# Routing-Based Delivery in Combination-Type Networks with Random Topology

Mozhgan Bayat, Kai Wan and Giuseppe Caire
Communications and Information Theory Group, Technische Universität Berlin, 10623 Berlin, Germany
E-mails: {bayat, kai.wan, caire}@tu-berlin.de



***Abstract*—The coded caching scheme proposed by Maddah-Ali and Niesen (MAN) transmits coded multicast messages to users equipped with caches and it is known to be optimal within a constant factor. This work extends this caching scheme to two-hop relay networks with one main server with access to a library of $N$ files, and $H$ relays communicating with $K$ users with cache, each of which is connected to a random subset of relays. This topology can be considered as a generalized version of a celebrated family of networks, referred to as *combination networks*. Our approach is simply based on routing MAN packets through the network. The optimization of the routing can be formulated as a Linear Program (LP). In addition, to reduce the computation complexity, a dynamic algorithm is proposed to approach the LP solution. Numerical simulations show that the proposed scheme outperforms the existing caching schemes for this class of networks.**

***Index Terms*—coded caching, random topology, linear optimization, uncoded placement, relay network, combination network.**

## I. Introduction

Due to the growing consumption of on-demand video and audio streaming services with prominent platforms like Youtube, Netflix and Spotify garnering billions of users on a daily basis, a clever usage of the low cost storage to cache data plays a key role in network design. One of the simplest methods is uncoded caching to duplicate popular files at the edge nodes in the networks. The authors in [1] introduced femto caching in heterogeneous wireless networks, where helpers posses high storage and are placed in a fixed position within the cell. The femto base stations cache popular files requested by mobile users. Recently, Fog Radio Access Network (F-RAN) has been proposed as a smart network based on cloud-RAN (C-RAN). In F-RAN the remote radio heads (RRHs) may posses a local cache as well as baseband processing units. In the prefetching phase RRHs store popular files in their cache memory. In [2], the authors study the delivery phase in F-RAN and design the cloud and edge processing jointly. Different transfer fronthaul strategies were proposed in [2] to maximize the delivery rate with limited fronthaul capacities and power constraints.

In parallel, coded caching strategy was orginally proposed by Maddah Ali and Niesen (MAN) in [3] for bottleneck networks with a shared error-free link. In the MAN setting, the server has a library of $N$ files of equal size and is connected to $K$ users through an error-free link. Each of the users is equipped with a cache memory. The MAN caching scheme consists of *prefetching* and *delivery* phases. The key idea is to treat the cache content as receiver side information and design the caches during the prefetching phase such that, during the delivery phase, the server can send coded multicast messages from which users can retrieve their desired packets with the aid of their own cache content. The MAN caching scheme which can lead to an additional coded caching gain compared to the conventional uncoded caching scheme, was proved in [4] to be optimal under the constraint of uncoded prefetching phase when $N \geq K$. By removing the redundant MAN multicast messages when $N < K$, the authors in [5] proposed an optimal caching scheme under the constraint of uncoded prefetching phase for any $N$ and $K$.

Recently, the coded caching approach was applied to different scenarios with specific topologies. [6]–[8] have investigated a class of two layered symmetric networks, referred to as *combination networks* (CN). The topology of these networks in determined by two parameters $H$ and $L$. There are a single server that connects to $H$ relays, and $\binom{H}{L}$ users each of which is connected to a different $L$-subset of relays. All links are error-free orthogonal. In [6] the authors applied the centralized MAN algorithm to create each coded multicast message, which is then encoded with a $(H, L)$ MDS code so that any user that receives $L$ out of the $H$ coded blocks will be able to decode the coded message. The server transmit one different MDS coded symbol for each MAN multicast message to each relay, which then forwards it to the connected users. In [7], the authors improved the coded caching scheme in [6] by using the network topology information. Notice that while the MDS coded scheme of [6] applies to any network topology as long as each user can receive from at least $L$ relays, the scheme of [7] critically hinges on the combination network topology and therefore does not generalize to networks with random topology, which is the focus of this work. In [8] an improvement of [7] obtained by removing redundancy is presented, and in [9] optimal use of transport and caching resources is achieved by using network coding on the set of elemental information objects being cached and transported over the network.

