Subspace Coding for Coded Caching: Decentralized and Centralized Placements Meet for Three Users

Hadi Reisizadeh
Department of ECE
University of Minnesota
Email: hadir@umn.edu

Mohammad Ali Maddah-Ali
Nokia Bell Labs
Holmdel, NJ
Email: mohammad.maddahali@nokia-bell-labs.com

Soheil Mohajer
Department of ECE
University of Minnesota
Email: soheil@umn.edu

***Abstract*—Coded caching is a new approach to decrease the communication load during the peak hours of the network. It provides a significant gain, that is maximized in the centralized setting, where the server controls the placement. In many situations, each user fills its cache without any information about the placement of other users. We show that subspace precoding for placement improves the delivery load of a decentralized caching system compared to uncoded placement. Surprisingly, the proposed scheme achieves the delivery load of the *centralized* placement for $K = 3$ users for the entire range of *cache size*.**

I. Introduction

Caching is a promising solution to reduce the network load during the peak traffic time. A caching system operates in two phases: (i) *placement phase* (*prefetching*), where each user has access to the dataset of server and duplicates some packets of each file in its memory when the server is in its off-peak traffic time, and (ii) *delivery phase* (*fetching*), where each user requests one file from the dataset and the server will broadcast a message so that each user can extract his desired file from the received message and the cache content. In *centralized placement*, the server has the privilege to place any packets in cache of end users, in order to minimize the load of the delivery phase. However, in practice, due to the dynamic behavior of the network, centralized placement might not be feasible. In this situation, a *decentralized placement* will be implemented, where each user decides about the packets to cache, independently from the other users.

In [1], it is shown that coded caching with centralized placement can substantially reduce the network traffic. While decentralized caching is proven to offer a significant gain, in general, there is a gain loss due to the absence of the centralized placement [2]. The exact trade-off between the cache size and the delivery load is fully characterized in [3].

All of the above-mentioned works focus on *uncoded placement*, where the placement performs on pure packets of the files, without any pre-processing. Precoding and prefetching of coded packets can improve the systems performance in general. In particular, a coded placement strategy with coding *across* files is presented in [4], for a system with small cache size. Alternatively, a coded placement based on *erasure coding over individual files* (such as Maximum Distance Separable (MDS) code) for decentralized coded caching is independently reported in [5] and [6]. It is shown that erasure-coded placement can improve the performance of the decentralized placement compared to uncoded prefetching for very small and very large cache sizes.

In this work, we introduces the notion of subspace coding for decentralized placement, which is indeed a generalization of erasure coding. We argue that by exploiting the vector spaces describing the files and cache contents we have the flexibility of serving a user in many different ways, among which some can achieve/approach the performance of centralized placement. In particular, we show that the delivery load achieved in [5] and [6] can be further improved. Specifically, for system with $K = 3$ users the (optimum) centralized delivery load is achievable for any cache size.

II. System Model and the Main Result

We consider a single server that is connected to K users through a shared and error-free link as shown in Fig. 1, i.e., the server can simultaneously broadcast to all users. The server has access to a dictionary of N files, namely $W_1, \ldots, W_N$ such that each file W_n consists of F symbols from some finite field F_q. Each user i is equipped with a cache (memory) C_i which can store up to MF symbols/packets. The system operates in two phases, namely, *placement phase* and *delivery phase*. In a *decentralized* placement phase, each user stores a set of μF , MF/N symbols from each file, selected randomly and independently across files and users.

After the completion of the placement phase, each user requests one of the N files, where all files are equally likely to be requested. We denote the request of user i by $d_i \in [N]$, and the sequence of all requests by $\boldsymbol{d} = (d_1, d_2, \ldots, d_K)$. Once the requests are revealed to the server, it forms a broadcasting message $X = X_{\boldsymbol{d}}(W_1, W_2, \ldots, W_N)$ and sends it to all the users over the common link during the delivery phase. The broadcast message X should be such that each user i be able to extract his desired file using the content of his cache C_i and the broadcast message X, i.e., $H(W_{d_i}|\mathsf{C}_i, X) = 0$. The *delivery load* of the system is defined as the *normalized* size of the broadcasting message X, which is denoted by R. Hence, we have that is $R = H(X)/F$. The goal is to design the two main phases of caching to minimize the delivery load of the system, R.

