Performance Limits of Coded Caching in Time Constrained Networks

Qitian Wu
Shanghai Jiao Tong University
Shanghai, China
echo740@sjtu.edu.com

Jinbei Zhang
Shanghai Jiao Tong University
Shanghai, China
abelchina@sjtu.edu.cn

ABSTRACT

Coded caching can dramatically reduce the transmission rate of multiple users by exploiting their multicasting opportunities, which has aroused wide attention recently. A common assumption of previous works is that the channels of all the users are of the same quality and in good condition. In practice, due to the interference of wireless networks or energy constraints, some users may have worse channels, leading to less active time. In this work, we focus on the performance limits of coded caching when users have distinct active time slots. To characterize the heterogeneity and variation of channel condition, we introduce time constrained model and two simplified models, depending on distributions of users' active time slots. For each model, we propose a class of caching and delivery schemes, aiming to minimize the transmission delay as well as guarantee the effectiveness of coded caching. We also derive the theoretical performance lower bounds using cut-set bound arguments. Furthermore, we prove that the gap between upper bound and lower bound will not vary by a factor of 24, and is independent of any model parameter, which indicates that our proposed schemes are order-optimal. Finally, the numerical results validate our theoretical findings and the efficiency of the proposed schemes.

KEYWORDS

Coded Caching, Channel Heterogeneity, Time Constrained Network, Performance Analysis

1 INTRODUCTION

With dramatically increasing demand for video streaming, the global data traffic has grown 18-fold over the past five years [16]. However, the speed of current mobile network is difficult to keep the same pace, leading to more congestions on shared link in the network. To solve this problem, caching schemes [14][13] are proposed to alleviate the network traffic pressure by prefetching some contents beforehand.

With traditional caching schemes, one node can only have the benefit brought by its own storage, which is called *local caching gain*. Recently, [11] introduced an innovative caching scheme, called *coded caching*, which can exploit multicasting opportunities over different users, even when they request different files. Hence, coded caching can significantly reduce the pressure on shared link during peak hours. For example, consider two files $F_1 = \{F_1^1, F_1^2\}$ and $F_2 = \{F_2^1, F_2^2\}$. User A caches F_1^1 and F_2^1 while user B caches F_1^2 and F_2^2. Then if user A requests file F_1 and user B requests file F_2, the server can send one XOR signal $F_2^1 \oplus F_1^2$ and each user can decode the requested file by doing XOR operation between the received signal and the cached contents. In other words, one transmission serves two users. As the number of users or the storage size becomes larger, this multicasting opportunities will grow almost linearly, which is called *global caching gain* [11].

Due to the introduction of the global caching gain, [11] has inspired a series of works. However, a common assumption of these works is that the channel of each user is of the same quality or users can always be active. Under this assumption, each user can receive intact signals from the server at any time. Unfortunately, in practice, some users in the network may suffer from weak links, making them unable to receive signals from the server. For example, in wireless networks, channels of users may be interfered by changeable surroundings, resulting in the 'on' and 'off' states alternately. Furthermore, if the channels of users possess distinguished power, users with deficient power may have fewer 'on' states compared with users with sufficient power. Under these circumstances, the schemes proposed in previous works may not be effective, since the broadcasted signals can not be received by all the users. The failure of receiving signals could disable users to recover their requested file and deprive the algorithm of effectiveness. Thus, a new caching and delivery scheme is needed.

Nevertheless, there are two challenges making this problem non-trivial. First, since the 'off' states make users fail to receive signals, the caching and delivery scheme must guarantee that each user can receive the needed contents and retrieve the requested file, which we call *effectiveness*. At the same time, the scheme should have a small delay to satisfy users' requests, which we call *efficiency*. It is challenging to achieve both effectiveness and efficiency when users suffer from distinct channel conditions. Second, the channel state of each user may vary as time goes by. The random variation of channel state is difficult to characterize, thus making it hard to analyze the performance of caching scheme.

In this study, we firstly propose a time constrained model to characterize the heterogeneity and variation of channel states from the perspective of active time slot. Specifically, we split the time interval into tiny time slots and each time slot has two states, active or inactive. Users at inactive time slot can not receive any signal. We use the number of active time slots for users to quantify channel heterogeneity and use different distributions of active time slots to characterize the state variation. Since it is challenging to directly probe into time constrained model, we propose deterministic overlapped model, in which the active time slots of users with weak channels are contained in that of users with strong channels, and deterministic non-overlapped model, in which users with different channel conditions share no active time slots. These two simplified models play as two extreme cases of the time constrained model and help to shed some light on further study.

To accommodate the channel state heterogeneity and variation,

we adaptively design new coded caching schemes for two simplified models and then generalize the algorithm to time constrained scenario. To put it specifically, during delivery stage, the signals are carefully separated into multiple sets, depending on the user set interested in the contents. If users with worse channel condition are involved, their interested signals will be allocated with time slots in priority, in order to fully utilize channel capacity. The signal set separation strategy does not violate the effectiveness of the algorithm, while the time slot allocation strategy guarantees the efficiency. To evaluate the performance of proposed schemes, we derive the achievable upper bound in the worst case and the theoretical lower bound of transmission delay by using cut-set bound arguments. Then we prove the gap between two performance limits within a constant 24 and independent of any model parameter as well as the distribution of active time slots, which indicates that the proposed schemes are order-optimal. Finally, we conduct numerical evaluation to validate our theoretical findings. The results reveal that the proposed scheme performs better than other straightforward schemes.

Our main contributions are summarized as follows:

- We propose a time constrained model with two simplified models to study the heterogeneity and variation of channel state from a perspective of active time slot.
- We adaptively design a new coded caching schemes for each scenario and prove its effectiveness mathematically.
- We provide a rigorous analysis on the achievable upper bounds of algorithm performance, derive the theoretical lower bounds, and then prove the constant gap between two limits, which demonstrates the order-optimality of our proposed schemes.

The remainder of the paper is organized as follows. In Section 2, we study some related literatures concerning coded caching. Then we formulate the problem and present our time constrained model in Section 3. In Section 4, we propose our new coded caching scheme for time constrained model. Then we illustrate the performance analysis of our proposed scheme in Section 5. Section 6 shows the numerical results. Finally, we conclude.

2 RELATED WORK

There are some related literatures concerning coded caching after [11]. Some of them conduct further analysis on performance upper bound [3] or lower bound [5] of coded caching. Some works extend the model to more practical settings, such as nonuniform demands[15], files with distinct sizes [7], files with arbitrary popularity [8], files with finite length [9], requests for partial files[21], finite file packetization [24] and online caching [17]. Moreover, some works adapt coded caching scheme to more complicated network architectures. [6] considers multi-level networks and [1] considers hybrid networks with mirrors between server and users. [18] introduces multi-server networks.

The above works all assume that the channel of each user is always in good condition and they can receive the signal at any time. By removing this assumption, some recent studies make an attempt to study the heterogeneity and variation of channel state from different perspectives. [10] considers random fading channel in multi-antenna wireless network and proposes a scheme that carefully combines the multicast and unicast capabilities offered by MIMO. [4] focuses on packet-erasure channel with the probability of information loss and presents the lower and upper bounds on the capacity-memory tradeoff of this network. However, these two studies both assume that the variation of channel condition is of the same probability distribution among the weak users. In reality, different weak users may suffer from different channel conditions, which makes it significant to probe into state variation with distinct probability distributions. Furthermore, [23] studies the situation where some users have higher link capacities than others, while [20] considers channel energy deficiency slowing transmission rate. Users with low link capacity or deficient energy may spend more time receiving signals. Differently, in our model, weak users may experience inactive time slots at which they can receive nothing from the server. We study the channel state variation from the persective of different active time slot distributions, instead of different transmission rates.

Moreover, [2] considers users with different link rates and packet loss rates and design a cache size allocation strategy, where a user with weaker channel condition needs to be allocated with more cache memory, to guarantee the transmission efficiency. Nevertheless, since the cache size allocation is conducted at caching stage, the effectiveness of this strategy requires an assumption that the channel state of each user remains unchanged from caching stage to delivery stage, or the channel state at delivery stage can be known in advance at caching stage, which may not be feasible in reality. To address the aforementioned limitations, in this work, we study the channel state heterogeneity and variation from the perspective of active time slot distributions. Our proposed model considers users with distinct channel conditions and weak users in the network may experience inactive time slots, at which they can receive nothing from the server. In our caching network, the channel conditions at delivery stage can be arbitrary and unknown when the server places the cached content.

3 PROBLEM FORMULATION AND TIME CONSTRAINED MODEL

In this section, we firstly formulate the problem and then propose our time constrained model.

