Content delivery in erasure broadcast channels with cache and feedback

Asma Ghorbel, Mari Kobayashi, Sheng Yang

Content Delivery in Erasure Broadcast Channels with Cache and Feedback

Asma Ghorbel, Mari Kobayashi, and Sheng Yang
LSS, CentraleSupélec
Gif-sur-Yvette, France
{asma.ghorbel, mari.kobayashi, sheng.yang}@centralesupelec.fr

***Abstract*—We study a content delivery problem in the context of a K-user erasure broadcast channel such that a content providing server wishes to deliver requested files to users, each equipped with a cache of a finite memory. Assuming that the transmitter has state feedback and user caches can be filled during off-peak hours reliably by decentralized cache placement, we characterize the achievable rate region as a function of the memory sizes and the erasure probabilities. The proposed delivery scheme, based on the broadcasting scheme proposed by Wang and Gatzianas et al., exploits the receiver side information established during the placement phase. Our results can be extended to centralized cache placement as well as multi-antenna broadcast channels with state feedback.**

I. Introduction

Today's exponentially growing mobile data traffic is mainly due to video applications such as content-based video streaming. The skewness of the video traffic together with the ever-growing cheap on-board storage memory suggests that the quality of experience can be boosted by caching popular contents at (or close to) the end-users in wireless networks. Most of existing works assume that caching is performed in two phases: *placement phase* to prefetch users' caches under their memory constraints (typically during off-peak hours) prior to the actual demands; *delivery phase* to transmit codewords such that each user, based on the received signal and the contents of its cache, is able to decode the requested file. In this work, we study the delivery phase based on a coded caching model where a server is connected to many users, each equipped with a cache of finite memory [1]. By carefully choosing the sub-files to be distributed across users, coded caching exploits opportunistic multicasting such that a common signal is simultaneously useful for all users even with distinct file requests. A number of extensions of coded caching have been developed, e.g. [1, Section VIII] including the case of decentralized placement phase, the case of non-uniform demands, the case in a more general network, as well as the performance analysis in different regime [11]. Further, very recent works have attempted to relax the unrealistic assumption of a perfect shared link by replacing it by wireless channels (e.g. [4], [5] and references therein). The works [4], [5] have studied also the role of channel state feedback in the context of coded caching.

We model the bottleneck link between the server with N files and K users equipped with a cache of a finite memory as an erasure broadcast channel (EBC). We assume that the EBC is memoryless, independently distributed across users with erasure probabilities $\{\delta_k\}$ and that user k has a cache of M_k files. Moreover, the server is assumed to acquire the channel states causally via feedback sent by users. Under this setting, we study the achievable rate region of the EBC for the case of decentralized cache placement [2]. Our contributions are three-fold and summarized together with the outline of this paper : 1) characterize the upper bound on the achievable rate region (section III) ; 2) propose a multi-phase delivery scheme extending the algorithm proposed by Wang and Gatzianas et al. [6], [7] to the case of receiver side information and prove that it achieves the optimal rate region for special cases of interest (section IV) ; 3) extend the results to centralized cache placement [1] as well as the multi-antenna broadcast channel with state feedback (section V). It should be remarked that the current work is a non-trivial extension of [4] which is restricted to the symmetric network (with equal erasure probabilities and memory sizes) because the achievability proof in [4], exploiting the polyhedron structure of the rate region, cannot be applied to a general network setting considered here. Finally, the numerical examples in section VI enable to quantify the benefit of state feedback, the relative merit of centralized caching to decentralized counterpart, as well as the gain due to the optimization of memory sizes as a function of other system parameters.

Throughout the paper, we use the following notations. X^n and $X_{\mathcal{I}}$ denote a sequence $(X_1, \ldots, X_n)$ and $\{X_i\}_{i\in\mathcal{I}}$, respectively. The entropy of X is denoted by $H(X)$. We let $[k] = \{1, \ldots, k\}$. Due to the space limitation, detailed proofs are omitted and will be deferred to the full version [8]. We let ϵ_n denote a constant which vanishes as $n \to \infty$.