This work is supported in part by the National Science Foundation under Grant CCF-1749981.

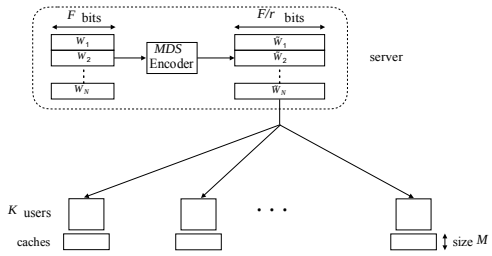

Fig. 1. Erasure-coded placement for coded caching system.

A. Subspace Coding

The delivery load of a caching systems with decentralized placement is in general larger than that of the centralized one [2]. However, in this paper we will argue that this load can be reduced by *subspace coding*. In particular, we show that this load matches with that of the centralized placement for a systems with $K = 3$ users, over the entire range of $\mu \in [0, 1]$

To this end, we map each packet of each file into a vector. More precisely, the m-th packet of file W_n with value $\alpha \in \mathsf{F}_q$ will be mapped to a vector of length F, which is all zero, except its m-th position which is α. With this mapping, the file will be mapped to set of vectors. We denote by W_n the vector space spanned by the vectors of file W_n, which has dimension F. Then, the cached part of the file will be a subspace of W_n of dimension μF for each user. The delivery phase using subspace coding is equivalent to providing each user with enough vectors, that are linearly independent from the cached ones (we refer to as *innovative*), so that the entire vector space can be spanned by the union of cached and delivered vectors. Hence, each user can be served by several different sets of vectors, in contrast to the unique possibility of sending the missing packets in an uncoded scenario. We will show that this flexibility can significantly improve the load of delivery.

Theorem 1. *Subspace coding over files reduces the delivery load of a coded caching problem with $N \geq 3$ files and $K = 3$ users and decentralized placement to that of the centralized placement, given by*[1]

$$R = \begin{cases} 3 - 6\mu & 0 \leq \mu < \frac{1}{3}, \\ \frac{5}{3} - 2\mu & \frac{1}{3} \leq \mu < \frac{2}{3}, \\ 1 - \mu & \frac{2}{3} \leq \mu \leq 1. \end{cases} \tag{1}$$

The proof of this theorem is presented in Section IV.

III. The Preliminaries

In the following, we define some notations that are used frequently throughout the paper. We denote the set of integers $\{1, 2, \ldots, K\}$ by $[K]$. We use capital letters (e.g. C) to denote a set of vectors, and script letters (e.g. $\mathcal{C}$) to denote the vector space spanned by the vectors in C. Moreover, with slightly abuse of notation, we use $|\cdot|$ to denote the cardinality of sets and the number of vectors in a vector space over a finite field. In this paper, we consider an underlying finite field F_q, where q is assumed to be very large. We have several statements that

[1]For all other (repeated) requests, we can show the centralized rate is achievable. We skip the details due to the page limit.

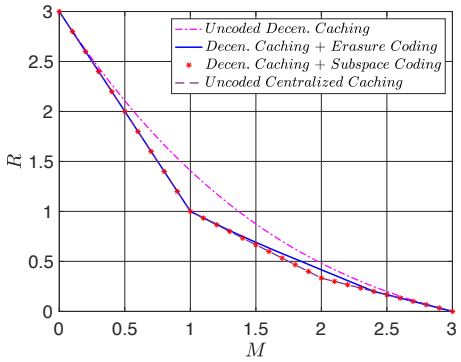

Fig. 2. Comparison of centralized, decentralized, erasure coding, and subspace coding for decentralized caching rate for the caching problem with parameters $N = K = 3$.

hold *with high probability* (w.h.p.) as $q \to \infty$. Throughout the mathematical derivations, we use "$\doteq$" to denote an equality which holds with high probability for large enough q.

