An Optimal Linear Error Correcting Delivery Scheme for Coded Caching with Shared Caches

Nujoom Sageer Karat *, Spandan Dey†, Anoop Thomas † and B. Sundar Rajan *
*Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru 560012, KA, India
E-mail: {nujoom,bsrajan}@iisc.ac.in
† School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar 752050, OD, India
E-mail: {anoopthomas,sd40}@iitbbs.ac.in

***Abstract*—Classical coded caching setting avails each user to have one dedicated cache. This is generalized to a more general shared cache scheme and the exact expression for the worst case rate was derived in [E. Parrinello, A. Unsal, P. Elia, " Fundamental Limits of Caching in Heterogeneous Networks with Uncoded Prefetching," available on arXiv:1811.06247 [cs.IT], Nov. 2018]. For this case, an optimal linear error correcting delivery scheme is proposed and an expression for the peak rate is established for the same. Furthermore, a new delivery scheme is proposed, which gives an improved rate for the case when the demands are not distinct.**

I. INTRODUCTION

The technique of coded caching introduced in [1] helps in reducing the peak traffic experienced by networks. This is achieved by making a part of the content locally available at the users during non-peak periods. In [1], it is shown that apart from the *local caching gain* obtained by placing contents at user caches before the demands are revealed, a *global caching gain* can be obtained by coded transmissions. The scheme in [1] is a centralized coded caching scheme, where all users are linked to a single fixed server. Since then there have been many extensions to this, like decentralized scheme [2], non-uniform demands [3] and online coded caching [4].

A coded caching scheme involves two phases: a placement phase and a delivery phase. In the placement phase or prefetching phase, each user can fill their local cache memory using the entire database. During this phase there is no bandwidth constraint as the network is not congested and the only constraint here is the memory. Delivery phase is carried out once the users reveal their demands. During the delivery phase only the server has access to the file database and the constraint here is the bandwidth as the network is congested in this phase. During placement phase some parts of files have to be judiciously cached at each user in such a way that the rate of transmission is reduced during the delivery phase. The prefetching can be done with or without coding. If during prefetching, no coding of parts of files is done, the prefetching scheme is referred to as uncoded prefetching [1], [5]. If coding is done during prefetching stage, then the prefetching scheme is referred to as coded prefetching [6]–[9].

An extension of the coded caching problem involving heterogeneous networks is considered in [10], where multiple users share a common cache. Each user has access to a helper cache, which is potentially accessed by multiple users. The scheme introduced in [10] is referred to as Shared Cache (SC) scheme throughout the paper. The corresponding prefetching scheme and delivery scheme are referred to as the SC prefetching scheme and SC delivery scheme respectively. In addition to the cache placement and delivery phase, there is an additional intermediate step which is the *user-to-cache association phase*. The expression for rate in this scenario under the assumption of uncoded placement is derived in [10]. The rate expression was under the assumption of worst case demand, which means that all the files are demanded. In our work, a new delivery scheme is proposed for the non-distinct demand case which provides improved rate compared to the SC scheme (Section IV).

Error correcting coded caching scheme was introduced in [11], [12]. In this set up, the delivery phase is assumed to be error-prone and placement is assumed to be error-free. A similar model in which the delivery phase takes place over a packet erasure broadcast channel was considered in [13]. In this work, shared cache systems in which the delivery phase is error prone is considered. An error correcting delivery scheme has to be designed to correct the required transmission errors. Each user has to decode their demands even in the presence of these errors. In our work, an optimal error correcting delivery scheme is proposed for the worst case demand in the shared cache system.

The main contributions of this paper are as follows:

- An optimal linear error correcting delivery scheme for coded caching problems with SC prefetching is proposed using techniques from index coding. For error correcting delivery scheme for coded caching problems with SC prefetching, a closed form expression for peak rate is established (Section III).
- A new delivery scheme for SC prefetching for all the demand cases having an improved rate compared to the scheme in [10] is proposed (Section IV).

Due to page constraints the proofs of all the theorems in this paper have been made available in [25].

In this paper $\mathbb{F}_q$ denotes the finite field with q elements, where q is a power of a prime, and $\mathbb{F}_q^*$ denotes the set of all non-zero elements of $\mathbb{F}_q$. For any integer K, let $[K]$ denote the set $\{1, 2, \ldots, K\}$. For a $K \times N$ matrix L, L_i denotes its ith row. Also, $\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$ and $\binom{n}{k} = 0$ if $n < k$. The

lower convex envelope of points $\{(i, f(i)) : i \in [n] \cup \{0\}\}$ for some natural number n is denoted by $Conv(f(i))$.