Cache-aided multiple relay networks with random topology was considered in [10], where each user is randomly connected to $L$ relays. By observing that for each relay, there are some MAN multicast messages which are not desired by any of its connected users, the authors in [10] presents an improved delivery strategy with respect to [6] by using the same MDS coding idea but only transmitting MDS coded symbol to the

relays which is connected to at least one user desiring this symbol.

### A. Our Contribution

This paper considers a generalized version of the cache-aided two-hop relay networks considered in [10], where a server is connected to multiple relays through unit capacity links and users randomly connect to an identical number of relays. In our work, we consider different capacity links and that the number of connected relays to each user is not necessarily identical. We use the centralized MAN cache placement and in the delivery phase, we first generate coded multicast messages and find the optimal routing for every coded multicast message through *Linear Programming* (LP). We also propose a dynamic programming which approaches the solution of the LP problem in a recursive manner with much less computational complexity than solving the original LP.

The distinguishing features of this paper are two-fold:

- We minimize the delivery latency and max-link load of the coded caching scheme through LP. We remove the transmission of non necessary messages and the parity blocks of the MDS coding.
- Our network topology is a general random relay network with links of different capacity and non-identical number of connected relay to each user. Even though the proposed LP-based delivery schemes require information about the network's topology, the prefetching phase is independent of the network topology.

## II. System Model and Related Results

### A. System Model

***Notation Convention:*** Calligraphic symbols denote sets, and bold symbols denote vectors. We use $|\cdot|$ to represent the cardinality of a set; $[a:b]:=\{a,a+1,\ldots,b\}$ and $[n]:=[1:n]$. $\mathcal{A}\setminus\mathcal{B}:=\{x\in\mathcal{A}:x\notin\mathcal{B}\}$;

Our scenario entails a relay network with a random topology and limited capacity error-free links. The server has access to a library consisting of $N$ files $\mathcal{F}=\{W_1,W_2,\ldots,W_N\}$ each of which contains $F$ bits. The server is connected to $H$ relays, each of which in turn is serving a random subset of $K$ users. All links are error-free and orthogonal. We focus on parallel transmission in which all links works in parallel. Notice that this is generalized setting of the case where every user is connected to exactly $L$ relays as described in [10], [11] and [6]. Additionally, each user is equipped with a cache memory capable of storing up to $MF$ bits, while relays do not posses any cache memory. The subset of users connected to relay $h\in[H]$ and the subset of relays connected to user $k\in[K]$ are denoted by $\mathcal{U}_h$ and $\mathcal{H}_k$, respectively. Similarly, for each subset of users $\mathcal{V}\subseteq[K]$, we define $\mathcal{H}_{\mathcal{V}}=\bigcup_{k\in\mathcal{V}}\mathcal{H}_k$ as the union of relays connected to users in $\mathcal{V}$. In other words, $\mathcal{H}_{\mathcal{V}}$ consists of relays that connect to at least one user in $\mathcal{V}$. In this paper, we consider the system with limited capacity. The relay nodes are connected to server via a fronthaul link has capacity $C_F$ bits per channel use, as well as they are connected to users through links with a given capacity $C_E$ bits per channel use.

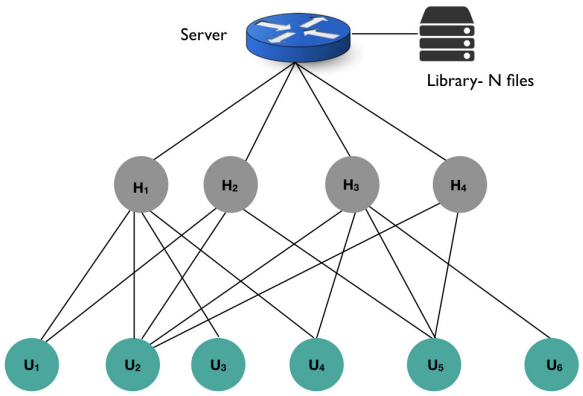


Fig. 1. A random topology with 4 relays and 6 users

In the prefetching phase, user $k\in[\mathsf{K}]$ stores information about the $N$ files in its cache of size $MF$ bits, where $M\in[0,N]$. This phase is done without knowledge of users' demands. We denote the content in the cache of user $k\in[K]$ by $Z_k$ and let $\mathbf{Z}:=(Z_1,\ldots,Z_K)$.