Definition 1. For two subspaces $V, U \subseteq \mathsf{F}_q$, we define the following relationships and operations:
(i) V is a subspace of U and denoted by $V \vee U$ if for every vector $v \in V$ we have $v \in U$.
(ii) Intersection: $U \cap V = \{u : u \in U, u \in V\}$.
(iii) Summation: $U + V = \{u + v : u \in U, v \in V\}$.
(iv) Dimension of $U + V$ is given by

$$\dim(U + V) = \dim(U) + \dim(V) - \dim(U \cap V). \tag{2}$$

(v) Size of U can be obtained from

$$|U| \doteq q^{\dim(U)}. \tag{3}$$

(vi) If V is a subspace of U, we define the *quotient* U/V as a vector space satisfying the three following properties
- $U/V \vee U$,
- $\dim(U/V) = \dim(U) - \dim(V)$,
- $U \vee U/V + V$.

Note that a quotient of a subspace is not unique.

Let W be a vector space of dimension F over a finite field F_q. For an *expansion parameter* $\tau \in (0, 1]$, consider the set of vectors $B = \{b_1, b_2, \ldots, b_{F/\tau}\}$, where each vector $b_\ell \in W$ is drawn *randomly and uniformly* from the subspace W for $\ell = 1, \ldots, F/\tau$. The following lemma can be easily proved using the facts that $q \to \infty$ and vectors in B are drawn uniformly at random. The formal proof is omitted for the sake of brevity.

Lemma 1. The set of vectors B drawn uniformly at random from W satisfies the MDS property approaching 1, as q grows, i.e., any subset of vectors in $B \subseteq U$ of size at most F includes *linearly independent* vectors. Hence, for $U = \mathrm{span}(U)$, the subspace spanned by vectors in U, we have

$$\dim(U) \doteq \min(F, |U|), \tag{4}$$

with high probability as $q \to \infty$. In particular, for any subset $U \subseteq B$ with $|U| = F$, we have $\mathrm{span}(U) \doteq W$.

In the following proposition, we evaluate the size of intersection of two subspaces.

Proposition 1. For subsets $U,V \subseteq B$ of size at most F, we have

$$|U \cap V| = \max \quad q^{|U|+|V|-F} \quad , q^{|U \cap V|} \quad . \tag{5}$$

Proof. First note that $U + V = \text{span}(U \cup V)$. Consequently, using (4) for $U \cup V \subseteq B$ we can write

$$\begin{aligned} \dim(U + V) &= \dim(\text{span}(U \cup V)) = \min(F, |U \cup V|) \\ &= \min(F, |U| + |V| - |U \cap V|). \end{aligned} \tag{6}$$

From (2) we have

$$\begin{aligned} \dim(U \cap V) &= \dim(U) + \dim(V) - \dim(U + V) \\ &= |U| + |V| - \min(F, |U| + |V| - |U \cap V|) \\ &= \max(|U| + |V| - F, |U \cap V|). \end{aligned} \tag{7}$$

This, together with (3), implies the claim. □

A. Subspace Precoding for Cache Placement

Consider a subspace W of dimension F, and a set of vectors $B = \{b_1, \ldots, b_{F/\tau}\}$, drawn uniformly at random from W. For each user $i \in [K]$, let C_i be a *random* subset of B of expected size MF/N, wherein each vector $b_\ell \in B$ is included in C_i with probability $(MF/N)/(F/\tau) = M\tau/N = \mu\tau$, independently across vectors, subsets, and users, and define $C_i = \text{span}(C_i)$. In the following, we define *direct* and *total* common spaces and derive their corresponding dimensions.

Definition 2. For any $\Gamma \subseteq [K]$ the *direct* common space D_Γ and the *total* common space T_Γ are defined as

$$D_\Gamma \triangleq \text{span}(D_\Gamma) = \text{span}\Big(\bigcap_{i \in \Gamma} C_i\Big), \tag{8}$$

$$T_\Gamma \triangleq \bigcap_{i \in \Gamma} \text{span}(G_i) = \bigcap_{i \in \Gamma} C_i, \tag{9}$$

We denote the expected dimensions of these spaces by

$$\delta_\Gamma = \mathsf{E}[\dim(D_\Gamma)], \quad \text{and} \quad \theta_\Gamma = \mathsf{E}[\dim(T_\Gamma)].$$

Note that for $\Gamma = \{i\}$ with $|\Gamma| = 1$, we have $D_\Gamma = T_\Gamma = C_i$, and hence $\delta_{\{i\}} = \theta_{\{i\}} = \mu F$. Moreover, from the definition of the direct and total common spaces, it is easy to verify that $D_\Gamma \subseteq T_\Gamma$ for any $\Gamma \subseteq [K]$. In the proposed schemes in [5] and [6], the server just exploits the direct common spaces in the construction of the coded sub-messages. In the following example, we show that in general $D_\Gamma \neq T_\Gamma$, i.e., D_Γ is a *proper* subspace of T_Γ.