3.1 Problem Formulation

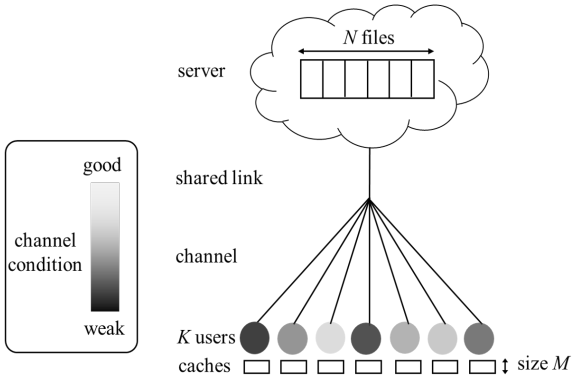

Figure 1: Caching network.

In the caching network, there are N files with unit file size, all of which are stored in a server. K users are connected to the server through a broadcast channel, each of which has a storage sized M. An illustration is presented in Fig. 1. Within off-peak hours, each user can cache part of contents in the local storage in order to reduce the pressure on the shared link during peak hours, which is called *caching stage*. The cached content for user k is denoted by M_k and its entropy is defined as $H(M_k)$, which represents the information contained in M_k [19]. Later within peak hours, each user requests one file and the K requested files are denoted by a K-dimension vector $\vec{F} = \{F_1, F_2, \cdots, F_K\}$, called request pattern, where F_k denotes the file requested by user k. For each request pattern $\vec{F}$, the server should send a set of signals R to satisfy users' requests. This is called *delivery stage*. In the sequel, user k will receive a set of signals $\hat{R}_k$, which is a subset of R. Note that different users may receive different signals since the channel conditions are different. With the received signals and locally cached contents, user k should be able to reconstruct the requested file, formally denoted as $H(F_k|\hat{R}_k, M_k) = 0$, which means information in $\hat{R}_k$ and M_k contains information in F_k. The formal definition of *effectiveness* is given as follows.

Definition 1. *(Effectiveness): For any given request pattern $\vec{F} = \{F_1, F_2, \cdots, F_K\}$, the caching and delivery scheme is effective if and only if*

$$H(F_k|\hat{R}_k, M_k) = 0, k = 1, 2, \cdots, K.$$

This definition is similar to that of the worst-case scenario in [11]. Then we consider transmission delay, which characterizes the *efficiency* in Section 1.

Definition 2. *(Transmission delay): For one request pattern $\vec{F}$, the transmission delay is from the time when users request files to the time when each user recovers the requested file, denoted by $D(\vec{F})$ or D.*

We now present the main objective of this paper. Our goal is to design a caching and delivery scheme, i.e., placement of M_k for each user and delivery of R for the server, such that the maximal transmission delay $D(\vec{F})$ among all possible request patterns $\vec{F}$ can be minimized. Within the delay $D(\vec{F})$, each user will receive content $\hat{R}_k$ and recover its requested file along with the cached contents M_k. The mathematical form can be written as:

$$\begin{aligned} \min \quad & \max_{\vec{F}} D(\vec{F}) \\ s.t. \quad & H(F_k|\hat{R}_k, M_k) = 0, \forall k, \forall F_k. \end{aligned}$$

3.2 Time Constrained Model

The above problem is non-trivial since the channel heterogeneity and variation do influence both the effectiveness and efficiency of coded caching algorithm. To overcome the challenge, we firstly model the channel heterogeneity from a new perspective. We split the time interval into time slots, each of which has a unit time size. For time slot t, use a_t to denote its state.[1] Each time slot has two states, *active* or *inactive* and here is the definition:

[1]We assume that the channel state can be known through Channel State Information Feedback (CSIF) system in the network [22].

Definition 3. *(Active time slot): If the signal sent by the server in time slot t can be received by user k, we call time slot t is active for user k or user k is active in time slot t, denoted as $a_t^k = 1$. On the contrary, if user k can not receive any signal from the server in time slot t, we call time slot t is inactive for user k or user k is inactive in time slot t, denoted as $a_t^k = 0$.*

For simplicity of analysis, we introduce a parameter, called time cycle T:

Definition 4. *(Time cycle): During a time cycle, there are T time slots in total. For user k, the states of time slots in one cycle are denoted as $A_k = \{a_1^k, \cdots, a_T^k\}$ and the number of active time slots is $|A_k| = \sum_{t=1}^{T} a_t^k$.*

Note that T is equal to transmission delay $D(\vec{F})$ if time slot state has no repetitiveness. The number of active time slots in one cycle may vary from user to user since different users have different channel conditions. To tackle this challenge, we can aggregate the users with similar active time slots into a group. Given K users with arbitrary active time slots, there exists an integer m satisfying $\frac{T}{2^{m-1}} \leq \min_{1 \leq k \leq K} |A_k|$. We divide K users into m groups: for i-th group, $1 \leq i \leq m$, there are K_i users ($K_i \geq 0$) satisfying $\frac{T}{2^{i-1}} \leq |A_k| < \frac{T}{2^{i-2}}$.

To further proceed, we decrease $|A_k|$ to its lower bound in the group, denoted as $|A_k| = \frac{T}{2^{i-1}}$. Note that after this approximation, the transmission delay will not vary by a factor of two, since if the number of active time slots for each user increases by two times, the transmission delay will at most reduce by half. Since the number of active time slots must be an integer, we set $T = 2^{m-1}$. To keep notation neat, we use A_i to denote time slot states for users in i-th group. $|A_i|$ characterizes the channel state heterogeneity.

To further investigate the variation of channel states, we assume A_i is uniformly random in every time cycle. We call this model as *Time Constrained Model* (TC model). Then the problem is to design an effective and efficient coded caching algorithm for TC model.

4 CACHING AND DELIVERY SCHEMES

In this section, we present our caching and delivery schemes. We firstly introduce the caching scheme and then propose our new delivery algorithm for TC model.

In the caching stage, we adopt decentralized coded caching in [12]: each user independently caches a subset of $\frac{M}{N}$ bits of file n chosen uniformly at random, $n = 1, \cdots, N$. Then in the delivery stage, each user requests one file forming a request pattern $\vec{F} = \{F_1, \cdots, F_K\}$. For F_k requested by user k, we divide it into two parts $F_k = \{F_k^m, F_k^r\}$, where F_k^m denotes the parts cached by user k and F_k^r denotes the parts needed to be transmitted by the server. Use $[K] = \{1, 2, \cdots, K\}$ to denote the set of users and consider a subset $S \subset [K]$ with size $|S| = s$. Use $F_{k,S}^r$ ($k \in S$) to denote the parts of F_k^r cached by $s-1$ users in $S \setminus \{k\}$ but not cached by any user in $[K] \setminus S$. Since $F_k^r = \bigcup_S F_{k,S}^r$, the server can transmit all $F_{k,S}^r$ to help user k recover F_k in delivery stage. The advantage of decentralized coded caching is the multicasting gain. Since for user $k \in S$, $F_{k,S}^r$ is needed by user k and cached by users in $S \setminus \{k\}$, the server can send one XOR signal $\oplus_{k \in S} F_{k,S}^r$ and each user k in S will obtain the

needed content by doing $\oplus_{k\in S} F^r_{k,S} \oplus_{k\in S\setminus\{k\}} F^r_{k,S} = F^r_{k,S}$. Traverse all possible S, send signal set $R = \{\oplus_{k\in S} F^r_{k,S} | S \in [K]\}$ and user k will obtain F^r_k, then recover F_k.

However, if one user k in S is inactive, the transmitted signal $\oplus_{k\in S} F^r_{k,S}$ will not be received and user k is unable to recover the requested file. The randomness of A_i makes it difficult to figure out the missing part of signals, so it is challenging to probe into TC model directly. Therefore, we first consider two simplified models, which are corresponding to two extreme cases of TC model, to gain some insights and then generalize the analysis to TC model.

- Deterministic overlapped model (DO model): For i-th group, A_i is fixed in every cycle and active time slots of different groups are overlapped as much as possible. Specifically, for $2 \le i \le m$, $a^i_t = 1$, $2^{m-1} - 2^{m-i} + 1 \le t \le 2^{m-1}$ and $a^i_t = 0$, $1 \le t \le 2^{m-1} - 2^{m-i}$. Particularly, $a^1_t = 1$, $1 \le t \le 2^{m-1}$.
- Deterministic non-overlapped model (DN model): For i-th group, A_i is fixed in every cycle and active time slots of different groups are overlapped as little as possible. For $2 \le i \le m$, $a^i_t = 1$, $2^{m-1} - 2^{m-i+1} + 2 \le t \le 2^{m-1} - 2^{m-i} + 1$ and $a^i_t = 0$, otherwise. Particularly, $a^1_t = 1$, $1 \le t \le 2^{m-1}$.

4.1 Proposed Scheme for DO model

In DO model, active time slots of each group are overlapped as much as possible. For example, consider three user groups and we have $A_1 = (1, 1, 1, 1)$, $A_2 = (0, 0, 1, 1)$, $A_3 = (0, 0, 0, 1)$. Time slot 4 is active for three groups and time slot 3 is active for group 1 and group 2. These overlaps create opportunities to accomplish global caching gain with multicast transmission. However, the inactive time slots will limit this multicast transmission since inactive users can not receive any signal. In order to solve this problem, we propose a new delivery scheme including two strategies, which we call *signal set separation* and *time slot allocation*. Let us consider an example.