Let $N_q[k, d]$ denote the length of the shortest linear code over $\mathbb{F}_q$ which has dimension k and minimum distance d.

II. Preliminaries and Background

To obtain the main results of this paper, we use results from error correcting index coding problems [14]. In this section we recall some results from this and also review the concepts of error correcting coded caching scheme [11]. Furthermore, we review the SC placement and delivery scheme [10].

A. Index Coding Problem

The index coding problem with side information was introduced in [15]. A single source has n messages $x_1, x_2 \ldots, x_n$ where $x_i \in \mathbb{F}_q$, $\forall i \in [n]$. There are K receivers, $R_1, R_2, \ldots, R_K$. Each receiver possesses a subset of messages as side information. Let $\mathcal{X}_i$ denote the set of indices of the messages belonging to the side information of receiver R_i. The map $f : [K] \to [n]$ assigns receivers to indices of messages demanded by them. Receiver R_i demands the message $x_{f(i)}$, $f(i) \notin \mathcal{X}_i$ [14]. The source knows the side information available to each receiver and has to satisfy the demand of each receiver in minimum number of transmissions. An instance of index coding problem can be completely characterized by a side information hypergraph [16]. Given an instance of the index coding problem, finding the best *scalar linear* binary index code is equivalent to finding the *min-rank* of the side information hypergraph [14], which is known to be an NP-hard problem in general [17]–[19].

An index coding problem with K receivers and n messages can be represented by a hypergraph $\mathcal{H}(V, E)$, where $V = [n]$ is the set of vertices and E is the set of hyperedges [16]. Vertex i represents the message x_i and each hyperedge represents a receiver. In [14], the min-rank of a hypergraph $\mathcal{H}$ over $\mathbb{F}_q$ is defined as,

$$\kappa(\mathcal{H}) \triangleq \min\{\text{rank}_q(\{\mathbf{v_i} + \mathbf{e_{f(i)}}\}_{i\in[K]}) : \mathbf{v_i} \in \mathbb{F}_q^n, \mathbf{v_i} \triangleleft \mathcal{X}_i\},$$

where $\mathbf{v_i} \triangleleft \mathcal{X}_i$ denotes that $\mathbf{v_i}$ is the subset of the support of $\mathcal{X}_i$; the support of a vector $\mathbf{u} \in \mathbb{F}_q^n$ is defined to be the set $\{i \in [n] : u_i \neq 0\}$. This min-rank defined above is the smallest length of scalar linear index code for the problem. A linear index code of length N can be expressed as XL, where L is an $n \times N$ matrix and $X = [x_1\ x_2 \ldots\ x_n]$. The matrix L is said to be the *matrix corresponding to the index code*.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, then a subset of vertices $\mathcal{S} \subseteq \mathcal{V}$ is called an independent set if $\forall u, v \in \mathcal{S}$, $\{u, v\} \notin \mathcal{E}$. The size of a largest independent set in the graph $\mathcal{G}$ is called the independence number of $\mathcal{G}$. Dau *et al.* in [14] extended the notion of independence number to the case of directed hypergraph corresponding to an index coding problem. For each receiver R_i, define the sets

$$\mathcal{Y}_i \triangleq [n] \setminus \Big(\{f(i)\} \cup \mathcal{X}_i\Big)$$

and

$$\mathcal{J}(\mathcal{H}) \triangleq \cup_{i\in[K]}\{\{f(i)\} \cup Y_i : Y_i \subseteq \mathcal{Y}_i\}.$$

A subset H of $[n]$ is called a generalized independent set in $\mathcal{H}$, if every nonempty subset of H belongs to $\mathcal{J}(\mathcal{H})$. The size of the largest independent set in $\mathcal{H}$ is called the generalized independence number and is denoted by $\alpha(\mathcal{H})$. It is proved in [11] that for any index coding problem,

