

# Benefits of Coded Placement for Networks with Heterogeneous Cache Sizes

Abdelrahman M. Ibrahim, Ahmed A. Zewail, and Aylin Yener

Wireless Communications and Networking Laboratory (WCAN)  
 Electrical Engineering and Computer Science Department  
 The Pennsylvania State University, University Park, PA 16802.  
 {ami137,zewail}@psu.edu      yener@engr.psu.edu

**Abstract**—In this work, we study coded placement in caching systems where the users have unequal cache sizes and demonstrate its performance advantage. In particular, we propose a caching scheme with coded placement for three-user systems that outperforms the best caching scheme with uncoded placement. In our proposed scheme, users cache both uncoded and coded pieces of the files, and the coded pieces at the users with large memories are decoded using the unicast/multicast signals intended to serve users with smaller memories. Furthermore, we extend the proposed scheme to larger systems and show the reduction in delivery load with coded placement compared to uncoded placement.

## I. INTRODUCTION

Coded caching [1] alleviates network congestion during peak-traffic hours, known as the *delivery phase*, by placing some of the data in the cache memories at the network edge during off-peak hours, known as the *placement phase*. Reference [1] has shown that the joint design of the two phases leads to significant reduction in the delivery load, which is achieved by designing the cache contents in the placement phase in a manner that allows serving the users using multicast transmissions in the delivery phase. In turn, there exists a fundamental trade-off between the delivery load on the server and the cache sizes in the network, which has been studied in several setups [1]–[10]. Recently, references [2]–[5] have studied the effect of heterogeneity in cache sizes at the users on the delivery load memory trade-off. In particular, we have shown that placement and delivery schemes jointly optimized with respect to given cache sizes provide significant improvement over schemes tailored to uniform cache sizes. Additionally, cache sizes at the end users may be optimized for further gain [6], [7]. All of these consider placement of partial files in the caches, i.e., uncoded placement.

As eluded to above coded caching schemes are often categorized according to whether in the placement phase coding over the files is utilized or not. In caching with uncoded placement, the server places uncoded pieces of each file in the cache memories of the users [1], [3]–[5]. Alternatively, in systems with *coded placement*, the server places coded pieces of the files in the users' caches which are decoded using the transmissions in the delivery phase [9], [10]. While uncoded placement is sufficient for some systems, clearly, coding over files in general has the potential to perform better.

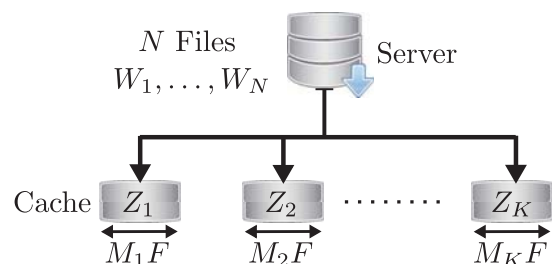


Fig. 1: Caching system with heterogeneous cache sizes.

For systems with equal cache sizes, references [9], [10] have shown that coded placement is beneficial in the small memory regime when the number of files is less than or equal the number of users. Recent reference [11] has shown that coded placement is essential in achieving the optimal delivery load in a two-user system when the cache sizes of the two users differ.

In this work, we show that coded placement in systems where the users are equipped with heterogeneous cache sizes, outperforms uncoded placement. We show that coded placement not only increases the local caching gain at the users, but also increases the multicast gain in the system. In particular, we propose a caching scheme with coded placement for three-user systems that illustrates the reduction in the worst-case delivery load compared to the best caching scheme with uncoded placement in [5]. In our proposed scheme, users cache both uncoded and coded pieces of the files, and users with large memories decode the cached coded pieces using the transmissions intended to serve users with smaller memories. We observe that the gain from coded placement increases as the difference between the cache sizes increases, and decreases with the number of files. We extend the proposed scheme to systems with more than three end-users and show the reduction in delivery load with coded placement. This work thus establishes the first result showing a clear benefit of coded placement in a heterogeneous caching systems with three or more users.

*Notation:* Vectors are represented by boldface letters,  $\oplus$  refers to the binary XOR operation,  $|W|$  denotes cardinality of  $W$ ,  $\mathcal{A} \setminus \mathcal{B}$  denotes the set of elements in  $\mathcal{A}$  and not in  $\mathcal{B}$ ,  $[K] \triangleq \{1, \dots, K\}$ , and  $\phi$  denotes the empty set.

