \frac{N}{M_2} \left(1 - \frac{M_2}{N}\right) - 1 \\
\geq \min \left\{ K_1K_2, \frac{N}{M_2} \left(1 - \frac{M_2}{N}\right), \frac{2(N - M_1)^2}{NM_2} \right\} - 1 \\
\geq R_1^{ib}(M_1, M_2) - 1. \tag{71}
\]

- E) \(\frac{N}{2} \leq M_1 < \frac{N}{4}\) and \(\frac{N-M_1}{2M_2} \leq M_2 \leq \frac{N-M_1}{2}\): we choose \(s_1 = 1\) and \(s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor\) in the lower bound \((15)\). This is a valid choice since
\[
1 \leq \left\lfloor \frac{N - M_1}{2M_2} \right\rfloor \leq \frac{N - M_1}{2M_2} \leq \frac{K_2}{2}. \tag{72}
\]
where \((a)\) follows from \(\frac{N-M_1}{2M_2} \leq M_2\).

Substitute \(s_1\) and \(s_2\) into the lower bound \((15)\), we have
\[
R_1^{ib}(M_1, M_2) \geq \frac{\left\lfloor \frac{N-M_1}{2M_2} \right\rfloor}{N + \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor} \left(1 - \frac{M_2}{N}\right) \cdot \frac{2(N - M_1)^2}{NM_2}
\]
\[
= \frac{(N - M_1)^2}{10NM_2}. \tag{73}
\]
where \((a)\) follows from \(x \geq \left\lfloor x \right\rfloor \geq \frac{x}{2}\) for any \(x \geq 1\), and
\[
\left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \leq \left\lfloor \frac{K_2}{2} \right\rfloor \leq \frac{K_2}{2} \leq \frac{K_1K_2}{2K_2} \leq \frac{K_1}{2}\)
using \(\frac{N-M_1}{2M_2} \leq M_2\), \(K_1K_2 \leq N\), and \(K_1 \geq 2\). Combining with \((25)\), we have
\[
R_1^{ib}(M_1, M_2) \geq \frac{(N - M_1)^2}{10NM_2}
\]
\[
\geq \frac{1}{20} \min \left\{ K_1K_2, \frac{N}{M_2} \left(1 - \frac{M_2}{N}\right), \frac{2(N - M_1)^2}{NM_2} \right\}
\]
\[
\geq \frac{1}{20} R_1^{ib}(M_1, M_2). \tag{74}
\]

- F) \(\frac{N}{4} \leq M_1 < \frac{N}{2}\) and \(\frac{N-M_1}{2} \leq M_2 \leq N\): We have
\[
R_1^{ib}(M_1, M_2) \geq 0
\]
\[
= \frac{2(N - M_1)^2}{NM_2} \geq \frac{2(N - M_1)^2}{M_2} \cdot \frac{N}{M_2}
\]
\[
\geq \frac{2(N - M_1)^2}{NM_2} \geq \frac{2(N - M_1)^2}{NM_2} - 2 \cdot 2
\]
\[
= \frac{2(N - M_1)^2}{NM_2} - 2 \cdot 2 = \frac{2(N - M_1)^2}{NM_2} - 4. \tag{75}
\]
where \((a)\) follows from \(\frac{N-M_1}{2} \leq M_2\) and \(\frac{N-M_1}{N} \leq 1\). Combining with \((20)\), we have
\[
R_1^{ib}(M_1, M_2) \geq \frac{2(N - M_1)^2}{NM_2} - 4
\]
\[
\geq \min \left\{ K_1K_2, \frac{N}{M_2} \left(1 - \frac{M_2}{N}\right), \frac{2(N - M_1)^2}{NM_2} \right\} - 4
\]
\[
\geq R_1^{ib}(M_1, M_2) - 4. \tag{76}
\]

2) Gap in Subregime II: In this subregime, i.e., \(M_1 + K_2M_2 \geq N\) and \(\frac{N}{2} \leq M_1 \leq N\), we have \(M_2 \geq \frac{N-M_1}{K_2}\). Then, we consider

A) \(\frac{N}{4} \leq M_1 < \frac{N}{2}\) \(\frac{N-M_1}{2M_2} \leq M_2 \leq \frac{N-M_1}{2}\): we choose \(s_1 = 1\) and \(s_2 = \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor\) in the lower bound \((15)\). This is a valid choice since
\[
1 \leq \left\lfloor \frac{N - M_1}{2M_2} \right\rfloor \leq \frac{N - M_1}{2M_2} \leq \frac{K_2}{2}. \tag{77}
\]
where \((a)\) follows from \(x \geq \left\lfloor x \right\rfloor \geq \frac{x}{2}\) for any \(x \geq 1\), \((b)\) follows from \(\frac{N-M_1}{2M_2} \leq M_2\).

