論文 / 著書情報
Article / Book Information

Title	Achievable Rate Regions of Cache-Aided Broadcast Networks for Delivering Content with a Multilayer Structure
Authors	Tetsunao Matsuta, Tomohiko Uyematsu
Citation	IEICE Trans. Fundamentals, Vol. E100-A, No. 12, pp. 2629-2640
Pub. date	2017, 12
URL	http://search.ieice.org/
Copyright	(c) 2017 Institute of Electronics, Information and Communication Engineers

PAPER Special Section on Information Theory and Its Applications

Achievable Rate Regions of Cache-Aided Broadcast Networks for Delivering Content with a Multilayer Structure*

Tetsunao MATSUTA[†a)], ***Member and*** **Tomohiko UYEMATSU**[†b)], ***Fellow***

SUMMARY This paper deals with a broadcast network with a server and many users. The server has files of content such as music and videos, and each user requests one of these files, where each file consists of some separated layers like a file encoded by a scalable video coding. On the other hand, each user has a local memory, and a part of information of the files is cached (i.e., stored) in these memories in advance of users' requests. By using the cached information as side information, the server encodes files based on users' requests. Then, it sends a codeword through an error-free shared link for which all users can receive a common codeword from the server without error. We assume that the server transmits some layers up to a certain level of requested files at each different transmission rate (i.e., the codeword length per file size) corresponding to each level. In this paper, we focus on the region of tuples of these rates such that layers up to any level of requested files are recovered at users with an arbitrarily small error probability. Then, we give inner and outer bounds on this region.

key words: *broadcast, caching, coded caching, scalable video coding, successive refinement*

1. Introduction

Because of the increase of a number of communication devices and the size of digital contents, network traffic has been significantly increasing in recent years. This increase causes network congestion during peak-traffic times. As an instance of a congested network, this paper deals with a broadcast network with a server and many users. The server has a lot of files of content such as music and videos, and each user requests one of these files. We assume that users are connected to the server through an error-free shared link for which all users can receive common symbols from the server without error. The content on demand system (such as YouTube, Netflix, Spotify, etc.) is an example of a practical system employing this network.

The network is usually not congested during *off-peak* times. Hence, one possible approach to reduce traffic during peak times is to use a local memory of each user. This approach consists of two distinct phases: the *placement phase* and the *delivery phase*. In the placement phase, a part of information of files is cached (i.e., stored) in users' local memories during off-peak times in advance of users' requests. Hence, the server only has to send remaining information of the files. In the delivery phase, by using the cached information as side information, the server encodes files based on users' requests and sends a codeword during peak times. Then, each user decodes the requested file from the codeword and the cached information. This caching system was introduced by Maddah-Ali and Niesen [2], where they assumed that files are the same size. In the context of the content on demand system, this assumption implies that each file is fixed-length bits (or packets) that are part of a music or video bit stream, which is handled by the server at a time. For this caching system, they [2], [3] gave upper and lower bounds to the infimum of transmission rates (i.e., the codeword length per file size) such that requested files are recovered at users with an arbitrarily small error probability. Their upper bound is given by a coding scheme called *coded caching* scheme using simple but effective XOR operations on files. This scheme achieves the significant improvement of performance compared with a traditional *uncoded caching* scheme (cf. [2]). Their bounds are nearly optimal in the sense that the gap between the upper bound and the lower bound is constant for any number of users and files. Further, there are many studies [4]–[7] to improve the upper bound and the lower bound.

In the above setting, the server must transmit the whole of requested files. However, we can consider the situation that the server only needs to transmit a part of requested files in the delivery phase. This may occur when files are encoded by scalable video coding (cf. e.g. [8]), where each file consists of a base layer and enhancement layers. Then, due to a heavy load and insufficient memory, the server may transmit only the base layer and enhancement layers up to a certain level for somewhat low-definition videos and music. In a more theoretical sense, this is the situation that files of the server are encoded by the successive refinement coding [9], [10]. Then each file consists of several codewords which have the same role of layers of the scalable video coding. This situation also may occur when each file consists of some types of files in order of importance. For example, if each file consists of a text file and a video file, the server only transmits the text file when the network is quite congested. In this situation, i.e., the situation that files consist of layers, the server may transmit some layers up to a certain level at each different transmission rate corresponding to each level during the delivery phase. Hence one of our main interest is, for a given local memory size, whether the server can transmit layers up to each level at each optimal minimum

Manuscript received January 30, 2017.
Manuscript revised June 20, 2017.

[†]The authors are with Dept. of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.

*Portions of this paper were presented at the 2016 IEEE International Symposium on Information Theory [1].

a) E-mail: tetsu@it.ce.titech.ac.jp
b) E-mail: uyematsu@ieee.org

DOI: 10.1587/transfun.E100.A.2629

rate without any redistribution of cached information.

To this end, we focus on the region of tuples of these rates such that layers up to any level of requested files are recovered at users with an arbitrarily small error probability. Here, in order to keep fairness among users, we assume that all users recover layers up to the same level. We also assume that layers of the same level have the same size in order to simplify a practical construction. Then, as in the original caching system, each layer can be regarded as fixed-length bits (or packets) of each layer bit stream. Yang and Gündüz [11] also deal with a caching system for content with a multilayer structure. However, unlike our system, they suppose that the server knows which layers are required in advance of users' request in the placement phase.

In this paper, we call the region of tuples of rates the *achievable rate region* and give an inner bound and an outer bound on this region. We also give a tighter outer bound for the case where each file consists of 2 layers. Then, we give some examples of inner and outer bounds for this 2-layer case. Interestingly, according to these examples, there exist both cases where the server can and cannot achieve optimal rates of the region simultaneously. The inner bound is derived by employing a memory dividing scheme in which a local memory of each user is divided into the same number of parts as the layers. Then, each part is employed to transmit each layer. Outer bounds are derived by employing a similar argument to the cut-set bound in [2] and Han's inequality [12] that is also used in [7].

The rest of this paper is organized as follows. In Sect. 2, we provide the setting of our caching system and define the achievable rate region. In Sect. 3, as our main result, we give an inner bound and outer bounds of the achievable rate region. In this section, we also give examples of inner and outer bounds for the case where each file consists of 2 layers. In Sect. 4, we prove our inner bound and give a practical construction method of a caching scheme. In Sect. 5, we prove our outer bounds for the multilayer case and the 2-layer case, respectively. In Sect. 6, we conclude the paper.

2. Problem Setting

Let $\mathbb{N}$, $\mathbb{Z}_+$ and $\mathbb{R}_+$ be sets of positive integers, non-negative integers, and non-negative real numbers, respectively. For $i, j \in \mathbb{Z}_+$, we will denote the consecutive integers $\{i, i+1, \cdots, j\}$ as $[i:j]$, where $[i:j] = \emptyset$ if $i > j$.

For $F, L \in \mathbb{N}$ and $p_l \in \mathbb{R}_+$ $(l \in [1:L])$ such that $\sum_{l=1}^{L} p_l = 1$, let

$$F_l = \begin{cases} \lfloor p_l F \rfloor & \text{if } l \in [1:L-1], \\ F - \sum_{l=1}^{L-1} F_l & \text{if } l = L, \end{cases}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. Hence, we have $p_l F - 1 \le F_l \le p_l F + L - 1$ and

$$F = \sum_{l=1}^{L} F_l.$$

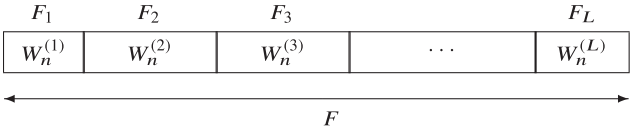

Fig. 1 A file W_n with L layers.

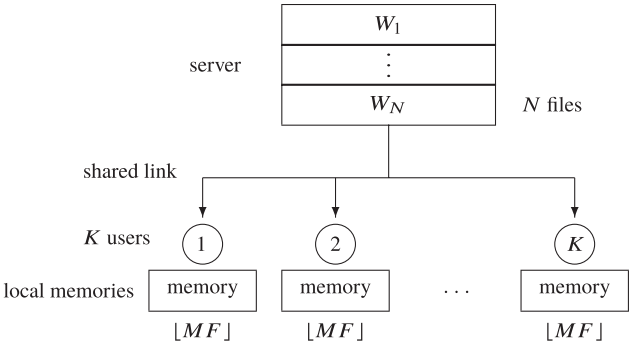

Fig. 2 Caching system.

For $N \in \mathbb{N}$ and $l \in [1:L]$, let $W_1^{(l)}, W_2^{(l)}, \cdots, W_N^{(l)}$ be N independent random variables each uniformly distributed over $[1:2^{F_l}]$ which represent l-th layers of N files. For $n \in [1:N]$, the concatenation of all layers $W_n^{(1)}, W_n^{(2)}, \cdots, W_n^{(L)}$ represents a file W_n of size $F = \sum_{l=1}^{L} F_l$ bits (see Fig. 1), i.e.,

$$W_n = (W_n^{(1)}, W_n^{(2)}, \cdots, W_n^{(L)}).$$

Hence, in our setting, there are N files with L layers. We will denote the sequence of layers $(W_n^{(1)}, W_n^{(2)}, \cdots, W_n^{(l)})$ up to the l-th layer as $W_n^{(1:l)}$. If each file in the server represents a video and encoded by a scalable video coding, $W_n^{(1)}$ can be regarded as a part of the base layer bit stream of a low-definition video, and $W_n^{(2)}$ can be regarded as a part of an enhancement layer bit stream, and so on. In this example, $W_n^{(1:l)}$ can be regarded as a part of a bit stream of a high-definition video using layers up to the l-th layer.

For $K \in \mathbb{N}$ and $M \in \mathbb{R}_+$, as shown in Fig. 2, we consider the situation that the server has all files $W_1, \cdots, W_N$, and K users are connected to the server through an error-free shared link. We assume that each of K users has a local memory of size $\lfloor MF \rfloor$ bits.

For $R_l \in \mathbb{R}_+$ $(l \in [1:L])$, we denote $(R_1, R_2, \cdots, R_L)$ as $\mathbf{R}$. Now, we describe an $(M, \mathbf{R})$ caching scheme for the placement and the delivery phases. This scheme consists of following K caching functions, L sets of N^K encoding functions, and L sets of KN^K decoding functions.