II. System Model and Main Results

A. System model and definitions

We consider a cache-enabled network depicted in Fig. 1 where a server is connected to K users through an erasure broadcast channel (EBC). The server has an access to N files $W_1, \ldots, W_N$ where file i, i.e. W_i, consists of F_i packets of L bits each (F_iL bits). Each user k has a cache memory Z_k of M_kF packets for $M_k \in [0, N]$, where $F \triangleq \frac{1}{N}\sum_{i=1}^{N} F_i$ is the average size of the files. Under such a setting, consider a discrete time communication system where a packet is sent in

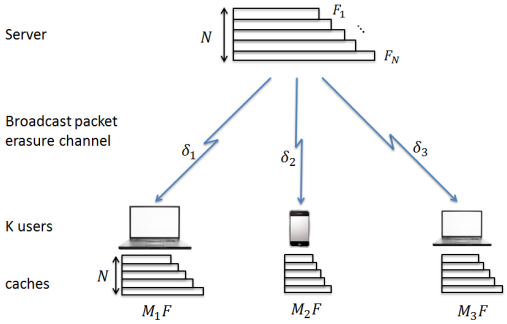

Fig. 1. A cached-enabled erasure broadcast channel with $K = 3$.

each slot over the K-user EBC. The channel input $X_l \in \mathbb{F}_q$ belongs to the alphabet of size $L = \log_2 q$ bits. The channel is assumed to be memoryless and independently distributed across users so that in a given slot we have

$$\Pr(Y_1, Y_2, \ldots, Y_K|X) = \prod_{k=1}^{K} \Pr(Y_k|X) \tag{1}$$

$$\Pr(Y_k|X) = \begin{cases} 1-\delta_k, & Y_k = X, \\ \delta_k, & Y_k = E \end{cases} \tag{2}$$

where Y_k denotes the channel output of receiver k, E stands for an erased output, δ_k denotes the erasure probability for user k. We let $S_l \in \mathcal{S} = 2^{\{1,\ldots,K\}}$ denote the state of the channel in slot l which indicates the users who received correctly the packet. We assume that the transmitter obtains the state feedback S^{l-1} at the end of slot l while all the receivers know S^n at the end of the transmission.

There are two phases for our content delivery problem: placement phase and delivery phase. In placement phase, the server fills the caches of all users $Z_1, \ldots, Z_K$ up to the memory constraint. As in most works in the literature, we assume that placement phase is performed over an error-free link and does not incur any resource overhead, since it takes place usually during off-peak traffic hours. Once each user k makes a request d_k, the server sends codewords so that each user can decode its requested file as a function of its cache content and received signals during delivery phase. We provide a more formal definition below. A $(M_1, \ldots, M_K, F_{d_1}, \ldots, F_{d_K}, n)$ caching scheme consists of the following components.

- N message files $W_1, \ldots, W_N$ are independently and uniformly distributed over $\mathcal{W}_1 \times \cdots \times \mathcal{W}_N$ with $\mathcal{W}_i = \mathbb{F}_q^{F_i}$ for all i.
- K caching functions are given by $\phi_k : \mathbb{F}_q^{\sum_{i=1}^{N} F_i} \to \mathbb{F}_q^{FM_k}$ map the files $W_1, \ldots, W_N$ into the cache contents
$$Z_k = \phi_k(W_1, \ldots, W_N) \tag{3}$$
for each user k.
- A sequence of encoding functions which transmit at slot l a symbol $X_l = f_l(W_{d_1}, \ldots, W_{d_K}, S^{l-1}) \in \mathbb{F}_q$, based on the requested file set and the channel feedback up to slot $l-1$ for $l = 1, \ldots, n$, where W_{d_k} denotes the message file requested by user k with $d_k \in \{1, \ldots, N\}$.
- A decoding function of user k is given by the mapping $\psi_k : \mathbb{F}_q^n \times \mathbb{F}_q^{FM_k} \times \mathcal{S}^n \to \mathbb{F}_q^{F_{d_k}}$ so that the decoded file is $\hat{W}_{d_k} = \psi_k(Y_k^n, Z_k, S^n)$ as a function of the received signals Y_k^n, the cache content Z_k, as well as the state information S^n.

A rate tuple $(R_1, \ldots, R_K)$ is said to be achievable if, for every $\epsilon > 0$, there exists a $(M_1, \ldots, M_K, F_{d_1}, \ldots, F_{d_K}, n)$ caching strategy that satisfies the reliability condition

$$\max_{(d_1,\ldots,d_K)\in\{1,\ldots,N\}^K} \max_k \Pr(\psi_k(Y_k^n, Z_k, S^n) \neq W_{d_k}) < \epsilon$$

as well as rate condition $R_k \leq \frac{F_{d_k}}{n}$. Throughout the paper, we express the entropy and the rate in terms of packets in oder to avoid the constant factor L.