## II. SYSTEM MODEL

We consider a caching system where a single server is connected to  $K$  users via a shared error-free multicast link [1], as shown in Fig. 1. The server has access to a library  $\{W_1, \dots, W_N\}$  of  $N$  independent files, each with size  $F$  symbols over the field  $\mathbb{F}_{2^r}$ . We consider a heterogeneous system, where user  $k$  is equipped with a cache memory of size  $M_k F$  symbols. Without loss of generality, we assume that  $M_1 \leq M_2 \leq \dots \leq M_K$ . Additionally, we define  $m_k = M_k/N$  to denote the memory size of user  $k$  normalized by the library size  $NF$ , i.e.,  $m_k \in [0, 1]$  for  $M_k \in [0, N]$ .

The system operates over two phases: placement phase and delivery phase. In the placement phase, the server populates the users' cache memories without the knowledge of users' demands that will be made in the delivery phase. The cached content by user  $k$  is denoted by  $Z_k$ , i.e.,  $|Z_k| \leq M_k F$ . In the delivery phase, user  $k$  requests a file  $W_{d_k}$  from the server. The users' demands are uniform and independent as in [1]. To serve the users' demands, the server transmits a sequence of unicast/multicast signals,  $X_{\mathcal{T},d}$ . At the end of the delivery phase, user  $k$  must be able to decode  $\hat{W}_{d_k}$  reliably. Formally, for a given normalized cache size vector  $\mathbf{m} \triangleq [m_1, \dots, m_K]$ , the worst-case delivery load  $R(\mathbf{m}) \triangleq \sum_{\mathcal{T}} |X_{\mathcal{T},d}|/F$  is said to be achievable if for every  $\epsilon > 0$  and large enough  $F$ , there exists a caching scheme such that  $\max_{\mathbf{d}, k \in [K]} \Pr(\hat{W}_{d_k} \neq W_{d_k}) \leq \epsilon$ . Our achievability scheme utilizes maximum distance separable (MDS) codes which are defined as follows.

**Definition 1.** [12] An  $(n, k)$  maximum distance separable (MDS) code is an erasure code that allows recovering  $k$  initial information symbols from any  $k$  out of the  $n$  coded symbols. In a systematic  $(n, k)$ -MDS code the first  $k$  symbols in the output codeword is the information symbols. That is, we have

$$\begin{aligned} [i_1, \dots, i_k] \mathbf{G}_{k \times n} &= [i_1, \dots, i_k] [\mathbf{I}_{k \times k} \ \mathbf{P}_{k \times n-k}] \\ &= [i_1, \dots, i_k, c_{k+1}, \dots, c_n], \end{aligned} \quad (1)$$

where  $\mathbf{G}_{k \times n}$  is the code generator matrix and  $\mathbf{I}_{k \times k}$  is an identity matrix. ■

For a systematic  $(2N - j, N)$  MDS-code, we define

$$\sigma_j([i_1, \dots, i_N]) \triangleq [i_1, \dots, i_N] \mathbf{P}_{N \times N-j} \quad (2)$$

to denote the  $N - j$  parity symbols in the output codeword. Note that  $\sigma_j([i_1, \dots, i_N])$  represents  $N - j$  independent equations in the information symbols  $[i_1, \dots, i_N]$ . For example,  $\sigma_1([i_1, \dots, i_N]) = [i_1 \oplus i_2, i_2 \oplus i_3, \dots, i_{N-1} \oplus i_N]$ .

## III. RESULTS

In this section, we present our results showing the reduction in the delivery load when coded placement is utilized. In Theorem 1, we characterize an achievable delivery load for three-user systems, that is lower than the minimum worst-case delivery load under uncoded placement, given by

$$R_{\text{uncoded}}^*(\mathbf{m}) = \max \left\{ 3 - 3m_1 - 2m_2 - m_3, \frac{5}{3} - \frac{3m_1 - 2m_2 - m_3}{3}, 2 - 2m_1 - m_2, 1 - m_1 \right\}, \quad (3)$$

which has been characterized in [5].

**Theorem 1.** For a three-user system with  $N \geq 4$  and  $m_1 \leq m_2 \leq m_3$ , the worst-case delivery load

$$R_{\text{coded}}(\mathbf{m}, N) = \max \left\{ 3 - 3m_1 - 2m_2 - m_3 - \frac{3(m_2 - m_1)}{N - 1} - \frac{2(m_3 - m_2)}{N - 2}, \frac{5}{3} - \frac{3m_1 - 2m_2 - m_3}{3} - \frac{m_2 - m_1}{3(N - 1)}, 2 - 2m_1 - m_2 - \frac{m_2 - m_1}{N - 1}, 1 - m_1 \right\}, \quad (4)$$

is achievable with coded placement. □

*Proof.* The reduction in the delivery load in (4) compared to (3) is achieved by placing coded pieces of the files at users 2 and 3, which are decoded in the delivery phase. For example, in order to achieve  $R_{\text{coded}}(\mathbf{m}, N) = R_{\text{uncoded}}^*(\mathbf{m}) - \frac{m_2 - m_1}{3(N - 1)}$ , part of the multicast signal to users  $\{1, 2\}$  is utilized in decoding the cached pieces at user 3. The proposed caching scheme is presented in Section IV. □

Next theorem characterizes the gain achieved by coded placement in the small memory regime, where the unicast signals intended for users  $\{1, \dots, k\}$  are utilized in decoding the cache content at users  $\{k + 1, \dots, K\}$ .