The K caching functions

$$\phi_k : \prod_{l=1}^{L} [1:2^{F_l}]^N \to [1:2^{\lfloor MF \rfloor}]$$

map the N files into the cache content

$$Z_k \triangleq \phi_k(W_1, W_2, \cdots, W_N)$$

for each user $k \in [1:K]$. The cache content Z_k is stored in

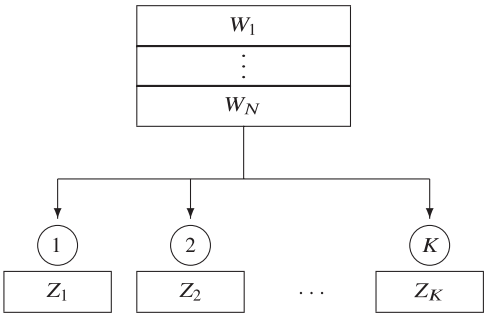

Fig. 3 Placement phase.

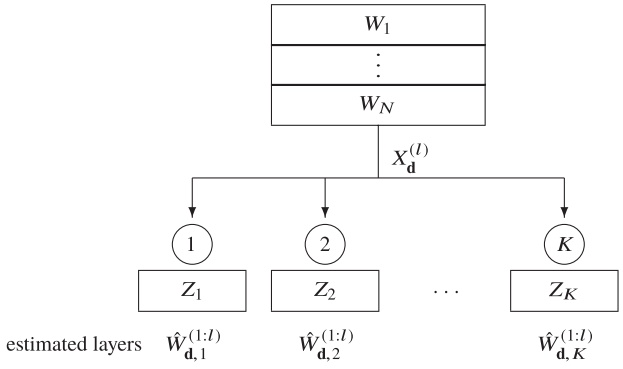

Fig. 4 Delivery phase.

the local memory of user k during the placement phase (see Fig. 3).

During the delivery phase, the server employs one of L sets of N^K encoding functions:

$$\psi_{\mathbf{d}}^{(l)} : \prod_{l=1}^{L}[1 : 2^{F_l}]^N \to [1 : 2^{\lfloor R_l F \rfloor}],$$

where $\mathbf{d} = (d_1, d_2, \cdots, d_K) \in [1 : N]^K$, d_k represents the requested file number of k-th user, and R_l represents the rate. The l-th set $\{\psi_{\mathbf{d}}^{(l)}\}$ of functions is employed to transmit requested sequences of layers in $W_1^{(1:l)}, W_2^{(1:l)}, \cdots, W_N^{(1:l)}$. For a given users' requests $\mathbf{d} \in [1 : N]^K$, the encoding function maps the N files into the codeword

$$X_{\mathbf{d}}^{(l)} \triangleq \psi_{\mathbf{d}}^{(l)}(W_1, \cdots, W_N).$$

Then, the server transmits $X_{\mathbf{d}}^{(l)}$ to all users by using the shared link (see Fig. 4).

Depending on the index $l \in [1 : L]$ of the encoding function, users employ one of L sets of KN^K decoding functions with the same index:

$$\mu_{\mathbf{d},k}^{(l)} : [1 : 2^{\lfloor R_l F \rfloor}] \times [1 : 2^{\lfloor MF \rfloor}] \to \prod_{i=1}^{l}[1 : 2^{F_i}].$$

For user k, the function maps the codeword $X_{\mathbf{d}}^{(l)}$ and the cache content Z_k to the estimate

$$\hat{W}_{\mathbf{d},k}^{(1:l)} \triangleq \mu_{\mathbf{d},k}^{(l)}(X_{\mathbf{d}}^{(l)}, Z_k)$$

of the sequence of layers $W_{d_k}^{(1:l)}$ of the requested file (see Fig. 4).

For the $(M, \mathbf{R})$ caching scheme, we define L types of error probabilities as follows:

$$\varepsilon_F^{(l)} \triangleq \max_{\mathbf{d}\in[1:N]^K} \max_{k\in[1:K]} \Pr\{\hat{W}_{\mathbf{d},k}^{(1:l)} \neq W_{d_k}^{(1:l)}\}.$$

Now, we define achievability and the achievable rate region.

Definition 1: A pair $(M, \mathbf{R})$ is achievable if and only if for every $\epsilon > 0$ and every large enough file size F, there exists an $(M, \mathbf{R})$ caching scheme such that $\varepsilon_F^{(l)} \le \epsilon$ for all $l \in [1 : L]$.

Definition 2: We define the achievable rate region $\mathcal{R}_L(M)$ as

$$\mathcal{R}_L(M) \triangleq \mathrm{cl}\{\mathbf{R} : (M, \mathbf{R}) \text{ is achievable}\},$$

where $\mathrm{cl}\{\cdot\}$ denotes the closure of the set.

Remark 1: Note that each user needs to know the actual request $\mathbf{d}$ and the index l of the layer to decode the sequence of layers. Hence, the server needs to send these information to all users. In practice, this can be realized by adding a *header* of size $\lceil K \log_2 N \rceil + \lceil \log_2 L \rceil$ bits representing the request $\mathbf{d} \in [1 : N]^K$ and the index $l \in [1 : L]$ to the codeword. Since this header does not depend on the file size F, it does not affect the achievable rate region. Hence, we omit the header in this study.

When files have only a single layer, i.e., $L = 1$, the server can use all users' memories to transmit files $W_1, \cdots, W_N$ of size F. This case is the same as the previous caching system [2], [3]. Hence, the memory-rate tradeoff $R^*(M)$ introduced in [2] can be defined as

$$R^*(M) \triangleq \inf \mathcal{R}_1(M).$$

In this single layer case, according to [2], the line connecting any two achievable points (M, R_1) and (M', R_1') is also achievable. Thus, it is easy to see that $R^*(M)$ is a convex and continuous function.

For this memory-rate tradeoff $R^*(M)$, Maddah-Ali and Niesen [2] gave the next upper bound using coded caching scheme.

Theorem 1: For N files and K users, it holds that

$$R^*(M) \le \sup\{f(M) : f \text{ is any real convex function}$$
$$\text{s.t. } f(\tilde{M}) \le K(1 - \tilde{M}/N) \min\left\{\frac{1}{1 + K\tilde{M}/N}, \frac{N}{K}\right\},$$
$$\forall \tilde{M} \in \{0, N/K, 2N/K, \cdots, N\}\}.$$

They also clarified a closed-form expression of $R^*(M)$ in the case where $N = K = 2$ as the next theorem, which can be achieved by the coded caching scheme and a specially designed caching scheme (see [2, Appendix]).

Theorem 2: For $N = 2$ files and $K = 2$ users, it holds that

$$R^*(M) = \max\{2 - 2M, 1 - M/2, 3/2 - M, 0\}.$$

3. Main Result

In this section, we give an inner bound and outer bounds on the achievable rate region. Proofs of these bounds are presented in Sects. 4 and 5.

The next theorem shows the inner bound.

Theorem 3: Let $\bar{R}(M)$ be an arbitrary upper bound on $R^*(M)$ (including $R^*(M)$ itself). Then, for N files with L layers, and K users with a local memory each of size M, we have

$$\mathcal{R}_L(M) \supseteq \bigcup_{\substack{(q_1,q_2,\cdots,q_L)\in[0,1]^L:\\ \sum_{l=1}^L q_l=1}} \left\{\mathbf{R}\in\mathbb{R}_+^L : R_l \geq \sum_{i=1}^{l} p_i \bar{R}\left(\frac{q_i}{p_i}M\right), \forall l \in [1:L]\right\}.$$

The next theorem shows the outer bound for multilayer cases.

Theorem 4: For N files with L layers, and K users with a local memory each of size M, we have

$$\begin{aligned}
&\mathcal{R}_L(M)\\
&\subseteq \bigcap_{\mathcal{L}\subseteq[1:L]} \left\{\mathbf{R}\in\mathbb{R}_+^L : \sum_{l\in\mathcal{L}} R_l \geq \max_{t\in[1:N]} \max_{\substack{s\in[0:K]:\\ st\leq N, |\mathcal{L}|s\leq\alpha}} \frac{1}{t}\right.\\
&\times\left(st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} p_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} p_i\right)\right.\\
&\left.\left.+\frac{|\mathcal{L}|s}{\alpha}\left|\min\{N,\alpha t\} - |\mathcal{L}|st\right|^+\left(\sum_{i=1}^{\bar{L}} p_i\right) - |\mathcal{L}|sM\right)\right\},
\end{aligned}$$

where $\bar{L} = \max\mathcal{L}$, $|x|^+ = \max\{0, x\}$, $\alpha = \min\{K, \lceil N/t\rceil\}$, and $\lceil\cdot\rceil$ denotes the ceiling function.

When $L = 1$, according to Theorem 4, it holds that

$$\mathcal{R}_1(M) \subseteq \{R_1 \in \mathbb{R}_+ : R_1 \geq R'(M)\},$$

where

$$\begin{aligned}
R'(M) = &\max_{t\in[1:N]} \max_{\substack{s\in[0:K]:\\ st\leq N, s\leq\alpha}}\\
&\times\left(s + \frac{s}{\alpha t}\left|\min\{N,\alpha t\} - st\right|^+ - \frac{s}{t}M\right).
\end{aligned}$$

This gives a lower bound on the memory-rate tradeoff:

$$R^*(M) \geq R'(M).$$

On the other hand, we have

$$\begin{aligned}
R'(M) &\geq \max_{s\in[0:K]} \max_{\substack{t\in[1:N]:\\ st\leq N, s\leq\min\{K,\lceil N/t\rceil\}}} \left(s - \frac{s}{t}M\right)\\
&\geq \max_{s\in[1:\min\{N,K\}]} \left(s - \frac{s}{\lfloor N/s\rfloor}M\right). \qquad (1)
\end{aligned}$$

Since (1) is the lower bound on $R^*(M)$ by Maddah-Ali and Niesen [2, Theorem 2], our bound gives a tighter bound compared with their bound. Especially, when $N = K = 2$, our lower bound coincides with the memory-rate tradeoff. In fact, when $N = K = 2$, we have

$$\begin{aligned}
R'(M) &= \max_{t\in[1:2]} \max_{\substack{s\in[0:2]:\\ st\leq 2}} \left(s + \frac{s}{\lceil 2/t\rceil t}(2 - st) - \frac{s}{t}M\right)\\
&= \max\{2 - 2M, 1 - M/2, 3/2 - M, 0\}.
\end{aligned}$$

For the 2-layer case, we can give the next outer bound which is tighter than the above bound.