B. Decentralized cache placement

We consider the decentralized cache placement proposed in [2]. Under the memory constraint of $M_k F$ packets, each user k independently caches a subset of $p_k F_i$ packets of file i, chosen uniformly at random for $i = 1, \ldots, N$, where $p_k = \frac{M_k}{N}$. Let $\mathcal{L}_{\mathcal{J}}(W_i)$ denote the sub-file of W_i stored exclusively in the cache memories (known) of the users in $\mathcal{J}$. The cache memory Z_k of user k after decentralized placement is given by

$$Z_k = \{\mathcal{L}_{\mathcal{J}}(W_i) : \mathcal{J} \subseteq [K],\ \mathcal{J} \ni k,\ i = 1, \ldots, N\}. \tag{4}$$

The size of each sub-file is given by

$$|\mathcal{L}_{\mathcal{J}}(W_k)| = \prod_{j\in\mathcal{J}} p_j \prod_{j\in[K]\setminus\mathcal{J}} (1-p_j) F_k + \epsilon_{F_k} \tag{5}$$

as $F_k \to \infty$. It can be easily verified that the memory constraint of each user is fulfilled.

C. Main results

In order to present the main results, we specify two special cases.

Definition 1. *The cache-enabled EBC (or the network) is symmetric if we have $\delta_1 = \cdots = \delta_K$ and $p_1 = \cdots = p_K$.*

Definition 2. *The rate vector is one-sided fair in the cache-enabled EBC if $\delta_k \geq \delta_j$ and for $k \neq j$ implies*

$$\frac{R_k}{R_j} \geq \max\left\{\frac{\delta_j}{\delta_k}, \frac{(1-p_j)/p_j}{(1-p_k)/p_k}\right\} \tag{6}$$

For the special case without cache memory ($p_k = 0, \forall k$), Definition 2 boils down to the original definition [6], i.e. $R_k/R_j \geq \delta_j/\delta_k$. Such rate vector corresponds to the sub-region where user k (with the worse channel) requires more transmission time than user j to receive a common packet (as it will be clear in subsection IV-B). In the presence of cache memory, such subregion depends also on the memory size. We focus on the most relevant case of $N \geq K$ and assume further that all demands are distinct.

Theorem 1. *For $K \leq 3$, or for the symmetric network with $K \geq 3$, or for the one-sided fair rate with $K > 3$, the achievable rate region of the cached-enabled EBC with state feedback under decentralized cache placement is given by*

$$\sum_{k=1}^{K} \frac{\prod_{j=1}^{k}(1-p_{\pi_j})}{1-\prod_{j=1}^{k}\delta_{\pi_j}} R_{\pi_k} \leq 1 \tag{7}$$

for any permutation π of $\{1,\dots,K\}$.

The proof of Theorem 1 is provided in upcoming sections. Theorem 1 covers existing results as special cases including the symmetric network [4] and the EBC without cache memory [6], [7].

Corollary 1. *For $K \le 3$, or for the symmetric network with $K \ge 3$, or for the one-sided fair rate with $K > 3$, the total transmission duration to deliver a distinct requested file to each user in the cached-enabled EBC under decentralized cache placement is given by, as $F \to \infty$,*

$$T_{\mathrm{tot}} = \max_{\pi} \left\{ \sum_{k=1}^{K} \frac{\prod_{j=1}^{k}(1-p_{\pi_j})}{1-\prod_{j=1}^{k}\delta_{\pi_j}} F_{d_{\pi_k}} \right\} + \Theta(1). \quad (8)$$

Corollary 1 yields existing results for the special cases without erasure. For the case of no erasure and equal file size ($\delta_k = 0, \forall k$, $F_i = F, \forall i$), (8) coincides with the rate-memory tradeoff[1] of decentralized coded caching under asymmetric memory sizes [3, Theorem 3]. Additionally, if we let the memory size be equal, (8) boils down to the rate-memory tradeoff of decentralized coded caching [2].

III. Converse

Since the converse of Theorem 1 follows the similar idea as the one in [4, Section III], we describe only the main steps. First we recall the two useful lemmas.