**Theorem 2.** For a  $K$ -user system with  $N \geq K + 1$ ,  $m_1 \leq m_2 \leq \dots \leq m_K$ , and

$$\sum_{i=1}^K m_i + \sum_{i=2}^K \frac{(i-1)(K-i+1)(m_i - m_{i-1})}{N - i + 1} \leq 1,$$

the worst-case delivery load

$$R_{\text{coded}}(\mathbf{m}, N, K) = R_{\text{uncoded}}^*(\mathbf{m}, K) - \sum_{i=2}^K \frac{(i-1)(K-i+1)(K-i+2)(m_i - m_{i-1})}{2(N - i + 1)}, \quad (5)$$

is achievable with coded placement, where

$$R_{\text{uncoded}}^*(\mathbf{m}, K) = K - \sum_{i=1}^K (K - i + 1)m_i, \quad (6)$$

is the minimum worst-case delivery load with uncoded placement for  $\sum_{i=1}^K m_i \leq 1$  [4], [5]. □

*Proof.* The caching scheme is detailed in Section V. □

**Remark 1.** For given  $K$  and  $\mathbf{m}$ ,  $\lim_{N \rightarrow \infty} R_{\text{coded}}(\mathbf{m}, N, K) = R_{\text{uncoded}}^*(\mathbf{m}, K)$ . That is, the gain due to coded placement decreases with  $N$  and is negligible for  $N \gg K$ . ■

## IV. PROOF OF THEOREM 1

In this section, we present our caching schemes for three-user systems. The achievable delivery load in Theorem 1 consists of the following regions

- Region I: If  $\sum_{i=1}^3 m_i + \frac{2(m_2 - m_1)}{N - 1} + \frac{2(m_3 - m_2)}{N - 2} \leq 1$ , then  $R_{\text{coded}} = 3 - 3m_1 - 2m_2 - m_3 - \frac{3(m_2 - m_1)}{N - 1} - \frac{2(m_3 - m_2)}{N - 2}$ .
- Region II: If  $\sum_{i=1}^3 m_i + \frac{2(m_2 - m_1)}{N - 1} + \frac{2(m_3 - m_2)}{N - 2} > 1$ ,  $Nm_3 \leq (N + 3)m_2 + 3(N - 2)m_1 - (N - 1)$ , and

$Nm_3 \leq 2(N-1) - (2N-3)m_2$ , then  $R_{\text{coded}} = 5/3 - m_1 - 2m_2/3 - m_3/3 - (m_2 - m_1)/(3(N-1))$ .

- **Region III:** If  $\sum_{i=1}^3 m_i + \frac{2(m_2 - m_1)}{N-1} + \frac{2(m_3 - m_2)}{N-2} > 1$ ,  $Nm_3 > (N+3)m_2 + 3(N-2)m_1 - (N-1)$ , and  $Nm_2 + (N-2)m_1 \leq N-1$ , then  $R_{\text{coded}} = 2 - 2m_1 - m_2 - (m_2 - m_1)/(N-1)$ .
- **Region IV:** If  $Nm_2 + (N-2)m_1 > N-1$ , and  $Nm_3 > 2(N-1) - (2N-3)m_2$ , then  $R_{\text{coded}} = 1 - m_1$ .

Region I is a special case of Theorem 2 which will be explained in Section V. Next, we consider regions II to IV.

#### A. Region II

1) *Placement Phase:* Each file  $W_n$  is split into subfiles  $W_{n,1}$ ,  $W_{n,2}$ ,  $W_{n,3}$ ,  $W_{n,\{1,2\}}$ ,  $\{W_{n,\{1,3\}}^{(1)}, W_{n,\{1,3\}}^{(2)}\}$ , and  $\{W_{n,\{2,3\}}^{(1)}, W_{n,\{2,3\}}^{(2)}, W_{n,\{2,3\}}^{(3)}, W_{n,\{2,3\}}^{(4)}\}$ , such that

$$|W_{n,\{1\}}| = \left(\frac{2}{3} - m_1 - \frac{N(m_3 - m_2)}{3(N-1)}\right)F, \quad (7)$$

$$|W_{n,\{2\}}| = |W_{n,\{3\}}| = |W_{n,\{1\}}| - (m_2 - m_1)F, \quad (8)$$

$$|W_{n,\{2,3\}}^{(3)}| = |W_{n,\{2,3\}}^{(4)}| = (m_2 - m_1)F, \quad (9)$$

$$\begin{aligned} |W_{n,\{1,2\}}| &= |W_{n,\{1,3\}}^{(1)}| = |W_{n,\{2,3\}}^{(1)}| \\ &= \left(m_1 - 1/3 - \frac{N(m_3 - m_2)}{3(N-1)}\right)F, \end{aligned} \quad (10)$$