During the delivery phase, user $k\in[K]$ demands file $d_k\in[N]$; the demand vector $\mathbf{d}:=(d_1,\ldots,d_K)$ is revealed to all nodes. Given $(\mathbf{d},\mathbf{Z})$, the server sends a message $X_{s\to h}$ of $R_hF$ bits to relay $h\in[H]$. Then, relay $h\in[H]$ transmits a message $X_{h\to k}$ of $R_{h\to k}F$ bits to user $k\in\mathcal{U}_h$. User $k\in[K]$ must recover its desired file $F_{d_k}$ from $Z_k$ and $(X_{h\to k}:h\in\mathcal{H}_k)$ with high probability for some $F$. The objective is to determine the minimum worst-case transmission time,

$$T^\star:=\min_{\mathbf{Z}}\max_{\mathbf{d}\in[N]^K}\left\{\max_{h\in[H]}\frac{R_hF}{C_F},\max_{h\in[H],k\in\mathcal{U}_h}\frac{R_{h\to k}F}{C_E}\right\}. \quad (1)$$

### B. MAN Caching Scheme

In the following, we introduce the MAN caching scheme for bottleneck caching systems. we assume that the library replication parameter $t=KM/N$ (how many times the library can be contained in the collective cache memory) is an integer in $[0:K]$ (for non-integer numbers, the memory sharing is used). Each file $W_i$ is divided into $\binom{K}{t}$ non-overlapping and equal-length subfiles $W_i=\{W_{i,\mathcal{T}}:\mathcal{T}\subseteq[K],|\mathcal{T}|=t\}$ with user $k$ caching the segments $W_{i,\mathcal{T}}$ for which $k\in\mathcal{T}$. In the delivery phase, The server broadcasts one MAN coded multicast message for each group of users $\mathcal{S}$ where $\mathcal{S}\subseteq[K]$ and $|\mathcal{S}|=t+1$ as

$$V_{\mathcal{S}}=\bigoplus_{k\in\mathcal{S}}W_{d_k,\mathcal{S}\setminus\{k\}}. \quad (2)$$

Each user $k\in\mathcal{S}$ requires $W_{d_k,\mathcal{S}\setminus\{k\}}$ and knows all other subfiles in $V_{\mathcal{S}}$ such that it can recover $W_{d_k,\mathcal{S}\setminus\{k\}}$. By considering all the group of users with cardinality $t+1$, each user can recover its desired file.

## III. MAIN RESULT

### A. General Optimization Formulation

In this section, we introduce our proposed optimization problem for the routing of the coded multicast messages in (2). In order to deliver each MAN multicast message $V_{\mathcal{S}}$, we encode $V_{\mathcal{S}}$ into $H$ linearly independent random linear combination messages $X_{\mathcal{S}}^h$, $h \in [H]$, with $X_{\mathcal{S}}^h$ denoting the coded message corresponding to the relay $h$. In the following step the server transmits $X_{\mathcal{S}}^h, h \in [H]$, to the related relays. After receiving $X_{\mathcal{S}}^h$, relay $h$ transmits this message to the users in $\mathcal{S} \cap \mathcal{U}_h$. We define the normalized length of the message $X_{\mathcal{S}}^h$ as $y_{\mathcal{S}}^h := \frac{|X_{\mathcal{S}}^h|}{|V_{\mathcal{S}}|}$ where $0 \leq y_{\mathcal{S}}^h \leq 1$. We let $y_{\mathcal{S}}^h = 0$ when $h \in [H] \setminus \mathcal{H}_{\mathcal{S}}$; i.e., the coded multicast message $V_{\mathcal{S}}$ is delivered only through the relays which are connected to at least one user in $\mathcal{S}$. When the file size $F \to \infty$, the messages $X_{\mathcal{S}}^h, h \in \mathcal{H}_k$ are linearly independent with high probability such that user $k \in \mathcal{S}$ can recover the message $V_{\mathcal{S}}$ through these coded messages if

$$\sum_{h \in \mathcal{H}_k} y_S^h \geq 1, \qquad \forall k \in S. \tag{3}$$

The normalized transmission load over the link between the server and each relay $h \in [H]$ is the sum of the loads of all multicast messages and is given by