Example 1. Let $W = \text{span}(\{e_1, e_2, e_3\}) \subseteq \mathsf{F}^3$ be vector space of dimension 3, and consider the set of vectors $B = \{e_1, e_1 + e_2, e_1 + 2e_2 + e_3, e_2 + e_3, e_3\}$ with $|B| = 5 = 3/0.6$, which implies $\tau = 0.6$. Note that B satisfies the MDS property. Let $\mu = \frac{2}{3}$, and $C_1 = \{e_1 + e_2, e_1 + 2e_2 + e_3\}$ and $C_2 = \{e_2 + e_3, e_3\}$ are drawn uniformly at random from B. Then, for $\Gamma = \{1, 2\}$, the direct common space of Γ is given by

$$D_{\{1,2\}} = \text{span}(C_1 \cap C_2) = \text{span}(\emptyset) = \{0\},$$

implying $\delta_\Gamma = 0$. However, the total common space T_Γ is

$$\begin{aligned} C_1 \cap C_2 &= \text{span}(\{e_1 + e_2, e_1 + 2e_2 + e_3\}) \cap \text{span}(\{e_2 + e_3, e_3\}) \\ &\overset{(a)}{=} \text{span}(\{e_2 + e_3\}). \end{aligned}$$

The equality (a) follows from the fact that for $u_1 = e_1 + e_2 \in C_1$, and $u_2 = e_1 + 2e_2 + e_3 \in C_2$, we have $u_1 + u_2 = e_2 + e_3 \in C_2$. This implies $\theta_{\{1,2\}} = 1$.

It turns out that as $q \to \infty$, the values of δ_Γ and θ_Γ only depend on $|\Gamma|$, but not the realization of Γ. Hence, we denote them by $\delta_{|\Gamma|}$ and $\theta_{|\Gamma|}$, respectively.

Proposition 2. For any $\Gamma \subseteq [K]$ and $\mu \leq 1$, we have

$$\delta_{|\Gamma|} \cup (\mu\tau)^{|\Gamma|} \frac{F}{\tau}, \quad \text{and} \quad |D_\Gamma| \cup q^{(\mu\tau)^{|\Gamma|} \frac{F}{\tau}}, \tag{10}$$

with high probability as $q \to \infty$.

Proof. Let $D_\Gamma = \bigcap_{i \in \Gamma} C_i$, and $D_\Gamma = \text{span}(D_\Gamma)$. Recall that the probability that each vector b_ℓ belongs to C_i is $\mu\tau$. Thus,

$$\begin{aligned} \mathsf{E}[|D_\Gamma|] &= \mathsf{E}\Big[\Big|\bigcap_{i \in \Gamma} C_i\Big|\Big] \cup \sum_{\ell=1}^{F/\tau} \mathsf{P}[b_\ell \in C_i, \forall i \in \Gamma] \\ &= \sum_{\ell=1}^{F/\tau} \prod_{i \in \Gamma} \mathsf{P}[b_\ell \in C_i] = (\mu\tau)^{|\Gamma|} \frac{F}{\tau}. \end{aligned}$$

As $q \to \infty$, the vectors in D_Γ are linear independent w.h.p., and random variable $|D_\Gamma|$ concentrates at its mean, i.e., we have $\delta_{|\Gamma|} = |D_\Gamma| \cup (\mu\tau)^{|\Gamma|} \frac{F}{\tau}$, and $|D_\Gamma| \cup q^{|D_\Gamma|}$, w.h.p. □

In the following lemma, we evaluate the dimension of total common subspace for any number of users.

