Decentralized Coded Caching with Distinct Cache Capacities

Mohammad Mohammadi Amiri, Qianqian Yang, Deniz Gündüz
Electrical and Electronic Engineering Department, Imperial College London
Email: {m.mohammadi-amiri15, q.yang14, d.gunduz}@imperial.ac.uk

***Abstract*—Decentralized coded caching is studied for a content server with N files, each of size F bits, serving K active users, each equipped with a cache of distinct capacity. It is assumed that the users' caches are filled in advance during the off-peak traffic period without the knowledge of the number of active users, their identities, or the particular demands. User demands are revealed during the peak traffic period, and are served simultaneously through an error-free shared link. A new decentralized coded caching scheme is proposed for this scenario, and it is shown to improve upon the state-of-the-art in terms of the required delivery rate over the shared link, when there are more users in the system than the number of files. Numerical results indicate that the improvement becomes more significant as the cache capacities of the users become more skewed.**

***Index Terms*—Decentralized coded caching, network coding, proactive caching.**

I. Introduction

The ever-increasing mobile data traffic has imposed a great challenge on the current network architectures. The growing demand is typically addressed by increasing the achievable data rates; however, moving content to the network edge has recently emerged as a promising alternative solution as it reduces both the bandwidth requirements and the delay. The use of edge caching is further motivated by the continuous drop in the cost of memory. In this paper, we consider an extreme form of edge caching, in which contents are stored directly at user terminals in a proactive manner. Proactive caching of popular content during off-peak traffic periods also helps flattening the high temporal variability of traffic. [1], [2].

Proactive caching is performed in two phases: The *placement phase* takes place during off-peak traffic hours, when the resources are abundant, and the users' caches are filled by the server without knowing the future user demands. When the user demands are revealed, the *delivery phase* is performed, in which a common message is transmitted from the server to all the users over the shared communication channel. Each user decodes its requested file by combining the bits received in the delivery phase with the contents of its local cache. The goal is to minimize the *delivery rate*, which guarantees that all the user demands are satisfied, independent of the demand combination of the users.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 690893, project TACTILENet: Towards Agile, effiCient, auTonomous and massIvely LargE Network of things.

Research on caching over the past decade has mainly focused on the placement phase in order to identify the most popular contents to be cached locally at user terminals [3], [4]. Recently, *coded caching* scheme was introduced in [5] for proactive caching, and it is shown that by storing and transmitting coded contents, and designing the placement and delivery phases jointly, it is possible to significantly reduce the delivery rate compared to uncoded caching.

A *centralized* caching scenario is studied in [5], in which the number and the identities of the users are known in advance by the server. This allows coordination of the cache contents across the users during both the placement and delivery phases; such that, by carefully placing pieces of contents in user caches, a maximum number of multicasting opportunities are created for tranmission during the delivery phase. Several other recent work has considered centralized coded caching, and the required delivery rate has been further reduced [6], [7], [8].

In practice, however, the number or the identity of active users that will participate in the delivery phase might not be known in advance during the placement phase. In such a scenario, called *decentralized coded caching*, coordination across users is not possible during the placement phase. However, Maddah-Ali and Niesen proposed a scheme that randomly caches parts of each content at each user, and can still exploit multicasting opportunities in the delivery phase, albeit limited compared to the centralized setting [9]. Decentralized coded caching has been studied in various other settings, e.g., files with different popularities [10], [11], and distinct lengths [12], online caching [13], etc.

Most of the existing literature on coded caching assume identical cache sizes across users. Recently, in [14] decentralized caching to users with heterogeneous cache sizes is studied, and by extending the scheme proposed in [9] to this scenarios, authors have shown that significant gains can still be obtained compared to uncoded caching. In this paper, we propose a novel decentralized caching algorithm for users with distinct cache capacities. We show that the proposed scheme requires a smaller delivery rate than the one achieved in [14]. The simulation results illustrate that the improvement in the delivery rate is more significant when the distribution of the cache capacities across users is more skewed.

The rest of this paper is organized as follows. The system model is introduced in Section II. In Section III, we introduce the proposed coded caching scheme, analyze its performance in terms of the delivery rate. The performance of the proposed

caching scheme is compared with the state-of-the-art result, and some numerical results are presented in Section IV. We conclude the paper in Section V.