$$R_h = \sum_{\mathcal{S} \subseteq [K]: |\mathcal{S}| = t+1} y_S^h. \tag{4}$$

We also define the vector $\mathbf{R}_{\mathcal{H}} = \{R_1, R_2, \ldots, R_H\}$. Similarly, the normalized transmission load between each relay $h \in [H]$ and each user $k \in \mathcal{U}_h$ is equal to sum of the loads of all messages as follows

$$R_{h \to k} = \sum_{\mathcal{S} \subseteq [K]: |\mathcal{S}| = t+1, k \in \mathcal{S}} y_{\mathcal{S}}^h. \tag{5}$$

We consider the relay nodes are connected to server via a fronthaul link with capacity $C_F$ bits per channel use, as well as they are connected to users though links with a given capacity $C_E$. We focus on parallel transmission in which fronthaul links and relay-user links works in parallel. The delivery time would be maximum of fronthaul and relays-users delivery time. For given messages $X_{\mathcal{S}}^h$ in each block, the required time for fronthaul transmission can be computed as $\frac{|X_{\mathcal{S}}^h|}{C_F}$ , so that the worst-case delivery time in fronthual can be writen as follows

$$T_F = \frac{\max_{h \in [H]} R_h F}{C_F} = \max_{h \in [H]} \sum_{\mathcal{S} \subseteq [K]: |\mathcal{S}| = t+1} \frac{y_{\mathcal{S}}^h F}{C_F}. \tag{6}$$

For transmission toward the users the worst-case delivery time from relays to users can be computed as follows

$$\begin{aligned} T_E &= \max_{h \in [H], k \in \mathcal{U}_h} \frac{R_{h \to k} F}{C_E} \\ &= \max_{h \in [H]} \max_{k \in \mathcal{U}_h} \sum_{\mathcal{S} \subseteq [K]: |\mathcal{S}| = t+1, k \in \mathcal{S}} \frac{y_{\mathcal{S}}^h F}{C_E}. \end{aligned} \tag{7}$$

We are interested to minimize the worst-case delivery time by finding optimal random linear combination messages length. So that our linear optimization problem can be formalized as follows

$$\begin{aligned} \underset{y_{\mathcal{S}}^h}{\text{minimize}} \quad & \max\{T_E, T_F\} \\ \text{subject to} \quad & \forall \mathcal{S}, \text{ such that } \mathcal{S} \subseteq [K], |\mathcal{S}| = t+1 : \\ & 0 \leq y_{\mathcal{S}}^h \leq 1, \text{ if } h \in \mathcal{H}_{\mathcal{S}}, \\ & y_{\mathcal{S}}^h = 0, \text{ if } h \notin \mathcal{H}_{\mathcal{S}}, \\ & \sum_{h \in \mathcal{H}_k} y_{\mathcal{S}}^h \geq 1, \ \forall k \in \mathcal{S}. \end{aligned} \tag{8}$$

### B. Optimization Formulation for Identical Capacity Links

It is straightforward to show that $R_{h \to k} \leq R_h, \ \forall k \in \mathcal{U}_h$. Therefore, the optimization in the case of identical capacity links (for the simplicity, we consider $C_F = 1, C_E = 1$) reduces to allocating the lengths of the messages $X_{\mathcal{S}}^h$, such that the $\max_{h \in [H]} R_h$ is minimized. Since for all S, the MAN messages $V_{\mathcal{S}}$ have all the same length, this optimization can be expressed in terms of the normalized loads $y_{\mathcal{S}}^h$ allocation as the following LP

$$\begin{aligned} \underset{y_S^h}{\text{minimize}} \quad & \max_{h \in [H]} R_h \\ \text{subject to} \quad & \forall \mathcal{S}, \text{ such that } \mathcal{S} \subseteq [K], |\mathcal{S}| = t+1 : \\ & 0 \leq y_{\mathcal{S}}^h \leq 1, \text{ if } h \in \mathcal{H}_{\mathcal{S}}, \\ & y_{\mathcal{S}}^h = 0, \text{ if } h \notin \mathcal{H}_{\mathcal{S}}, \\ & \sum_{h \in \mathcal{H}_k} y_{\mathcal{S}}^h \geq 1, \ \forall k \in \mathcal{S}. \end{aligned} \tag{9}$$

