Coded Caching With User Grouping Over Wireless Channels

Busra Tegin and Tolga M. Duman, *Fellow, IEEE*

***Abstract*—We study coded caching over non-ergodic fading channels. As the multicast capacity of a broadcast channel is restricted by the user experiencing the worst channel conditions, we formulate an optimization problem to minimize the transmission time by grouping users based on their channel conditions, and transmit coded messages according to the worst channel conditions in the group, as opposed to the worst among all. We develop two algorithms to determine the user groups: a locally optimal iterative algorithm and a numerically more efficient solution through a shortest path problem, and we illustrate the effectiveness of developed solutions via numerical examples.**

***Index Terms*—Coded caching, wireless fading channels.**

I. Introduction

CACHING is a strategy to prefetch server's contents at individual user caches during off-peak hours, i.e., when the network is not congested, and to exploit the cache contents during the delivery phase where communication is more expensive. In [1], a novel centralized coded caching scheme which provides a global caching gain by jointly optimizing the placement and delivery phases along with the usual local caching gain is introduced. In [2], a decentralized coded caching scheme is developed outperforming the traditional caching strategies without any coordination in the placement phase.

Coded caching has drawn significant attention and various extensions have been proposed over the last few years. References [3]–[4] consider variants of coded caching schemes over fading channels. MIMO extensions of the problem are considered in [5]–[6], while [7] applies interference management to alleviate the negative effects of link quality difference among users. Closely related to our present set-up, in [8], the authors aim to overcome detrimental effects of weak users by designing opportunistic scheduling policies using a long-term average rate utility function. They also propose a threshold-based scheduling algorithm for asymmetric channel statistics to achieve fairness among users. Both of these approaches focus on long-term averages and ignore the users whose channel gains are below a threshold, and hence, are not served.

In this letter, we follow a coded caching model where the placement phase is performed in a decentralized manner and the delivery phase takes place over a wireless fading channel. Different from [8], which considers long-term average rates, our interest is to study non-ergodic channels and minimize the transmission time by letting some of the weak users to be in outage. With a fixed outage probability, we formulate an optimization problem to reduce the total transmission time by grouping the participating users to overcome the detrimental effects of channel fading. We also develop a locally optimal iterative algorithm to compute the signal to noise ratio (SNR) thresholds. Furthermore, we quantize the SNR thresholds, and model the optimization process as a shortest path problem, and obtain a reduced complexity solution.

This letter is organized as follows. Section II introduces the system model and preliminaries. Proposed user grouping approaches are described in Sections III and IV. Performance of the developed approaches are studied via simulations in Section V, and the letter is concluded in Section VI.

Throughout the letter, we will use the notation $[a\ b]$ to indicate the integer set $\{a, \ldots, b\}$ where $a \leq b$, a and b are positive integers, and simply $[b] = [1\ b]$.

II. System Model and Preliminaries

We consider a system which contains a server with N files each of size F bits connected through a fading channel to K users. Users are equipped with local caches which are able to store MF bits. The normalized cache size for each user is defined as $m = M/N$. We consider the decentralized coded caching framework introduced in [2], where the placement phase is performed during the off-peak hours over an error-free shared link. However, the delivery phase takes place over a wireless (fading) channel.

We model the channel between the server and the users during the delivery phase as a quasi-static fading channel. We examine the coded caching system where the server only knows the channel statistics to determine the SNR thresholds of user groups. During the delivery phase, it receives limited feedback from the users indicating their groups with a low overhead (a few bits of feedback). Since the channels are non-ergodic, the server chooses not to serve the users with low SNRs, and puts them in outage. If user $k \in [K]$ is not in outage, it receives $\mathbf{y}_k = \sqrt{\rho_k} h_k \mathbf{X} + \mathbf{n}$ where the coded message $\mathbf{X}$ is constructed according to [2], components of $\mathbf{n}$ are independent and identically distributed (i.i.d.) zero mean circularly symmetric complex Gaussian random variables, i.e., each follows $\sim \mathcal{CN}(0, 1)$. h_k's are fading coefficients which are independent complex random variables. For instance, if h_k's are zero mean circularly symmetric complex Gaussian with variance 1/2 per dimension, then the channels are Rayleigh fading with ρ_k denoting the average SNR for user k.