Notations: The set of integers $\{1, ..., K\}$ is denoted by $[1:K]$. Notation $\oplus$ illustrates the bitwise XOR operation. For two sets Q and P, $Q\backslash P$ is a set including the members of Q and excluding the members of P. Notation $|.|$ represents cardinality of a set or length of a file. We use the notation $\bar{\oplus}$ to represent the bitwise XOR operation between binary sequences with different lengths. The arguments of $\bar{\oplus}$ are first zero-padded to have the same length as the longest argument, and then they are bitwise XOR-ed.

II. System Model

A server with N independent F-bit files, $W_1, ..., W_N$, is considered, where each file is assumed to be uniformly distributed over $\left[1 : 2^F\right]$. There are K active users, $U_1, ..., U_K$, where user U_k is equipped with a cache of capacity $M_k F$ bits, with $M_k \leq N$, $\forall k$. We denote the cache capacities by vector $\mu \triangleq (M_1, \ldots, M_K)$. Let Z_k denote the contents of U_k's cache at the end of the *placement phase*. Unlike in centralized coded caching [5], cache contents are independent of the number of users, their identities, or the user requests. User requests are revealed after the placement phase, where $d_k \in [1 : N]$ denotes the file requested by user U_k, $k = 1, \ldots, K$. User demands are served simultaneously through an error-free shared link. Let X denote the RF-bit message transmitted over the shared link by the server to enable each user U_k to decode its requested file W_{d_k}, together with its local cache content. Our goal is to characterize the minimum rate $R(\mu)$; such that, each user can decode its desired file with arbitrarily small probability of error, independent of the particular demand combination.

III. Decentralized Coded Caching

We first illustrate our decentralized coded caching scheme on the following example.

Example 1. Consider a caching system with $N = 2$ files W_1 and W_2, and $K = 4$ users. It is assumed that the cache capacity of user U_k is $M_k = (1/2)^{4-k} M$, $\forall k \in [1 : 4]$.

In the placement phase, user U_k caches a random $M_k F/2$ bits of each file independently. Since there are $N = 2$ files in the database, a total of $M_k F$ bits are cached by user U_k.

When $N < K$, it can be shown that the worst-case user demands happens when N users with the smallest cache capacities have different requests. For this particular example, we have $M_1 \leq \cdots \leq M_4$, and the worst-case happens when users U_1 and U_2 request distinct files. Hence, we can assume the worst-case demand combination of $d_k = 1$, if $k = 1, 3$, and $d_k = 2$, otherwise.

The contents served in the delivery phase are divided into three distinct parts, where X_i is delivered in part i, for $i = 1, 2, 3$. Thus, the common message is $X = (X_1, X_2, X_3)$. We further divide the message X_2 into three pieces X_2^1, X_2^2, and X_2^3. Below, we explain the purpose of each part in detail.

Part 1: In the first part of the delivery phase, the bits of each requested file which have not been cached by any user are directly delivered by the server. The following contents are delivered in this part. $X_1 = \left(W_{1,\{\emptyset\}}, W_{2,\{\emptyset\}}\right)$.

Part 2: The bits of the file requested by a user having been cached by another user are transmitted in the second part of the delivery phase. The server first delivers each user the bits of its requested file which are in the cache of one user with the same request. Then, each user receives the bits of its requested file which are in the cache of a single user with different request. By delivering the following contents, user U_k can obtain the bits of file W_{d_k} having been cached in user U_l, for $k, l \in [1 : 4]$, such that $l \neq k$. $X_2^1 = \left(W_{1,\{3\}} \bar{\oplus} W_{1,\{1\}}, W_{2,\{4\}} \bar{\oplus} W_{2,\{2\}}\right)$, $X_2^2 = \left(W_{1,\{4\}} \bar{\oplus} W_{1,\{2\}}, W_{2,\{3\}} \bar{\oplus} W_{2,\{1\}}\right)$, $X_2^3 = \left(W_{1,\{2\}} \bar{\oplus} W_{2,\{1\}}\right)$.