Lemma 2. For large enough field size q, a set $\Gamma \subseteq [K]$ with $|\Gamma| = \ell$, and $p \in [K] \setminus \Gamma$, with high probability we have

$$\theta_{\ell+1} = \theta_{|\Gamma \cup \{p\}|} = \theta_{\Gamma \cup \{p\}} \cup \max(\theta_\ell + \theta_1 - F, \delta_{\ell+1}). \tag{11}$$

Proof. First note that $D_{\Gamma \cup \{p\}} \subseteq T_\Gamma$. Therefore, for each vector $v \in T_\Gamma$ and a given quotient space $T_\Gamma / D_{\Gamma \cup \{p\}}$, there exist vectors $v_1 \in D_{\Gamma \cup \{p\}}$ and $v_2 \in T_\Gamma / D_{\Gamma \cup \{p\}}$ such that $v = v_1 + v_2$. Hence, we have

$$\begin{aligned} \mathsf{E}[|T_{\Gamma \cup \{p\}}|] &= \mathsf{E}[|T_\Gamma \cap C_p|] = \mathsf{E}\Big[\sum_{v \in T_\Gamma} 1[v \in C_p]\Big] \\ &= \sum_{v \in T_\Gamma} \mathsf{P}[v \in C_p] = \sum_{v_1 \in D_{\Gamma \cup \{p\}}} \sum_{v_2 \in T_\Gamma / D_{\Gamma \cup \{p\}}} \mathsf{P}[v_1 + v_2 \in C_p] \\ &\overset{(a)}{=} \sum_{v_1 \in D_{\Gamma \cup \{p\}}} \sum_{v_2 \in T_\Gamma / D_{\Gamma \cup \{p\}}} \mathsf{P}[v_2 \in C_p] \\ &= |D_{\Gamma \cup \{p\}}| \Big(1 + \sum_{0 \neq v_2 \in T_\Gamma / D_{\Gamma \cup \{p\}}} \mathsf{P}[v_2 \in C_p]\Big) \\ &\overset{(b)}{\cup} q^{\dim(D_{\Gamma \cup \{p\}})} \Big(1 + \frac{q^{\mu F}}{q^F}\Big)^{\dim(T_\Gamma / D_{\Gamma \cup \{p\}})} - 1 \\ &= q^{\delta_{\ell+1}} + \frac{q^{\mu F}}{q^F} \times (q^{\theta_\ell} - q^{\delta_{\ell+1}}) \\ &= q^{\delta_{\ell+1}} + q^{\theta_\ell + \theta_1 - F} - q^{\delta_{\ell+1} + \theta_1 - F}, \end{aligned}$$

where (a) holds since $v_1 \in D_{\Gamma\cup\{p\}} \setminus C_p$, (b) follows from the fact that vectors in $T_\Gamma/D_{\Gamma\cup\{p\}}$ and those in C_p are drawn identically and independently from each other which implies $\mathsf{P}[v \in C_p]$ is the same for every (non-zero) $v_2 \in T_\Gamma/D_{\Gamma\cup\{p\}}$. We also used the facts that subspace $|C_p| = q^{\mu F}$ and $|W| = q^F$. When $q \to \infty$, the size of the total common space converges to its expectation, and hence the dimension of this subspace can be found from

$$\begin{aligned}\theta_{\Gamma\cup\{p\}} &= \dim T_{\Gamma\cup\{p\}} = \lim_{q\to\infty} \log_q |T_\Gamma \cap C_p| \\ &= \lim_{q\to\infty} \log_q \left(q^{\delta_{\Gamma\cup\{p\}}} + q^{\theta_\Gamma + \theta_p - F} - q^{\delta_{\Gamma\cup\{p\}} + \theta_p - F}\right) \\ &= \max(\theta_\Gamma + \theta_p - F, \delta_{\Gamma\cup\{p\}}).\end{aligned}$$

This completes the proof. □

Remark 1. It is worth noting that all the above claims only hold *with high probability*, as $q \to \infty$, but they are not necessarily true for every deterministic set B of vectors that satisfy the MDS property. In the following, we present a counter-example that shows the statements above may fail for carefully chosen subsets of B.