Theorem 5: Let $L = 2$, $p_1 = p$, and $p_2 = 1 - p$ for a certain constant $p \in [0, 1]$. Then, for N files with 2 layers, and K users with a local memory each of size M, we have

$$\begin{aligned}
\mathcal{R}_2(M) \subseteq \{&(R_1, R_2) \in \mathbb{R}_+^2 : R_1 \geq r_1(p, K, M, N),\\
&R_2 \geq r_2(p, K, M, N, R_1)\}, \qquad (2)
\end{aligned}$$

where

$$\begin{aligned}
&r_1(p, K, M, N)\\
&= \max_{t_1\in[1:N]} \max_{\substack{s_1\in[0:K]\\ s_1t_1\leq N, s_1\leq\beta}} \frac{1}{t_1}\\
&\quad\times\left(s_1t_1p + \frac{s_1}{\beta}\left|\min\{N,\beta t_1\} - s_1t_1\right|^+ p - s_1M\right),\\
&r_2(p, K, M, N, R_1)\\
&= \max_{t_1\in[0:N], t_2\in[1:N]} \max_{\substack{s_1,s_2\in[0:K]:s_1+s_2\leq\gamma_1,\\ t_1s_1\leq N, t_2s_2\leq N}} \frac{1}{t_2}\\
&\quad\times\left(s_2t_2 + s_1t_1p + \left|\min\{N,\gamma_2\} - \max\{s_1t_1, s_2t_2\}\right|^+ p\right.\\
&\quad+ \left|\min\{N, s_2t_2 + s_1t_2\} - s_2t_2\right|^+ (1 - p)\\
&\quad+ \frac{s_1 + s_2}{\gamma_1}\left(\left|\min\{N,\gamma_3\} - \gamma_2\right|^+ p\right.\\
&\quad\left.+ \left|\min\{N,\gamma_1t_2\} - s_2t_2 - s_1t_2\right|^+ (1 - p)\right)\\
&\quad\left.- (s_1 + s_2)M - t_1R_1\right),
\end{aligned}$$

$$\beta = \min\{K, \lceil N/t_1\rceil\}, \qquad (3)$$

$$\gamma_1 = \begin{cases}\min\{K, \lceil N/t_2\rceil\} & \text{if } t_2 \neq 0,\\ \min\{K, \lceil N/t_1\rceil\} & \text{if } t_2 = 0,\end{cases} \qquad (4)$$

$$\gamma_2 = \max\{s_2t_2 + s_1t_2, s_1t_1\} + s_2t_1, \qquad (5)$$

$$\gamma_3 = \max\{\gamma_1t_2, \gamma_2\} + (\gamma_1 - s_1 - s_2)t_1. \qquad (6)$$

When $L = 2$ and $p = 1$, due to the definition of the memory-rate tradeoff, one can easily check that

$$R^*(M) = \inf\{R_1 : \exists(R_1, R_2) \in \mathcal{R}_2(M)\}.$$

Then, according to Theorem 5, we have

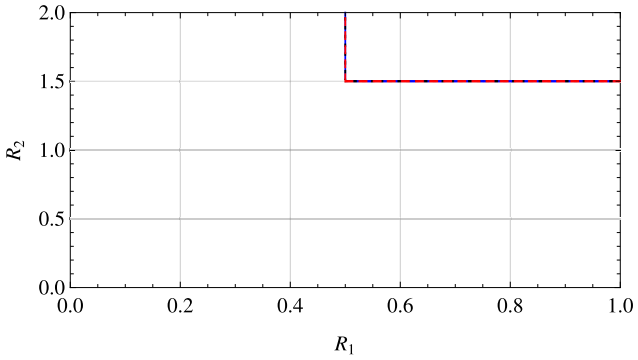

Fig. 5 p=0.5, K=2, M=0.25, N=2.

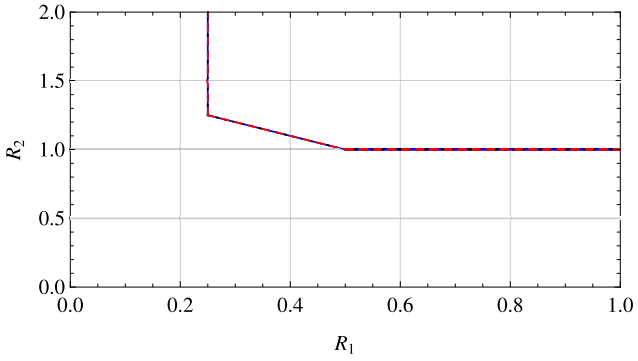

Fig. 6 p=0.5, K=2, M=0.5, N=2.

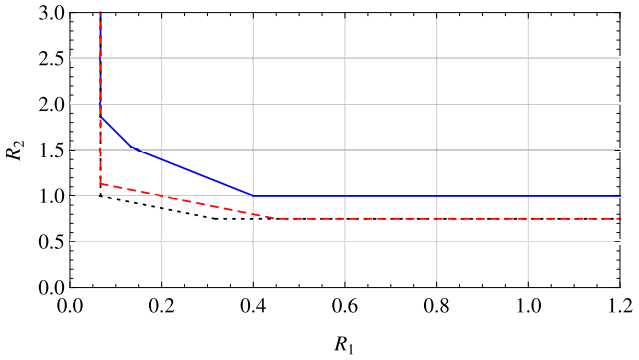

Fig. 7 p=0.4, K=3, M=1, N=3.

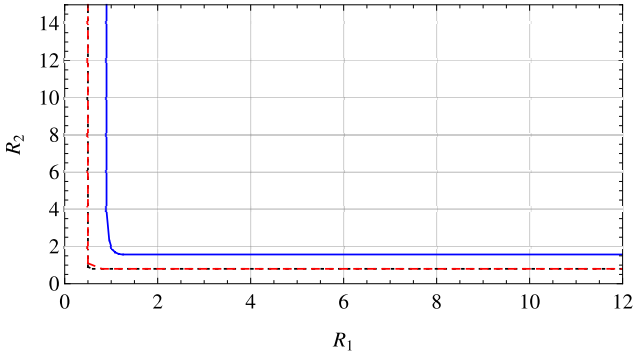

Fig. 8 p=0.8, K=15, M=7, N=20.

$$
\begin{aligned}
R^*(M) &= \inf\{R_1 : \exists (R_1, R_2) \in \mathcal{R}_2(M)\} \\
&\geq \inf\{R_1 : \exists (R_1, R_2) \in \mathbb{R}_+^2,\ R_1 \geq r_1(1, K, M, N), \\
&\quad R_2 \geq r_2(1, K, M, N, R_1)\} \\
&= r_1(1, K, M, N) \\
&= \max_{t_1 \in [1:N]} \max_{\substack{s_1 \in [0:K] \\ s_1 t_1 \leq N, s_1 \leq \beta}} \\
&\quad \times \left(s_1 + \frac{s_1}{\beta t_1} |\min\{N, \beta t_1\} - s_1 t_1|^+ - \frac{s_1}{t_1} M \right) \\
&= R'(M).
\end{aligned}
$$

Hence, Theorem 5 gives the same lower bound on $R^*(M)$ as Theorem 4.

Remark 2: Since we prove Theorems 4 and 5 by referring to the argument by Sengupta et al. [7] using Han's inequality, the lower bound $R'(M)$ is very similar to their lower bound [7, Theorem 1]. However, since our proof is not exactly the same as their proof (also in the sense that we deal with the multilayer case), the form of $R'(M)$ does not coincide with their bound. Nevertheless, according to our numerical calculation, $R'(M)$ gives the same lower bound as their bound.

Figures 5–8 show numerical examples of our bounds in the case where $L = 2$. Solid curves are inner bounds obtained by Theorem 3. In Figs. 5 and 6, we employ Theorem 2 which is the memory-rate tradeoff for $N = K = 2$ as $\bar{R}(M)$ in Theorem 3. On the other hand, in Figs. 7 and 8, we employ Theorem 1 which is the upper bound using the coded caching scheme as $\bar{R}(M)$ in Theorem 3. Dotted curves are outer bounds of Theorem 4. Dashed curves are outer bounds of Theorem 5.

In Figs. 5 and 6, our inner and two outer bounds coincide. Hence, there exist some cases where our bounds are optimal. However, as shown in Figs. 7 and 8, our bounds need to be improved.

Figure 5 shows that there exists a case where we can simultaneously achieve the pair of optimal rates (i.e., $(R_1, R_2) = (0.5, 1.5)$). This means that cache contents are optimal to transmit both first layer $W_n^{(1)}$ and all layers $W_n^{(1:2)}$. As shown in the next section, this case is achieved by the memory dividing scheme. Interestingly, Figs. 6 and 7 show that we cannot always achieve the pair of optimal rates. Hence, in general, it is impossible to make cache contents to be useful to transmit the both a part and the whole of the files. Note that this fact can be seen from a slightly loose outer bound of Theorem 4 only in Fig. 6.

4. Inner Bound

In this section, we prove Theorem 3 and give a practical construction method of a caching scheme.

4.1 Proof of Theorem 3

In order to prove Theorem 3, we employ a memory dividing scheme. For an arbitrarily fixed sequence of constants

$(q_1, \cdots, q_L) \in [0,1]^L$ satisfying $\sum_{l=1}^{L} q_l = 1$, we divide each local memory into L parts, each with $\lfloor q_l MF \rfloor$ bits ($l \in [1:L]$). We do not use remaining bits. Then, l-th layers $W_1^{(l)}, \cdots, W_N^{(l)}$ of size F_l will be transmitted by using local memories of the size $\lfloor q_l MF \rfloor$.