Part 3: In the last part, the server delivers the users the bits of their requested files which have been cached by more than one another user. Accordingly, each user U_k, $\forall k \in [1 : 4]$, can obtain all the bits of file W_{d_k} which are in the cache of users in any set $S \subset [1 : 4] \setminus \{k\}$, where $|S| \geq 2$, after receiving the following contents. $X_3 = \big(W_{1,\{2,3\}} \bar{\oplus} W_{2,\{1,3\}} \bar{\oplus} W_{1,\{1,2\}}, W_{1,\{2,4\}} \bar{\oplus} W_{2,\{1,4\}} \bar{\oplus} W_{2,\{1,2\}}, W_{1,\{3,4\}} \bar{\oplus} W_{1,\{1,4\}} \bar{\oplus} W_{2,\{1,3\}}, W_{2,\{3,4\}} \bar{\oplus} W_{1,\{2,4\}} \bar{\oplus} W_{2,\{2,3\}}, W_{1,\{2,3,4\}} \bar{\oplus} W_{2,\{1,3,4\}} \bar{\oplus} W_{1,\{1,2,4\}} \bar{\oplus} W_{2,\{1,2,3\}}\big)$.

After these parts, each user can decode all the missing bits of its desired file. To find the delivery rate, we first note that, by the law of large number, the length of the subfile $W_{k,V}$, for any set $V \subset [1 : K]$, is approximately given by

$$|W_{k,V}| \approx \prod_{i \in V} \left(\frac{M_i}{2}\right) \prod_{j \in [1:4] \setminus V} \left(1 - \frac{M_j}{2}\right) F, \quad \forall k \in [1 : K]\,. \tag{1}$$

For the example under consideration, when $M = 1$, i.e., $\mu = \{1/8, 1/4, 1/2, 1\}$, the delivery rate is 1.758, while the delivery rate of the scheme proposed in [14] for this setting is 2.681. Hence, the proposed scheme provides 34.43% reduction in the delivery rate compared to the state-of-the-art result for this example. □

A. Placement Phase

Since the active users are not known in advance in the decentralized setting, cache contents cannot be coordinated among the users. Similarly to the placement phases of the the decentralized coded schemes in the literature [9], [14], user U_k caches a random $M_k F/N$ bits of each file independently, for $k = 1, ..., K$. Since N files are hosted in the database, a total of $M_k F$ bits are cached by each user U_k, and hence, the corresponding cache-capacity constraint is satisfied.

For any set $V \subset [1 : K]$, let $W_{i,V}$ represent the bits of file W_i that have been *exclusively* cached by the users in set V at the end of the placement phase, i.e., $W_{i,V} \subset Z_k$, $\forall k \in V$, and $W_{i,V} \cap Z_k = \emptyset$, $\forall k \in [1 : K] \setminus V$.

B. Delivery Phase

User demands are revealed at the beginning of the delivery phase. Without loss of generality, we re-label the users such

that the first K_1 users, referred to as group G_1, have the same request W_1, the next K_2 users, group G_2, request file W_2, and so on so forth. For notational convenience, we define $S_i \triangleq \sum_{l=1}^{i} K_l$. Therefore, the user demands are as follows:

$$d_k = i, \quad \text{for } i = 1, ..., N, \text{ and } k = S_{i-1} + 1, ..., S_i, \qquad (2)$$

where we set $S_0 = 0$. We further order the users within a group according to their cache sizes, and assume, without loss of generality, that $M_{S_{i-1}+1} \leq M_{S_{i-1}+2} \leq \cdots \leq M_{S_i}$, for $i = 1, \ldots, N$.

The proposed delivery phase is presented in Algorithm 1. For any general demand combination described above, the delivery phase presented in Algorithm 1 contains two procedures CODED DELIVERY and RANDOM DELIVERY, and in each case the server chooses the one with the smaller delivery rate. Below, we explain these two procedures in detail.

The CODED DELIVERY procedure includes three distinct parts, where the content delivered in part i is denoted by X_i, $i = 1, 2, 3$, and the common message $X = (X_1, X_2, X_3)$ is sent to all the users during the delivery phase. The message transmitted in part 2, X_2, is further divided into three pieces X_2^1, X_2^2, and X_2^3, i.e., $X_2 = \left(X_2^1, X_2^2, X_2^3\right)$. Based on the aforementioned placement phase, the main motivation of the CODED DELIVERY procedure is to enable each user to recover the missing bits of its requested file which have been cached by i other users, $\forall i \in \{0, ..., K-1\}$.

In Part 1 of the this procedure, each user receives the bits of its requested file which have not been cached by any user.