$$|W_{n,\{1,3\}}^{(2)}| = |W_{n,\{2,3\}}^{(2)}| = \frac{N(m_3 - m_2)}{(N-1)}F, \quad (11)$$

where all subfiles are cached uncoded except for  $W_{n,\{2,3\}}^{(2)}, \forall n$  are encoded before being placed at user 3. More specifically, the cache contents are given as

$$Z_1 = \bigcup_{n=1}^N \left( W_{n,\{1\}} \cup W_{n,\{1,2\}} \cup W_{n,\{1,3\}}^{(1)} \cup W_{n,\{1,3\}}^{(2)} \right), \quad (12)$$

$$Z_2 = \bigcup_{n=1}^N \left( W_{n,\{2\}} \cup W_{n,\{1,2\}} \cup \left( \bigcup_{i=1}^4 W_{n,\{2,3\}}^{(i)} \right) \right), \quad (13)$$

$$\begin{aligned} Z_3 &= \bigcup_{n=1}^N \left( W_{n,\{3\}} \cup \left( \bigcup_{i=1}^2 W_{n,\{1,3\}}^{(i)} \right) \cup \left( \bigcup_{i=1, i \neq 2}^4 W_{n,\{2,3\}}^{(i)} \right) \right) \\ &\quad \cup \sigma_1([W_{1,\{2,3\}}^{(2)}, \dots, W_{N,\{2,3\}}^{(2)}]), \end{aligned} \quad (14)$$

2) *Delivery Phase:* The server sends the following multicast signals

$$X'_{\{1,2\},\mathbf{d}} = W_{d_2,\{1\}} \oplus \left( W_{d_1,\{2\}} \cup W_{d_1,\{2,3\}}^{(3)} \right), \quad (15)$$

$$X_{\{1,3\},\mathbf{d}} = W_{d_3,\{1\}} \oplus \left( W_{d_1,\{3\}} \cup W_{d_1,\{2,3\}}^{(4)} \right), \quad (16)$$

$$X_{\{2,3\},\mathbf{d}} = W_{d_3,\{2\}} \oplus W_{d_2,\{3\}}, \quad (17)$$

$$X_{\{1,2,3\},\mathbf{d}} = W_{d_3,\{1,2\}} \oplus W_{d_2,\{1,3\}}^{(1)} \oplus W_{d_1,\{2,3\}}^{(1)}, \quad (18)$$

$$X''_{\{1,2\},\mathbf{d}} = W_{d_2,\{1,3\}}^{(2)} \oplus W_{d_1,\{2,3\}}^{(2)}. \quad (19)$$

3) *Achievability:* The proposed placement scheme is valid since the cache sizes constraints are satisfied. In the delivery phase, the users retrieve the requested pieces from the multicast signals using the cached subfiles. Additionally, using the

multicast signal  $X''_{\{1,2\},\mathbf{d}}$  and the cached piece  $W_{d_2,\{1,3\}}^{(2)}$ , user 3 decodes  $W_{d_1,\{2,3\}}^{(2)}$ , which is used in retrieving  $W_{d_3,\{2,3\}}^{(2)}$  from  $\sigma_1([W_{1,\{2,3\}}^{(2)}, \dots, W_{N,\{2,3\}}^{(2)}])$ .

#### B. Region III

1) *Placement Phase:* Each file  $W_n$  is split into  $W_{n,1}$ ,  $\{W_{n,\{2\}}^{(1)}, W_{n,\{2\}}^{(2)}\}$ ,  $\{W_{n,\{3\}}^{(1)}, W_{n,\{3\}}^{(2)}, W_{n,\{3\}}^{(3)}\}$ ,  $W_{n,\{1,3\}}$ , and  $\{W_{n,\{2,3\}}^{(1)}, W_{n,\{2,3\}}^{(2)}\}$ , such that

$$|W_{n,\{1\}}| + |W_{n,\{1,3\}}| = |W_{n,\{2\}}^{(1)}| + |W_{n,\{2,3\}}^{(1)}| = m_1F, \quad (20)$$

$$|W_{n,\{1,3\}}| = |W_{n,\{2,3\}}^{(1)}| = m_1F - |W_{n,\{3\}}^{(1)}|, \quad (21)$$

$$|W_{n,\{2\}}^{(2)}| = |W_{n,\{3\}}^{(2)}| = \frac{N(m_2 - m_1)}{N-1}F - |W_{n,\{2,3\}}^{(2)}|. \quad (22)$$

In particular, we have the following three cases.