### C. Dynamic Programming with Low Complexity

Each of the LPs in (8) and (9) contains $\binom{K}{t+1} \times H$ variables and $\binom{K}{t+1} \times (t+1) + H$ constraints, which grow exponentially with the number of users. For large value of $K$, it eventually becomes infeasible to solve the optimization problem. In order to overcome this difficulty, we used dynamic programming. In the following, we explain dynamic programming for the optimization with identical and unit capacities in (9). Similarly, the dynamic programming can be easily apply to the worst-case transmission time optimization problem in (8).

Consider the optimization problem in (9). We randomly divide all the $\binom{K}{t+1}$ MAN multicast messages into $G$ non-overlapping and 'equal-length' groups.[1] We also let $g := \lceil \binom{K}{t+1}/G \rceil$, representing the maximum number of MAN multicast messages included in one group. In addition, the set of the indices of the MAN multicast messages in the $i$-th group is denoted by $\mathcal{G}_i$. The optimization problem in (9) can be broken into smaller optimizations by iterating through the groups and using the results from the previous optimization as initial

[1] If $G$ does not divide $\binom{K}{t+1}$, each of the first $\binom{K}{t+1} - G\lfloor \binom{K}{t+1}/G \rfloor$ groups contains $\lceil \binom{K}{t+1}/G \rceil$ MAN multicast messages and each of the remaining groups contains $\lceil \binom{K}{t+1}/G \rceil - 1$ MAN multicast messages.

values. This allows us to apply minimization of the delivery load in each group individually in a sequential manner. This means that in the $i$-th step we take the previous optimal load allocations as a fixed initial load into the next $i+1$-th step's optimization problem. We define the load on relay $h$ in group $\mathcal{G}_i$ as

$$R_{h,i} = \sum_{\mathcal{S} \in \mathcal{G}_i} y_{\mathcal{S}}^h. \tag{10}$$

In the first step, we optimize the load allocation in $\mathcal{G}_1$ with initial normalized load $\mathbf{R}_1^0 = 0$. The initial normalized load for step $i > 1$ is obtained by summing all optimal normalized load allocation from previous steps as following

$$R_{h,i}^0 = \sum_{j=1}^{i-1} R_{h,j}^*, \tag{11}$$

where $R_{h,j}^*$ is optimal normalized load allocation in step $j$ for relay $h$ and also we denote the vector $\mathbf{R}_{\mathcal{H},i}^0 = \{R_{1,i}^0, R_{2,i}^0, \cdots, R_{H,i}^0\}$ as the initial normalized load allocation burden on relays in the $i$-th step. Accordingly, the total normalized load on relay $h$ until step $1 \leq i \leq G$ being equal to $R_{h,i} + R_{h,i}^0$. Having explained the dynamic programming aspect of Algorithm 1, we will now address the max-link load minimization problem for the $i$-th group with an initial normalized load allocation $\mathbf{R}_{\mathcal{H},i}^0$. Similary, the dynamic programming can be easily apply to the delivery time optimization problem in (8).

$$\begin{aligned} \underset{y_{\mathcal{S}}^h}{\text{minimize}} \quad & \max_{h \in [H]} R_{h,i} + R_{h,i}^0 \\ \text{subject to} \quad & \forall \mathcal{S}, \text{ such that } \mathcal{S} \in \mathcal{G}_i : \\ & 0 \leq y_{\mathcal{S}}^h \leq 1, \\ & y_{\mathcal{S}}^h = 0, \text{ if } h \notin \mathcal{H}_{\mathcal{S}}, \\ & \sum_{h \in \mathcal{H}_k} y_{\mathcal{S}}^h \geq 1, \ \forall k \in \mathcal{S}. \end{aligned} \tag{12}$$

Consider that $R_{h,i}^*$ is the optimal value for optimization problem in step $i$. We denote $\mathbf{R}_{\mathcal{H},i}^* = \{R_{h,i}^* | h \in [H]\}$ as the total optimal normalized load vector for step $i$.