The purpose of Part 2 is to enable each user to obtain all the missing bits of its request that have been cached by another single user. First, consider the message X_2^1. For $i = 1, ..., N$, each user $k \in [S_{i-1} + 1 : S_i]$ (i.e., $U_k \in G_i$), has access to bits $W_{i,\{k\}}$ locally in its cache, and with X_2^1 it can decode all the pieces $W_{i,\{l\}}$, $\forall l \in [S_{i-1} + 1 : S_i]$, i.e., the bits of its demand W_i, which are in the cache of another user in the same group, and no other user. Delivering the messages X_2^2 and X_2^3 together helps the users to decode the bits of their requested files having been cached by a single user in other groups; that is, after receiving $\bigcup_{k=S_{j-1}+1}^{S_j-1} \left(W_{i,\{k\}} \bar{\oplus} W_{i,\{k+1\}}\right)$, $\bigcup_{k=S_{i-1}+1}^{S_i-1} \left(W_{j,\{k\}} \bar{\oplus} W_{j,\{k+1\}}\right)$, and $W_{i,\{S_{j-1}+1\}} \bar{\oplus} W_{j,\{S_{i-1}+1\}}$, the users in both groups G_i and G_j can obtain the missing bits of their requested files that have been cached by a user in another group, for $i = 1, ..., N-1$ and $j = i+1, ..., N$ (and no other user). Note that, having received X_2^2, the third message $W_{i,\{S_{j-1}+1\}} \bar{\oplus} W_{j,\{S_{i-1}+1\}}$ is the smallest number of bits (based on the assumption $M_{S_{l-1}+1} \leq M_{S_{l-1}+2} \leq \cdots \leq M_{S_l}$, $\forall l \in [1 : N]$) that enable all the users in both groups G_i and G_j to obtain the missing bits of their desired files that are in the cache of users in the other group, for $i = 1, ..., N-1$ and $j = i+1, ..., N$.

Part 3 of our algorithm is the same as the delivery phase proposed in [14, Algorithm 2], and it is performed to send the users the missing bits of their requests that have been cached by more than one user.

Finally, in the RANDOM DELIVERY procedure, as in the DELIVERY procedure of [9], the server transmits enough random linear combinations of the bits of file W_i to the users in group G_i to make sure they all can decode the file, for $i = 1, \ldots, N$.

Algorithm 1 Coded Delivery Phase

procedure CODED DELIVERY
  Part 1: Delivering bits that are not in the cache of any user
  for $i = 1, 2, \ldots, N$ **do**
    $X_1 = \left(W_{d_{S_{i-1}+1},\{\emptyset\}}\right)$
  end for

  Part 2: Delivering bits that are in the cache of only one user
  $X_2^1 = \left(\bigcup_{i=1}^{N} \bigcup_{k=S_{i-1}+1}^{S_i-1} \left(W_{i,\{k\}} \bar{\oplus} W_{i,\{k+1\}}\right)\right)$
  $X_2^2 = \bigcup_{i=1}^{N-1} \bigcup_{j=i+1}^{N} \left(\bigcup_{k=S_{j-1}+1}^{S_j-1} \left(W_{i,\{k\}} \bar{\oplus} W_{i,\{k+1\}}\right), \bigcup_{k=S_{i-1}+1}^{S_i-1} \left(W_{j,\{k\}} \bar{\oplus} W_{j,\{k+1\}}\right) \right)$
  $X_2^3 = \left(\bigcup_{i=1}^{N-1} \bigcup_{j=i+1}^{N} W_{i,\{S_{j-1}+1\}} \bar{\oplus} W_{j,\{S_{i-1}+1\}}\right)$

  Part 3: Delivering bits that are in the cache of more than one user
  for $i = 1, 2, \ldots, K-2$ **do**
    for $j = 2, 3, \ldots, K-i$ **do**
      for $V \subset [i+1 : K] : |V| = j$ **do**
        $X_3 = \left(\left(\bar{\bigoplus}_{v \in V} W_{d_v,\{V,i\}\setminus\{v\}}\right) \bar{\oplus} W_{d_i,V}\right)$
      end for
    end for
  end for
end procedure

procedure RANDOM DELIVERY
  for $i = 1, 2, \ldots, N$ **do**
    server sends enough random linear combinations of the bits of file W_i to enable the users demanding it to dedcode it.
  end for
end procedure