- For  $m_3 \leq \frac{N-2}{N} + \frac{m_2}{N-1} - \frac{(2N-3)(N-2)m_1}{(N-1)(N)}$

$$|W_{n,\{2,3\}}^{(2)}| = \left( \sum_{i=1}^3 m_i + \frac{2(m_2 - m_1)}{N-1} + \frac{2(m_3 - m_2)}{N-2} - 1 \right)F, \quad (23)$$

$$|W_{n,\{3\}}^{(3)}| = \frac{N(m_3 - m_2)}{N-2}F, \quad |W_{n,\{3\}}^{(1)}| = m_1F. \quad (24)$$

- For  $\frac{m_2}{N-1} - \frac{(2N-3)(N-2)m_1}{(N-1)(N)} < m_3 - \frac{N-2}{N} \leq \frac{m_2}{N-1} - \frac{m_1}{(N-1)(N)}$

$$|W_{n,\{2,3\}}^{(2)}| = \frac{N(m_2 - m_1)}{N-1}F, \quad (25)$$

$$|W_{n,\{3\}}^{(3)}| = \left( 1 - 2m_1 - \frac{N(m_2 - m_1)}{N-1} \right)F - |W_{n,\{3\}}^{(1)}|, \quad (26)$$

$$\begin{aligned} |W_{n,\{3\}}^{(1)}| &= \left( \frac{N-2}{2N-3} \right) \left( 1 - \frac{(N-3)m_1}{(N-2)} - \frac{N(m_2 - m_1)}{N-1} \right. \\ &\quad \left. - \frac{N(m_3 - m_2)}{N-2} \right)F. \end{aligned} \quad (27)$$

- For  $m_3 > \frac{N-2}{N} + \frac{m_2}{N-1} - \frac{m_1}{(N-1)(N)}$

$$|W_{n,\{3\}}^{(3)}| = \left( 1 - 2m_1 - \frac{N(m_2 - m_1)}{N-1} \right)F, \quad (28)$$

$$|W_{n,\{3\}}^{(1)}| = 0, \quad |W_{n,\{2,3\}}^{(2)}| = \frac{N(m_2 - m_1)}{N-1}F, \quad (29)$$

The cache contents are defined as

$$Z_1 = \bigcup_{n=1}^N \left( W_{n,\{1\}} \cup W_{n,\{1,3\}} \right), \quad (30)$$

$$\begin{aligned} Z_2 &= \bigcup_{n=1}^N \left( W_{n,\{2\}}^{(1)} \cup W_{n,\{2,3\}}^{(1)} \right) \cup \sigma_1([W_{1,\{2\}}^{(2)}, \dots, W_{N,\{2\}}^{(2)}]) \\ &\quad \cup \sigma_1([W_{1,\{2,3\}}^{(2)}, \dots, W_{N,\{2,3\}}^{(2)}]), \end{aligned} \quad (31)$$

$$\begin{aligned} Z_3 &= \bigcup_{n=1}^N \left( W_{n,\{3\}}^{(1)} \cup W_{n,\{2,3\}}^{(1)} \right) \cup \sigma_1([W_{1,\{3\}}^{(2)}, \dots, W_{N,\{3\}}^{(2)}]) \\ &\quad \cup \sigma_1([W_{1,\{2,3\}}^{(2)}, \dots, W_{N,\{2,3\}}^{(2)}]) \cup \sigma_1([W_{1,\{1,3\}}, \dots \\ &\quad , W_{N,\{1,3\}}]) \cup \sigma_2([W_{1,\{3\}}^{(3)}, \dots, W_{N,\{3\}}^{(3)}]). \end{aligned} \quad (32)$$

2) *Delivery Phase*:: The server sends the following multicast signals

$$X_{\{1,2\},\mathbf{d}} = (W_{d_2,\{1\}} \cup W_{d_2,\{1,3\}}) \oplus (W_{d_1,\{2\}}^{(1)} \cup W_{d_1,\{2,3\}}^{(1)}), \quad (33)$$

$$X_{\{2,3\},\mathbf{d}} = (W_{d_3,\{2\}}^{(1)} \cup W_{d_3,\{2\}}^{(2)}) \oplus (W_{d_2,\{3\}}^{(1)} \cup W_{d_2,\{3\}}^{(2)}), \quad (34)$$

$$X_{\{1,3\},\mathbf{d}} = W_{d_3,\{1\}} \oplus W_{d_1,\{3\}}^{(1)}. \quad (35)$$

The following unicast signals complete the requested files and help users  $\{2, 3\}$  in decoding their cache contents.