Receiver $k \in [K]$ reconstructs its demanded content based on $\mathbf{y}_k$, cache content and the demand vector, and an error occurs for the user when the reconstructed content is different than the requested one.

Manuscript received December 19, 2019; revised February 9, 2020; accepted February 13, 2020. Date of publication February 21, 2020; date of current version June 10, 2020. The work of Busra Tegin was supported by Huawei through a Graduate Fellowship Program. The associate editor coordinating the review of this article and approving it for publication was M. Nafie. *(Corresponding author: Tolga M. Duman.)*

The authors are with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: btegin@ee.bilkent.edu.tr; duman@ee.bilkent.edu.tr).

Digital Object Identifier 10.1109/LWC.2020.2975661

III. Grouping Users Using Channel Statistics

In this section, we examine the optimal user grouping problem for the case where the server only has access to the channel statistics of users. First, we note that for multicast transmission to a group of users, the capacity of the channel is restricted by the user with the worst channel condition [5], and conditioned on the channel gains, it is given by $R(\mathbf{\Lambda}_{n\in[N_g]}) = \log_2(1+\min_{n\in[N_g]}\Lambda_n)$ where N_g is the number of users in consideration, and Λ_n is the instantaneous SNR of user *n*. Therefore, the rate $R(\mathbf{\Lambda}_{n\in[N_g]})$ dictates the reliable transmission limit. The corresponding transmission takes $T_{\text{req}} = \frac{T(m,N_g)}{R(\mathbf{\Lambda}_{n\in[N_g]})}$ units of time [8] where the normalized length of the coded message to serve N_g users each equipped with a normalized cache size of *m* is $T(m,N_g) = \frac{1-m}{m}(1-(1-m)^{N_g})$ [5].

By taking into consideration the limitation due to the user with the worst channel condition, creating a single coded message for all the users which are to be served may increase the total transmission time dramatically. We argue that this can be alleviated by grouping users and creating specific coded messages to different groups experiencing instantaneous SNRs close to each other. Based on this intuition, we formulate an optimization problem to minimize the total transmission time. Also, since the server does not have access to the instantaneous SNR values as it only receives limited feedback from users indicating their group, we use a single transmission rate for every user in the same group.

A. Optimization Problem for Threshold Determination

Since the individual links between the server and the users are modeled as non-ergodic channels, Shannon type capacity is zero, hence we adopt an outage capacity formulation. For a given rate *R*, the outage probability for user $k \in [K]$ is $P_{\text{out}} = P(C(\lambda_k) < R)$ with $C(\lambda_k) = \log_2(1+\lambda_k)$ where $\lambda_k = |h_k|^2\rho_k$ is the effective SNR of user *k* with cumulative distribution function (CDF) $F_k(\cdot)$.

Let us denote the SNR thresholds for user groups by x_j, $j = 0,1,\ldots,t-1$ where *t* is the number of groups and x_0 is the SNR threshold that determines the users in outage, i.e., users with a lower instantaneous SNR than x_0 are not served. The number of users in group *j* is denoted by K_j.

Proposition 1: For independent fading links, the outage SNR threshold x_0 with a given expected ratio of users that are not served (P_{out}) can be obtained by solving

$$P_{\text{out}} = \frac{1}{K}\left(K - \sum_{k=1}^{K} 1 - F_k(x_0)\right), \tag{1}$$

where $F_k(\cdot)$ is the CDF of the effective SNR for user $k \in [K]$, and *K* is the total number of users.

Proposition 2: For independent fading links, the expected number of users in a group formed by users with SNR $\in [x_{j-1}, x_j)$ (denoted by $\bar{K}_j$) is given by

$$\bar{K}_j = \sum_{k=1}^{K}\big(F_k(x_j) - F_k(x_{j-1})\big), \tag{2}$$

where $j = 0,1,\ldots,t$, $x_{-1} = 0$, $x_t = \infty$, $\bar{K}_0$ is the expected number of users in outage.