C. Delivery Rate Analysis

In the following, we evaluate the delivery rate of the proposed caching scheme for the worst-case user demands. Consider first the case $N \geq K$. It can be argued in this case that the worst-case user demands happens if each file is requested by at most one user. Hence, by re-ordering the users,

for the worst-case user demands, we have $K_i = 1$, for $1 \leq i \leq N$, and $K_i = 0$, otherwise. In this case, it can be shown that the CODED DELIVERY procedure requires a lower delivery rate than the RANDOM DELIVERY procedure; hence, the server uses the former. In this case, it is possible to simplify the CODED DELIVERY procedure such that, only message X_2^3 is transmitted in Part 2, when $N \geq K$, i.e., $X_2 = X_3^2$. The corresponding common message, $X = \left(X_1, X_2^3, X_3\right)$, transmitted over the CODED DELIVERY procedure, reduced to the delivery phase of [14, Algorithm 2]. Thus, the proposed scheme achieves the same delivery rate as [14, Algorithm 2] when $N \geq K$.

Next, we consider the case $N < K$. It is possible to show that the worst-case user demands in this case happens when N users with the smallest cache capacities all request different files, i.e., they end up in different groups. The delivery rate of the proposed delivery phase when $N < K$ is presented in the following theorem. The proof of the worst-case demand distribution as well as Theorem 1 are skipped due to space limitations; however, they can be found in the longer version of the paper in [15].

Theorem 1. *In a decentralized caching system with N files in the database, each of size F bits, and K users with cache capacities $\mu = \{M_1, ..., M_K\}$, such that $M_1 \leq M_2 \leq \cdots \leq M_K$, the following delivery rate-cache capacity trade-off is achievable when $N < K$:*

$$R_c(\mu) = \min\left\{\sum_{i=1}^{K}\left[\prod_{j=1}^{i}\left(1-\frac{M_j}{N}\right)\right] - \Delta R_1(\mu) - \Delta R_2(\mu), \sum_{i=1}^{N}\left(1-\frac{M_i}{N}\right)\right\}, \quad (3)$$

where

$$\Delta R_1(\mu) = (K-N)\prod_{l=1}^{K}\left(1-\frac{M_l}{N}\right), \quad (4a)$$

$$\Delta R_2(\mu) = \left[\sum_{k=1}^{K-N}\left(\frac{(k-1)M_{k+N}}{N-M_{k+N}}\right)\right]\prod_{l=1}^{K}\left(1-\frac{M_l}{N}\right). \quad (4b)$$

IV. Comparison with the State-of-the-Art and Numerical Results

In this section, the proposed caching scheme is compared with the scheme proposed in [14] both analytically and numerically. We note that, although the scheme presented in [14] is for $N \geq K$, it can also be applied to the case $N < K$, and the same delivery rate as [14, Theorem 2], denoted here by $R_b(\mu)$, can be achieved. Hence, in the following, when we refer to the scheme stated in [14, Algorithm 2] for $N < K$, we consider its generalization to this scenario. When $N < K$, according to [14, Theorem 2] and (3), we have

$$R_b(\mu) - R_c(\mu) \geq \Delta R_1(\mu) + \Delta R_2(\mu) \overset{(a)}{>} 0. \quad (5)$$

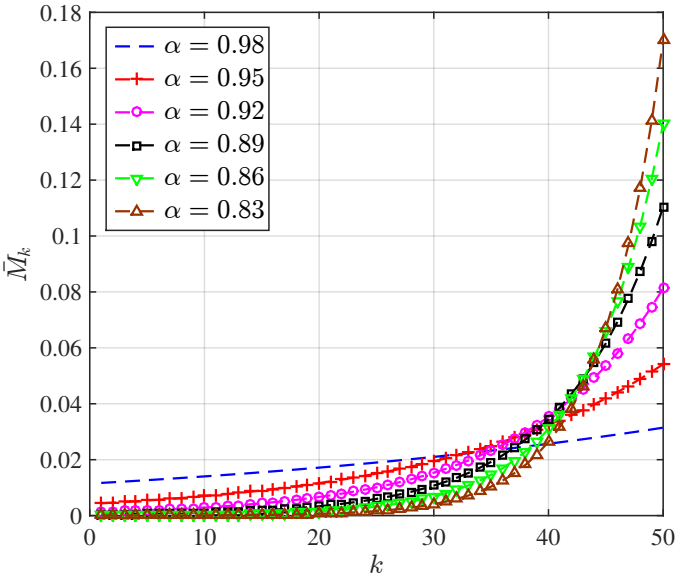

Fig. 1. Illustration of cache capacity distribution normalized by $\sum_{k=1}^{K} M_k$ for different α values, and $K = 50$. The x-axis corresponds to the user index k.