$$X_{\{1\},\mathbf{d}} = W_{d_1,\{2\}}^{(2)} \cup W_{d_1,\{2,3\}}^{(2)} \cup W_{d_1,\{3\}}^{(2)} \cup W_{d_1,\{3\}}^{(3)}, \quad (36)$$

$$X_{\{2\},\mathbf{d}} = W_{d_2,\{3\}}^{(3)}. \quad (37)$$

3) *Achievability*: User 2 decodes its cache using  $W_{d_1,\{2\}}^{(2)}, W_{d_1,\{2,3\}}^{(2)}$  from  $X_{\{1\},\mathbf{d}}$ . Similarly, user 3 decodes its cache using  $W_{d_1,\{2,3\}}^{(2)}, W_{d_1,\{3\}}^{(2)}, W_{d_1,\{3\}}^{(3)}$  from  $X_{\{1\},\mathbf{d}}$  and  $W_{d_2,\{3\}}^{(3)}$  from  $X_{\{2\},\mathbf{d}}$ .

### C. Region IV

Next, we consider the case where  $m_1 + m_2 \leq 1$ , since uncod placement is optimal for  $m_1 + m_2 > 1$  [5].

1) *Placement Phase*: Each file  $W_n$  is split into  $W_{n,\{1,2\}}, W_{n,\{1,3\}}$ , and  $\{W_{n,\{2,3\}}^{(1)}, W_{n,\{2,3\}}^{(2)}, W_{n,\{2,3\}}^{(3)}\}$ , such that

$$|W_{n,\{1,2\}}| = m_1 F - |W_{n,\{1,3\}}|, \quad (38)$$

$$|W_{n,\{1,2\}}| = |W_{n,\{2,3\}}^{(2)}|, \quad |W_{n,\{1,3\}}| = |W_{n,\{2,3\}}^{(1)}|, \quad (39)$$

$$|W_{n,\{2,3\}}^{(3)}| = (1 - 2m_1)F. \quad (40)$$

In particular, we have

$$|W_{n,\{1,2\}}| = \begin{cases} \left( \frac{Nm_2 + (N-2)m_1}{N-1} - 1 \right) F, & \text{if } m_3 \leq \frac{(N-1)+m_1}{N}, \\ 0, & \text{otherwise.} \end{cases} \quad (41)$$

The cache contents are defined as

$$Z_1 = \bigcup_{n=1}^N (W_{n,\{1,2\}} \cup W_{n,\{1,3\}}), \quad (42)$$

$$Z_2 = \bigcup_{n=1}^N (W_{n,\{2,3\}}^{(1)} \cup W_{n,\{2,3\}}^{(2)}) \cup \sigma_1([W_{1,\{1,2\}}, \dots, W_{N,\{1,2\}}]) \cup \sigma_1([W_{1,\{2,3\}}^{(3)}, \dots, W_{N,\{2,3\}}^{(3)}]), \quad (43)$$

$$Z_3 = \bigcup_{n=1}^N (W_{n,\{2,3\}}^{(1)} \cup W_{n,\{2,3\}}^{(2)}) \cup \sigma_1([W_{1,\{1,3\}}, \dots, W_{N,\{1,3\}}]) \cup \sigma_1([W_{1,\{2,3\}}^{(3)}, \dots, W_{N,\{2,3\}}^{(3)}]). \quad (44)$$

2) *Delivery Phase*: The server sends the following signals

$$X_{\{1,2\},\mathbf{d}} = W_{d_2,\{1,3\}} \oplus W_{d_1,\{2,3\}}^{(1)}, \quad (45)$$

$$X_{\{1,3\},\mathbf{d}} = W_{d_3,\{1,2\}} \oplus W_{d_1,\{2,3\}}^{(2)}, \quad (46)$$

$$X_{\{1\},\mathbf{d}} = W_{d_1,\{2,3\}}^{(3)}. \quad (47)$$

3) *Achievability*: User 2 retrieves  $W_{d_2,\{1,2\}}$  from its cache using  $W_{d_3,\{1,2\}}$  which is extracted from  $X_{\{1,3\},\mathbf{d}}$ . Similarly, user 3 retrieves  $W_{d_3,\{1,3\}}$  by utilizing  $X_{\{1,2\},\mathbf{d}}$ .

## V. PROOF OF THEOREM 2

In this section, we explain the caching schemes that achieves the delivery load in Theorem 2.