Example: Consider three user groups with user number K_1, K_2, K_3. Note that $[K_1]$, $[K_2]$, $[K_3]$, which represent the user set of group 1, 2, 3 respectively, are disjoint subsets of $[K]$ and $[K_1] \cup [K_2] \cup [K_3] = [K]$.

Signal Set Separation. We separate the transmitted signal set R into three parts, $R_1 = \{\oplus_{k\in S} F^r_{k,S} | S \subset [K_1]\}$, $R_2 = \{\oplus_{k\in S} F^r_{k,S} | S \cap [K_2] \neq \emptyset, S \cap [K_3] = \emptyset\}$, $R_3 = \{\oplus_{k\in S} F^r_{k,S} | S \cap [K_3] \neq \emptyset\}$. In this way, R_1 only contains the content needed by users in group 1 (i.e., $[K_1]$), R_2 is needed by at least one user in $[K_2]$ but no user in $[K_3]$, and R_3 is needed by at least one user in $[K_3]$.

Time Slot Allocation. Then we focus on how to deliver these signal sets. Inspired by time-sharing strategy, we can transmit different signal sets in different time slots in one cycle. The transmission of R_1, R_2 and R_3 includes three steps and we allocate four time slots in one cycle in the following way. At the first step, allocate time slot 4, which is active for three groups, to R_3, allocate time slot 3, which is active for group 1, 2, to R_2, and allocate the remaining two time slots to R_1. Then there will be one signal set finishing transmission and we call the current step is over. Assume R_2 finishes transmission and in the second step, we allocate time slot 4 to R_3 and the remaining three time slots to R_1. Then if R_3 is finished, we will allocate all time slots to R_1 in the last step. The intuitive idea behind this allocation is to give priority to signal set needed by users with the least active time slots, since they play as the bottleneck throughout the transmission. This strategy can help to fully utilize active time slots and improve the efficiency. In fact, there are totally six cases according to the order of signal set finishing transmission. Fig. 2 shows specific time slot allocation in each step and the six arrow paths are corresponding to six cases.

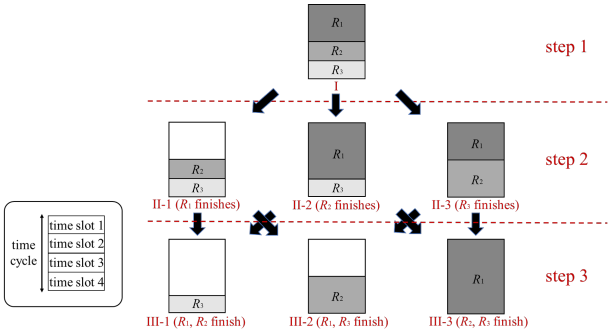

Figure 2: An example of time slot allocation in DO model with three user groups.

Now, we present the formal caching and delivery algorithm. In

Algorithm 1: Achievable Scheme in DO model

1 **for** $k = 1 : K, n = 1 : N$ **do**
2 User k caches $\frac{M}{N}$ parts of file n uniformly at random.
3 **end**
4 **for** $i = 1 : m$ **do**
5 $R_i \leftarrow \emptyset$;
6 **for** *every* $S \subset \cup_{j=1}^{i} [K_j]$ **do**
7 $R_i \leftarrow R_i \cup \{\oplus_{k\in S} F^r_{k,S} | S \cap [K_i] \neq \emptyset\}$
8 **end**
9 **end**
10 **repeat**
11 **for** $t = 1 : 2^{m-1}$ **do**
12 $a = \max_{A^i_t = 1, R_i \neq \emptyset} i$;
13 **transmit** one XOR signal F_{XOR} in R_a;
14 $R_a \leftarrow R_a \setminus \{F_{XOR}\}$
15 **end**
16 **until** $\cup_{i=1}^{m} R_i = \emptyset$;

Alg. 1, line 1-3 is decentralized coded caching, line 4-9 is *signal set separation* and line 10-16 is *time slot allocation*. Next, we prove our proposed scheme is effective.

Proposition 1. *(Effectiveness): Achievable scheme in Alg. 1 is effective in DO model.*

Proof. We first prove that each user can receive every needed XOR signal. This statement is true if for each user $k \in [K_i]$, $\hat{R}_k = R_j$ is satisfied for $j \ge i$. We prove it by contradiction and assume that one user k, $k \in [K_i]$, fails to receive signal $\oplus_{k\in S} F^r_{k,S} \in R_j$,

$i \le j \le m$. According to line 12 and line 13 in Alg. 1, if $\oplus_{k\in S} F^r_{k,S}$ is transmitted in time slot t, there must exist p satisfying $A^p_t = 1$ and $j \le p \le m$. Then according to the definition of DO model and $i \le j \le p$, we have $A^i_t = 1$. Since $k \in [K_i]$, user k is active in time slot t and thus can receive the transmitted signal, which is contradictory to the assumption. Thus, for each user k, $\oplus_{k\in S} F^r_{k,S} \in \hat{R}_k$.

Then, receiving $\oplus_{k\in S} F^r_{k,S}$, user k can obtain $F^r_{k,S}$. Since $\bigcup_S F^r_{k,S} = F^r_k$ and $F^r_k \cup F^m_k = F_k$, we have $H(F_k|\hat{R}_k, M_k) \le H(F_k|F^r_k, F^m_k) = 0$. So $H(F_k|\hat{R}_k, M_k) = 0$. The effectiveness is proved. □

4.2 Proposed Scheme for DN model

Different from DO model, in DN model, active time slots of each group are overlapped as little as possible. Specifically, there only exist overlaps between A_1 and $A_i, i = 2, \cdots, m$. For example, consider three user groups and we have $A_1 = (1, 1, 1, 1), A_2 = (0, 1, 1, 0), A_3 = (0, 0, 0, 1)$. Since no time slot is active for all the users, multicasting among all groups may not be feasible in this model. On the other hand, since active time slots of each group are disjoint except the first group, the signal sets serving different groups can be transmitted exclusively, which can shorten the delay to a certain degree.

The key idea of caching and delivery scheme in DN model is similar to that in DO model. We combine the first group and one of other groups as a clique for multicasting and serve the remaining groups independently. To formulate the combination, we define a new concept.

Definition 5. *(Clique combination $C[1, g]$): If we adopt clique combination $C[1, g]$, we combine $[K_1]$ and $[K_g]$ as a clique and in delivery stage the signal sets to be transmitted are composed of 1)$R'_{1,g} = \{\oplus_{k\in S} F^r_{k,S} | S \subset [K_1] \cup [K_g]\}$, 2)$R'_i = \{\oplus_{k\in S} F^r_{k,S} | S \subset [K_i]\}$, $i = 2, \cdots, g-1, g+1, \cdots, m$.*

There are $m - 1$ different clique combinations in total. Next we conduct *signal set separation* and *time slot allocation*. Also, we start from an instance.

Example: Consider three user groups and there are two clique combination $C[1, 2]$, $C[1, 3]$. We take $C[1, 3]$ as an example.

Signal Set Separation. We separate $R'_{1,3}$ into two parts, $R'_1 = \{\oplus_{k\in S} F^r_{k,S} | S \subset [K_1]\}$, $R'_3 = \{\oplus_{k\in S} F^r_{k,S} | S \cap [K_3] \neq \emptyset, S \cap [K_2] = \emptyset\}$. Here, R'_1 only contains the contents needed by users in $[K_1]$ and R_3 is needed by at least one user in $[K_3]$ but no user in $[K_2]$. Besides, R'_2 is only needed by user in $[K_2]$ because we serve group 2 independently.

Time Slot Allocation. At the first step, we do allocation in this way: allocate time slot 4, which is active for group 1 and group 3, to R'_3, allocate time slot 2 and 3, which are active for group 2, to R'_2 and allocate time slot 1 to R'_1. Fig. 3 shows specific time slot allocation in the following steps and the six arrow paths are corresponding to six cases.

For $C[1, 2]$, the delivery scheme is similar. But there is a problem: which clique combination is better between $C[1, 2]$ and $C[1, 3]$? We can estimate the achievable transmission delay and pick more efficient one. The derivation of achievable transmission delay will be presented in Section 5. Now we formulate the achievable scheme in Alg. 2. In Alg. 2, line 1-3 is decentralized coded caching, line 4-14 is *signal set separation* and line 15-21 is *time slot allocation*. Next, we prove the effectiveness of the proposed scheme.