Example 2. Let $F = 3$ and $W = \text{span}(e_1, e_2, e_3)$. For $\tau = 0.5$, we can select the set of $F/\tau = 6$ vectors given by

$$\begin{aligned}B = \{&e_1 + e_2 + e_3, e_1 + e_2 + 2e_3, 2e_1 + e_2 + 3e_3, \\ &2e_1 + e_2 + 4e_3, 3e_1 + e_2 + 17e_3, 3e_1 + e_2 + 25e_3\}.\end{aligned}$$

It is easy to verify that the MDS property is satisfied by B, i.e., any collection of up to 3 vectors in B are linearly independent. For $\mu = \frac{2}{3}$, we can pick th following subsets of B:

$$\begin{aligned}C_1 &= \{e_1 + e_2 + e_3, e_1 + e_2 + 2e_3\}, \\ C_2 &= \{2e_1 + e_2 + 3e_3, 2e_1 + e_2 + 4e_3\}, \\ C_3 &= \{3e_1 + e_2 + 17e_3, 3e_1 + e_2 + 25e_3\}.\end{aligned}$$

First note that no vector from B is directly cached in more than one cache, which implies $\delta_{\{1,2\}} = \delta_{\{1,2,3\}} = 0$. Let G_i denote the span of vectors in C_i, for $i = 1, 2, 3$, i.e,

$$G_1 = \text{span}(C_1) = \text{span}(\{e_1 + e_2, e_3\}) \quad (12)$$

$$G_2 = \text{span}(C_2) = \text{span}(\{2e_1 + e_2, e_3\}) \quad (13)$$

$$G_3 = \text{span}(C_3) = \text{span}(\{3e_1 + e_2, e_3\}) \quad (14)$$

It is easy to verify that vector e_3 is the only common vector in all three subspaces. This implies $\theta_{\{1,2\}} = \theta_{\{1,2,3\}} = 1$. Then for $\Gamma = \{1, 2\}$ and $p = 3$, we have $\theta_{\{3\}} = \mu F = 2$, and thus, $\theta_{\{1,2\}} + \theta_{\{3\}} - F = 1 + 2 - 3 = 0$, $\delta_{\{1,2,3\}} = 0$, $\theta_{\{1,2,3\}} = 1$, which is in contradiction with Lemma 2. This shows that the claim of the lemma does not necessarily hold for every collection of vectors from B, and only holds true with high probability, as mentioned earlier.

IV. Proof of Theorem 1

Before going through the details of the proof, we define the concept of *utility* which is helpful in our analysis.

Definition 3. For a given placement scheme, a coded transmit vector/packet has utility m (for $m = 1, \ldots, K$), if it can provide innovative information for m users. Such a packet intended for users in M with $|M| = m$ will be generated by a linear combination of m file packets, each intended for one user $i \in M$, and cached by every user in $M \setminus \{i\}$.

Let us denote by $C_{i,n}$ the part of the cache of user i filled by random vectors of the vector space W_n, for $i \in [K]$ and $n \in [N]$. Hence, the cache of User i is given by $C_i = \bigcup_{n\in[N]} C_{i,n}$. We assume that Users 1, 2, and 3, request distinct files W_{d_1}, W_{d_2}, and W_{d_3}, respectively. Each User i has $\mu F = \theta_1$ linearly independent vectors (w.h.p.) from his desired vector space W_{d_i}, and can reconstruct the file if he collects a total of F linearly independent vectors. Denoting by U_j the number of coded packets of utility j sent by the server which are useful for each user, the decodability condition can be written as

$$\theta_1 + U_1 + U_2 + U_3 = F. \quad (15)$$

Since packets of utility j serve j users simultaneously, the total load of delivery can be obtained as

$$R = \frac{1}{F}\left(\frac{K}{1}U_1 + \frac{K}{2}U_2 + \frac{K}{3}U_3\right) = \frac{3U_1 + 3U_2/2 + U_3}{F}. \quad (16)$$

From (15) and (16) it is clear that the central server should intend to transmit packets with higher utilities in order to minimizing the delivery load. The following proposition determines the number of packets of utility 1 needed to be sent for each user.

Proposition 3. The number of packets/vectors of utility 1 the server sends to each user is given by

$$U_1 = F - \min(F, 3(\theta_1 - \delta_2) + \delta_3). \quad (17)$$

Proof. A vector of file W_n intended for user i has utility 1 when it cannot be combined with any other packet, i.e., such packet is not cached at any user in the network. The number of packets/vectors of file n cached across all users can be found as follows. As $q \to \infty$, with high probability we have

$$\begin{aligned}|C_{1,n} \cup C_{2,n} \cup C_{3,n}| &= |C_{1,n}| + |C_{2,n}| + |C_{3,n}| - |C_{1,n} \cap C_{2,n}| \\ &\quad - |C_{1,n} \cap C_{3,n}| - |C_{2,n} \cap C_{3,n}| + |C_{1,n} \cap C_{2,n} \cap C_{3,n}| \\ &= 3\mu F - 3\delta_2 + \delta_3 = 3\theta_1 - 3\delta_2 + \delta_3.\end{aligned}$$