Lemma 1. *[4, Lemma 1] For the erasure broadcast channel, if U is such that $X_l \leftrightarrow UY_{\mathcal{J}}^{l-1}S^{l-1} \leftrightarrow (S_{l+1},\dots,S_n)$, $\forall \mathcal{J}$,*

$$\frac{1}{1-\prod_{j\in\mathcal{J}}\delta_j} H(Y_{\mathcal{J}}^n \,|\, U, S^n) \le \frac{1}{1-\prod_{j\in\mathcal{J}}\delta_j} H(Y_{\mathcal{J}}^n \,|\, U, S^n),$$

for any sets $\mathcal{I},\mathcal{J}$ such that $\mathcal{J}\subseteq\mathcal{I}\subseteq\{1,\dots,K\}$.

As a straightforward extension of [4, Lemma 2], we have the following lemma.

Lemma 2. *Considering the cache placement in [2], the following equality holds for any i and $\mathcal{K}\subseteq[K]$*

$$H(W_i \,|\, \{Z_k\}_{k\in\mathcal{K}}) \ge \prod_{k\in\mathcal{K}} (1-p_k)\, H(W_i). \quad (9)$$

Let us focus on the case without permutation and the demand $(d_1,\dots,d_K) = (1,\dots,K)$ without loss of generality. We create a degraded broadcast channel, by providing (W_k, Y_k, Z_k) to receivers $k+1,\dots,K$, and obtain

$$\begin{aligned} n\prod_{j=1}^{k}(1-p_j)R_k &= \prod_{j=1}^{k}(1-p_j)H(W_k) && (10)\\ &\le H(W_k|Z^kS^n) && (11)\\ &\le I(W_k; Y_{[k]}^n \,|\, W^{k-1}Z^kS^n) + n\epsilon'_{n,k} && (12)\end{aligned}$$

where the second inequality is by applying Lemma 2 and noting that S^n is independent of others; the last inequality is from Fano's inequality and from the fact that $I(W_k; W^{k-1} \,|\, Z^kS^n) = 0$. Defining $\epsilon_{n,k} \triangleq \epsilon'_{n,k} / \prod_{j=1}^{k}(1-p_j)$, we obtain

$$\begin{aligned} &n\prod_{j=1}^{k}(1-p_j)(R_k-\epsilon_{n,k})\\ &\quad\le H(Y_{[k]}^n \,|\, W^{k-1}Z^kS^n) - H(Y_{[k]}^n \,|\, W^kZ^kS^n). \quad (13)\end{aligned}$$

Summing up (13) with different weights and applying Lemma 1 for $K-1$ times, it readily follows that

$$\begin{aligned} &\sum_{k=1}^{K} \frac{\prod_{j\in[k]}(1-p_j)}{1-\prod_{j\in[k]}\delta_j}(R_k-\epsilon_{n,k})\\ &\quad\le \frac{H(Y_1^n \,|\, Z_1S^n)}{n(1-\delta_1)} - \frac{H(Y_{[K]}^n \,|\, W^KZ^KS^n)}{n(1-\prod_{j\in[k]}\delta_j)} && (14)\\ &\quad\le \frac{H(Y_1^n)}{n(1-\delta_1)} \le 1. && (15)\end{aligned}$$

This establishes the converse part, after letting $n\to\infty$.

IV. Achievability

A. Revisiting the broadcasting scheme [6], [7]

We provide a high-level description of the broadcasting scheme [6], [7] by assuming the number of private packets $\{F_k\}$ is arbitrarily large so that the length of each phase becomes deterministic. The broadcasting algorithm has two main roles: 1) broadcast new information packets and 2) multicast side information or overheard packets thanks to state feedback. From this reason, we can call phase 1 *broadcasting phase* and phases 2 to K *multicasting phase*. Phase j consists of $\binom{K}{j}$ subphases in each of which the transmitter sends packets intended to a subset of users $\mathcal{J}$ for $|\mathcal{J}| = j$. We let $\mathcal{L}_{\mathcal{J}}(V_{\mathcal{K}})$ denote the part of packet $V_{\mathcal{K}}$ received by users in $\mathcal{J}$ and erased by $[K]\setminus\mathcal{J}$. Here is a high-level description of the broadcasting algorithm:

1) Broadcasting phase (phase 1): send each message $V_k = W_k$ of F_k packets sequentially for $k = 1,\dots,K$. This phase generates overheard symbols $\{\mathcal{L}_{\mathcal{J}}(V_k)\}$ to be transmitted via linear combination in multicasting phase, where $\mathcal{J}\subseteq[K]\setminus k$ for all k.
2) Multicasting phase (phases $2-K$): for a subset $\mathcal{J}$ of users, generate $V_{\mathcal{J}}$ as a linear combination of overheard packets such that

$$V_{\mathcal{J}} = \mathcal{F}_{\mathcal{J}}\left(\{\mathcal{L}_{\mathcal{J}\setminus\mathcal{I}\cup\mathcal{J}'}(V_{\mathcal{I}})\}_{\mathcal{J}'\subset\mathcal{I}\subset\mathcal{J}}\right) \quad (16)$$

where $\mathcal{F}_{\mathcal{J}}$ denotes a linear function. Send $V_{\mathcal{J}}$ sequentially for all $\mathcal{J}\subseteq[K]$ of the cardinality $|\mathcal{J}| = 2,\dots,K$.

In order to determine the total transmission duration, we need to introduce further some notions and parameters.

- A packet intended to $\mathcal{J}$ is consumed *for a given user* $k\in\mathcal{J}$ if this user or at least one user in $[K]\setminus\mathcal{J}$ receives it. The probability of such event is $1-\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}}\delta_j$.
- A packet intended to $\mathcal{I}$ creates a packet intended to users in $\mathcal{J}$ for user $k\in\mathcal{I}\subset\mathcal{J}\subseteq[K]$ if erased by user k and all users in $[K]\setminus\mathcal{J}$ but received by $\mathcal{J}\setminus\mathcal{I}$.

[1] In [2], [3] the "rate" is defined as the number of files to deliver over the shared link, which corresponds here to T_{tot}.

The probability of such event is denoted by $\alpha_{\mathcal{I}\to\mathcal{J}}^{\{k\}} = \prod_{j'\in[K]\setminus\mathcal{J}\cup\{k\}}\delta_{j'}\prod_{j\in\mathcal{J}\setminus\mathcal{I}}(1-\delta_j)$. We let

$$N_{\mathcal{I}\to\mathcal{J}}^{\{k\}} = t_{\mathcal{I}}^{\{k\}}\alpha_{\mathcal{I}\to\mathcal{J}}^{\{k\}} \tag{17}$$

denote the number of such packets, where $t_{\mathcal{I}}^{\{k\}}$ is the duration needed by user k during sub-phase $\mathcal{J}$.

- The duration $t_{\mathcal{J}}$ of sub-phase $\mathcal{J}$ is given by: $t_{\mathcal{J}} = \max_{k\in\mathcal{J}} t_{\mathcal{J}}^{\{k\}}$ where

$$t_{\mathcal{J}}^{\{k\}} = \frac{\sum_{k\in\mathcal{I}\subset\mathcal{J}} N_{\mathcal{I}\to\mathcal{J}}^{\{k\}}}{1-\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}}\delta_j}. \tag{18}$$

The total transmission length is obtained by summing up all sub-phases, i.e. $T_{\text{tot}} = \sum_{\mathcal{J}\subseteq[K]} t_{\mathcal{J}}$.

B. Proposed delivery scheme

We describe the proposed delivery scheme for the case[2] of $K > 3$ assuming that user k requests file W_k of size F_k packets for $k = 1,\ldots,K$. Compared to the algorithm [6], [7] revisited previously, our scheme must convey packets generated in the placement phase as well as all previous phases in a given phase.

Placement phase: This phase creates equivalently the "overheard" packets $\{\mathcal{L}_{\mathcal{J}}(W_k)\}$ for $\mathcal{J}\subset[K]\setminus k$ and for $k = 1,..,K$.

Phase 1 : The transmitter sends $V_1,..,V_K$ sequentially until at least one user receives it, where $V_k = \mathcal{L}_{\emptyset}(W_k)$ corresponds to the order-1 packets created by placement phase.

Phases $2\ldots K$: For a subset $\mathcal{J}$ of users, generate $V_{\mathcal{J}}$ as a linear combination of overheard packets during the placement phase as well as during phases 1 to $j-1$.

$$V_{\mathcal{J}} = \mathcal{F}_{\mathcal{J}}\left(\{\mathcal{L}_{\mathcal{J}\setminus\mathcal{I}\cup\mathcal{I}'}(V_{\mathcal{I}})\}_{\mathcal{I}'\subset\mathcal{I}\subset\mathcal{J}}, \mathcal{L}_{\mathcal{J}\setminus\{k\}}(W_k)\right) \tag{19}$$

The rest of the subsection is dedicated to the achievability proof of Theorem 1 for the case of one-sided fair rate vector defined in Definition 2.