$$\alpha(\mathcal{H}) \leq \kappa(\mathcal{H}). \tag{1}$$

The quantities $\alpha(\mathcal{H})$ and $\kappa(\mathcal{H})$ decide the bounds on the optimal length of error correcting index codes. The error correcting index coding problem with side information was defined in [14]. An index code is said to correct δ errors if after receiving at most δ transmissions in error, each receiver is able to decode its demand. A δ-error correcting index code is represented as $(\delta, \mathcal{H})$-ECIC. An optimal linear $(\delta, \mathcal{H})$-ECIC over $\mathbb{F}_q$ is a linear $(\delta, \mathcal{H})$-ECIC over $\mathbb{F}_q$ of the smallest possible length $\mathcal{N}_q[\mathcal{H}, \delta]$. Lower and upper bounds on $\mathcal{N}_q[\mathcal{H}, \delta]$ were established in [14]. The Lower bound is known as the α-bound and the upper bound is known as the κ-bound. The length of an optimal linear $(\delta, \mathcal{H})$-ECIC over $\mathbb{F}_q$ satisfies

$$\underbrace{N_q[\alpha(\mathcal{H}), 2\delta + 1]}_{\alpha\text{-bound}} \leq \mathcal{N}_q[\mathcal{H}, \delta] \leq \underbrace{N_q[\kappa(\mathcal{H}), 2\delta + 1]}_{\kappa\text{-bound}}. \tag{2}$$

The κ-bound is achieved by concatenating an optimal linear classical error correcting code and an optimal linear index code. Thus for any index coding problem, if $\alpha(\mathcal{H})$ is same as $\kappa(\mathcal{H})$, then concatenation scheme would give optimal error correcting index codes [20]–[23].

B. Error Correcting Coded Caching Scheme

Error correcting coded caching scheme was proposed in [11]. The server is connected to K users through a shared link which is error prone. The server has access to N files $X^1, X^2, \ldots, X^N$, each of size F bits. Every user has an isolated cache with memory MF bits, where $M \in [0, N]$. A prefetching scheme is denoted by $\mathcal{M}$. During the delivery phase, only the server has access to the database. Every user demands one of the N files. The demand vector is denoted by $\mathbf{d} = (d_1, \ldots, d_K)$, where d_i is the index of the file demanded by user i. The number of distinct files requested in $\mathbf{d}$ is denoted by $N_e(\mathbf{d})$. During the delivery phase, the server informed of the demand $\mathbf{d}$, transmits a function of $X^1, \ldots, X^N$, over a shared link. Using the cache contents and the transmitted data, each user i needs to reconstruct the requested file X^{d_i} even if δ transmissions are in error.

For the δ-error correcting coded caching problem, a communication rate $R(\delta)$ is *achievable* for demand $\mathbf{d}$ if and only if there exists a transmission of $R(\delta)F$ bits such that every user i is able to recover its desired file X^{d_i} even after at most δ transmissions are in error. Rate $R^*(\mathbf{d}, \mathcal{M}, \delta)$ is the minimum achievable rate for a given $\mathbf{d}$, $\mathcal{M}$ and δ. The average rate $R^*(\mathcal{M}, \delta)$ is defined as the expected minimum average rate given $\mathcal{M}$ and δ under uniformly random demand. Thus $R^*(\mathcal{M}, \delta) = \mathbb{E}_{\mathbf{d}}[R^*(\mathbf{d}, \mathcal{M}, \delta)]$.

The average rate depends on the prefetching scheme $\mathcal{M}$. The minimum average rate $R^*(\delta) = \min_{\mathcal{M}} R^*(\mathcal{M}, \delta)$ is the minimum rate of the delivery scheme over all possible

$\mathcal{M}$. The rate-memory trade-off for average rate is finding the minimum average rate $R^*(\delta)$ for different memory constraints M. Another quantity of interest is the peak rate, denoted by $R^*_{\text{worst}}(\mathcal{M},\delta)$, which is defined as $R^*_{\text{worst}}(\mathcal{M},\delta) = \max_{\mathbf{d}} R^*(\mathbf{d},\mathcal{M},\delta)$. The minimum peak rate is defined as $R^*_{\text{worst}}(\delta) = \min_{\mathcal{M}} R^*_{\text{worst}}(\mathcal{M},\delta)$.

C. Shared Cache Scheme

The coded caching system with shared cache [10] is described as follows. There are N files, K users and $\Lambda \leq K$ caches, with normalized memory of each cache being M. Parameter γ is defined to be $\gamma = \frac{M}{N}$. Each cache $\lambda = 1, 2, \ldots, \Lambda$, is assigned to a set of users $\mathcal{U}_\lambda$, and all these disjoint sets,

$$\mathcal{U} \triangleq \{\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_\Lambda\}$$

form a partition of the set of users $\{1, 2, \ldots, K\}$, describing the overall association of the users to the caches. For any given $\mathcal{U}$, we consider the association profile

$$\mathcal{L} = (\mathcal{L}_1, \ldots, \mathcal{L}_\Lambda)$$

where $\mathcal{L}_\lambda$ is the number of users assigned to the λth most populated helper node/cache.