Note that, if $h_k \sim \mathcal{CN}(0,1)$, we have Rayleigh fading channels, which result in $P_{\text{out}} = \frac{1}{K}(K - \sum_{k=1}^{K} e^{-\frac{x_0}{\rho_k}})$, and $\bar{K}_j = \sum_{k=1}^{K}(e^{-\frac{x_{j-1}}{\rho_k}} - e^{-\frac{x_j}{\rho_k}})$.

To calculate the transmission time to serve the requests of group *j*, we focus on the coded message constructed for that group with a normalized length of $T(m,K_j)$, and the worst case rate of $\log_2(1+x_{j-1})$. We emphasize that the server does not have access to the instantaneous SNR values as each user only sends $\lceil \log_2(t) \rceil$ bits of feedback indicating their group along with their demands at the beginning of the delivery phase. This limited feedback enables the server to create coded messages for different user groups separately. The total required time to satisfy all the requests except for those in outage can be written as

$$T_{\text{req}} = \sum_{i=1}^{t} \frac{T(m,K_i)}{\log_2(1+x_{i-1})}. \tag{3}$$

Since $T(m,K_i)$ is a concave function of K_i, using Jensen's inequality, we have $\mathbb{E}[T_{\text{req}}] \le \sum_{i=1}^{t} \frac{T(m,\bar{K}_i)}{\log_2(1+x_{i-1})}$.

Given the number of groups *t* and the outage SNR threshold x_0, we are interested in minimizing $\sum_{i=1}^{t} \frac{T(m,\bar{K}_i)}{\log_2(1+x_{i-1})}$ which is a bound for expected transmission time $\mathbb{E}[T_{\text{req}}]$ over the SNR threshold vector $\boldsymbol{x} = [x_1 \ \cdots \ x_{t-1}]$, i.e.,

$$\begin{aligned} \underset{x_1\cdots x_{t-1}}{\text{minimize}} \ & \sum_{i=1}^{t} \frac{T(m,\bar{K}_i)}{\log_2(1+x_{i-1})} \\ \text{subject to} \ & x_{i-1} \le x_i, \ i = 1,\ldots,t-1. \end{aligned} \tag{$\mathbf{P}_1$}$$

Note that for a large number of users, $\sum_{i=1}^{t} \frac{T(m,\bar{K}_i)}{\log_2(1+x_{i-1})}$ converges to $\mathbb{E}[T_{\text{req}}]$ as $K_j \to \bar{K}_j$ by the law of large numbers.

Since $\mathbf{P}_1$ does not have any appealing structure, it requires a brute force search to determine the continuous valued parameters $x_1, x_2, \ldots, x_{t-1}$, which is infeasible for nontrivial values of *t*. To reduce the computational complexity, we propose an iterative algorithm to find the threshold values. The proposed approach is practical, and it is guaranteed to converge to a locally optimal solution.

B. An Efficient Algorithm for Threshold Determination

We notice that each parameter to be optimized in $\mathbf{P}_1$ is only present in two different terms of the objective function. Let us now focus on the terms that involve a given threshold value x_j of the summation. That is, for the SNR threshold x_j, we consider

$$\bar{T}_{x_j} = \frac{T(m,\bar{K}_j)}{\log_2(1+x_{j-1})} + \frac{T(m,\bar{K}_{j+1})}{\log_2(1+x_j)}. \tag{4}$$

Recall that $\bar{K}_j$'s can be calculated using (2). It is clear that if *j* is even, the only other dependence is to the adjacent odd indexes and vice versa. Hence, to simplify the optimization process, we can exploit this structure to obtain sub-problems which only depend on a single parameter and solve the original problem by using an iterative approach. Our proposed solution proceeds as follows: we first select arbitrary initial values for x_j's. We minimize (4) over the odd indexed thresholds

by fixing the even indexed ones. And then, we apply same procedure by reversing the roles of odd and even indices. We continue these iterations for a predetermined number of times to find the final set of SNR thresholds.