Then, we obtain
\[
R_1^{ib}(M_1, M_2) \geq \frac{\left\lfloor \frac{N-M_1}{2M_2} \right\rfloor}{N + \left\lfloor \frac{N-M_1}{2M_2} \right\rfloor} \left(1 - \frac{M_2}{N}\right) \cdot \frac{2(N - M_1)^2}{NM_2}
\]
\[
= \frac{(N - M_1)^2}{10NM_2}. \tag{78}
\]
where \((a)\) follows from \(x \geq \left\lfloor x \right\rfloor \geq \frac{x}{2}\) for any \(x \geq 1\) and
\[
\left\lfloor \frac{N-M_1}{2M_2} \right\rfloor \leq \left\lfloor \frac{K_2}{2} \right\rfloor \leq \frac{K_2}{2} \leq \frac{K_1K_2}{2K_2} \leq \frac{K_1}{2}\)
using \(\frac{N-M_1}{2M_2} \leq M_2\), \(K_1K_2 \leq N\), and \(K_1 \geq 2\). Combining with \((30)\), we have
\[
R_1^{ib}(M_1, M_2) \geq \frac{(N - M_1)^2}{10NM_2}
\]
\[
\geq \frac{1}{20} \min \left\{ K_1K_2, \frac{N}{M_2} \left(1 - \frac{M_2}{N}\right), \frac{2(N - M_1)^2}{NM_2} \right\}
\]
\[
\geq \frac{1}{20} R_1^{ib}(M_1, M_2). \tag{79}
\]

- B) \(\frac{N}{2} \leq M_1 < \frac{N}{4}\) \(\frac{N-M_1}{2} \leq M_2 \leq N\): We have
\[
R_1^{ib}(M_1, M_2) \geq 0
\]
\[
= \frac{2(N - M_1)^2}{NM_2} \geq \frac{2(N - M_1)^2}{M_2} \cdot \frac{N}{M_2}
\]
\[
\geq \frac{2(N - M_1)^2}{NM_2} \geq \frac{2(N - M_1)^2}{NM_2} - 2 \cdot 2
\]
\[
= \frac{2(N - M_1)^2}{NM_2} - 2 \cdot 2 = \frac{2(N - M_1)^2}{NM_2} - 4. \tag{80}
\]
where \((a)\) follows from \(\frac{N-M_1}{2} \leq M_2\) and \(\frac{N-M_1}{N} \leq 1\). Combining with \((30)\), we have
\[
R_1^{ib}(M_1, M_2) \geq \frac{2(N - M_1)^2}{NM_2} - 4
\]
\[
\geq \min \left\{ K_1K_2, \frac{N}{M_2} \left(1 - \frac{M_2}{N}\right), \frac{2(N - M_1)^2}{NM_2} \right\} - 4
\]
\[
\geq R_1^{ib}(M_1, M_2) - 4. \tag{81}
\]
Combining the results in Subsections A and B, we have that the upper bound $R_1^{ub}$ and the lower bound $R_1^{lb}$ are within a constant multiplicative and additive gap for all pairs of $M_1$ and $M_2$. More specifically, we have

$$R_1^{lb}(M_1, M_2) \geq \frac{1}{48} R_1^{ub}(M_1, M_2) - 4. \quad (82)$$

### Appendix C

**Gap between the upper bound $R_2^{ub}(M_2)$ and lower bound $R_2^{lb}(M_2)$**

In this section, we will characterize the gap between the upper bound $R_2^{ub}(M_2)$ and the lower bound $R_2^{lb}(M_2)$. We also consider $K_1 \geq 2$, $K_2 \geq 2$, and $N \geq K_1 K_2$. Recall that lower bound and upper bound of the achievable rate $R_2(\alpha^*, \beta^*)$ are

$$R_2^{lb}(M_2) \triangleq \max_{t \in [K_2]} \frac{t(N - t M_2)}{N + t} \quad (83)$$

and

$$R_2^{ub}(M_2) = 2 \min \left\{ K_2, \frac{N}{M_2} \right\}. \quad (84)$$

respectively.

To discuss the gap between the lower and the upper bound, we consider two regimes as follows,

A) $0 \leq M_2 < \frac{N}{2}$,

B) $\frac{N}{2} \leq M_2 \leq N$.