We show functions employing these divided memories to transmit layers. For an arbitrarily fixed $\delta > 0$, and $\tilde{M} \in \mathbb{R}_+$, let $\tilde{R} = R^*(\tilde{M}) + \delta$. Then, from the definition of the memory-rate tradeoff, for every $\epsilon > 0$, and every large enough file size $\tilde{F}$, there exist functions

$$\phi_k : [1 : 2^{\tilde{F}}]^N \to [1 : 2^{\lfloor \tilde{M}\tilde{F} \rfloor}], \tag{7}$$

$$\psi_{\mathbf{d}} : [1 : 2^{\tilde{F}}]^N \to [1 : 2^{\lfloor \tilde{R}\tilde{F} \rfloor}], \tag{8}$$

$$\mu_{\mathbf{d},k} : [1 : 2^{\lfloor \tilde{R}\tilde{F} \rfloor}] \times [1 : 2^{\lfloor \tilde{M}\tilde{F} \rfloor}] \to [1 : 2^{\tilde{F}}] \tag{9}$$

such that

$$\max_{\mathbf{d} \in [1:N]^K} \max_{k \in [1:K]} \Pr\{\hat{W}_{\mathbf{d},k} \neq \tilde{W}_{d_k}\} \leq \epsilon,$$

where

$$\hat{W}_{\mathbf{d},k} = \mu_{\mathbf{d},k}(\psi_{\mathbf{d}}(\tilde{W}_1, \cdots, \tilde{W}_N), \phi_k(\tilde{W}_1, \cdots, \tilde{W}_N)),$$

and $\tilde{W}_1, \cdots, \tilde{W}_N$ are N independent random variables each uniformly distributed over $[1 : 2^{\tilde{F}}]$. For convenience, we call these functions $(\tilde{M}, \tilde{R}, \epsilon, \tilde{F})$-functions. We will use these functions (7)–(9) to transmit layers.

Now, for an arbitrarily fixed $\delta > 0$, let

$$\tilde{R}^{(l)} = R^*\left(\frac{q_l}{p_l + \delta} M\right) + \delta, \quad (l \in [1:L]).$$

Then, for every $\epsilon > 0$, and every large enough file size F, there exist $(\frac{q_l}{p_l+\delta} M, \tilde{R}^{(l)}, \epsilon, F_l)$-functions. On the other hand, since $\frac{F_l}{F} \leq p_l + \delta$ for every large enough file size F, we have

$$\frac{q_l}{p_l + \delta} MF_l \leq q_l MF.$$

This means that the memory size used by $(\frac{q_l}{p_l+\delta} M, \tilde{R}^{(l)}, \epsilon, F_l)$-functions is smaller than the size $\lfloor q_l MF \rfloor$ of the divided memories. Hence, we can employ $(\frac{q_l}{p_l+\delta} M, \tilde{R}^{(l)}, \epsilon, F_l)$-functions to transmit l-th layers $W_1^{(l)}, \cdots, W_N^{(l)}$ of size F_l. Then, from the definition of $(\frac{q_l}{p_l+\delta} M, \tilde{R}^{(l)}, \epsilon, F_l)$-functions, the probability that l-th layers cannot be transmitted is less than or equal to ϵ for any $l \in [1:L]$. Hence, by using this caching scheme, we have

$$\varepsilon_F^{(l)} \leq l\epsilon.$$

We note that this scheme totally uses $\sum_{l=1}^{L} \lfloor \frac{q_l}{p_l+\delta} MF_l \rfloor$ bits of a local memory, and it satisfies the memory size:

$$\sum_{l=1}^{L} \left\lfloor \frac{q_l}{p_l + \delta} MF_l \right\rfloor \leq \sum_{l=1}^{L} \lfloor q_l MF \rfloor \leq \lfloor MF \rfloor.$$

On the other hand, the length $\sum_{i=1}^{l} \lfloor \tilde{R}^{(i)} F_i \rfloor$ of a codeword for transmitting sequences of layers $W_1^{(1:l)}, \cdots, W_N^{(1:l)}$ satisfies

$$\begin{aligned}
&\sum_{i=1}^{l} \lfloor \tilde{R}^{(i)} F_i \rfloor \\
&\leq \sum_{i=1}^{l} \lfloor (p_i + \delta)\, \tilde{R}^{(i)} F \rfloor \\
&= \sum_{i=1}^{l} \left\lfloor (p_i + \delta) \left(R^*\left(\frac{q_i}{p_i + \delta} M\right) + \delta \right) F \right\rfloor \\
&\leq \sum_{i=1}^{l} \left\lfloor \left(p_i R^*\left(\frac{q_i}{p_i + \delta} M\right) + \min\{N, K\}\delta + p_i\delta + \delta^2 \right) F \right\rfloor \\
&\leq \lfloor \tilde{R}_l F \rfloor,
\end{aligned} \tag{10}$$

where

$$\tilde{R}_l = \sum_{i=1}^{l} p_i R^*\left(\frac{q_i}{p_i + \delta} M\right) + L \min\{N, K\}\delta + \delta + L\delta^2,$$

and the second inequality comes from the fact that $R^*(M) \leq \min\{N, K\}$ for any $M \in \mathbb{R}_+$ because the server transmits at most $\min\{N, K\}$ files to users (it also immediately follows from Theorem 1).

Hence, by letting $\tilde{\mathbf{R}} = (\tilde{R}_1, \tilde{R}_2, \cdots, \tilde{R}_L)$, a pair $(M, \tilde{\mathbf{R}})$ is achievable for any sequence of constants $(q_1, \cdots, q_L) \in [0,1]^L$ satisfying $\sum_{l=1}^{L} q_l = 1$. Since $\delta > 0$ is arbitrary, and $R^*(M)$ is a continuous function such that $R^*(M) \leq \bar{R}(M)$ for any $M \in \mathbb{R}_+$, we have Theorem 3.

Remark 3: In this caching scheme, when the server transmits sequences of layers $W_1^{(1:l)}, \cdots, W_N^{(1:l)}$, it only uses the same sequences of layers. Hence, the inner bound does not change even if we define encoding functions $\{\psi_{\mathbf{d}}^{(l)}\}$ as $\psi_{\mathbf{d}}^{(l)} : \prod_{i=1}^{l} [1 : 2^{F_i}]^N \to [1 : 2^{\lfloor R_l F \rfloor}]$.

4.2 A Caching Scheme Using the Memory Dividing Scheme

We end this section by giving a practical construction method of a caching scheme using the memory dividing scheme.

$(\tilde{M}, \tilde{R}, \epsilon, \tilde{F})$-functions in the proof of Theorem 3 are equivalent to a caching scheme for the *single* layer case (e.g., coded-caching schemes in [2] and [3]). This implies that, once a caching scheme for the single layer case is given, we can construct an $(M, \mathbf{R})$ caching scheme for finite file size F as follows:

1. Divide each local memory into L parts, each with $\lfloor q_l MF \rfloor$ bits ($l \in [1:L]$).
2. Let $\delta_F = \frac{L-1}{F}$. Then, by using a caching scheme for the single layer case and noting that $\frac{F_l}{F} \leq p_l + \delta_F$, the server can generate a cache content $Z_{k,l}$ of size $\lfloor \frac{q_l}{p_l+\delta_F} MF_l \rfloor (\leq \lfloor q_l MF \rfloor)$ bits for l-th layers $W_1^{(l)}, \cdots, W_N^{(l)}$ of size F_l.

Repeat this generation for each $l \in [1:L]$ and store generated cache contents $(Z_{k,1}, Z_{k,2}, \cdots, Z_{k,L})$ in the divided local memory of the user $k \in [1:K]$.

3. By using the same caching scheme, the server can generate a codeword $X_{\mathbf{d},i}$ for i-th layers. Repeat this generation for each $i \in [1:l]$ up to a desired l-th layer. Then, the server transmits codewords $(X_{\mathbf{d},1}, X_{\mathbf{d},2}, \cdots, X_{\mathbf{d},l})$ to all users.
4. By using the same caching scheme, each user $k \in [1:K]$ can recover the i-th layer $W_{d_k}^{(i)}$ of the requested file from the codeword $X_{\mathbf{d},i}$ and the cache content $Z_{k,i}$. By repeating this recovery for each $i \in [1:l]$ up to the desired l-th layer, each user k can recover layers $W_{d_k}^{(1:l)}$ up to the l-th layer.

We note that since Maddah-Ali and Niesen [2], [3] give concrete algorithms of several caching schemes, the above procedure using the memory-dividing scheme can be constructed quite explicitly when we employ their caching schemes.

Let $\bar{R}\left(\frac{q_i}{p_i+\delta_F}M\right)$ be the rate of a given caching scheme for the single layer case for the memory size $\lfloor \frac{q_i}{p_i+\delta_F} M F_i \rfloor$. Then, the length of a codeword $X_{\mathbf{d},i}$ is $\lfloor \bar{R}\left(\frac{q_i}{p_i+\delta_F}M\right) F_i \rfloor$. Hence, in the same way as (10), it is sufficient to set the rate R_l for transmitting codewords $(X_{\mathbf{d},1}, X_{\mathbf{d},2}, \cdots, X_{\mathbf{d},l})$ to satisfy

$$R_l = \sum_{i=1}^{l} (p_i + \delta_F)\, \bar{R}\left(\frac{q_i}{p_i+\delta_F}M\right),$$

where $\delta_F \to 0$ as $F \to \infty$. Hence, by employing the coded caching scheme [2] that achieves the upper bound in Theorem 1, we can explicitly construct an $(M, \mathbf{R})$ caching scheme that achieves any tuple of rates in the inner bound of Theorem 3 in which we employ the upper bound in Theorem 1. Especially, by employing the specially designed caching scheme in [2] that achieves the memory-rate tradeoff when $K = N = 2$ (see Theorem 2), we can construct an $(M, \mathbf{R})$ caching scheme that achieves any pair of rates arbitrarily close to the boundary shown in Figs. 5 and 6.

5. Outer Bound

In this section, we prove Theorems 4 and 5.