The proposed approach divides the initial problem into two sub-problems each containing $t/2$ minimization problems with a single unknown parameter in each step, hence simplifying the search considerably. We also note that, the overall cost function reduces at each iteration, and since it is bounded from below, the algorithm is guaranteed to converge by the Monotone Convergence Theorem [9]. However, since the problem is not convex, the algorithm is only guaranteed to be locally optimal. Therefore, we repeat the above described steps for different number of thresholds, and the solution which gives the minimum transmission time is selected as the final result.

IV. A Reduced Complexity Grouping Approach

In this section, we propose a reduced complexity approach for user grouping. We first quantize the SNR threshold values using a large number of possible groups, denoted by q. Although different quantizers can be used, we utilize a uniform quantizer where τ_i is the i^{th} boundary with $i = 1, \ldots, q-1$. With this discretization, the problem $\mathbf{P}_1$ can be converted into the following integer program which minimizes the bound on the total transmission time over the set of SNR thresholds $\boldsymbol{x}$:

$$\begin{aligned} \underset{x_1 \cdots x_{q-1}}{\text{minimize}} \quad & \sum_{i=1}^{q} \frac{T(m, \bar{K}_i)}{\log_2(1 + x_{i-1})} \\ \text{subject to} \quad & x_{i-1} \leq x_i, \; i = 1, \ldots, q-1, \\ & x_j \in \{\tau_1, \ldots, \tau_{q-1}\}, j = 1, \ldots, q-1, \end{aligned} \quad (\mathbf{P}_2)$$

where x_0 is predetermined using (1) for a given P_{out}, and $\bar{K}_i$'s are calculated using (2). Note that, even though we start with a large number of possible groups q, the solution determines the optimal number of groups which minimizes the overall transmission time.

Although formulating $\mathbf{P}_2$ as an integer program and utilizing quantized thresholds decreases the complexity of the brute force search based solution, the problem still does not have any appealing structure to be exploited. Therefore, to interpret the optimization problem further, we construct a directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ which comprises of a set $\mathbf{V}$ of vertices and a set $\mathbf{E}$ of edges. In this model, each vertex represents a quantization level for the SNR thresholds and each edge carries the corresponding transmission time by choosing the thresholds according to its incident vertices. With this, the minimization problem given in $\mathbf{P}_2$ becomes an instance of a well-known graph theory problem, namely, the shortest path problem which can be solved in polynomial time if there are no negative dicycles [10]. In this way, the worst case complexity can be reduced to $O(q^2)$ using existing approaches in the literature. In the shortest path problem, the aim is to minimize the length of the path (transmission time) between the starting vertex and the terminating one. It is clear that determination of shortest path automatically finds the optimal number of user groups, and hence there is no need to try different number of groups as needed for the case of the iterative approach proposed in Section III.

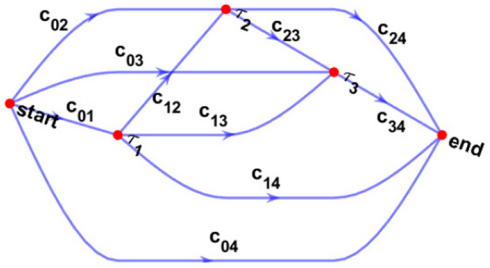

Fig. 1. Sample of a directed graph with 3 quantization levels and edge costs c_{ij}.

To compose the corresponding graph and determine the cost of each edge, the expected number of users whose SNRs are between τ_i and τ_j with $\tau_i \leq \tau_j$ can be found as $\bar{K}_{ij} = \sum_{k=1}^{K}(F_k(\tau_j) - F_k(\tau_i))$. And, the cost (bound on the required transmission time) of each edge (for the users whose SNRs are between τ_i and τ_j) can be calculated as $c_{ij} = T(m, \bar{K}_{ij})/\log_2(1 + \tau_i)$.