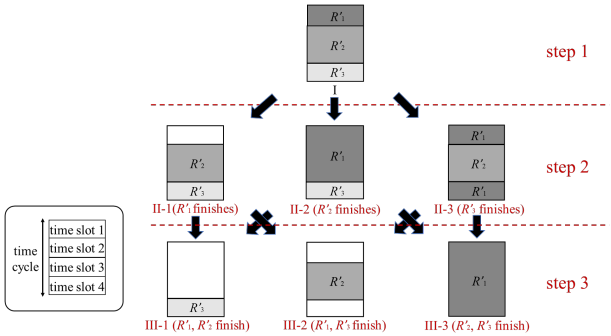

Figure 3: An example of time slot allocation in DN model with three user groups.

Algorithm 2: Achievable Scheme in DN model

1 **for** $k = 1 : K, n = 1 : N$ **do**
2 | User k caches $\frac{M}{N}$ parts of file n uniformly at random.
3 **end**
4 Select $C[1, g]$ with lowest achievable transmission delay;
5 **for** $i = 1 : m$ *and* $i \neq g$ **do**
6 | $R_i \leftarrow \emptyset$;
7 | **for** *every* $S \subset [K_i]$ **do**
8 | | $R_i \leftarrow R_i \cup \{\oplus_{k\in S} F^r_{k,S}\}$
9 | **end**
10 **end**
11 $R_g \leftarrow \emptyset$;
12 **for** *every* $S \subset [K_1] \cup [K_g]$ **do**
13 | $R_g \leftarrow R_g \cup \{\oplus_{k\in S} F^r_{k,S} | S \cap [K_g] \neq \emptyset\}$
14 **end**
15 **repeat**
16 | **for** $t = 1 : 2^{m-1}$ **do**
17 | | $a = \max_{A^i_t = 1, R_i \neq \emptyset} i$;
18 | | **transmit** one XOR signal F_{XOR} in R_a;
19 | | $R_a \leftarrow R_a \setminus \{F_{XOR}\}$
20 | **end**
21 **until** $\cup^m_{i=1} R_i = \emptyset$;

Proposition 2. *(Effectiveness): Achievable scheme in Alg. 2 is effective in DN model.*

Proof. The proof is also based on contradiction, which is similar to that in Proposition 1. □

4.3 Generalization to TC model

We have presented solutions to DO and DN model. Now we can apply the proposed schemes to a more practical scenario, i.e., time constrained model. According to the definition, in TC model, active time slots of $[K_i]$ may vary uniformly and randomly in different cycle. Despite the variation, there always exist overlaps among

users in $[K_1]$ and one of other groups. Therefore, we can adopt the scheme in Alg. 2.

5 PERFORMANCE ANALYSIS

To evaluate our proposed schemes in time constrained network, we now determine the performance limits of the transmission delay of the algorithms. In this section, we aim to provide a theoretical analysis on the achievable performance upper bound and theoretical performance lower bound. Then by proving the gap between two limits remains a constant factor, which is independent of any model parameter, we can prove that our proposed schemes are order-optimal. For the analysis on two performance limits and the order-optimality, we firstly handle DO and DN model and then generalize the conclusions to TC model.

5.1 Upper Bound of Transmission Delay

We first probe into DO model. Also consider three user groups as an example. According to line 2 in Alg. 1, one bit of file F_k is in $F^r_{k,S}$ with probability $(\frac{M}{N})^{s-1}(1-\frac{M}{N})^{K-s+1}$. If one file has B bits and B is large enough, then by the Law of Large Number, the expected bits of $F^r_{k,S}$ will be $B(\frac{M}{N})^{s-1}(1-\frac{M}{N})^{K-s+1}$. For simplicity, we normalize the size of $F^r_{k,S}$ as $(\frac{M}{N})^{s-1}(1-\frac{M}{N})^{K-s+1}$. According to the constitution of R_1, R_2, R_3 and summing over all s as well as all subsets S, we have

$$\begin{aligned} |R_1| &= \sum_{s=1}^{K_1} C^s_{K_1} \cdot \left(\frac{M}{N}\right)^{s-1} \left(1-\frac{M}{N}\right)^{K-s+1} \\ &= (1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_1}}{\frac{M}{N}} \cdot \left(1-\frac{M}{N}\right)^{K_2+K_3}, \end{aligned} \tag{1}$$

$$\begin{aligned} |R_2| &= \sum_{s=1}^{K_1+K_2} C^s_{K_1+K_2} \cdot \left(\frac{M}{N}\right)^{s-1} \left(1-\frac{M}{N}\right)^{K-s+1} - |R_1| \\ &= (1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_2}}{\frac{M}{N}} \cdot \left(1-\frac{M}{N}\right)^{K_3}, \end{aligned} \tag{2}$$

$$\begin{aligned} |R_3| &= \sum_{s=1}^{K} C^s_{K} \cdot \left(\frac{M}{N}\right)^{s-1} \left(1-\frac{M}{N}\right)^{K-s+1} - |R_1| - |R_2| \\ &= (1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_3}}{\frac{M}{N}}. \end{aligned} \tag{3}$$

Then we derive the transmission delay. For simplicity, we normalize the bandwidth of broadcast channel as 1. According to Fig. 2, there are six arrow paths from step 1 to step 3 corresponding to six cases. For path I→II-1→III-1 and path I→II-2→III-1, R_3 occupies time slot 4 throughout the transmission, so the transmission delay $D = \frac{|R_3|}{1/4} = 4|R_3|$. For path I→II-1→III-2 and path I→II-3→III-2, R_2 and R_3 share time slot 3 and 4 throughout the transmission, so $D = \frac{|R_2|+|R_3|}{1/2} = 2(|R_2|+|R_3|)$. Finally, for path I→II-2→III-3 and path I→II-3→III-3, R_1, R_2, R_3 share all time slots throughout the transmission, so $D = \frac{|R_1|+|R_2|+|R_3|}{1} = |R_1|+|R_2|+|R_3|$. The performance will be limited by the worst case,

$$D \le \max\{|R_1|+|R_2|+|R_3|, 2(|R_2|+|R_3|), 4|R_3|\}. \tag{4}$$

(4) gives the upper bound for 3 groups, and we generalize the results to m groups in the following lemma.

Lemma 1. *For m user groups $[K_1]$, $[K_2]$,$\cdots$, $[K_m]$ in DO model, the transmission delay D should be upper bounded by*

$$D \le D_{DO} = \max\{\sum_{i=1}^{m}|R_i|, 2\sum_{i=2}^{m}|R_i|, \cdots, 2^{m-1}|R_m|\} \tag{5}$$

where $R_1, R_2, \cdots, R_m$ are signal sets constituted by Alg. 1.

Proof. We adopt mathematical induction to prove this lemma since we have obtained the results for 3 groups, i.e., $m = 3$. First, we assume achievable upper bound for $(m-1)$ groups is

$$\begin{aligned} D &\le D^{(m-1)}_{DO} \\ &= \max\{\sum_{i=1}^{m-1}|R^{(m-1)}_i|, 2\sum_{i=2}^{m-1}|R^{(m-1)}_i|, \cdots, 2^{m-2}|R^{(m-1)}_{m-1}|\}. \end{aligned} \tag{6}$$

Then we aim to prove the achievable upper bound for m groups is (6). Note that for m groups, time cycle $T^{(m)} = 2^{m-1}$, compared with $T^{(m-1)} = 2^{m-2}$ for $(m-1)$ groups.

Since there are m signal sets to be transmitted, the time slot allocation includes m steps. In each step, one signal set finishes transmission. We focus on the last step and consider two situations. One is that there remains R_1 in the last step, the other is that there remains one of R_i, $2 \le i \le m$. For the former situation, since R_1 can be transmitted at any time slot according to line 12 and line 13 in Alg. 1, these m signal sets can share all time slots throughout transmission. Hence, we obtain the delay

$$D^{(m)}_{DO} = \frac{\sum_{i=1}^{m} R^{(m)}_i}{1}. \tag{7}$$

For the latter situation, we make use of the particularity of active times slots distribution in DO model. The time slot states of i-th group, $2 \le i \le m$, in the second half of time cycle for m groups can be mapped into the time slot states of j-th group, $1 \le j \le m-1$, in one time cycle for $(m-1)$ groups. The similarity of active time slots leads to the similarity of signal set transmission according to time slot allocation in our algorithm. Thus, we can treat $R^{(m)}_2, R^{(m)}_3, \cdots, R^{(m)}_m$ as $R^{(m-1)}_1, R^{(m-1)}_2, \cdots, R^{(m-1)}_{m-1}$ in (6) and since $T^{(m)} = 2T^{(m-1)}$, the total transmission delay will be two times the delay for $(m-1)$ groups,

$$\begin{aligned} D^{(m)}_{DO} &\le 2D^{(m-1)}_{DO} \\ &= 2\cdot\max\{\sum_{i=2}^{m}R^{(m)}_i, 2\sum_{i=3}^{m}R^{(m)}_i, \cdots, 2^{m-2}R^{(m)}_m\}. \end{aligned} \tag{8}$$