Therefore, from (4) we have

$$\begin{aligned}\dim(C_{1,n} + C_{2,n} + C_{3,n}) &= \dim(\text{span}(C_{1,n} \cup C_{2,n} \cup C_{3,n})) \\ &= \min(F, |C_{1,n} \cup C_{2,n} \cup C_{3,n}|) = \min(F, 3(\theta_1 - \delta_2) + \delta_3).\end{aligned}$$

Consequently, the number of vectors of utility 1 is given by $U_1 = F - \dim(C_{1,n} + C_{2,n} + C_{3,n})$, which reduces to (17). □

Proposition 4. For large enough q, the number of packets of utility 3 to simultaneously serve all users, is given by (w.h.p)

$$U_3 = \theta_2 - \theta_3. \quad (18)$$

Proof. A coded packet of utility 3 is of the form of

$$v_j = b_j^{d_1} + b_j^{d_2} + b_j^{d_3}, \qquad j = 1, \ldots, U_3,$$

where $b_j^{d_i}$ is a vector from B_{d_i} (describing W_{d_i}), such that

$$\begin{aligned} b_j^{d_1} &\in C_{2,d_1} \cap C_{3,d_1}, \quad b_j^{d_1} \notin C_{1,d_1}, \\ b_j^{d_2} &\in C_{1,d_2} \cap C_{3,d_2}, \quad b_j^{d_2} \notin C_{2,d_2}, \\ b_j^{d_3} &\in C_{1,d_3} \cap C_{2,d_3}, \quad b_j^{d_3} \notin C_{3,d_3}. \end{aligned} \tag{19}$$

In order for all such $b_j^{d_1}$ to be informative, they should be linearly independent from each other. Let us count the number of $b_j^{d_1}$ satisfying (19). Let $\{u_1, \ldots, u_{\theta_3}\}$ be a basis for the subspace $C_{1,d_1} \cap C_{2,d_1} \cap C_{3,d_1}$. Since $C_{1,d_1} \cap C_{2,d_1} \cap C_{3,d_1} \nu C_{2,d_1} \cap C_{3,d_1}$, we can extend it to form a basis for $C_{2,d_1} \cap C_{3,d_1}$, given by $\{u_1, \ldots, u_{\theta_3}, u_{\theta_3+1}, \ldots, u_{\theta_2}\}$. Therefore, the subset $\{u_{\theta_3+1}, \ldots, u_{\theta_2}\}$ contains $\theta_2 - \theta_3$ linearly independent vectors that *only* belong to C_{2,d_1} and C_{3,d_1}, but not to C_{1,d_1}. By symmetry, a similar argument holds for vectors $b_j^{d_2}$ and $b_j^{d_3}$, which proves the claim of the proposition. □

Finally, the number of packets of utility 2 can be obtained from plugging (17) and (18) into (15). That is,

$$U_2 = F - \theta_1 - U_1 - U_3 \tag{20}$$

Therefore, the delivery load in (16) can be evaluated as

$$\begin{aligned} RF &= 3U_1 + \frac{3}{2}U_2 + U_3 \qquad (21) \\ &= 3F - \frac{3}{2}\theta_1 - \frac{3}{2}\min(3(\theta_1 - \delta_2) + \delta_3, F) - \frac{1}{2}(\theta_2 - \theta_3). \end{aligned}$$

A. Corner Points

Next, we simplify the delivery load in (21) for the four corner points associated with $\mu \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$. To this end, we choose $\tau \to 0$. Note that since τ is vanishing, the size of B is growing unboundedly, and hence the probability that a vector is directly cached at more than one user is negligible. This is indicated by (10), where $\delta_j = 0$ for $j > 1$.

(i) $\underline{\mu_0 = 0}$: In this case no vector in cached, and hence $\delta_j = \theta_j = 0$ for $j = 1, 2, 3$. Hence, the delivery load in (21) can be evaluated as $RF = 3F$, which implies $R(0) = 3$.