- A) $0 \leq M_2 < \frac{N}{2}$: we choose $t = \left\lfloor \frac{1}{2} \min \{ K_2, \frac{N}{M_2} \} \right\rfloor$ in the lower bound (83). This is a valid choice since

$$1 = \left\lfloor \frac{1}{2} \min \{ K_2, \frac{N}{M_2} \} \right\rfloor \leq \frac{K_2}{2}. \quad (85)$$

Then, we have

$$R_2^{lb}(M_2) \geq \frac{1}{2} \min \{ K_2, \frac{N}{M_2} \} \left( N - \frac{1}{2} \min \{ K_2, \frac{N}{M_2} \} \right) \geq \frac{1}{4} \min \{ K_2, \frac{N}{M_2} \} \frac{N}{4} = \frac{1}{10} \min \{ K_2, \frac{N}{M_2} \}, \quad (86)$$

where (a) follows from $x \geq |x| \geq \frac{x}{2}$ for any $x \geq 1$ and $\frac{1}{2} \min \{ K_2, \frac{N}{M_2} \} \leq \frac{K_2}{2} \leq \frac{K_2^2}{4} \leq \frac{N}{4}$ using $N \geq K_1 K_2$ and $K_1 \geq 2$. Combining with (84), we have

$$R_2^{lb}(M_2) \geq \frac{1}{10} \min \{ K_2, \frac{N}{M_2} \} \geq \frac{1}{20} \cdot 2 \min \{ K_2, \frac{N}{M_2} \} \geq \frac{1}{20} R_2^{ub}(M_2) \geq \frac{1}{20} R_2^{ub}(M_2) \quad (87)$$

- B) $\frac{N}{2} \leq M_2 < N$: we have

$$R_2^{lb}(M_2) \geq 0 = \frac{N}{M_2} - \frac{N}{M_2}. \quad (88)$$

Combining with (84), we have

$$R_2^{lb}(M_2) \geq 2 \frac{N}{M_2} - 2 \frac{N}{M_2} \geq 2 \min \{ K_2, \frac{N}{M_2} \} - 4 \geq R_2^{ub}(M_2) - 4. \quad (89)$$

Combining (87) and (89), we have that the upper bound $R_2^{ub}$ and the lower bound $R_2^{lb}$ are within a constant multiplicative and additive gap for all pairs of $M_1$ and $M_2$. More specifically, we have

$$R_2^{lb}(M_2) \geq \frac{1}{20} R_2^{ub}(M_2) - 4. \quad (90)$$

**References**

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” IEEE J. Select. Areas Commun., vol. 32, no. 6, pp. 1065-1082, Jun. 2014.

[2] L. W. Dowdy and D. V. Foster, “Comparative models of the file assignment problem,” ACM Computing Survey, vol. 14, pp. 287C313, 1982.

[3] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, “Placement algorithms for hierarchical cooperative caching,” Journal of Algorithms, vol. 38, pp. 260C302, 2001.

[4] J. Kangasharju, J. W. Roberts, and K. W. Ross, “Object replication strategies in content distribution networks,” Comp. Commun. Joumal, vol. 25, pp. 376C383, 2002.

[5] I. D. Baev, R. Rajaraman, and C. Swamy, “Approximating algorithms for data placement problems,” SIAM J. Computing, vol. 38, pp. 1411C1429, 2008.

[6] S. Borst, V. Gupta, and A. Walid, “Distributed caching algorithms for content distribution networks,” in Proceedings of IEEE INFOCOM, 2010.

[7] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Trans. Inform. Theory, vol. 60, no. 5, pp. 2856C2867, May 2014.

[8] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” IEEE/ACM Transactions on Networking (TON), vol. 23, pp. 1029C1040, Aug. 2015.

[9] S. Wang, W. Li, X. Tian, and H. Liu, “Coded caching with heterogeneous cache sizes,” 2015. [Online]. Available: [http://arxiv.org/pdf/1504.01123v3.pdf](http://arxiv.org/pdf/1504.01123v3.pdf).

[10] U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands,” in IEEE Conference on Computer Communication Workshops (INFOCOM WKSHPS), 2014.

[11] M. Ji, A. M. Tulino, J. Lloca, and G. Caire, “Order-optimal rate of caching and coded multicasting with random demands,” 2015. [Online]. Available: [arXiv:1502.03124](http://arxiv.org/pdf/1405.5336v3.pdf).

[12] J. Hachem, N. Karamchandani, and S. Diggavi, “Coded caching for multi-level popularity and access,” IEEE Trans. Inform. Theory, Dec. submitted. [Online]. Available: [https://dl.dropboxusercontent.com/u/689613/Papers/MultilevelJournal.pdf](https://dl.dropboxusercontent.com/u/689613/Papers/MultilevelJournal.pdf).

[13] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, “Online coded caching,” 2013. [Online]. Available: [arXiv:1311.3646](http://arxiv.org/pdf/1311.3646).

[14] M. Ji, G. Caire, and A. F. Molisch, “Fundamental limits of caching in wireless d2d networks,” 2014. [Online]. Available: [arXiv:1405.5336](http://arxiv.org/pdf/1405.5336).

[15] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. Diggavi, “Hierarchical coded caching,” IEEE Trans. Inf. Theory, vol. 62, no. 6, pp. 3212C3229, Jun. 2016.

[16] A. E. Gamal and Y. H. Kim, Network Information Theory, Cambridge University Press, 2011.