1) *SC Prefetching Phase*: Each file X^n is split into $\binom{\Lambda}{\Lambda\gamma}$ disjoint subfiles $X^n_{\mathcal{T}}$, for each $\mathcal{T} \subset [\Lambda]$, $|\mathcal{T}| = \Lambda\gamma$, and then each cache stores a fraction γ of each file. For instance, the λth cache stores subfiles in the set $\{X^n_{\mathcal{T}} : \lambda \in \mathcal{T}, \forall n \in [N]\}$. This prefetching scheme is denoted by $\mathcal{M}_{\text{SC}}$.

2) *SC Delivery Phase*: Without loss of generality assume $|\mathcal{U}_1| \geq |\mathcal{U}_2| \geq \ldots |\mathcal{U}_\Lambda|$ (any other case can be handled by simple relabeling of the caches) and $\mathcal{L}_\lambda = |\mathcal{U}_\lambda|$. With a slight abuse of notation, each $\mathcal{U}_\lambda$ denotes an ordered vector describing the users associated to cache λ. Delivery phase consists of $\mathcal{L}_1$ rounds, where each round $j \in [\mathcal{L}_1]$ serves users $\mathcal{R}_j = \bigcup_{\lambda\in[\Lambda]}(\mathcal{U}_\lambda(j) : \mathcal{L}_\lambda \geq j)$, where $\mathcal{U}_\lambda(j)$ is the jth user in the set $\mathcal{U}_\lambda$. For each round j, the sets $\mathcal{Q} \subseteq [\Lambda]$ of size $|\mathcal{Q}| = \Lambda\gamma + 1$ are considered and for each set $\mathcal{Q}$. The set of receiving users are $\mathcal{E}_\mathcal{Q} = \bigcup_{\lambda\in\mathcal{Q}}(\mathcal{U}_\lambda(j) : \mathcal{L}_\lambda \geq j)$. If $\mathcal{E}_\mathcal{Q} \neq \Phi$, the server transmits, $T_{\mathcal{E}_\mathcal{Q}} = \oplus_{\lambda\in\mathcal{Q}:\mathcal{L}_\lambda\geq j} X^{d_{\mathcal{U}_\lambda(j)}}_{\mathcal{Q}\setminus\{\lambda\}}$. If $\mathcal{E}_\mathcal{Q} = \Phi$, there is no transmission. The decoding is possible for each user using these transmissions [10]. The optimal worst case rate for the SC scheme is obtained in [10] as

$$R^*_{\text{worst}}(\mathcal{M}_{\text{SC}}, 0) = Conv\left(\frac{\sum_{i=1}^{\Lambda-\Lambda\gamma} \mathcal{L}_i \binom{\Lambda-i}{\Lambda\gamma}}{\binom{\Lambda}{\Lambda\gamma}}\right)$$

at points $\gamma = \{\frac{1}{\Lambda}, \frac{2}{\Lambda}, \ldots, 1\}$.

For a fixed prefetching $\mathcal{M}$ and for a fixed demand $\mathbf{d}$, the delivery phase of a coded caching problem is an index coding problem [1]. In fact, for fixed prefetching, a coded caching scheme consists of N^K parallel index coding problems one for each of the N^K possible user demands. Thus finding the minimum achievable rate for a given demand $\mathbf{d}$ is equivalent to finding the min-rank of the equivalent index coding problem induced by the demand $\mathbf{d}$.

Consider the SC prefetching scheme $\mathcal{M}_{\text{SC}}$. The index coding problem induced by the demand $\mathbf{d}$ for SC prefetching is denoted by $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d})$. Each subfile $X^n_{\mathcal{T}}$ corresponds to a message in the index coding problem. The corresponding generalized independence number and min-rank are represented as $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d})$ and $\kappa(\mathcal{M}_{\text{SC}}, \mathbf{d})$ respectively.