In what follows, for a subset $\mathcal{K} \subseteq [1:K]$ of users, let $Z_{\mathcal{K}}$ be a tuple of cache contents of users in $\mathcal{K}$, i.e., $Z_{\mathcal{K}} = (Z_k : k \in \mathcal{K})$. For example, $Z_{\{1,2,4\}} = (Z_1, Z_2, Z_4)$. For a set $\mathcal{D} \subseteq [1:N]^K$ of K-tuples of requests, let $X_{\mathcal{D}}^{(l)}$ be a tuple of codewords for requests in $\mathcal{D}$ of sequences of layers $W_1^{(1:l)}, \cdots, W_N^{(1:l)}$, i.e., $X_{\mathcal{D}}^{(l)} = (X_{\mathbf{d}}^{(l)} : \mathbf{d} \in \mathcal{D})$, where $\mathbf{d} = (d_1, \cdots, d_K) \in [1:N]^K$. For example, $X_{\{(1,2),(2,1)\}}^{(l)} = (X_{(1,2)}^{(l)}, X_{(2,1)}^{(l)})$.

5.1 Proof of Theorem 4: Multilayer Cases

In order to show Theorem 4 which is an outer bound for multilayer cases, we use the next lemma.

Lemma 1: For a subset $\mathcal{K}$ of users, and a set $\mathcal{D}$ of requests, let

$$\mathcal{A} = \{d_k \in [1:N] : \mathbf{d} \in \mathcal{D},\ k \in \mathcal{K}\}. \tag{11}$$

Then, for any subset $\mathcal{B} \subseteq [1:N]$ of file numbers, $\epsilon > 0$, $l \in [1:L]$, and caching scheme such that $\varepsilon_F^{(l)} \le \epsilon$, we have

$$\begin{aligned} &H(X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}} | W_{\mathcal{B}}^{(1:l)}) \\ &\ge |\mathcal{A} \setminus \mathcal{B}| \sum_{i=1}^{l} F_i + H(X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}} | W_{\mathcal{A}\cup\mathcal{B}}^{(1:l)}) \\ &\quad - \epsilon\Theta(F) - \Theta(1), \end{aligned} \tag{12}$$

where $W_{\mathcal{B}}^{(1:l)} = (W_n^{(1:l)} : n \in \mathcal{B})$ and $\Theta(x)$ denotes the big theta notation, i.e., it is bounded asymptotically both above and below by x.

Proof. According to the assumption of this lemma, for any $k \in [1:K]$ and any $\mathbf{d} \in [1:N]^K$, we have

$$\Pr\{\mu_{\mathbf{d},k}^{(l)}(X_{\mathbf{d}}^{(l)}, Z_k) \ne W_{d_k}^{(1:l)}\} \le \epsilon,$$

where d_k is the file number in $\mathbf{d}$ of user k. Hence, we have

$$\begin{aligned} &H(W_{\mathcal{A}}^{(1:l)} | X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}}, W_{\mathcal{B}}^{(1:l)}) \\ &\le \sum_{\mathbf{d}\in\mathcal{D}} \sum_{k\in\mathcal{K}} H(W_{d_k}^{(1:l)} | X_{\mathbf{d}}^{(l)}, Z_k) \\ &\le |\mathcal{D}||\mathcal{K}|\,(1 + \epsilon F) \\ &= \epsilon\Theta(F) + \Theta(1), \end{aligned} \tag{13}$$

where the first inequality comes from (11), and the second inequality comes from Fano's inequality [13]. Then, for any $\mathcal{B} \subseteq [1:N]$, we have

$$\begin{aligned} &H(X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}} | W_{\mathcal{B}}^{(1:l)}) \\ &= H(W_{\mathcal{A}}^{(1:l)} | W_{\mathcal{B}}^{(1:l)}) - H(W_{\mathcal{A}}^{(1:l)} | X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}}, W_{\mathcal{B}}^{(1:l)}) \\ &\quad + H(X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}} | W_{\mathcal{A}}^{(1:l)}, W_{\mathcal{B}}^{(1:l)}) \\ &\ge H(W_{\mathcal{A}}^{(1:l)} | W_{\mathcal{B}}^{(1:l)}) + H(X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}} | W_{\mathcal{A}}^{(1:l)}, W_{\mathcal{B}}^{(1:l)}) \\ &\quad - \epsilon\Theta(F) - \Theta(1) \\ &= |\mathcal{A} \setminus \mathcal{B}| \sum_{i=1}^{l} F_i + H(X_{\mathcal{D}}^{(l)}, Z_{\mathcal{K}} | W_{\mathcal{A}\cup\mathcal{B}}^{(1:l)}) \\ &\quad - \epsilon\Theta(F) - \Theta(1), \end{aligned}$$

where the inequality comes from (13). □

As we will see later, the coefficient of $\sum_{i=1}^{l} F_i$ in (12) should be large in order to effectively employ this lemma. Since this coefficient depends on assignments of the subset of users and the set of requests, we should properly define these sets.

To this end, for $\mathcal{L} \subseteq [1:L]$ and $t \in [0:N]$, let $s \in [0:K]$ satisfy $st \le N$ and $|\mathcal{L}|s \le \alpha$, where $\alpha = \min\{K, \lceil N/t \rceil\}$.

Table 1 An example of Assignment 1.

	k_1	k_2	$\cdots$	k_s
$\mathbf{d}_1$	1	2	$\cdots$	s
$\mathbf{d}_2$	$s+1$	$s+2$	$\cdots$	$2s$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{d}_t$	$s(t-1)+1$	$s(t-1)+2$	$\cdots$	st

Table 2 An example of Assignment 2.

	k_{s+1}	k_{s+2}	$\cdots$	$k_{\lvert\mathcal{L}\rvert s}$
$\mathbf{d}_1$	$st+1$	$st+2$	$\cdots$	$s(t-1)+\lvert\mathcal{L}\rvert s$
$\mathbf{d}_2$	$s(t-1)+\lvert\mathcal{L}\rvert s+1$	$\cdots$	$\cdots$	$s(t-2)+2\lvert\mathcal{L}\rvert s$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{d}_t$	$s+\lvert\mathcal{L}\rvert s(t-1)+1$	$\cdots$	$\cdots$	$\lvert\mathcal{L}\rvert st$

Then, for a subset $\bar{\mathcal{K}} \subseteq [1:\alpha] (\subseteq [1:K])$ of users with cardinality $|\mathcal{L}|s$, we divide it into subsets $\mathcal{K}_l$ ($l \in \mathcal{L}$) with the same size without overlap, i.e.,

$$\begin{aligned}\bar{\mathcal{K}} &= \bigcup_{l\in\mathcal{L}} \mathcal{K}_l,\\ |\mathcal{K}_l| &= s, \quad \forall l \in \mathcal{L},\\ \mathcal{K}_l \cap \mathcal{K}_{l'} &= \emptyset, \quad \forall l, l' \in \mathcal{L} \text{ s.t. } l \neq l'.\end{aligned}$$

By using subset $\mathcal{K}_l$, we assign integers (i.e., requesting file numbers) to all elements in set $\mathcal{D}_l \subseteq [1:N]^K$ of requests satisfying $|\mathcal{D}_l| = t$ and all the following three conditions (Assignments 1–3):

- **Assignment 1:** The following assignments tighten later inequalities (17) and (18).

$$\begin{aligned}&\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_l,\ k \in \mathcal{K}_l\} = [1:st],\\ &\forall l \in \mathcal{L}.\end{aligned} \tag{14}$$

  This assignment is valid because $|\mathcal{D}_l||\mathcal{K}_l| = st (\leq N)$, and hence we can assign integers to $\mathbf{d} \in \mathcal{D}_l$ without overlap. Table 1 gives an example of this assignment, where $\mathcal{D}_l = \{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_t\}$ and $\mathcal{K}_l = \{k_1, k_2, \cdots, k_s\}$. The i-th column and the j-th row of the table denotes a requesting file number of a user k_i in a request $\mathbf{d}_j$ ($i \in [1:s], j \in [1:t]$).
- **Assignment 2:** The following assignment tightens the later inequality (19).

$$\begin{aligned}&\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_{\bar{L}},\ k \in \cup_{l\in\mathcal{L}\setminus\{\bar{L}\}}\mathcal{K}_l\}\\ &\quad = [st+1 : \min\{N, |\mathcal{L}|st\}],\end{aligned} \tag{15}$$

  where $\bar{L} = \max \mathcal{L}$. This is valid because $|\bar{\mathcal{K}}||\mathcal{D}_{\bar{L}}| = |\mathcal{L}|st$. Table 2 gives an example of this assignment when $|\mathcal{L}|st \leq N$, where $\mathcal{D}_{\bar{L}} = \{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_t\}$, $\mathcal{K}_{\bar{L}} = \{k_1, k_2, \cdots, k_s\}$ and $\cup_{l\in\mathcal{L}\setminus\{\bar{L}\}}\mathcal{K}_l = \bar{\mathcal{K}} \setminus \mathcal{K}_{\bar{L}} = \{k_{s+1}, k_{s+2}, \cdots, k_{|\mathcal{L}|s}\}$. We note that integers are already assigned for $\mathcal{D}_{\bar{L}}$ and $\mathcal{K}_{\bar{L}}$ by Assignment 1.
- **Assignment 3:** The following assignment tightens the later inequality (22).

$$\begin{aligned}&\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_{\bar{L}},\ k \in [1:\alpha]\setminus\bar{\mathcal{K}}\}\\ &\quad = [|\mathcal{L}|st+1 : \min\{N, \alpha t\}].\end{aligned} \tag{16}$$

  This is valid because $|[1:\alpha]||\mathcal{D}_{\bar{L}}| = \alpha t$. Table 3 gives an example of this assignment when $\alpha t \leq N$, where $\mathcal{D}_{\bar{L}} = \{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_t\}$, $\bar{\mathcal{K}} = \{k_1, k_2, \cdots, k_{|\mathcal{L}|s}\}$, and $[1:\alpha] \setminus \bar{\mathcal{K}} = \{k_{|\mathcal{L}|s+1}, k_{|\mathcal{L}|s+2}, \cdots, k_\alpha\}$. We note that integers are already assigned for $\mathcal{D}_{\bar{L}}$ and $\bar{\mathcal{K}}$ by Assignments 1 and 2.