We assume without loss of generality $\delta_1 \geq \cdots \geq \delta_K$, $\delta_1 R_1 \geq \cdots \geq \delta_K R_K$, and $\frac{1-p_1}{p_1}R_1 \geq \cdots \geq \frac{1-p_2}{p_2}R_K$. By incorporating the packets generated during placement phase in (18), we have for $k\in\mathcal{J}\subseteq[K]$

$$t_{\mathcal{J}}^{\{k\}} = \frac{\sum_{\mathcal{I}:k\in\mathcal{I}\subset\mathcal{J}} t_{\mathcal{I}}^{\{k\}}\alpha_{\mathcal{I}\to\mathcal{J}}^{\{k\}} + |\mathcal{L}_{\mathcal{J}\setminus\{k\}}(W_k)|}{1-\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}}\delta_j}. \tag{20}$$

We calculate T_{tot} in three main steps:

Step 1 We express $t_{\mathcal{J}}^{\{k\}}$ as a function of key parameters $\{\delta_k\},\{p_k\},\{F_k\}$ in two different ways. By following similar steps as in [7, Appendix C], we obtain the total duration needed for user k for a system with a fixed subset $\mathcal{J}$ of users by

$$\sum_{\mathcal{I}:k\in\mathcal{I}\subseteq\mathcal{J}} t_{\mathcal{I}}^{\{k\}} = \frac{\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}}(1-p_j)}{1-\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}}\delta_j}F_k. \tag{21}$$

[2]For K=3 a different algorithm was proposed in [6] that we adapt it with the presence of cache and readily prove the achievability.

As proved in [8], the above expression can be rewritten as

$$t_{\mathcal{J}}^{\{k\}} = \sum_{\mathcal{H}\subseteq\mathcal{J}\setminus\{k\}}(-1)^{|\mathcal{H}|}\frac{\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}\cup\mathcal{H}}(1-p_j)}{1-\prod_{j\in[K]\setminus\mathcal{J}\cup\{k\}\cup\mathcal{H}}\delta_j}F_k. \tag{22}$$

Step 2 The length of a sub-phase $\mathcal{J}$ is determined by the user which requires the maximum length, i.e. $\arg\max_{k\in\mathcal{J}} t_{\mathcal{J}}^{\{k\}}$. For the special case of one-sided fair rate vector, by using (22) it is possible to prove, as in [8], that

$$\arg\max_{k\in\mathcal{J}} t_{\mathcal{J}}^{\{k\}} = \min\{\mathcal{J}\} \qquad \forall\mathcal{J}\subseteq[K]. \tag{23}$$

This means that the user permutation (which determines the sub-phase length) is preserved in all sub-phases for the one-sided fair rate vector.

Step 3 By combining the two previous steps, the total transmission length can be derived as follows

$$T_{\text{tot}} = \sum_{\mathcal{J}\subseteq[K]}\max_{k\in\mathcal{J}} t_{\mathcal{J}}^{\{k\}} \tag{24}$$

$$= \sum_{\mathcal{J}\subseteq[K]} t_{\mathcal{J}}^{\{\min\mathcal{J}\}} \tag{25}$$

$$= \sum_{i=1}^{K}\sum_{i\in\mathcal{J}\subseteq\{i..K\}} t_{\mathcal{J}}^{\{i\}} \tag{26}$$

$$= \sum_{i=1}^{K} F_i\frac{\prod_{j=1}^{i}(1-p_j)}{1-\prod_{j=1}^{i}\delta_j} \tag{27}$$

where (25) is obtained from (23); the last equality follows from (21). Dividing both sides by T_{tot} and letting $R_k = \frac{F_{d_k}}{T_{\text{tot}}}$, we readily obtain the RHS of (7) for the identity permutation. Since under the one-sided fair rate constraint the inequality corresponding to the identity permutation implies all the other $K!-1$ inequalities of the rate region as proved in [8], this establishes the achievability.

V. Extensions

A. Centralized cache placement

We consider centralized cache placement proposed in [1] for the special case of the symmetric network. Each file is split into $\binom{K}{b}$ disjoint sub-files of equal size, where $b \triangleq \lfloor pK\rfloor$. Each sub-file is cached in a subset of users $\mathcal{J}$ of cardinality $|\mathcal{J}| = b$. The size of any sub-file of file i is given by

$$|\mathcal{L}_{\mathcal{J}}(W_i)| = \frac{1}{\binom{K}{b}}F_i. \tag{28}$$

As a rather straightforward extension of Theorem 1 in the symmetric network, we obtain the following result.