III. Optimal Error Correcting Delivery Scheme for SC Prefetching for Worst Case Demand

In this section we find a closed form expression for generalized independence number $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d})$ of the index coding problem $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d})$ for the case when all the files are demanded. Using this, we give an expression for the worst case rate. We denote the worst case demand vector as $\mathbf{d}_{\text{worst}}$. Hence our aim is to find an expression for $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$. In $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d})$ each subfile corresponds to a message. The side information sets of all the receivers in the index coding problem is completely decided by the placement scheme in [10]. We assume a unicast index coding problem for convenience (if there is a receiver demanding multiple messages, we split that receiver into multiple receivers each demanding one file). Hence there are $N\binom{\Lambda}{\Lambda\gamma}$ messages and $K\binom{\Lambda}{\Lambda\gamma}$ receivers in $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d})$. From the delivery scheme and the expression for rate in [10], we get an upper bound for $\kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$ as

$$\kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) \leq \sum_{i=1}^{\Lambda-\Lambda\gamma} \mathcal{L}_i \binom{\Lambda-i}{\Lambda\gamma}. \tag{3}$$

A. Generalized Independence Number for $\mathcal{I}(\mathcal{M}_{SC}, \mathbf{d})$

In this subsection we find the closed form expression for $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$. In the proof of the theorem below, we give a technique to find a generalized independent set for $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$ by intelligently picking messages to the set. Using this we get a lower bound for the generalized independence number, $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$. From this we conclude that $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) = \kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$.

Theorem 1: For the index coding problems $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$ for the case when all the files are demanded, we have

$$\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) = \kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) = \sum_{i=1}^{\Lambda-\Lambda\gamma} \mathcal{L}_i \binom{\Lambda-i}{\Lambda\gamma}.$$

Example 1: Consider a scenario with $K = N = 8$, $M = \Lambda = 4$ and $\mathcal{L} = (3, 2, 2, 1)$. In the placement phase, each file X^n is first split into $\binom{\Lambda}{\Lambda\gamma} = 6$ equally-sized subfiles[1]: $X^n_{1,2}, X^n_{1,3}, X^n_{1,4}, X^n_{2,3}, X^n_{2,4}, X^n_{3,4}$ and then each cache λ stores $X^n_{\mathcal{T}} : \lambda \in \mathcal{T}, \forall n \in [8]$. For example, cache 1 stores subfiles $X^n_{1,2}, X^n_{1,3}, X^n_{1,4}$. In the cache assignment, users $\mathcal{U}_1 = \{1, 2, 3\}, \mathcal{U}_2 = \{4, 5\}, \mathcal{U}_3 = \{6, 7\}$ and $\mathcal{U}_4 = \{8\}$ are assigned to caches $1, 2, 3$ and 4 respectively, so that the association profile is $\mathcal{L} = (3, 2, 2, 1)$. Without loss of generality we assume that the demand vector $\mathbf{d}_{\text{worst}} = (1, 2, 3, 4, 5, 6, 7, 8)$.

We consider the index coding problem $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$. Each of the subfiles correspond to a message in the index coding problem. Hence for this example, the corresponding $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$ will have 48 messages and 48 receivers (each user demanding more than one message is split into multiple

[1]For simplicity we use $X^n_{1,2}$ instead of $X^n_{\{1,2\}}$.

receivers demanding one message each). We construct a set $B(\mathbf{d}_{\text{worst}})$, whose elements are messages of $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$ such that the set of indices of the messages in $B(\mathbf{d}_{\text{worst}})$ forms a generalized independent set. The set $B(\mathbf{d}_{\text{worst}})$ for this case can be constructed as

$$B(\mathbf{d}_{\text{worst}}) = \{X^1_{2,3}, X^1_{2,4}, X^1_{3,4}, X^2_{2,3}, X^2_{2,4}, X^2_{3,4}, X^3_{2,3}, X^3_{2,4}, X^3_{3,4}, X^4_{3,4}, X^5_{3,4}\}.$$

Hence $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) \geq 11$. From the transmission scheme in [10], there are 11 transmissions which satisfy the demands of all the users. Hence $\kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) \leq 11$. Thus from (1) we have for this case, $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) = \kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) = 11$.

B. Expression for Rate

For the worst case demand, we have proved in Theorem 1 that $\alpha(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}}) = \kappa(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$. Hence for this case, the optimal linear error correcting delivery scheme can be constructed by concatenating the worst case delivery scheme in [10] with an optimal error correcting code which corrects the required number of errors. Based on this we give an expression for the worst case rate for SC prefetching in the theorem below.

Theorem 2: For a shared cache system with SC prefetching scheme, we have

$$R^*_{\text{worst}}(\mathcal{M}_{\text{SC}}, \delta) = \frac{N_q[\sum_{i=1}^{\Lambda-\Lambda\gamma} \mathcal{L}_i \binom{\Lambda-i}{\Lambda\gamma}, 2\delta+1]}{\binom{\Lambda}{\Lambda\gamma}}$$

at points $\gamma = \{\frac{1}{\Lambda}, \frac{2}{\Lambda}, \ldots, 1\}$.