Combining (7) and (8) we obtain Proposition 2. □

Then we focus on the upper bound for DN model. We consider three user groups with $C[1,3]$ as an example. Firstly, we calculate the size of each signal set,

$$|R'_1| = (1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_1}}{\frac{M}{N}} \cdot \left(1-\frac{M}{N}\right)^{K_3}, \tag{9}$$

$$|R'_2| = (1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_2}}{\frac{M}{N}}, \tag{10}$$

$$|R_3'| = (1 - \frac{M}{N})\frac{1 - (1 - \frac{M}{N})^{K_3}}{\frac{M}{N}}. \tag{11}$$

For $C[1,2]$, the size of signal sets can be calculated similarly. Then we derive the transmission delay. According to Fig. 3, there are six cases. For path I→II-1→III-1 and path I→II-2→III-1, R_3' occupies time slot 4 throughout the transmission, so $D = \frac{|R_3'|}{1/4} = 4|R_3'|$. For path I→II-1→III-2 and path I→II-3→III-2, R_2' occupies time slot 2 and 3 throughout the transmission, so $D = \frac{|R_2'|}{1/2} = 2|R_2'|$. Finally, for path I→II-2→III-3 and path I→II-3→III-3, R_1', R_2', R_3' share all time slots throughout the transmission, so $D = \frac{|R_1'|+|R_2'|+|R_3'|}{1} = |R_1'| + |R_2'| + |R_3'|$. The performance will be limited by the worst case,

$$D \le \max\{|R_1'| + |R_2'| + |R_3'|, 2|R_2'|, 4|R_3'|\}. \tag{12}$$

Finally, the achievable delay should be the minimum among two clique combinations:

$$D \le \min_{C[1,g]} \max\{|R_1'| + |R_2'| + |R_3'|, 2|R_2'|, 4|R_3'|\}. \tag{13}$$

Next, we generalize the upper bound to m user groups in the next lemma.

Lemma 2. *For m user groups $[K_1]$, $[K_2]$,$\cdots$, $[K_m]$ in DN model, the transmission delay D should be upper bounded by*

$$D \le D_{DN} = \min_{C[1,g]} \max\{\sum_{i=1}^{m} |R_i'|, 2|R_2'|, \cdots, 2^{m-1}|R_m'|\}$$

where $R_1', R_2', \cdots, R_m'$ are signal sets constituted by clique combination $C[1,g]$ and $2 \le g \le m$.

Proof. Based on Alg. 2, there will be at most m steps in time slot allocation. Consider the last step. If R_i', $2 \le i \le m$, is the last signal set to be transmitted, the total delay will be dominated by this signal set and R_i' occupies 2^{m-i} time slots in one cycle throughout the transmission, i.e., $D_{DN} = \frac{2^{m-1}|R_i'|}{2^{m-i}} = 2^{i-1}|R_i|$. If there remains R_1' in the last step, these m signal sets share all time slots throughout transmission and the total delay should be $D_{DN} = \frac{\sum_{i=1}^{m}|R_i'|}{1} = \sum_{i=1}^{m}|R_i'|$. In summary, we can obtain the result in this proposition.

□

Based on the analysis of DO and DN model, then we aim to determine the achievable upper bound for TC model.

Theorem 1. *(Upper bound): For m user groups $[K_1]$, $[K_2]$,$\cdots$, $[K_m]$ in TC model, the transmission delay D should be upper bounded by*

$$D \le D_{TC} = \min_{C[1,g]} \max\{\sum_{i=1}^{m} |R_i'|, 2\sum_{i=2}^{m} |R_i'|, \cdots, 2^{m-1}|R_m'|\}$$

where $R_1', R_2', \cdots, R_m'$ are signal sets constituted by clique combination $C[1,g]$ and $2 \le g \le m$.

Proof. According to line 17 to line 19 in Alg. 2, transmission of R_i' is prior to R_j', $1 \le j < i$, but inferior to R_k', $i < j \le m$.

First we consider R_m' with the highest priority. Since $|A_m| = 1$, R_m' has only one time slot to transmit in each cycle, and the delay is

$$D_{TC}^{(m)} = 2^{m-1}|R_m'|. \tag{14}$$

Second, we consider R_{m-1}' as well as R_m'. There are totally $|A_{m-1}| = 2$ active time slots for $[K_{m-1}]$. Assume the probability of A_m and A_{m-1} overlapping is equal to α ($0 < \alpha < 1$). If there exists overlap, R_{m-1}' will have only one time slot to transmit since R_m' occupies another time slot. Or else, it will have two time slots. On condition that the transmission of R_{m-1}' is finished before R_m', the total delay will be $D_{TC}^{(m)}$. If R_m' is finished before R_{m-1}', the total delay should be the sum of delay spent in transmission of R_m' and delay spent in transmitting the remainder of R_{m-1}'. So we have

$$\begin{aligned} D_{TC}^{(m-1)} &= D_{TC}^{(m)} + \frac{2^{m-1}[|R_{m-1}'| - \frac{1}{2^m}(2-\alpha)D_{TC}^{(m)}]}{2} \\ &= 2^{m-2}(|R_{m-1}'| + \alpha|R_m'|) \\ &\le 2^{m-2}(|R_{m-1}'| + |R_m'|). \end{aligned} \tag{15}$$

Combining (14) and (15), we have

$$D_{TC}^{(m-1)} \le \max\{2^{m-2}(|R_{m-1}'| + |R_m'|), 2^{m-1}|R_m'|\}.$$

Next, we use mathematical induction to conclude the proof. Assume the following inequality is correct:

$$\begin{aligned} D_{TC}^{(m-k)} \le \max\{&2^{m-k-1}\sum_{i=m-k}^{m}|R_i'|, \\ &2^{m-k}\sum_{i=m-k+1}^{m}|R_i'|, \cdots, 2^{m-1}|R_m'|\}. \end{aligned} \tag{16}$$

Then we aim to prove (17):

$$\begin{aligned} D_{TC}^{(m-k-1)} \le \max\{&2^{m-k-2}\sum_{i=m-k-1}^{m}|R_i'|, \\ &2^{m-k-1}\sum_{i=m-k}^{m}|R_i'|, \cdots, 2^{m-1}|R_m'|\}. \end{aligned} \tag{17}$$

Consider two cases in the last step in time slot allocation. First, if the remaining signal set is any one of R_i', $m-k \le i \le m$, the delay will be upper bounded by $D_{TC}^{(m-k)}$. Second, if R_{m-k-1}' remains in the last step, we divide R_i', $m-k \le i \le m$, into two parts

$$R_i' = R_{i,1}' \cup R_{i,2}',$$

where transmission of $R_{i,1}'$ occupies time slots which overlap A_{m-k-1} and transmission of $R_{i,2}'$ occupies time slots which do not overlap A_{m-k-1}. Throughout transmission, $R_{m-k-1}', R_{m-k,1}', R_{m-k+1,1}', \cdots, R_{m,1}'$ share $|A_{m-k-1}| = 2^{k+1}$ active time slots of $[K_{m-k-1}]$. Suppose that a time cycle consists of 2^{k+1} time slots, the total delay will be $|R_{m-k-1}'| + \sum_{i=m-k}^{m}|R_{i,1}'|$. But, for m user groups, $T = 2^{m-1}$, so

$$\begin{aligned} D_{TC}^{(m-k-1)} &= \frac{2^{m-1}}{2^{k+1}} \cdot (|R_{m-k-1}'| + \sum_{i=m-k}^{m}|R_{i,1}'|) \\ &\le 2^{m-k-2}\sum_{i=m-k-1}^{m}|R_i'|. \end{aligned}$$

Combining the two cases we have

$$D_{TC}^{(m-k-1)} \leq \max\{2^{m-k-2} \sum_{i=m-k-1}^{m} |R_i'|, D_{TC}^{(m-k)}\} \tag{18}$$

Substituting (16) into (18) we obtain (17). Thus, we conclude the proof. □

5.2 Lower Bound of Transmission Delay

In order to further evaluate the performance of proposed schemes, we need to determine the theoretical optimal transmission delay, i.e. the performance lower bound. Similarly, we first consider two simplified cases, DO and DN models. The results are formulated as the following two lemmas.

Lemma 3. *For m user groups $[K_1]$, $[K_2]$, $\cdots$, $[K_m]$ in DO model, the transmission delay D is lower bounded by*

$$D \geq D_{OD}^* = \max_{1\leq i\leq m} 2^{m-i} \cdot \max_{s_{m-i+1},\cdots,s_m} \{ \sum_{j=m-i+1}^{m} s_j - \frac{\sum_{j=m-i+1}^{m} s_j M}{\lfloor \frac{N}{\sum_{j=m-i+1}^{m} s_j} \rfloor} \},$$

where $s_i \in \{1, 2, \cdots, K_i\}, i = 1, \cdots, m$.