Table 3 An example of Assignment 3.

	$k_{\lvert\mathcal{L}\rvert s+1}$	$k_{\lvert\mathcal{L}\rvert s+2}$	$\cdots$	k_α
$\mathbf{d}_1$	$\lvert\mathcal{L}\rvert st+1$	$\lvert\mathcal{L}\rvert st+2$	$\cdots$	$\lvert\mathcal{L}\rvert s(t-1)+\alpha$
$\mathbf{d}_2$	$\lvert\mathcal{L}\rvert s(t-1)+\alpha+1$	$\cdots$	$\cdots$	$\lvert\mathcal{L}\rvert s(t-2)+2\alpha$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{d}_t$	$\lvert\mathcal{L}\rvert s+\alpha(t-1)+1$	$\cdots$	$\cdots$	αt

We now turn to the derivation of the outer bound. For any $\mathcal{L} \subseteq [1:L]$, any $t \in [0:N]$, and any $s \in [0:K]$ such that $|\mathcal{L}|s \leq \alpha$ and $st \leq N$, achievable pair $(M, \mathbf{R})$, and large enough file size $F \in \mathbb{N}$, we have

$$\begin{aligned}&\left(\sum_{l\in\mathcal{L}} tR_lF\right) + |\mathcal{L}|sMF\\ &\geq \sum_{l\in\mathcal{L}} H(X^{(l)}_{\mathcal{D}_l}, Z_{\mathcal{K}_l})\\ &\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + \left(\sum_{l\in\mathcal{L}} H(X^{(l)}_{\mathcal{D}_l}, Z_{\mathcal{K}_l}|W^{(1:l)}_{[1:st]})\right)\\ &\quad - \epsilon\Theta(F) - \Theta(1) \qquad (17)\\ &\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + \left(\sum_{l\in\mathcal{L}} H(X^{(l)}_{\mathcal{D}_l}, Z_{\mathcal{K}_l}|W^{(1:\bar{L})}_{[1:st]})\right)\\ &\quad - \epsilon\Theta(F) - \Theta(1)\\ &\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + H(X^{(\bar{L})}_{\mathcal{D}_{\bar{L}}}, Z_{\bar{\mathcal{K}}}|W^{(1:\bar{L})}_{[1:st]})\\ &\quad - \epsilon\Theta(F) - \Theta(1) \qquad (18)\\ &\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} F_i\right)\\ &\quad + H(X^{(\bar{L})}_{\mathcal{D}_{\bar{L}}}, Z_{\bar{\mathcal{K}}}|W^{(1:\bar{L})}_{[1:\min\{N,|\mathcal{L}|st\}]})\\ &\quad - \epsilon\Theta(F) - \Theta(1), \qquad (19)\end{aligned}$$

where (17) comes from Assignment 1 and Lemma 1, and (19) comes from Assignment 2 and Lemma 1.

We give a more precise explanation for Assignments 1 and 2. In Assignment 1, consecutive positive integers are assigned without duplication for each of sets in (14). This tightens the inequality (17) because this makes the coefficient of $\sum_{i=1}^{l} F_i$ in (12) large. Further, these integers are assigned so as to overlap among these sets. This avoids the loss of the inequality (18) because the entropy does not greatly increase by the conditioning due to this overlapping. In Assignment 2, positive integers are assigned for the set (15) without

overlapping with the set (14). This is to avoid overlapping with the indices of layers $W_{[1:st]}^{(1:\bar{L})}$ in the condition of the entropy of (18). This tightens the inequality (19) because this makes the coefficient of $\sum_{i=1}^{l} F_i$ in (12) large.

The above inequality holds for any subset $\bar{\mathcal{K}} \subseteq [1:\alpha]$ such that $|\bar{\mathcal{K}}| = |\mathcal{L}|s$. By combining all these inequalities, we have

$$\begin{aligned}
&\sum_{l\in\mathcal{L}} tR_l F + |\mathcal{L}|sMF \\
&\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} F_i\right) \\
&\quad + \frac{1}{\binom{\alpha}{|\mathcal{L}|s}} \sum_{\substack{\bar{\mathcal{K}}\subseteq[1:\alpha]:\\ |\bar{\mathcal{K}}|=|\mathcal{L}|s}} \Big(H(Z_{\bar{\mathcal{K}}}|W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})}) \\
&\quad + H(X_{\mathcal{D}_{\bar{L}}}^{(\bar{L})}|Z_{[1:\alpha]}, W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})})\Big) \\
&\quad - \epsilon\Theta(F) - \Theta(1) \qquad (20) \\
&\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} F_i\right) \\
&\quad + \frac{|\mathcal{L}|s}{\alpha}\frac{1}{\binom{\alpha}{|\mathcal{L}|s}} \sum_{\substack{\bar{\mathcal{K}}\subseteq[1:\alpha]:\\ |\bar{\mathcal{K}}|=|\mathcal{L}|s}} \\
&\quad H(X_{\mathcal{D}_{\bar{L}}}^{(\bar{L})}, Z_{[1:\alpha]}|W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})}) \\
&\quad - \epsilon\Theta(F) - \Theta(1) \qquad (21) \\
&\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} F_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} F_i\right) \\
&\quad + \frac{|\mathcal{L}|s}{\alpha}\Big|\min\{N, \alpha t\} - |\mathcal{L}|st\Big|^{+}\left(\sum_{i=1}^{\bar{L}} F_i\right) \\
&\quad - \epsilon\Theta(F) - \Theta(1), \qquad (22)
\end{aligned}$$

where (20) comes from the fact that

$$\begin{aligned}
&H(X_{\mathcal{D}_{\bar{L}}}^{(\bar{L})}, Z_{\bar{\mathcal{K}}}|W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})}) \\
&\geq H(Z_{\bar{\mathcal{K}}}|W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})}) \\
&\quad + H(X_{\mathcal{D}_{\bar{L}}}^{(\bar{L})}|Z_{[1:\alpha]}, W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})}),
\end{aligned}$$

(21) comes from the fact that $\frac{|\mathcal{L}|s}{\alpha} \leq 1$ and Han's inequality [12] (or see [13, Theorem 17.6.1]):

$$\begin{aligned}
&\frac{1}{\binom{\alpha}{|\mathcal{L}|s}} \sum_{\bar{\mathcal{K}}\subseteq[1:\alpha]:|\bar{\mathcal{K}}|=|\mathcal{L}|s} \frac{H(Z_{\bar{\mathcal{K}}}|W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})})}{|\mathcal{L}|s} \\
&\geq \frac{H(Z_{[1:\alpha]}|W_{[1:\min\{N,|\mathcal{L}|st\}]}^{(1:\bar{L})})}{\alpha},
\end{aligned}$$

and (22) comes from Assignment 3 and Lemma 1. Here, for the same reason as the assignment of (15), positive integers are assigned for the set (16) in Assignment 3.

Since $\lim_{F\to\infty} \frac{F_l}{F} = p_l$ and $\epsilon > 0$ can be arbitrarily small, we have for any achievable pair $(M, \mathbf{R})$,

$$\begin{aligned}
&\sum_{l\in\mathcal{L}} tR_l + |\mathcal{L}|sM \\
&\geq st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} p_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} p_i\right) \\
&\quad + \frac{|\mathcal{L}|s}{\alpha}\Big|\min\{N, \alpha t\} - |\mathcal{L}|st\Big|^{+}\left(\sum_{i=1}^{\bar{L}} p_i\right). \qquad (23)
\end{aligned}$$

Thus, for any $\mathcal{L} \subseteq [1:L]$, any $t \in [1:N]$, and any $s \in [0:K]$ such that $|\mathcal{L}|s \leq \alpha$ and $st \leq N$, we have

$$\begin{aligned}
&\sum_{l\in\mathcal{L}} R_l \\
&\geq \frac{1}{t}\left(st\left(\sum_{l\in\mathcal{L}}\sum_{i=1}^{l} p_i\right) + (\min\{N, |\mathcal{L}|st\} - st)\left(\sum_{i=1}^{\bar{L}} p_i\right)\right. \\
&\quad \left. + \frac{|\mathcal{L}|s}{\alpha}\Big|\min\{N, \alpha t\} - |\mathcal{L}|st\Big|^{+}\left(\sum_{i=1}^{\bar{L}} p_i\right) - |\mathcal{L}|sM\right).
\end{aligned}$$

This completes the proof of Theorem 4.

The left-hand side of the inequality (23) implies the total size of $|\mathcal{L}|t$ codewords and $|\mathcal{L}|s$ cache contents. On the other hand, the right-hand side of the inequality (23) implies a lower bound on the total size of layers that can be recovered by these multiple codewords and cache contents. These facts justify the inequality (23) because the size of these codewords and cache contents transmitted by the server must be larger than the total size of the recovered layers. This approach is basically the same as the cut-set bound in [2]. Here we extend their approach to multilayer cases by considering a set $\mathcal{L}$ of indices of layers.

Remark 4: In the inequality (23), we do not restrict $\mathcal{L}$ to one layer, i.e., $\mathcal{L} = \{l\}$ for $l \in [1:L]$. If we only consider one layer $\mathcal{L} = \{l\}$, we only have a lower bound on the rate R_l independently of other rates. Hence, for example, we cannot give a lower bound on $R_1 + R_2$. Then, we cannot draw the sloping line in Fig. 6. This means that considering a set of indices of layers is quite important in multilayer cases.

5.2 Proof of Theorem 5: The 2-Layer Case

In this section, suppose that $L = 2$, and hence it holds that $W_n^{(1:1)} = W_n^{(1)}$ and $W_n^{(1:2)} = (W_n^{(1)}, W_n^{(2)}) = W_n$.

In the previous section, the number s of requests and the number t of cache contents are the same for any given $\mathcal{L}$, respectively. Instead of using the same parameters s and t, we will use different parameters in the 2-layer case. This causes rather difficult assignments but gives a tighter bound. Although the following argument for the outer bound of the 2-layer case is similar to that in the previous section, we give

it precisely because it is rather complicated.

First of all, we show the next key lemma for our outer bound.