Since α and κ bounds meet for $\mathcal{I}(\mathcal{M}_{\text{SC}}, \mathbf{d}_{\text{worst}})$, the optimal linear error correcting delivery scheme here would be concatenation of SC delivery scheme with an optimal classical error correcting delivery scheme which corrects δ errors. Decoding can be done by syndrome decoding for error correcting index codes proposed in [14]. We give an example for which we construct optimal error correcting delivery scheme for coded caching problems with SC prefetching.

Example 2: Consider the coded caching problem with shared caches which we considered in Example 1. For this we know that the α and κ bounds meet and hence the concatenation scheme is optimal. For this case, the SC delivery scheme is as follows. There are 3 rounds with each round serving the following sets of users: $\mathcal{R}_1 = \{1, 4, 6, 8\}, \mathcal{R}_2 = \{2, 5, 7\}, \mathcal{R}_3 = \{3\}$. In the first round, the server transmits the following symbols, $T_1 : X^1_{2,3} \oplus X^4_{1,3} \oplus X^6_{1,2}$, $T_2 : X^1_{2,4} \oplus X^4_{1,4} \oplus X^8_{1,2}$, $T_3 : X^1_{3,4} \oplus X^6_{1,4} \oplus X^8_{1,3}$ and $T_4 : X^4_{3,4} \oplus X^6_{2,4} \oplus X^8_{2,3}$. In the second round the transmissions are: $T_5 : X^2_{2,3} \oplus X^5_{1,5} \oplus X^7_{1,2}$, $T_6 : X^2_{2,4} \oplus X^5_{1,4}$, $T_7 : X^2_{3,4} \oplus X^7_{1,4}$ and $T_8 : X^5_{3,4} \oplus X^7_{2,4}$. The transmissions in the third round are: $T_9 : X^3_{2,3}$, $T_{10} : X^3_{2,4}$ and $T_{11} : X^3_{3,4}$. If we need to correct $\delta = 1$ error, we need to concatenate SC transmission scheme with a classical error correcting code with optimal length. From [24], we have $N_2[11, 3] = 15$. Hence the optimal concatenation can be done with a $[15, 11, 3]_2$ code.

IV. Improvement on SC Scheme for Non-distinct Demands

In this section, we consider the case when the demands are non-distinct. We give a delivery scheme which clearly has an advantage over the scheme in [10]. We give an expression for the achievable rate for any demand vector $\mathbf{d}$ which meets the expression for achievable rate in the case of [10] for the worst case demand. Before formally describing the proposed delivery scheme, we demonstrate the main ideas of the scheme through a motivating example.

A. Motivating Example

Consider the same system which we explained in Example 1. The placement scheme and user assignments are the same as in Example 1. We assume here that the demand vector $\mathbf{d} = (1, 2, 3, 1, 1, 1, 1, 1)$. Thus, here $N_e(\mathbf{d}) = 3$. Before the delivery scheme starts, we eliminate some demands which are redundant. If multiple users which are connected to the same cache demand the same file, the delivery scheme need to satisfy the demand of one of them and the others also get what they want. Hence we can eliminate the repeated demand among the users which are connected to the same cache. Thus in the example, we can modify the association profile as $\mathcal{L} = (3, 1, 1, 1)$ and $\mathbf{d} = (1, 2, 3, 1, 1, 1)$. After this, the delivery scheme is done in rounds as in [10], but with a modification. Delivery takes place in 3 rounds, with each round respectively serving the following sets of users: $\mathcal{R}_1 = \{1, 1, 1, 1\}, \mathcal{R}_2 = \{2\}$ and $\mathcal{R}_3 = \{3\}$. In the first round, the server transmits

$$T_{\{1,1,1\}_1} = X^1_{2,3} \oplus X^1_{1,3} \oplus X^1_{1,2}$$
$$T_{\{1,1,1\}_2} = X^1_{2,4} \oplus X^1_{1,4} \oplus X^1_{1,2}$$
$$T_{\{1,1,1\}_3} = X^1_{3,4} \oplus X^1_{1,4} \oplus X^1_{1,3}.$$