**Algorithm 1** Optimization with dynamic programming

**Require:** geometry of network ,$g$, $G$, $K$, $t$, and $H$

1: Calculate $g := \lceil \binom{K}{t+1} / G \rceil$
2: Divide the set of all coded multicast messages randomly to $G$ group with g members
3: Set the initial value $\mathbf{R}_{\mathcal{H},i}^0 = \mathbf{0}$
4: **for** $i = 1 : G$ **do**
5: Solve the optimization problem in (12)
6: $\mathbf{R}_{\mathcal{H},i}^* = \arg\min \max_{h \in [H]} R_{h,i} + R_{h,i}^0$
7: Calculate the initial load $R_{H,i+1}^0$ for next optimization
8: **end for**

## IV. RESULTS AND DISCUSSIONS

In this section, we focus on the case that each user is connected to $L$ relays and compare the worst-case loads or the delivery times achieved by the proposed scheme and the existing schemes. As previously discussed, our scenario is not constrained to have the same number of connections on the user's side. In order to compare our scheme with previous works based on MDS coding [6], [10], we assume that the users have the same number of connections. In our scheme, the transmitted messages to relays have different length but in the works [6], [10] the transmitted messages have same length.

In all of the figures, the curves labeled "MDS" refer to coded caching scheme for combination networks as described in [6]. In "MDS" work, first the coded multicast messages $V_{\mathcal{S}}$ are created by centralized MAN caching scheme. In the next step, each message is divided into $L$ subfiles which are then encoded by $(H, L)$ MDS code. Lastly, each of these $H$ MDS coded messages is transmitted to the related relays, which in broadcast the received coded messages to its users. The scenario proposed in [10] has been modified to our case by considering that the relays are not equipped with cache memory and curves for this scenario are labeled by "MGL ". The "MGL" scheme removed some redundancy from [6] by avoiding the transmission of coded multicast messages to relays connected to none of users interested in that specific message. The curves labeled by "LP " refer to the result of our work obtained through the optimization in (12). The numerical results are obtained by using Monte Carlo simulations that rely on repeated random topologies in which each user select L relays at random over all possible relays for 500 times. Fig. 2 and 3 show the worst-case load versus $g$ (maximum number of the MAN multicast messages in one group). Notice that the schemes "MGL" and " MDS" are independent from $g$. As seen from these numerical results, compared to the previous schemes our scenario dramatically reduces the worst-case load. For an increasing $g$ each optimization in iterative method will include more messages and as a consequence the load association of the relays performs better. In Fig. 4 we compare the delivery time of the proposed scheme for different capacities $C_E$ and number of connections on the user side $L$ with a unit fronthaul capacity $C_F = 1$. For smaller capacities, the relay-user links are the bottleneck of our scenario. After increasing $C_E$ up to a certain value, the server-relay links become the bottleneck of our scenario.

We used the simplex and interior point methods to solve the optimization problems. The complexity of simplex on average is $O(n^3)$ and $O(n^2 2^n)$ in the worst-case, while the complexity of the interior point method is $O(n^{3.5})$, where $n$ is the number of variables [12], [13]. The complexity order of the interior point method for a scenario is shown in Fig. 5.

## V. CONCLUSION

In this paper we investigated relay networks where the end users are equipped with caches. We extended the MAN caching scheme for the bottleneck networks to any cache-aided relay networks. For different scenarios with and without

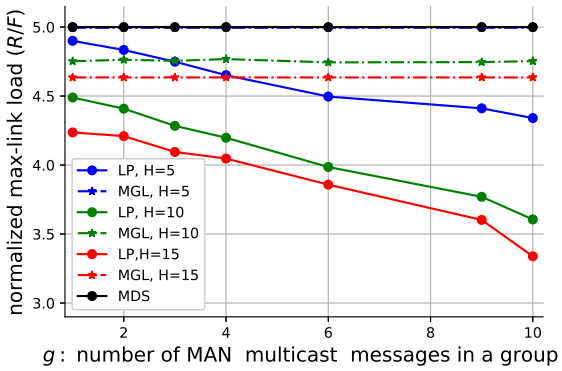


Fig. 2. The normalized max-link load versus $g$ for a scenario with $K = 5$ and $H = 5, 10, 15$ and $t = 2$