Lemma 2: For a subset $\mathcal{K}$ of users, and subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ of requests, let

$$\mathcal{A}_l = \{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_l,\ k \in \mathcal{K}\},\ (l \in \{1,2\}). \tag{24}$$

Then, for any subsets $\mathcal{B}_1, \mathcal{B}_2 \subseteq [1:N]$ of file numbers, $\epsilon > 0$, and caching scheme such that $\varepsilon_F^{(1)} \le \epsilon$ and $\varepsilon_F^{(2)} \le \epsilon$, we have

$$\begin{aligned} &H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\ge |\{\mathcal{A}_1 \cup \mathcal{A}_2\} \backslash \{\mathcal{B}_1 \cup \mathcal{B}_2\}| F_1 + |\mathcal{A}_2 \backslash \mathcal{B}_2| F_2 \\ &\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2}, W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad - \epsilon\Theta(F) - \Theta(1). \end{aligned} \tag{25}$$

where $W_{\mathcal{B}_1}^{(1)} = (W_n^{(1)} : n \in \mathcal{B}_1)$ and $W_{\mathcal{B}_2} = (W_n : n \in \mathcal{B}_2)$.

Proof. According to the assumption of this lemma, for any $k \in [1:K]$ and any $\mathbf{d} \in [1:N]^K$, we have

$$\Pr\{\mu_{\mathbf{d},k}^{(l)}(X_{\mathbf{d}}^{(l)}, Z_k) \ne W_{d_k}^{(1:l)}\} \le \epsilon, \quad \forall l \in \{1,2\}.$$

Hence by recalling that $W_n^{(1:1)} = W_n^{(1)}$ and $W_n^{(1:2)} = W_n$, we have

$$\begin{aligned} &H(W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2} | X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}}) \\ &\le H(W_{\mathcal{A}_1}^{(1)} | X_{\mathcal{D}_1}^{(1)}, Z_{\mathcal{K}}) + H(W_{\mathcal{A}_2} | X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}}) \\ &\le |\mathcal{D}_1||\mathcal{K}|\,(1+\epsilon F) + |\mathcal{D}_2||\mathcal{K}|\,(1+\epsilon F) \\ &= \epsilon\Theta(F) + \Theta(1), \end{aligned} \tag{26}$$

where the second inequality comes from Fano's inequality [13] and (24). Then, for any $\mathcal{B}_1 \subseteq [1:N]$ and $\mathcal{B}_2 \subseteq [1:N]$, we have

$$\begin{aligned} &H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &= H(W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2} | W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad - H(W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2} | X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}}, W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2}, W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\ge H(W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2} | W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2}, W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad - \epsilon\Theta(F) - \Theta(1) \\ &= H(W_{\mathcal{A}_1 \cup \mathcal{A}_2}^{(1)}, W_{\mathcal{A}_2}^{(2)} | W_{\mathcal{B}_1 \cup \mathcal{B}_2}^{(1)}, W_{\mathcal{B}_2}^{(2)}) \\ &\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2}, W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad - \epsilon\Theta(F) - \Theta(1) \\ &= |\{\mathcal{A}_1 \cup \mathcal{A}_2\} \backslash \{\mathcal{B}_1 \cup \mathcal{B}_2\}| F_1 + |\mathcal{A}_2 \backslash \mathcal{B}_2| F_2 \\ &\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}} | W_{\mathcal{A}_1}^{(1)}, W_{\mathcal{A}_2}, W_{\mathcal{B}_1}^{(1)}, W_{\mathcal{B}_2}) \\ &\quad - \epsilon\Theta(F) - \Theta(1), \end{aligned}$$

where the inequality comes from (26). □

As in the previous section, coefficients of F_1 and F_2 in (25) should be large in order to effectively employ this lemma. To this end, for $t_1, t_2 \in [0:N]$, let $s_1, s_2 \in [0:K]$ satisfy $s_1 t_1 \le N$, $s_2 t_2 \le N$ and $s_1 + s_2 \le \gamma_1$, where γ_1 is the constant of (4). For a subset $\bar{\mathcal{K}} \subseteq [1:\gamma_1] (\subseteq [1:K])$ of users with cardinality $s_1 + s_2$, we divide it into $\mathcal{K}_1$ and $\mathcal{K}_2$ without overlap such that $|\mathcal{K}_1| = s_1$ and $|\mathcal{K}_2| = s_2$. By using these subsets $\mathcal{K}_1$ and $\mathcal{K}_2$, we assign integers (i.e., requesting file numbers) to all elements in sets $\mathcal{D}_1 \subseteq [1:N]^K$ and $\mathcal{D}_2 \subseteq [1:N]^K$ of requests satisfying $|\mathcal{D}_1| = t_1$ and $|\mathcal{D}_2| = t_2$. Integers are assigned to satisfy the following three conditions (Assignments 1–3). Tables 1 – 3 would also be helpful in the following conditions.

- **Assignment 1:** The following assignments tighten the later inequalities (33) and (34).

$$\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_2,\ k \in \mathcal{K}_2\} = [1 : s_2 t_2], \tag{27}$$

$$\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_1,\ k \in \mathcal{K}_1\} = [1 : s_1 t_1]. \tag{28}$$

- **Assignment 2:** The following assignments tighten the later inequality (35).

$$\begin{aligned} &\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_2,\ k \in \mathcal{K}_1\} \\ &\quad = [s_2 t_2 + 1 : \min\{N, s_2 t_2 + s_1 t_2\}], \end{aligned} \tag{29}$$

$$\begin{aligned} &\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_1,\ k \in \mathcal{K}_2\} \\ &\quad = [\max\{s_2 t_2 + s_1 t_2, s_1 t_1\} + 1 : \min\{N, \gamma_2\}], \end{aligned} \tag{30}$$

  where we note that $\gamma_2 = \max\{s_2 t_2 + s_1 t_2, s_1 t_1\} + s_2 t_1$.
- **Assignment 3:** The following assignments tighten the later inequality (38).

$$\begin{aligned} &\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_2,\ k \in [1:\gamma_1] \backslash \bar{\mathcal{K}}\} \\ &\quad = [s_2 t_2 + s_1 t_2 + 1 : \min\{N, \gamma_1 t_2\}], \end{aligned} \tag{31}$$

$$\begin{aligned} &\{d_k \in [1:N] : \mathbf{d} \in \mathcal{D}_1,\ k \in [1:\gamma_1] \backslash \bar{\mathcal{K}}\} \\ &\quad = [\max\{\gamma_1 t_2, \gamma_2\} + 1 : \min\{N, \gamma_3\}], \end{aligned} \tag{32}$$

  where we note that $\gamma_3 = \max\{\gamma_1 t_2, \gamma_2\} + (\gamma_1 - s_1 - s_2) t_1$. The assignment (32) is valid because $|[1:\gamma_1] \backslash \bar{\mathcal{K}}| = \gamma_1 - s_1 - s_2$.

We now turn to the derivation of the outer bound. For any $t_1, t_2 \in [0:N]$, any $s_1, s_2 \in [0:K]$ such that $s_1 t_1 \le N$, $s_2 t_2 \le N$ and $s_1 + s_2 \le \gamma_1$, achievable pair $(M, \mathbf{R})$, and large enough file size $F \in \mathbb{N}$, we have

$$\begin{aligned} &t_1 R_1 F + s_1 M F + t_2 R_2 F + s_2 M F \\ &\ge H(X_{\mathcal{D}_1}^{(1)}, Z_{\mathcal{K}_1}) + H(X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}_2}) \\ &\ge s_1 t_1 F_1 + H(X_{\mathcal{D}_1}^{(1)}, Z_{\mathcal{K}_1} | W_{[1:s_1 t_1]}^{(1)}) \\ &\quad + s_2 t_2 F + H(X_{\mathcal{D}_2}^{(2)}, Z_{\mathcal{K}_2} | W_{[1:s_2 t_2]}) \\ &\quad - \epsilon\Theta(F) - \Theta(1) \end{aligned} \tag{33}$$

$$
\begin{aligned}
&\geq H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\bar{\mathcal{K}}}|W_{[1:s_2t_2]}, W_{[s_2t_2+1:s_1t_1]}^{(1)}) \\
&\quad + s_2t_2F + s_1t_1F_1 - \epsilon\Theta(F) - \Theta(1) \qquad (34)\\
&\geq s_2t_2F + s_1t_1F_1 \\
&\quad + |\min\{N, \gamma_2\} - \max\{s_1t_1, s_2t_2\}|^+ F_1 \\
&\quad + |\min\{N, s_2t_2 + s_1t_1\} - s_2t_2|^+ F_2 \\
&\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\bar{\mathcal{K}}}|\tilde{W}) - \epsilon\Theta(F) - \Theta(1), \qquad (35)
\end{aligned}
$$

where

$$
\tilde{W} = (W_{[1:\min\{N, s_2t_2+s_1t_2\}]}, W_{[s_2t_2+s_1t_2+1:\min\{N,\gamma_2\}]}^{(1)}),
$$

(33) comes from Assignment 1 and Lemma 2, and (35) comes from Assignment 2 and Lemma 2

We give a more precise explanation for Assignments 1 and 2. In Assignment 1, consecutive positive integers are assigned without duplication for each of two sets (28) and (27). This tightens the inequality (33), because this makes coefficients of F_1 and F_2 in (25) large. Further, these integers are assigned so as to overlap among these two sets. This avoids the loss of the inequality (34), because the entropy does not greatly increase by the conditioning due to this overlapping. In Assignment 2, positive integers are assigned for the set (29) without overlapping with the set (28). This is to avoid overlapping with the indices of layers $W_{[1:s_2t_2]}$ in the condition of the entropy of (34). This tightens the inequality (35), because this makes coefficients of F_1 and F_2 in (25) large. For a similar reason as the assignment of (29), positive integers are assigned for the set (30).