### A. Placement Phase

File  $W_n$  is divided into  $K(K+1)/2+1$  subfiles,  $W_{n,\phi}, W_{n,1}^{(1)}, \{W_{n,2}^{(1)}, W_{n,2}^{(2)}\}, \dots, \{W_{n,K}^{(1)}, \dots, W_{n,K}^{(K)}\}$ , such that

$$|W_{n,k}^{(1)}| = m_1 F, \quad \forall n, \quad (48)$$

$$|W_{n,k}^{(i)}| = \frac{N(m_i - m_{i-1})}{(N-i+1)} F, \quad i = 2, \dots, k, \forall n, \quad (49)$$

$$|W_{n,\phi}| = \left( 1 - \sum_{i=1}^K m_i - \sum_{i=2}^K \frac{(i-1)(K-i+1)(m_i - m_{i-1})}{N-i+1} \right) F, \quad (50)$$

where  $W_{n,\phi}$  is available only at the server. User  $k$  caches subfiles  $W_{1,k}^{(1)}, \dots, W_{N,k}^{(1)}$  uncod and the MDS encoded pieces  $\sigma_{i-1}([W_{1,k}^{(i)}, \dots, W_{N,k}^{(i)}])$  for  $i = 2, \dots, k$ . In turn, the cache contents of user  $k$  is defined as

$$Z_k = \left( \bigcup_{n=1}^N W_{n,k}^{(1)} \right) \cup \left( \bigcup_{i=2}^k \sigma_{i-1}([W_{1,k}^{(i)}, \dots, W_{N,k}^{(i)}]) \right), \quad (51)$$

### B. Delivery Phase

The server sends the following unicast signals

$$X_{\{K\},\mathbf{d}} = W_{d_K,\phi}, \quad (52)$$

$$X_{\{k\},\mathbf{d}} = \bigcup_{i=k+1}^K \bigcup_{j=i}^K W_{d_k,j}^{(i)}, \quad \forall k \in [K-1], \quad (53)$$

where the unicast signals  $X_{\{1\},\mathbf{d}}, \dots, X_{\{k\},\mathbf{d}}$  are used by users  $\{k+1, \dots, K\}$  in decoding their cache contents.

Next, the server sends the pairwise multicast signals

$$X_{\{k,j\},\mathbf{d}} = \left( \bigcup_{i=1}^k W_{d_j,k}^{(i)} \right) \oplus \left( \bigcup_{i=1}^k W_{d_k,j}^{(i)} \right), \quad (54)$$

for  $k = 1, \dots, K$  and  $j = k+1, \dots, K$ . In turn, the delivery load is given as

$$R_{\text{coded}} = \sum_{k=1}^K |X_{\{k\},\mathbf{d}}|/F + \sum_{k=1}^K \sum_{j=k+1}^K |X_{\{k,j\},\mathbf{d}}|/F, \quad (55)$$

$$= K - \sum_{i=1}^K (K-i+1)m_i - \sum_{i=2}^K \frac{(i-1)(K-i+1)(K-i+2)(m_i - m_{i-1})}{2(N-i+1)}. \quad (56)$$

### C. Achievability

First, the cache size constraints are satisfied, since

$$|Z_k| = Nm_1 F + \sum_{i=2}^k (N-i+1)|W_{n,k}^{(i)}|, \quad (57)$$

$$= M_1 F + \sum_{i=2}^k (M_i - M_{i-1})F = M_k F. \quad (58)$$

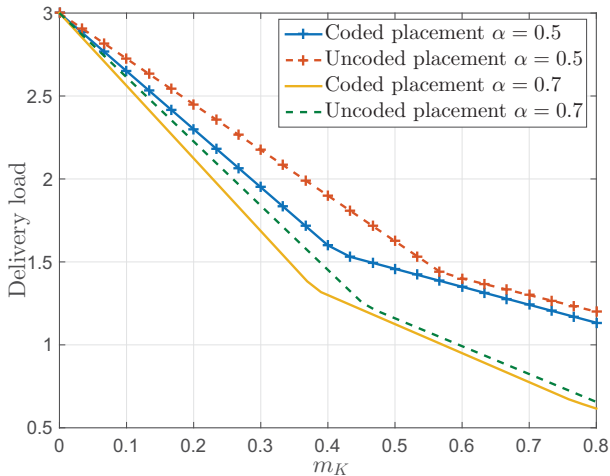


Fig. 2: Comparing the achievable delivery load assuming coded placement with the minimum delivery load under uncoded placement, for  $K = 3$ ,  $N = 4$ , and  $m_k = \alpha m_{k+1}$ .

In the delivery phase, user  $k$  reconstructs the requested file  $W_{d_k}$  by going through the following steps.

- Subfile  $W_{d_k,k}^{(1)}$  is uncoded and in turn can be directly retrieved from the cache memory.
- Subfiles  $W_{d_1,k}^{(i)}, \dots, W_{d_{i-1},k}^{(i)}$  are extracted from the unicast signals  $X_{\{1\},d}, \dots, X_{\{i-1\},d}$ .
- Subfile  $W_{d_k,k}^{(i)}$  is retrieved from the encoded pieces  $\sigma_{i-1}([W_{1,k}^{(i)}, \dots, W_{N,k}^{(i)}])$  using  $W_{d_1,k}^{(i)}, \dots, W_{d_{i-1},k}^{(i)}$ .
- Subfiles  $W_{d_k,j}^{(k+1)}, \dots, W_{d_k,j}^{(j)}$  where  $j \in \{k+1, \dots, K\}$  and  $W_{d_k,\phi}$  are retrieved from the unicast signal  $X_{\{k\},d}$ .
- Subfiles  $W_{d_k,j}^{(1)}, \dots, W_{d_k,j}^{(k)}$  are retrieved from the multicast signals  $X_{\{k,j\},d}, j \in [K] \setminus \{k\}$ .