Here the decoding is done as in [5]. For instance, user 1, upon receiving $T_{\{1,1,1\}_1}$, can decode $X^1_{2,3}$ using the helper cache contents $X^1_{1,3}$ and $X^1_{1,2}$. Similarly using other transmissions, other subfiles can be decoded. In the second round, we have the following set of transmissions:

$$T_{2_1} = X^2_{2,3}, \quad T_{2_2} = X^2_{2,4}, \quad T_{2_3} = X^2_{3,4}.$$

In the last round the server serves user 3 with three more transmissions given by:

$$T_{3_1} = X^3_{2,3}, \quad T_{3_2} = X^3_{2,4}, \quad T_{3_3} = X^3_{3,4}.$$

Hence there are a total of 9 transmissions, which means that the rate achieved is $\frac{9}{6} = \frac{3}{2}$. This is a smaller rate compared to the rate $\frac{11}{6}$ achieved by the scheme in [10].

B. General Delivery Phase

We follow the assumptions and most of the notations as in [10] to describe the scheme. Let the demand vector be $\mathbf{d}$ and let the number of distinct files requested be $N_e(\mathbf{d})$. We use the notation $N_e(\mathcal{U}_\lambda)$ for the number of distinct files demanded by the users in $\mathcal{U}_\lambda$. We need to consider only $N_e(\mathcal{U}_\lambda)$ users

which request distinct files and satisfy their demand. This is because, any other user in $\mathcal{U}_\lambda$ can get its requested file from the transmissions. Hence before the delivery starts, we eliminate the users with repeated demand from each $\mathcal{U}_\lambda$. After eliminating such users, let the modified association profile be $\mathcal{L}'$. The remaining users associated to cache λ is denoted by $\mathcal{U}'_\lambda$. Moreover, let $\mathcal{L}'{}_\lambda \triangleq |\mathcal{U}'_\lambda|$. Without loss of generality, we assume that $\mathcal{L}'_1 \geq \mathcal{L}'_2 \geq \ldots \mathcal{L}'_\Lambda$. Delivery phase consists of $\mathcal{L}'_1$ rounds, where each round $j \in [\mathcal{L}'{}_1]$ serves users

$$\mathcal{R}'_j = \bigcup_{\lambda \in [\Lambda]} (\mathcal{U}'_\lambda(j) : \mathcal{L}'{}_\lambda \geq j),$$

where $\mathcal{U}'_\lambda(j)$ is the jth user in the set $\mathcal{U}'_\lambda$. Let the number of distinct files in $\mathcal{R}'_j$ be $N_e(\mathcal{R}'_j)$. For each round j, the server selects a subset of $N_e(\mathcal{R}'_j)$ users, denoted by $\mathcal{P}_j$ that requests $N_e(\mathcal{R}'_j)$ different files. These users are considered as leaders. For each round j, we create sets $\mathcal{Q} \subseteq [\Lambda]$ of size $|\mathcal{Q}| = \Lambda\gamma+1$, and for each set $\mathcal{Q}$ which satisfy $\mathcal{A} \cap \mathcal{P}_j \neq \Phi$, we pick the set of receiving users as $\mathcal{E}_\mathcal{Q} = \bigcup_{\lambda \in \mathcal{Q}} (\mathcal{U}'_\lambda(j) : \mathcal{L}'{}_\lambda \geq j)$. If $\mathcal{E}_\mathcal{Q} \neq \Phi$, the server transmits, $T_{\mathcal{E}_\mathcal{Q}} = \oplus_{\lambda \in \mathcal{Q}: \mathcal{L}'{}_\lambda \geq j} X^{d_{\mathcal{U}'_\lambda(j)}}_{\mathcal{Q}\setminus\{\lambda\}}$. If $\mathcal{E}_\mathcal{Q} = \Phi$, there is no transmission. Since this transmission scheme uses scheme in [5] for each round, the decoding at each receiver is ensured. The theorem below gives an expression for rate in this scheme.

Theorem 3: For coded caching problems with SC prefetching scheme,

$$R(\mathcal{M}_{\text{SC}}, \delta = 0) = Conv\left(\mathbb{E}_{\mathbf{d}}\left[\frac{\sum_{j=1}^{\mathcal{L}_1} \binom{\Lambda}{\Lambda\gamma+1} - \binom{\Lambda - N_e(\mathcal{R}'_j)}{\Lambda\gamma+1} - \binom{\Lambda - |R'_j|}{\Lambda\gamma+1}}{\binom{\Lambda}{\Lambda\gamma}}\right]\right)$$

at points $\gamma = \{\frac{1}{\Lambda}, \frac{2}{\Lambda}, \ldots, 1\}$.