The above inequality holds for any subset $\bar{\mathcal{K}} \subseteq [1:\gamma_1]$ such that $|\bar{\mathcal{K}}| = s_1 + s_2$. By combining all these inequalities, we have

$$
\begin{aligned}
&t_1R_1F + s_1MF + t_2R_2F + s_2MF \\
&\geq s_2t_2F + s_1t_1F_1 \\
&\quad + |\min\{N, \gamma_2\} - \max\{s_1t_1, s_2t_2\}|^+ F_1 \\
&\quad + |\min\{N, s_2t_2 + s_1t_2\} - s_2t_2|^+ F_2 \\
&\quad + \frac{1}{\binom{\gamma_1}{s_1+s_2}} \sum_{\substack{\bar{\mathcal{K}}\subseteq[1:\gamma_1]:\\ |\bar{\mathcal{K}}|=s_1+s_2}} \Big(H(Z_{\bar{\mathcal{K}}}|\tilde{W}) \\
&\quad + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}|Z_{[1:\gamma_1]}, \tilde{W})\Big) \\
&\quad - \epsilon\Theta(F) - \Theta(1) \qquad (36)\\
&\geq s_2t_2F + s_1t_1F_1 \\
&\quad + |\min\{N, \gamma_2\} - \max\{s_1t_1, s_2t_2\}|^+ F_1 \\
&\quad + |\min\{N, s_2t_2 + s_1t_2\} - s_2t_2|^+ F_2 \\
&\quad + \frac{s_1+s_2}{\gamma_1} \frac{1}{\binom{\gamma_1}{s_1+s_2}} \sum_{\substack{\bar{\mathcal{K}}\subseteq[1:\gamma_1]:\\ |\bar{\mathcal{K}}|=s_1+s_2}} H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{[1:\gamma_1]}|\tilde{W}) \\
&\quad - \epsilon\Theta(F) - \Theta(1) \qquad (37)\\
&\geq s_2t_2F + s_1t_1F_1 \\
&\quad + |\min\{N, \gamma_2\} - \max\{s_1t_1, s_2t_2\}|^+ F_1 \\
&\quad + |\min\{N, s_2t_2 + s_1t_2\} - s_2t_2|^+ F_2 \\
&\quad + \frac{s_1+s_2}{\gamma_1} \left(|\min\{N, \gamma_3\} - \gamma_2|^+ F_1\right. \\
&\quad \left. + |\min\{N, \gamma_1t_2\} - s_2t_2 - s_1t_2|^+ F_2\right) - \epsilon\Theta(F) - \Theta(1), \qquad (38)
\end{aligned}
$$

where (36) comes from the fact that

$$
\begin{aligned}
&H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}, Z_{\bar{\mathcal{K}}}|\tilde{W}) \\
&\geq H(Z_{\bar{\mathcal{K}}}|\tilde{W}) + H(X_{\mathcal{D}_1}^{(1)}, X_{\mathcal{D}_2}^{(2)}|Z_{[1:\gamma_1]}, \tilde{W}),
\end{aligned}
$$

(37) comes from the fact that $\frac{s_1+s_2}{\gamma_1} \leq 1$ and Han's inequality [12] (or see [13, Theorem 17.6.1]):

$$
\frac{1}{\binom{\gamma_1}{s_1+s_2}} \sum_{\bar{\mathcal{K}}\subseteq[1:\gamma_1]:|\bar{\mathcal{K}}|=s_1+s_2} \frac{H(Z_{\bar{\mathcal{K}}}|\tilde{W})}{s_1+s_2} \geq \frac{H(Z_{[1:\gamma_1]}|\tilde{W})}{\gamma_1},
$$

and (38) comes from Assignment 3 and Lemma 2. Here, for a similar reason as the assignment of (29), positive integers are assigned for sets (31) and (32) in Assignment 3.

Since $\lim_{F\to\infty} \frac{F_1}{F} = p$, $\lim_{F\to\infty} \frac{F_2}{F} = (1-p)$, and $\epsilon > 0$ can be arbitrarily small, we have for any achievable pair $(M, \mathbf{R})$,

$$
\begin{aligned}
&t_1R_1 + s_1M + t_2R_2 + s_2M \\
&\geq s_2t_2 + s_1t_1p + |\min\{N, \gamma_2\} - \max\{s_1t_1, s_2t_2\}|^+ p \\
&\quad + |\min\{N, s_2t_2 + s_1t_2\} - s_2t_2|^+ (1-p) \\
&\quad + \frac{s_1+s_2}{\gamma_1} \left(|\min\{N, \gamma_3\} - \gamma_2|^+ p\right. \\
&\quad \left. + |\min\{N, \gamma_1t_2\} - s_2t_2 - s_1t_2|^+ (1-p)\right).
\end{aligned}
$$

Thus, for any $t_1 \in [0:N]$, $t_2 \in [1:N]$, and any $s_1, s_2 \in [0:K]$ such that $t_1s_1 \leq N$, $t_2s_2 \leq N$, and $s_1 + s_2 \leq \gamma_1$, we have

$$
\begin{aligned}
R_2 \geq \frac{1}{t_2}\Big(&s_2t_2 + s_1t_1p + |\min\{N, \gamma_2\} - \max\{s_1t_1, s_2t_2\}|^+ p \\
&+ |\min\{N, s_2t_2 + s_1t_2\} - s_2t_2|^+ (1-p) \\
&+ \frac{s_1+s_2}{\gamma_1} \left(|\min\{N, \gamma_3\} - \gamma_2|^+ p\right. \\
&\left. + |\min\{N, \gamma_1t_2\} - s_2t_2 - s_1t_2|^+ (1-p)\right) \\
&- (s_1+s_2)M - t_1R_1\Big).
\end{aligned}
$$

This implies that

$$
R_2 \geq r_2(p, K, M, N, R_1). \qquad (39)
$$

On the other hand, when $t_2 = 0$ and $s_2 = 0$, we have $\gamma_1 = \beta$, $\gamma_2 = s_1t_1$, and $\gamma_3 = \beta t_1$. Hence, for any $t_1 \in [1, N]$ and any $s_1 \in [0:K]$ such that $s_1t_1 \leq N$ and $s_1 \leq \beta$, we have

$$
R_1 \geq \frac{1}{t_1}\left(s_1t_1p + \frac{s_1}{\beta}|\min\{N, \beta t_1\} - s_1t_1|^+ p - s_1M\right).
$$

This implies that

$$
R_1 \geq r_1(p, K, M, N). \qquad (40)
$$

According to (39) and (40), the achievable rate region must satisfy (2). This completes the proof of Theorem 5.

6. Conclusion

In this paper, we have dealt with the caching system for content with a multilayer structure. We gave an inner bound (Theorem 3) and outer bounds (Theorems 4 and 5) on the achievable rate region. The inner bound was derived by employing the memory dividing scheme. Outer bounds were derived by employing the detailed assignments of parameters and Han's inequality [12]. We gave numerical examples of our bounds and showed that our bounds are optimal for some cases. We also showed that the server cannot always achieve the optimal rates of the region simultaneously in Figs. 5–7.

References

[1] T. Matsuta and T. Uyematsu, "Caching-aided multicast for partial information," Proc. IEEE Int. Symp. on Inform. Theory, pp.600–604, July 2016.
[2] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," IEEE Trans. Inf. Theory, vol.60, no.5, pp.2856–2867, May 2014.
[3] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," IEEE/ACM Trans. Net., vol.23, no.4, pp.1029–1040, Aug. 2015.
[4] N. Ajaykrishnan, N.S. Prem, V.M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," arXiv preprint arXiv:1501.02549.
[5] C. Tian, "A note on the fundamental limits of coded caching," arXiv preprint arXiv:1503.00010, 2015.
[6] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," arXiv preprint arXiv:1501.06003, 2015.
[7] A. Sengupta, R. Tandon, and T.C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," Proc. IEEE Int. Symp. on Inform. Theory, June 2015.
[8] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard," IEEE Trans. Circuits Syst. Video Technol., vol.17, no.9, pp.1103–1120, Sept. 2007.
[9] W. Equitz and T. Cover, "Successive refinement of information," IEEE Trans. Inf. Theory, vol.37, no.2, pp.269–275, March 1991.
[10] C. Tian and S.N. Diggavi, "On multistage successive refinement for Wyner–Ziv source coding with degraded side informations," IEEE Trans. Inf. Theory, vol.53, no.8, pp.2946–2960, Aug. 2007.
[11] Q. Yang and D. Gündüz, "Centralized coded caching for heterogeneous lossy requests," Proc. IEEE Int. Symp. on Inform. Theory, pp.405–409, July 2016.
[12] T.S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," Information and Control, vol.36, no.2, pp.133–156, Feb. 1978.
[13] T.M. Cover and J.A. Thomas, Elements of Information Theory, 2 ed., Wiley, New York, 2006.

Tetsunao Matsuta was born in Fukui, Japan, on March 17, 1985. He received the B.E. degree from Utsunomiya University in 2007. He received the M.E. degree, and the Dr.Eng. degree from Tokyo Institute of Technology, Japan, in 2009, 2012, respectively. Since 2012, he has been an assistant professor at Tokyo Institute of Technology, and currently he is in the Department of Information and Communications Engineering of Tokyo Institute of Technology. He received the Best Paper Award in 2016 from IEICE. His current research interests are in the area of multi-terminal information theory and non-asymptotic analysis of coding problems.

Tomohiko Uyematsu received the B.E., M.E. and Dr.Eng. degrees from Tokyo Institute of Technology in 1982, 1984 and 1988, respectively. From 1984 to 1992, he was with the Department of Electrical and Electronic Engineering of Tokyo Institute of Technology, first as research associate, next as lecturer, and lastly as associate professor. From 1992 to 1997, he was with School of Information Science of Japan Advanced Institute of Science and Technology as associate professor. Since 1997, he returned to Tokyo Institute of Technology as associate professor, and currently he is with the Department of Information and Communications Engineering as professor. In 1992 and 1996, he was a visiting researcher at the Supelec (Ecole superieure d'electricite), France and Delft University of Technology, Netherlands, respectively. He was an associate editor of IEEE Trans. Information Theory from 2010 to 2013. He received the Achievement Award in 2008, and the Best Paper Award in 1993, 1996, 2002, 2007, 2011, 2014 and 2016 both from IEICE. His current research interests are in the areas of information theory, especially Shannon theory and multi-terminal information theory. Dr. Uyematsu is a senior member of IEEE.