An illustration is given in Fig. 1 where c_{ij}'s represent edge costs. If the optimal solution contains only the starting and terminating vertices, there will only be a single edge in the solution. Hence, we need to create a single coded message for all the users using the algorithm introduced in [2]. If the optimal solution contains vertices other than the starting and terminating ones, the users will be placed in multiple groups, and separate coded messages will be created.

V. Numerical Examples

We consider Rayleigh fading, i.e., h_k's are independent zero mean circularly symmetric complex Gaussian random variables with variance 1/2 per dimension. In the first example, the total number of users is $K = 400$, the mean values of user SNRs are taken as -3, 0, 3, and 6 dB each for a quarter of users. The server contains $N = 1000$ files, and the outage probability is set to $P_{\text{out}} = 0.05$. For the iterative algorithm of Section III-B, the maximum number of thresholds t_{max} is set to 8. The number of quantization levels in the shortest path algorithm is determined by quantizing the SNR range from the outage SNR threshold x_0 to 30 dB uniformly with a step size of 0.3 dB.

The normalized required transmission time obtained by solving $\mathbf{P}_1$ with brute force search and different cache capacities are illustrated in Fig. 2a. It is observed that coded caching is effective, and allowing a greater number of user groups results in lower normalized transmission times. Also, as expected, all of the proposed algorithms outperform the coded caching solution without any user grouping. For this specific example, when $m = 0.025$, the shortest path algorithm achieves a 44% decrease in the normalized transmission time compared to sending without grouping, and a 38% decrease compared to uncoded caching.

As another example, we consider a network whose parameters are the same as the one above, except that we set $N = K = 5000$, and the user SNRs are exponentially distributed with means -6, 0, 6, and 12 dB, each for a quarter of them. The normalized transmission times are depicted as a function of the normalized cache size in Fig. 2b. We note

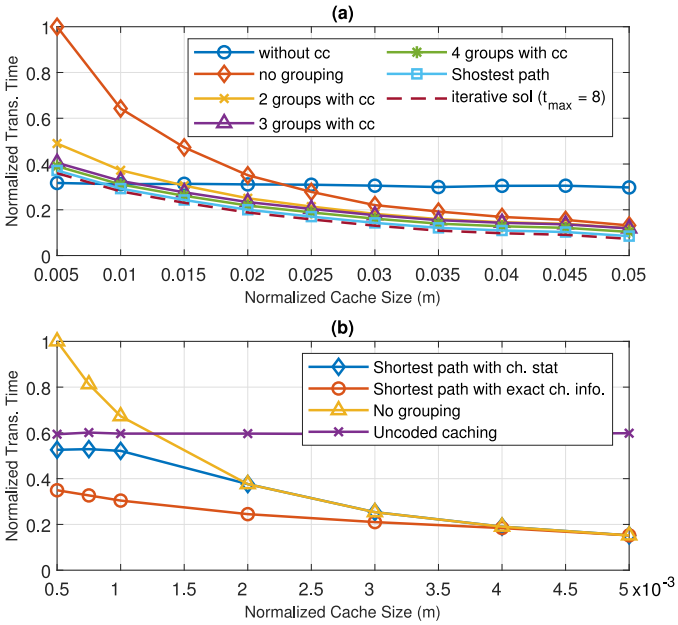

Fig. 2. (a) Simulation results with uncoded caching, coded caching with groups, iterative and shortest path solutions, (b) Effect of normalized cache size (m) on shortest path solution for different cases.