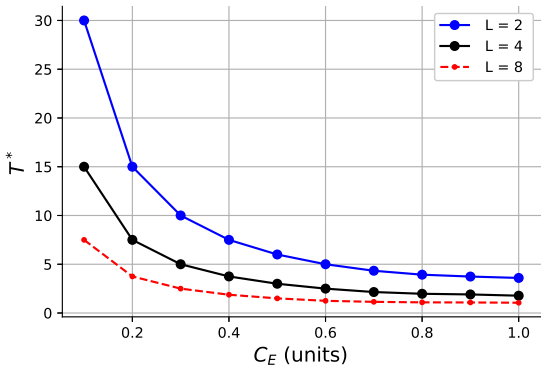


Fig. 4. The delivery time versus $C_E$ for a scenario with $K = 5$ and $H = 10$ and $t = 2$ and $C_F = 1$ unit

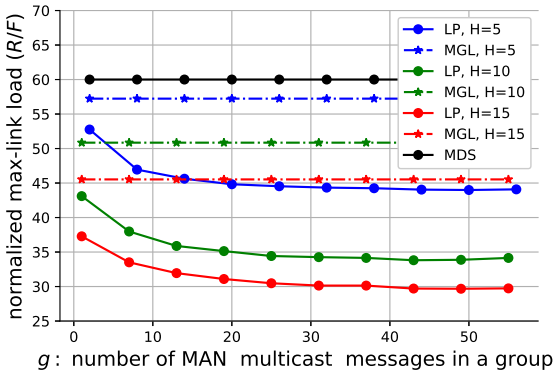


Fig. 3. The normalized max-link load versus $g$ for a scenario with $K = 10$ and $H = 5, 10, 15$ and $t = 2$ with interior point method

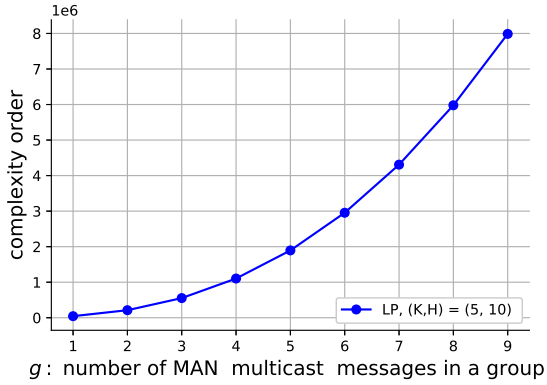


Fig. 5. Complexity order of whole scenario optimization with interior point method

unit/identical link capacities, we design a delivery scheme by solving linear optimization problems. It was concluded in the numerical results that the proposed schemes outperform the state-of-the-arts schemes in both scenarios. In addition, we also proposed a dynamic algorithm to approach the solution of each optimization problem with a lower computation complexity.

## REFERENCES


[1] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 1107–1115.

[2] S. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 315–319.

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

[4] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," *in IEEE Infor. Theory Workshop*, Sep. 2016.

[5] Q. Yu, M. A. Maddah-Ali, and S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *in IEEE Int. Symp. Inf. Theory*, Jun. 2017.

[6] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Caching in combination networks," in *Signals, Systems and Computers, 2015 49th Asilomar Conference on*. IEEE, 2015, pp. 1269–1273.

[7] A. A. Zewail and A. Yener, "Coded caching for combination networks with cache-aided relays," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2017, pp. 2433–2437.

[8] K. Wan, M. Ji, P. Piantanida, and D. Tuninetti, "Caching in combination networks: Novel multicast message generation and delivery by leveraging the network topology," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[9] J. Llorca, A. M. Tulino, K. Guan, and D. C. Kilper, "Network-coded caching-aided multicast for efficient content delivery," in *2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 3557–3562.

[10] N. Mital, D. Gündüz, and C. Ling, "Coded caching in a multi-server system with random topology," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[11] M. Bayat, R. K. Mungara, and G. Caire, "Coded caching in a cell-free simo network," in *WSA 2018; 22nd International ITG Workshop on Smart Antennas*, March 2018, pp. 1–8.

[12] V. Klee and G. J. Minty, "How good is the simplex algorithm," WASHINGTON UNIV SEATTLE DEPT OF MATHEMATICS, Tech. Rep., 1970.

[13] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.