(ii) $\underline{\mu_1 = 1/3}$: From Lemma 2, we have

$$\begin{aligned} \theta_2 &= \max(2\theta_1 - F, \delta_2) = \max(-F/3, 0) = 0 \\ \theta_3 &= \max(\theta_2 + \theta_1 - F, \delta_3) = 0. \end{aligned}$$

Plugging these into (21) we get

$$RF = 3F - \frac{3}{2}\theta_1 - \frac{3}{2}\min(3\theta_1, F) = 3F - \frac{3}{2}\frac{1}{3}F - \frac{3}{2}F = F,$$

and therefore, $R(\frac{1}{3}) = 1$.

(iii) $\underline{\mu_2 = 2/3}$: We have $\theta_1 = 2F/3$, and Lemma 2 implies

$$\begin{aligned} \theta_2 &= \max(2\theta_1 - F, \delta_2) = F/3 \\ \theta_3 &= \max(\theta_2 + \theta_1 - F, \delta_3) = 0. \end{aligned}$$

Therefore from (21) we get $RF = F/3$, which implies $R(\frac{2}{3}) = \frac{1}{3}$.

(iv) $\underline{\mu_3 = 1}$: Finally, for $\theta_1 = \mu F = F$, all users have cached the entire vector space. It is easy to verify that $\theta_2 = \theta_3 = F$. This implies $R(1) = 0$.

Note that the memory-load trade-off obtained for $\mu \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ matches with that of the centralized coded caching derived in [1].

B. Non-Corner Points

We use memory sharing to achieve the memory-load trade-off of centralized caching for the middle values of μ. This consist of splitting the files and caches into two sub-files and two sub-caches, and treating each sub-file and the corresponding sub-cache as a separate system. For $\mu = \lambda\mu_{t} + (1-\lambda)\mu_{t+1}$, where $t \in \{0, 1, 2\}$, and $\lambda \in [0, 1]$, we divide each file W_n into two parts $W_n = W_n^{(1)}, W_n^{(2)}$, with sizes $F_1 = \lambda F$ and $F_2 = (1-\lambda)F$. Furthermore, we divide the cache of each user $i \in [K]$ into $C_i = C_i^{(1)}, C_i^{(2)}$, where

$$|C_i^{(1)}| = \frac{t\lambda}{t + 1 - \lambda} MF, \qquad |C_i^{(2)}| = \frac{(t+1)(1-\lambda)}{t + 1 - \lambda} MF.$$

Then we apply the proposed coding scheme for each collection of sub-files, with $\tau_1, \tau_2 \to 0$. It is easy to verify that this leads to a delivery load given by $R = \lambda R(\mu_t) + (1 - \lambda)R(\mu_{t+1})$. The details of derivation are skipped due to the page limit.

Remark 2. The partitioning of files and caches with the properly chosen sizes is critical to achieve the performance of the centralized placement, otherwise (as shown in the example below) the subspace coding without file partitioning does not offer the performance of centralized placement.

Example 3. Let us consider a caching problem with $K = 3$ users and $\mu = 0.5$, i.e., $MF = NF/2$. Without file partitioning, for $\tau \to 0$, we have $\theta_1 = F/2$, and $\theta_2 = \theta_3 = 0$, leading to no packet of utility 3. However, using file partitioning, we have $F_1 = F_2 = F/2$, $|C_i^{(1)}| = NF/6$, and $|C_i^{(1)}| = NF/3$. Now, it is easy to see that for the system with $(F_2, |C_i^{(2)}|) = (F/2, NF/3)$ we have $\theta_2^{(2)} = F/6$ and $\theta_3^{(2)} = 0$, implying that $\theta_2^{(2)} - \theta_3^{(2)} = F/6$ of the packets will be sent at utility 3. This reduces the delivery load compared to a non-partitioned system, to that of the centralized placement.

REFERENCES

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

[2] ——, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, 2015.

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

[4] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for small buffer users," *arXiv preprint arXiv:1407.1935*, 2014.

[5] Y.-P. Wei and S. Ulukus, "Novel decentralized coded caching through coded prefetching," in *Proc. Inf. Theory Workshop (ITW), Kaohsiung, Taiwan*, 2017, pp. 1–5.

[6] H. Reisizadeh, M. A. Maddah-Ali, and S. Mohajer, "Erasure coding for decentralized coded caching," in *IEEE Int. Symp. Inf. Theory*. IEEE, 2018.