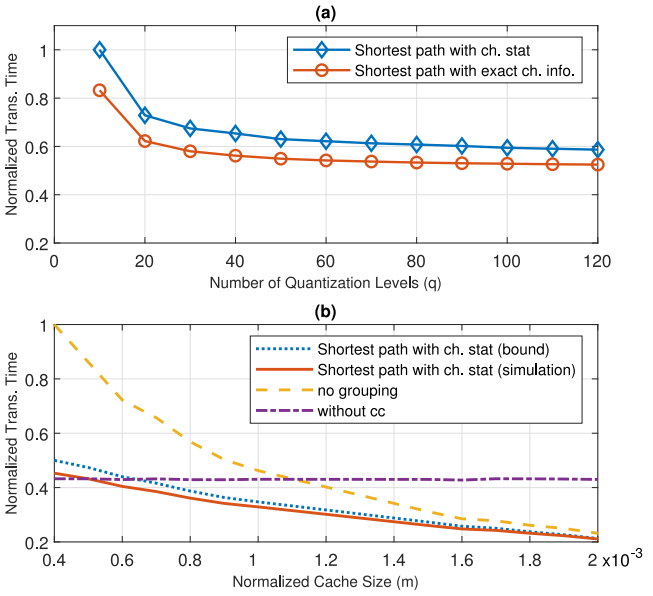

Fig. 3. (a) Effects of number of quantization levels on the shortest path solution, (b) Effects of normalized cache size (m) on the shortest path solution for i.i.d. user statistics.

that, for relatively low normalized cache sizes, grouping gain is particularly higher.

Next, without changing the other parameters of the network, we set the normalized cache size to $m = 0.0025$, and consider the number of quantization levels as a variable by changing the quantization step size. Fig. 3a shows the performance results for both cases of exact SNR knowledge and using channel statistics only. It is obvious that increasing the number of quantization levels decreases the normalized transmission times as it allows for further grouping opportunities for the users, however, the optimal solution may not contain all of the threshold values. For this specific example, when $q = 60$, the shortest path approach with the exact SNR knowledge divides the users into 20 groups while the one with the channel statistics forms 16 groups. We also observe that the performance with only the channel statistics is inferior to the one with the exact SNR knowledge, but it still performs better than no grouping and uncoded caching as shown in Figs. 2b and 3a. Note that, for this specific example, users send only $\lceil \log_2 60 \rceil = 6$ bits of feedback each instead of a heavy feedback required with instantaneous user SNRs.

Finally, in Fig. 3b, we consider a network whose parameters are the same as the one above, except all the user SNRs are exponentially distributed with mean -6 dB. We remark that, even when all of the users have same channel statistics, as the instantaneous SNRs are different due to channel fading, the proposed grouping approach attains lower transmission time than no grouping and uncoded caching. For instance, when $m = 10^{-3}$, the transmission time of the shortest path algorithm is 28.9% and 24.6% less than no grouping and uncoded caching, respectively. For further illustration, we also depict the direct result of the optimization problem which is a bound on the average transmission time on the same figure.

VI. Conclusion

We propose grouping of users for coded caching over non-ergodic fading channels to minimize the total transmission time by utilizing only the channel statistics of users for threshold determination, and relying on a few bits of feedback from each user indicating the group it belongs to. The first proposed approach for the threshold determination is a locally optimal iterative solution, while in the second one, we quantize the possible threshold values and obtain a shortest path problem enabling highly efficient solutions with a slight sacrifice in performance. The results demonstrate that user grouping for coded caching over wireless channels is highly advantageous, particularly, when the cache sizes are small.

References

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

[2] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[3] M. Ji and R.-R. Chen, "Caching and coded multicasting in slow fading environment," in *Proc. IEEE Wirel. Commun. Netw. Conf. (WCNC)*, Dec. 2017, pp. 1–6.

[4] A. Ghorbel, K.-H. Ngo, R. Combes, M. Kobayashi, and S. Yang, "Opportunistic content delivery in fading broadcast channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[5] K.-H. Ngo, S. Yang, and M. Kobayashi, "Cache-aided content delivery in MIMO channels," in *Proc. 54th Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, 2016, pp. 93–100.

[6] K.-H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.

[7] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 401–405.

[8] R. Combes, A. Ghorbel, M. Kobayashi, and S. Yang, "Utility optimal scheduling for coded caching in general topologies," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1692–1705, Aug. 2018.

[9] S. Abbott, *Understanding Analysis*. New York, NY, USA: Springer, 2001.

[10] L. Turner, "Variants of shortest path problems," *Alg. Oper. Res.*, vol. 6, no. 2, pp. 91–104, 2011.