Proof. We mainly use cut-set bound arguments to prove this lower bound. First, consider two user groups, $[K_1]$ with all time slots being active and $[K_2]$ with only half of active time slots in one cycle.

If we only consider $[K_2]$, choose s_2 users in $[K_2]$ and construct $\lfloor \frac{N}{s_2} \rfloor$ request patterns, such that in each request pattern, each user requests a distinct file. For every request pattern, the server will transmit signal R^* to help users recover files. Combining all the received contents sized $\lfloor \frac{N}{s_2} \rfloor |R^*|$ with users' storage, all the requested files should be recovered. Therefore, we have

$$\left\lfloor \frac{N}{s_2} \right\rfloor |R^*| + s_2 M \geq \left\lfloor \frac{N}{s_2} \right\rfloor s_2.$$

Since only half of time slots are active for these s_2 users, transmission delay for one request pattern should be 2 times the normalized size of transmitted signal, i.e.,

$$D \geq D_{DO}^* = 2|R^*| \geq 2\left(s_2 - \frac{s_2 M}{\lfloor \frac{N}{s_2} \rfloor}\right). \tag{19}$$

Besides, if we consider both groups, choose s_1 users in $[K_1]$ and s_2 users in $[K_2]$. Construct $\lfloor \frac{N}{s_1+s_2} \rfloor$ request patterns, such that in each request pattern, each user will request a distinct file. Similarly, by cut-set bound we have

$$\left\lfloor \frac{N}{s_1+s_2} \right\rfloor |R^*| + (s_1+s_2)M \geq \left\lfloor \frac{N}{s_1+s_2} \right\rfloor (s_1+s_2).$$

For transmission, suppose that $s_1 + s_2$ users are active in all time slots, delay for one request pattern will be equal to the normalized size of transmitted signal. Delay in this situation can be the lower bound, i.e.,

$$D \geq D_{DO}^* = |R^*| \geq s_1 + s_2 - \frac{(s_1+s_2)M}{\lfloor \frac{N}{s_1+s_2} \rfloor}. \tag{20}$$

Combining (19) and (20) and maximizing over all possible value of s_1 and s_2 we obtain

$$D_{DO}^* \geq \max(\max_{s_1,s_2} s_1 + s_2 - \frac{(s_1+s_2)M}{\lfloor \frac{N}{s_1+s_2} \rfloor}, 2\max_{s_2}(s_2 - \frac{s_2 M}{\lfloor \frac{N}{s_2} \rfloor})),$$

where $1 \leq s_1 \leq K_1$ and $1 \leq s_2 \leq K_2$.

In the sequel, we generalize the proof to m groups with $T = 2^{m-1}$. We divide the cut-sets into m categories as follows. When the cut-set includes users in $[K_i]$,$[K_{i+1}]$,$\cdots$,$[K_m]$, $i = 1, \cdots, m$, choose s_j users in $[K_j]$, $j = i, i+1, \cdots, m$. Construct $\lfloor \frac{N}{\sum_{j=i}^{m} s_j} \rfloor$ request patterns, such that in each request pattern, each user will request a distinct file. By cut-set bound, we have

$$\left\lfloor \frac{N}{\sum_{j=i}^{m} s_i} \right\rfloor |R^*| + \sum_{j=i}^{m} s_i M \geq \left\lfloor \frac{N}{\sum_{j=i}^{m} s_i} \right\rfloor \sum_{j=i}^{m} s_i.$$

For transmission, suppose that $A_j = A_i, j = i+1, \cdots, m$, i.e., $s_i + \cdots + s_m$ users share the same 2^{m-i} active time slots in one cycle, delay for one request pattern will be 2^{i-1} times the transmitted signal size. This delay can be the lower bound, i.e.,

$$D_{DO}^* = 2^{i-1}|R^*| \geq 2^{i-1} \cdot \left(\sum_{j=i}^{m} s_j - \frac{(\sum_{j=i}^{m} s_j)M}{\lfloor \frac{N}{\sum_{j=i}^{m} s_j} \rfloor} \right).$$

Combining m categories and maximizing over all possible values of $s_1, s_2, \cdots, s_m$, we will obtain the lower bound in Proposition 3. □

Lemma 4. *For m user groups $[K_1]$, $[K_2]$, $\cdots$, $[K_m]$ in DN model, the transmission delay D is lower bounded by*

$$D \geq D_{DN}^* = \max\{ \max_{s_1,\cdots,s_m} (\sum_{i=1}^{m} s_i - \frac{\sum_{i=1}^{m} s_i M}{\lfloor \frac{N}{\sum_{i=1}^{m} s_i} \rfloor}),$$
$$2\max_{s_2}(s_2 - \frac{s_2 M}{\lfloor \frac{N}{s_2} \rfloor}), \cdots, 2^m \max_{s_m}(s_m - \frac{s_m M}{\lfloor \frac{N}{s_m} \rfloor}) \},$$

where $s_i \in \{1, 2, \cdots, K_i\}, i = 1, \cdots, m$.

Proof. We consider two situations.

First, if the cut-set does not include users in $[K_1]$, since there is no overlap among other groups, delivery for users in one group will be independent of each other. Consider s_i users in $[K_i]$, $i = 2, 3, \cdots, m$. By cut-set bound we have,

$$\left\lfloor \frac{N}{s_i} \right\rfloor |R_i'^*| + s_i M \geq \left\lfloor \frac{N}{s_i} \right\rfloor s_i. \tag{21}$$

Note that even users in different groups request the same files, (24) is still true. And we have,

$$|R_i'^*| \geq s_i - \frac{s_i M}{\lfloor \frac{N}{s_i} \rfloor}.$$

Since $|A_i| = 2^{m-i}$, delay for one request pattern should be 2^{i-1} times the size of transmitted signals, i.e.,

$$D^*_{DN} = 2^{i-1}|R'^*_i| \geq 2^{i-1}(s_i - \frac{s_i M}{\lfloor \frac{N}{s_i} \rfloor}). \tag{22}$$

Second, if the cut-set includes users in $[K_1]$, we consider s_i users in $[K_i]$, $i = 1, 2, \cdots, m$, and the delay should be lower bounded by the delay when time slots are all active for each user, in which situation delay for one request pattern is equal to the size of transmitted signals, i.e.,

$$D^*_{DN} = |R'^*| \geq s_1 + \cdots + s_m - \frac{(s_1 + \cdots + s_m)M}{\lfloor \frac{N}{s_1 + \cdots + s_m} \rfloor}. \tag{23}$$

Combining (25)(26) and maximizing over all possible values of $s_1, s_2 \cdots, s_m$ we conclude the proof.

□

Then we turn to a more complicated problem, determining the theoretical lower bound for TC model. Interestingly, the performance lower bound in TC model is equal to the lower bound in DN model.

Theorem 2. *(Lower bound): For m user groups $[K_1]$, $[K_2]$,$\cdots$, $[K_m]$ in TC model, the transmission delay D is lower bounded by*

$$D \geq D^*_{TC} = D^*_{DN}.$$

Proof. We also use cut-set bound to conduct the proof and consider two situations.

In the first case, if the cut-set does not include users in $[K_1]$, since there is no overlap among other groups, delivery for users in one group will be independent of each other. Consider s_i users in $[K_i]$, $i = 2, 3, \cdots, m$. By cut-set bound we have,

$$\left\lfloor \frac{N}{s_i} \right\rfloor |R'^*_i| + s_i M \geq \left\lfloor \frac{N}{s_i} \right\rfloor s_i. \tag{24}$$

Note that even users in different groups request the same files, (24) is still true. And we have,

$$|R'^*_i| \geq s_i - \frac{s_i M}{\lfloor \frac{N}{s_i} \rfloor}.$$

Since $|A_i| = 2^{m-i}$, the delay spent in transmitting one request pattern should be 2^{i-1} times the size of transmitted signals, i.e.,

$$D^*_{TC} = 2^{i-1}|R'^*_i| \geq 2^{i-1}(s_i - \frac{s_i M}{\lfloor \frac{N}{s_i} \rfloor}). \tag{25}$$

Second, if the cut-set includes users in $[K_1]$, we consider s_i users in $[K_i]$, $i = 1, 2, \cdots, m$, and the delay should be lower bounded by the delay when time slots are all active for each user, in which situation delay spent in transmitting one request pattern is equal to the size of transmitted signals, i.e.,

$$D^*_{TC} = |R'^*| \geq s_1 + \cdots + s_m - \frac{(s_1 + \cdots + s_m)M}{\lfloor \frac{N}{s_1 + \cdots + s_m} \rfloor}. \tag{26}$$

Combining (25)(26) and maximizing over all possible values of $s_1, s_2 \cdots, s_m$, we conclude the proof. □

5.3 Order-Optimality of Proposed Scheme

The upper bound of transmission delay shows the performance limit of the proposed algorithm in the worst case, while the lower bound of transmission delay reveals the optimal performance that an ideal solution can achieve. To validate the efficiency of our algorithm, we need to compare the two performance limits and discuss the order difference by mathematical justification.