V. Conclusion

We considered the SC scheme and for worst case demand, we proved that for all the corresponding index coding problems, the α and κ bounds meet. This makes the concatenation of SC delivery scheme with an optimal classical error correcting code which corrects the required number of errors to be optimal. Moreover, for the case of non-distinct demands, we proposed an improved scheme which has clear advantage over the scheme in [10].

Acknowledgment

This work was supported partly by the Science and Engineering Research Board (SERB) of Department of Science and Technology (DST), Government of India, through J.C. Bose National Fellowship to B. Sundar Rajan.

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. Networking*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[3] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.

[4] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Trans. Networking*, vol. 24, no. 2, pp. 836–845, Apr. 2016.

[5] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 1613–1617.

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

[7] J. Gómez-Vilardebó, "Fundamental Limits of Caching: Improved Rate-Memory Tradeoff With Coded Prefetching," in *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4488-4497, Oct. 2018.

[8] C. Tian and J. Chen, "Caching and delivery via interference elimination," *IEEE Trans. on Information Theory*, vol. 64, no. 3, pp. 1548-1560, 2018.

[9] K. Zhang and C. Tian, "From Uncoded Prefetching to Coded Prefetching in Coded Caching Systems," 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, 2018, pp. 2087-2091.

[10] E. Parrinello, A. Unsal, P. Elia, " Fundamental Limits of Caching in Heterogeneous Networks with Uncoded Prefetching," Available on arXiv:1811.06247 [cs.IT], Nov. 2018.

[11] N. S. Karat, A. Thomas and B. S. Rajan, "Optimal Error Correcting Delivery Scheme for Coded Caching with Symmetric Batch Prefetching," *2018 IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, 2018, pp. 2092-2096.

[12] N. S. Karat, A. Thomas and B. S. Rajan,"Optimal Error Correcting Delivery Scheme for an Optimal Coded Caching Scheme with Small Buffers," *2018 IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, 2018, pp. 1710-1714.

[13] S. S. Bidokhti, M. Wigger and R. Timo, "Noisy Broadcast Networks With Receiver Caching," in *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6996-7016, Nov. 2018.

[14] S. H. Dau, V. Skachek, and Y. M. Chee, "Error correction for index coding with side information," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1517–1531, Mar. 2013.

[15] Y. Birk and T. Kol, "Coding-on-demand by an informed source (ISCOD) for efficient broadcast of different supplemental data to caching clients," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2825-2830, Jun. 2006.

[16] N. Alon, A. Hassidim, E. Lubetzky, U. Stav, and A. Weinstein, "Broadcasting with side information," in *Proc. 49th Annu. IEEE Symp. Found. Comput. Sci.*, Oct. 2008, pp. 823-832.

[17] Z. Bar-Yossef, Y. Birk, T. S. Jayram, and T. Kol, "Index coding with side information," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2006, pp. 197-206.

[18] R. Peeters, "Orthogonal representations over finite fields and the chromatic number of graphs," *Combinatorica*, vol. 16, no. 3, pp. 417–431, 1996.

[19] S. H. Dau, V. Skachek and Y. M. Chee, "Optimal Index Codes With Near-Extreme Rates," in IEEE Transactions on Information Theory, vol. 60, no. 3, pp. 1515-1527, March 2014.

[20] S. Samuel and B. S. Rajan, "Optimal linear error-correcting index codes for single-prior index-coding with side information," in *Proc. 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[21] N. S. Karat and B. S. Rajan, "Optimal linear error correcting index codes for some index coding problems," in *Proc. 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[22] S. Samuel, N. S. Karat, and B. S. Rajan, "Optimal linear error correcting index codes for some generalized index-coding problems," in Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Montreal, Canada, Oct. 2017.

[23] N. S. Karat, S. Samuel, and B. S. Rajan, "Optimal error correcting index codes for some generalized index coding problems," in *IEEE Transactions on Communications* early access: doi: 10.1109/TCOMM.2018.2878566.

[24] M. Grassl, "Bounds on the minimum distance of linear codes and quantum codes," Online available at http://www.codetables.de, 2007.

[25] Nujoom Sageer Karat, Spandan Dey, Anoop Thomas and B. Sundar Rajan, "An Optimal Linear Error Correcting Delivery Scheme for Coded Caching with Shard Caches," Available on arXiv:1811.06247 [cs.IT], Jan.09, 2019.