The inequality (a) holds as long as $N < K$. Therefore, when the number of files in the database is smaller than the number of active users in the delivery phase, the proposed coded caching scheme requires a smaller delivery rate than the one presented in [14].

For the numerical results, we consider an exponential cache distribution among users, such that the cache capacity of user U_k is given by

$$M_k = \alpha^{K-k} M, \quad (6)$$

where $0 \leq \alpha \leq 1$, for $k = 1, \ldots, K$, and M denote the maximum cache capacity in the system. Thus, we have $\mu = \left\{\alpha^{K-1}M, \alpha^{K-2}M, \ldots, M\right\}$, such that $M_1 \leq M_2 \leq \cdots \leq M_K$. The distribution of cache capacities normalized by $\sum_{k=1}^{K} M_k$, i.e., $M_k / \sum_{k=1}^{K} M_k$ denoted by $\bar{M}_K$, $\forall k \in [1:K]$, is demonstrated in Fig. 1 for different values of α, when $K = 50$. Observe that, the smaller the value of α, the more skewed the cache capacity distribution across users become. In the special case of $\alpha = 1$, we obtain the homogeneous cache capacity model studied in [9].

In Fig. 2, the delivery rate of the proposed scheme, $R_c(\mu)$, is compared with that of the coded scheme proposed in [14], i.e., $R_b(\mu)$, when $N = 50$, $K = 70$, and $\alpha = 0.97$. The delivery rate is plotted in this figure versus the largest cache capacity in the system, M. As expected the performance improves, i.e., the delivery rate reduces as M increases. We also clearly observe that the proposed scheme outperforms the scheme presented in [14]. The improvement is particularly significant for lower values of M. The cut-set lower bound for this setting is also included in the figure. Although the delivery rate of the proposed scheme approaches the lower bound for relatively small values of M, there is still a gap for large values of M,

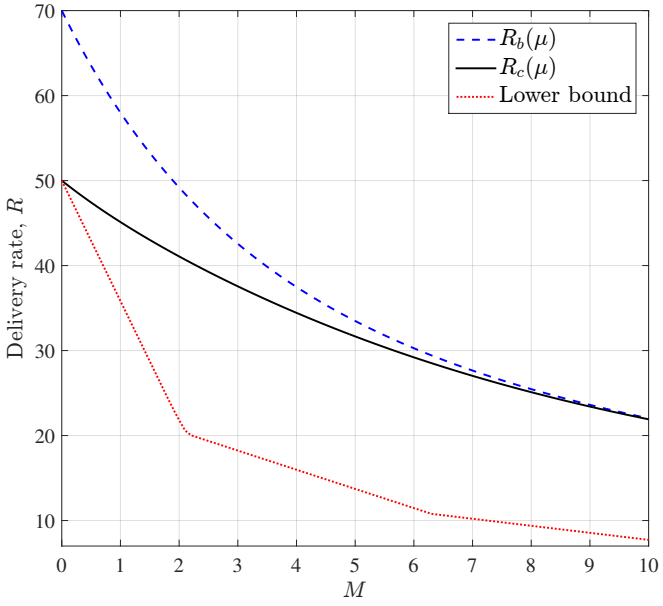

Fig. 2. Delivery rate versus M, where the cache capacity of user k is $M_k = \alpha^{K-k}M$, $k = 1, \ldots, K$, when $\alpha = 0.97$, $N = 50$, and $K = 70$.