Proposition 3. *In DO model, the gap between upper bound and the lower bound is limited by*

$$1 \leq \frac{D_{DO}}{D^*_{DO}} \leq 24$$

Proof. First, recalling from [12], for K users with all time slots being active, i.e., $m = 1$ in our model, the gap between fundamental upper bound $D_F(K) = (1 - \frac{M}{N})\frac{1-(1-\frac{M}{N})^K}{\frac{M}{N}}$ and fundamental lower bound $D^*_F(K) = \max_{1 \leq s \leq K}(s - \frac{sM}{\lfloor \frac{N}{s} \rfloor})$ is $\frac{D_F(K)}{D^*_F(K)} \leq 12$. Note that $D_(K)$ and $D^*_F(K)$ are function of K.

According to signal set separation and (1)–(3), by mathematical induction we can easily obtain

$$\begin{aligned} |R_i| &= (1 - \frac{M}{N})\frac{1 - (1 - \frac{M}{N})^{K_i}}{\frac{M}{N}} \cdot \left(1 - \frac{M}{N}\right)^{K_{i+1} + \cdots + K_m} \\ &\leq (1 - \frac{M}{N})\frac{1 - (1 - \frac{M}{N})^{K_i}}{\frac{M}{N}} \\ &= D_F(K_i), \end{aligned}$$

where $D_F(K_i)$ denotes the transmission delay spent in applying decentralized scheme to K_i users who are active all the time. So, we have

$$\begin{aligned} D_{DO} &= \max\{\sum_{i=1}^{m} |R_i|, 2\sum_{i=2}^{m} |R_i|, \cdots, 2^{m-1}|R_m|\} && \text{(27a)} \\ &\leq \max\{\sum_{i=1}^{m} D_F(K_i), 2\sum_{i=2}^{m} D_F(K_i), \cdots, 2^{m-1}D_F(K_m)\}. && \text{(27b)} \end{aligned}$$

(27a) is based on Lemma 1. Since there are m items in the right-hand-side of (27b), we should consider m different cases. If $D_{DO} \leq 2^{m-1}D_F(K_m)$, we have

$$\begin{aligned} D_{DO} &\leq 2^{m-1}D_F(K_m) \\ &\leq 2^{m-1} \cdot 12D^*_F(K_m) && \text{(28a)} \\ &= 12 \cdot 2^{m-1} \cdot \max_{1 \leq s_m \leq K_m} \left(s_m - \frac{s_m M}{\lfloor \frac{N}{s_m} \rfloor}\right) && \text{(28b)} \\ &\leq 12D^*_{DO}. && \text{(28c)} \end{aligned}$$

(28a) is obtained by the gap given in [11], (28b) is according to the lower bound given in [11] and (28c) is based on Lemma 3. If the maximum of (27b) equals to $2^{i-1}\sum_{j=i}^{m} D_F(K_j), i = 2, 3, \cdots, m$, we have the following inequality due to maximum condition,

$$2^{i-1}\sum_{j=i}^{m} D_F(K_j) \geq 2^i \sum_{j=i+1}^{m} D_F(K_j). \tag{29}$$

Rearranging (29), we have

$$2^{i-1}D_F(K_i) \geq 2^{i-1}\sum_{j=i+1}^{m} D_F(K_j).$$

Therefore,

$$\begin{aligned} D_{DO} &= 2^{i-1}\sum_{j=i}^{m} D_F(K_j) \\ &\leq 2^i D_F(K_i) \\ &\leq 12\cdot 2^i D_F^*(\sum_{j=i}^{m} K_j) \\ &= 12\cdot 2^i \max_{s_i,\cdots,s_m}\left(\sum_{j=i}^{m} s_j - \frac{\sum_{j=i}^{m} s_j M}{\lfloor \frac{N}{\sum_{j=i}^{m} s_j}\rfloor}\right) \\ &\leq 24 D_{D0}^*. \end{aligned}$$

We thus conclude this proposition. □

Also, we compare the gap between upper bound and lower bound in DN model, which is shown in the next proposition.

Proposition 4. *In DN model, the gap between upper bound and the lower bound is limited by*

$$1 \leq \frac{D_{DN}}{D_{DN}^*} \leq 24$$

Proof. First, for $\forall g$ $(2 \leq g \leq m)$, consider clique combination $C[1,g]$. According to signal set separation in Alg. 2 and (9)–(11), by mathematical induction can easily get

$$\begin{aligned} |R_1'| &= F(1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_1}}{\frac{M}{N}}\cdot\left(1-\frac{M}{N}\right)^{K_g} \\ &\leq F(1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_1}}{\frac{M}{N}} \\ &= D_F(K_1), \end{aligned}$$

where $D_F(K_1)$ denotes the fundamental transmission delay given in [12] when time slots are all active for these K_1 users and

$$|R_i'| = F(1-\frac{M}{N})\frac{1-(1-\frac{M}{N})^{K_i}}{\frac{M}{N}} = D_F(K_i),$$

where $i = 2,\cdots,m$. Then, we have

$$D_{DN} = \max\{\sum_{i=1}^{m}|R_i'|, 2|R_2'|,\cdots,2^{m-1}|R_m'|\} \tag{30a}$$

$$\leq \max\{\sum_{i=1}^{m} D_F(K_i), 2D_F(K_2),\cdots,2^{m-1}D_F(K_m)\}. \tag{30b}$$

(30a) is based on Lemma 2. There are m items in the right-hand-side in (30b). If the maximum value is $2^{i-1}D_F(K_i), 2 \leq i \leq m$, we have

$$\begin{aligned} D_{DN} &\leq 2^{i-1}D_F(K_i) \\ &\leq 2^{i-1}\cdot 12 D_F^*(K_i) \\ &= 2^{i-1}\cdot 12\max_{s_i}\left(s_i - \frac{s_i M}{\lfloor \frac{N}{s_i}\rfloor}\right) \\ &\leq 12 D_{DN}^*. \end{aligned}$$

The last inequality is based on Lemma 4. Besides, if the maximum value of (30b) is $\sum_{i=1}^{m} D_F(K_i)$, we do

$$\begin{aligned} D_{DN} &\leq \sum_{i=1}^{m} D_F(K_i) \\ &\leq 12 D_F^*(\sum_{i=1}^{m} K_i) + 12\sum_{i=2}^{m} D_F^*(K_i) \\ &\leq 12\cdot(1+1/2+1/4+\cdots+(1/2)^{m-1}) D_{DN}^* \\ &\leq 24 D_{DN}^*. \end{aligned}$$

The third inequality is due to Lemma 4 and the fundamental lower bound given by [11].

In conclusion, we finish the proof. □

Furthermore, the gap between D_{TC} and D_{TC}^* can be derived similarly.

Theorem 3. *(Order-optimality): In TC model, the gap between upper bound and the lower bound is limited by*

$$1 \leq \frac{D_{TC}}{D_{TC}^*} \leq 24$$

The proof of the above theorem is similar to Proposition 3, and we omit it here for space limitation. The constant gap indicates that the difference between achievable upper bound and theoretical lower bound is independent of any model variable, such as user number K, storage size M, file number N and, most importantly, group number m which is dominated by the user with the worst channel condition in the network. In other words, with arbitrary network size and active time slot distributions, the performance will not vary by more than a factor of 24 compared with the optimal performance, which indicates that our proposed scheme is order-optimal in TC model.

6 NUMERICAL EVALUATION

In this section, we conduct numerical simulations to verify the results obtained in previous sections and compare with some heuristic solutions.

Influence of Inactive Time Slots. We consider an example of $N = 2000$ files with a same size and four groups with $K_1 = 500$, $K_2 = 50$, $K_3 = 20$, $K_4 = 10$. We consider each time cycle has 8 time slots. The active time slots are uniformly and randomly distributed in each cycle. We set the memory size M from 100 to 1500 and the numerical results is shown in Fig. 4, where the *time constrained delay* is obtained by adopting our proposed scheme and *time constrained lower bound* is the theoretical optimal delay D_{TC}^* given by Theorem 2. Then, we let all time slots be active, adopt coded caching scheme and we obtain *Delay (full active time)* in Fig.

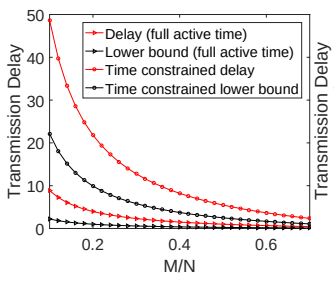

Figure 4: Influence of inactive time slot.

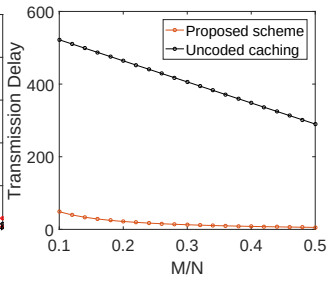

Figure 5: Comparison with uncoded caching.