## VI. NUMERICAL RESULTS

In Fig. 2, we compare the worst-case delivery load achieved by exploiting coded placement with the minimum worst-case delivery load assuming uncoded placement in a three-user system where  $N = 4$  and  $m_k = \alpha m_{k+1}$ . Fig. 2 shows that the gain achieved by coded placement increases with the heterogeneity in cache sizes. We also observe that the gain is higher when the total memory is small.

The delivery load achieved by utilizing coded placement in Theorem 2 is compared to the best uncoded placement scheme in Fig. 3, for  $K = 10$  and  $m_k = 0.7m_{k+1}$ . From Fig. 3, we observe that the reduction in the delivery load due to coded placement decreases with the number of files  $N$ . In turn, for a system where  $N \gg K$ , the delivery load achieved with our coded placement scheme is approximately equal to the minimum delivery load under uncoded placement. That is, the coded placement gain is negligible when  $N \gg K$ .

## VII. CONCLUSION

In this paper, we have shown that coded placement leads to significant reduction in the delivery load in systems where

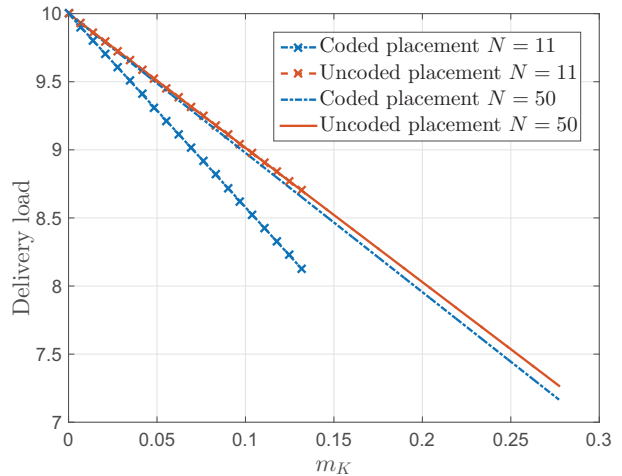


Fig. 3: Comparing the achievable delivery load assuming coded placement with the minimum delivery load under uncoded placement, for  $K = 10$ ,  $\alpha = 0.7$ , and  $m_k = \alpha m_{k+1}$ .

the users have different cache sizes. In particular, we have proposed novel coded placement schemes that outperform the best uncoded placement schemes for three-user systems with arbitrary cache sizes and  $K$ -user systems where the cache sizes satisfy  $\sum_{i=1}^K m_i + \sum_{i=2}^K (i-1)(K-i+1)(m_i - m_{i-1}) / (N-i+1) \leq 1$ .

Our proposed schemes illustrate that the signals intended to serve users with small cache sizes can be used in decoding the cache contents of users with larger cache sizes. Furthermore, we have shown that the gain due to coded placement increases with the heterogeneity in cache sizes and decreases with the number of files.

## REFERENCES

- [1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, Mar. 2014.
- [2] Q. Yang and D. Gündüz, "Coded caching and content delivery with heterogeneous distortion requirements," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4347–4364, 2018.
- [3] A. Sengupta, R. Tandon, and T. C. Clancy, "Layered caching for heterogeneous storage," in *Proc. IEEE Asilomar*, 2016.
- [4] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Centralized coded caching with heterogeneous cache sizes," in *Proc. IEEE WCNC*, 2017.
- [5] —, "Coded caching for heterogeneous systems: An optimization perspective," *arXiv:1810.08187*, 2018.
- [6] —, "Optimization of heterogeneous caching systems with rate limited links," in *Proc. IEEE ICC*, 2017.
- [7] S. S. Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," *arXiv:1702.08044*, 2017.
- [8] A. M. Ibrahim, A. A. Zewail, and A. Yener, "On coded caching with heterogeneous distortion requirements," in *Proc. IEEE ITA*, 2018.
- [9] C. Tian and K. Zhang, "From uncoded prefetching to coded prefetching in coded caching," *arXiv:1704.07901*, 2017.
- [10] J. Gómez-Vilardebó, "Fundamental limits of caching: Improved bounds with coded prefetching," *arXiv:1612.09071*, 2016.
- [11] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gündüz, "Coded caching with heterogeneous cache sizes and link qualities: The two-user case," *arXiv:1802.02706*, 2018.
- [12] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*. Elsevier, 1977.