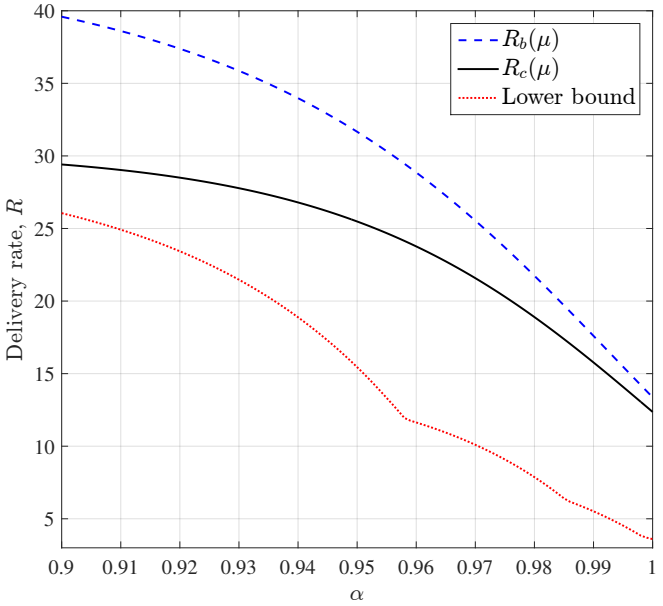

Fig. 3. Delivery rate versus $\alpha \in [0.9, 1]$, where $M_k = \alpha^{K-k}M$, $N = 30$, $K = 45$, and $M = 2$.

which may as well be due to the looseness of the lower bound.

In order to see the effect of skewness of the cache capacities on the delivery rate, in Fig. 3, the delivery rate of different schemes are plotted as a function of $\alpha \in [0.9, 1]$, for $N = 30$, $K = 45$, and the largest cache capacity of $M = 2$. The delivery rate of the proposed decentralized coded caching scheme is lower than the one presented in [14] for the whole range of α values under consideration, while the gain is more pronounced for smaller values of α, i.e., as the distribution of cache capacities becomes more skewed. We also observe the gap to the cut-set lower bound also diminishes in this regime.

V. Conclusions

In this paper, we have studied coded caching to users with distinct cache capacities, and proposed a novel decentralized coded caching scheme that improves upon the best known delivery rate in the literature. The improvement is achieved by improving the delivery of bits that have been cached by none of the users, or by only a single user. In particular, the proposed scheme exploits the group-based coded caching scheme we have introduced previously for centralized caching in a system with homogeneous cache capacities [16]. Our numerical results show that the improvement upon the scheme proposed in [14] is even more pronounced when the cache capacities of the users are more skewed.

We are currently aiming to improve the delivery rate for larger values of cache capacities by finding a more efficient coded delivery scheme which delivers the missing bits of the requested files that have been cached by more than one user.

References

[1] L. W. Dowdy and D. V. Foster, "Comparative models of the file assignment problem," *ACM Comput. Surv.*, vol. 14, pp. 287–313, Jun. 1982.

[2] K. C. Almeroth and M. H. Ammar, "The use of multicast delivery to provide a scalable and interactive video-on-demand service," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 6, pp. 1110–1122, Aug. 1996.

[3] I. Baev, R. Rajaraman, and C. Swamy, "Approximation algorithms for data placement problems," *SIAM Journal on Computing*, vol. 38, no. 4, pp. 1411–1429, 2008.

[4] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE INFOCOM*, San Diego, CA, Mar. 2010, pp. 1–9.

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

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

[7] M. Mohammadi Amiri and D. Gündüz, "Fundamental limits of caching: Improved delivery rate-cache capacity trade-off," *arXiv:1604.03888v1 [cs.IT]*, Apr. 2016.

[8] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," *arXiv: 1601.063834v2 [cs.IT]*, Jan. 2016.

[9] M. A. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw*, vol. 23, no. 4, pp. 1029–1040, Apr. 2014.

[10] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Toronto, ON, Apr. 2014, pp. 221–226.

[11] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv: 1502.03124v1 [cs.IT]*, Feb. 2015.

[12] J. Zhang, X. Lin, C. C. Wang, and X. Wang, "Coded caching for files with distinct file sizes," in *Proc. IEEE Int'l Symp. on Inform. Theory*, Hong Kong, Jun. 2015, pp. 1686–1690.

[13] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Jun. 2014, pp. 1878–1883.

[14] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," *arXiv:1504.01123v1 [cs.IT]*, Apr. 2015.

[15] M. Mohammadi Amiri, Q. Yang, and D. Gündüz, "Decentralized coded caching with distinct cache capacities," *arXiv:1601.05690v1 [cs.IT]*, Nov. 2016.

[16] ——, "Coded caching for a large number of users," *arXiv:1605.01993v1 [cs.IT]*, May 2016.