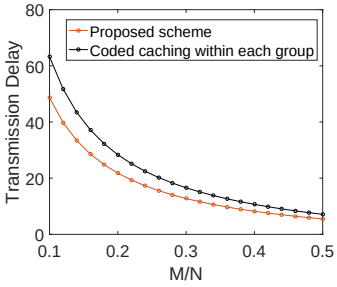

Figure 6: Comparison with coded caching within each group.

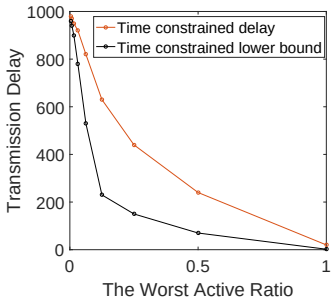

Figure 7: Impact of Different Active Time Slot Distributions.

4. For comparison, we calculate the lower bound of fundamental delay based on [11] and illustrate it as *Lower bound (full active time)* in Fig. 4. As is depicted in Fig. 4, the time constrained delay is much larger than delay (full active time) while time constrained lower bound is larger than lower bound (full active time) as well, which means the inactive time slots in time constrained networks do limit the performance of coded caching.

Comparison with Uncoded Caching. We compare our scheme with uncoded caching. For uncoded caching, the caching stage is the same as decentralized coded caching: each user k caches $\frac{M}{N}$ bits of file n uniformly at random. But in delivery stage, the server needs to send content sized $1-\frac{M}{N}$ to each user instead of aggregating into XOR signals. The time slot allocation strategy in Alg. 2 is also adopted to fully utilize the active time slots. Fig. 5 shows the numerical results. It is obvious that our proposed adaptive coded caching scheme performs much better and the additional caching gain compared to uncoded caching is approximately 10 times on average.

Comparison with Coded Caching within Each Group. There is a straightforward delivery scheme. Since users in one group share the same active time slots in every cycle, the delivery for each group can be independent, which we call *coded caching within each group*. Specifically, the signal sets to be transmitted are composed of $R_i = \{\oplus_{k\in S} F^r_{k,S} | S \subset [K_i]\}$, $i = 1, \cdots, m$. Fig. 6 shows the simulation results. According to Fig. 6, our proposed scheme outperforms coded caching within each group especially when storage size of users is not large enough. The reason is that as storage size increases, the transmitted content will reduce and the additional gain by multicasting will decrease as well.

Impact of Different Active Time Slot Distributions. We now evaluate the order-optimality of the proposed scheme when considering different active time slot distributions. We define *active ratio* as the ratio of the average number of active time slots in one cycle to the cycle size. Here, we set $N = 4000$, $M = 30$ and each time cycle has $2^8 = 256$ time slots. We conduct 9 experiments. In the first experiment, we consider $K_1 = 1024$ users who are active for all time slots (active ratio equals to 1). Then in the following experiment, we add $K_i = \frac{1024}{2^{i-1}}$ $(i = 2, \cdots, 9)$ users with active ratio $\frac{1}{2^{i-1}}$ to the network one by one. We organize the results in Fig. 7. As we can see, the transmission delay increases significantly as the worst active ratio decreases, which indicates that users with the worst channel condition in the network dominate the transmission performance. However, it is interesting that the difference between practical delay and theoretical optimal delay is within a certain range. This result verifies the order-optimality of our scheme under different active time distributions.

7 CONCLUSION

In this work, considering practical networks where users may have communication links of distinct qualities, we introduce active time slots and time constrained model to characterize the heterogeneity and variation of channel states. Deterministic overlapped model and deterministic non-overlapped model are proposed to simplify the analysis. For each model, we adaptively apply coded caching scheme incorporated with strategies of signal set separation and time slot allocation, trying to minimize transmission delay as well as guarantee the effectiveness. We also derive the performance upper bound and theoretical lower bound of transmission delay and prove the gap between two limits will not vary by more than a factor of 24. The constant gap reveals that the difference between the worst performance of our proposed scheme and the theoretical optimum is independent of any model parameter and the distribution of active time slots. We also conduct a series of numerical evaluations to verify our theoretical findings. We hope our work can shed more light on the study of channel heterogeneity in caching network.

REFERENCES

[1] Nikhil Karamchandani Abhinav Sridhar and Vinod M. Prabhakaran. 2016. Coded caching in hybrid networks. (2016), 1–6.

[2] Xudong Wang Aimin Tang and Sumit Roy. 2017. Centralized Coded Caching for Wireless Networks with Heterogeneous Channel Conditions. (2017), 1–6.

[3] Mohammad Mohammadi Amiri and Deniz Gündüz. 2017. Fundamental Limits of Coded Caching: Improved Delivery Rate-Cache Capacity Tradeoff. *IEEE Trans. Communications* 65, 2 (2017), 806–815.

[4] Georgios S. Paschos Apostolos Destounis, Mari Kobayashi and Asma Ghorbel. 2017. Alpha fair coded caching. (2017), 1–8.

[5] Hooshang Ghasemi and Aditya Ramamoorthy. 2017. Improved Lower Bounds for Coded Caching. *IEEE Trans. Information Theory* 63, 7 (2017), 4388–4413.

[6] Jad Hachem, Nikhil Karamchandani, and Suhas N. Diggavi. 2014. Multi-level coded caching. (2014), 56–60.

[7] Chih-Chun Wang Jinbei Zhang, Xiaojun Lin and Xinbing Wang. 2015. Coded caching for files with distinct file sizes. (2015), 1686–1690.

[8] Xiaojun Lin Jinbei Zhang and Xinbing Wang. 2018. Coded Caching Under Arbitrary Popularity Distributions. *IEEE Trans. Information Theory* 64, 1 (2018), 349–366.

[9] Antonia Maria Tulino Jaime Llorca Karthikeyan Shanmugam, Mingyue Ji and Alexandros G. Dimakis. 2016. Finite-Length Analysis of Caching-Aided Coded

Multicasting. *IEEE Trans. Information Theory* 62, 10 (2016), 5524–5537.
[10] Sheng Yang Khac-Hoang Ngo and Mari Kobayashi. 2016. Cache-aided content delivery in MIMO channels. (2016), 93–100.
[11] Mohammad Ali Maddah-Ali and Urs Niesen. 2014. Fundamental Limits of Caching. *IEEE Trans. Information Theory* 60, 5 (2014), 2856–2867.
[12] Mohammad Ali Maddah-Ali and Urs Niesen. 2015. Decentralized Coded Caching Attains Order-Optimal Memory-Rate Tradeoff. *IEEE/ACM Trans. Netw.* 23, 4 (2015), 1029–1040.
[13] Javier Matamoros Maria Gregori, Jesüs Gömez-Vilardebö and Deniz Gündüz. 2016. Wireless Content Caching for Small Cell and D2D Networks. *IEEE Journal on Selected Areas in Communications* 34, 5 (2016), 1222–1234.
[14] Alexandros G. Dimakis Negin Golrezaei, Andreas F. Molisch and Giuseppe Caire. 2013. Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution. *IEEE Communications Magazine* 51, 4 (2013), 142–149.
[15] Urs Niesen and Mohammad Ali Maddah-Ali. 2017. Coded Caching With Nonuniform Demands. *IEEE Trans. Information Theory* 63, 2 (2017), 1146–1158.
[16] White paper. 2017. Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021. (2017). http://goo.gl/1XYhqY
[17] Mohammad Ali Maddah-Ali Ramtin Pedarsani and Urs Niesen. 2016. Online Coded Caching. *IEEE/ACM Trans. Netw.* 24, 2 (2016), 836–845.
[18] Abolfazl Seyed Motahari Seyed Pooya Shariatpanahi and Babak Hossein Khalaj. 2016. Multi-Server Coded Caching. *IEEE Trans. Information Theory* 62, 12 (2016), 7253–7271.
[19] Shannon and Claude E. 1948. *A Mathematical Theory of Communication.* 27 (3): 379423 pages.
[20] Michële Angela Wigger Shirin Saeedi Bidokhti and Roy Timo. 2016. Noisy Broadcast Networks with Receiver Caching. *CoRR* abs/1605.02317 (2016).
[21] Qianqian Yang, Mohammad Mohammadi Amiri, and Deniz Gündüz. 2017. Audience retention rate aware coded video caching. (2017), 1189–1194.
[22] Jingjing Zhang and Petros Elia. 2017. Fundamental Limits of Cache-Aided Wireless BC: Interplay of Coded-Caching and CSIT Feedback. *IEEE Trans. Information Theory* 63, 5 (2017), 3142–3160.
[23] Jingjing Zhang and Petros Elia. 2017. Wireless coded caching: A topological perspective. (2017), 401–405.
[24] Naifu Zhang and Meixia Tao. 2018. Fitness-Aware Coded Multicasting for Decentralized Caching with Finite File Packetization. (2018), 1–4.