Theorem 2. *For the symmetric network, the optimal rate region of the cached-enabled BEC with state feedback under centralized cache placement is given by*

$$\sum_{k=1}^{K-b}\frac{\binom{K-k}{b}/\binom{K}{b}}{1-\delta^k}R_{\pi_k} \leq 1 \tag{29}$$

for any permutation π *of* $\{1,\ldots,K\}$.

As a direct consequence, the total transmission length to deliver a distinct requested file of size F to each user is

$$T_{\text{tot}} = \sum_{k=1}^{K-b} \frac{\binom{K-k}{b}/\binom{K}{b}}{1-\delta^k} F + \Theta(1), \tag{30}$$

as $F \to \infty$. It is worth mentioning that (30) boils down to the rate-memory tradeoff under centralized cache placement [1] for special case of no erasure.

B. MISO broadcast channel

We consider the K-user multi-input single-output broadcast channel (MISO-BC) with M antennas at the transmitter and K users each equipped with a singel antenna. The channel state in slot l is given by the $M \times K$ channel matrix. We define $\mathsf{DoF}_k = \lim_{\mathsf{snr}\to\infty} \frac{R_k}{\log_2 \mathsf{snr}}$ as the pre-log factor of the rate of user k. Exploiting the explicit connection between the EBC and the MISO broadcast channel in [9], the following theorem can be proved [8].

Theorem 3. *The optimal DoF region of the cached-enabled MISO broadcast channel with $M \geq K-b$ antennas and under centralized cache placement is given by*

$$\sum_{k=1}^{K-b} \frac{\binom{K-k}{b}/\binom{K}{b}}{k} \mathsf{DoF}_{\pi_k} \leq 1 \tag{31}$$

for any permutation π of $\{1,\ldots,K\}$.

It can be observed that (31) has the same structure as (29) where we replace $1-\delta^k$ by k and replace R_k by DoF_k. The achievability can be proved by modifying the scheme in [10] to the case of receiver side information. It can be shown that the optimal total transmission length $T_{\text{tot}} = \sum_{k=b+1}^{K} \frac{1}{k}$ obtained from Theorem 3 coincides with [5, Corollary 2b].

VI. Numerical Examples

In this section we provide some numerical examples to show the performance of our proposed delivery scheme. Fig. 2 plots the normalized total transmission length T_{tot}/F versus the memory size M in the symmetric network with $N = 100, K = 10$. We compare the performance with and without feedback under decentralized and centralized caching for $\delta = 0$ and $\delta = 0.6$. The state feedback is found useful especially for small memory size. The relative merit of centralized placement compared to decentralized the counterpart can be observed. Fig. 3 plots the normalized total transmission length T_{tot}/F versus *average* memory size M in the asymmetric network with $N = 20$ and $K = 4$. We let erasure probabilities $\delta_k = \frac{k}{5}$ for $k = 1,\ldots,4$ and consider equal file size. We compare "symmetric memory" ($M_k = M, \forall k$), "asymmetric memory" obtained by optimizing over all possible sets of $\{M_k\}$ using our delivery scheme, as well as "lower bound" obtained by optimizing all possible of $\{M_k\}$ based on (8). This result shows the advantage (in terms of delivery time) of optimally allocating cache sizes across users, whenever possible, according to the condition of the delivery channels.

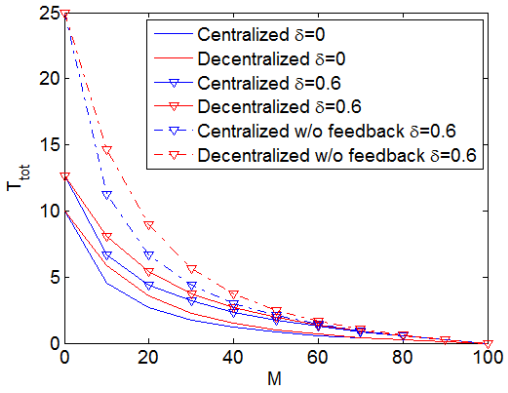

Fig. 2. T_{tot} vs. memory size M for $N = 100, K = 10$.

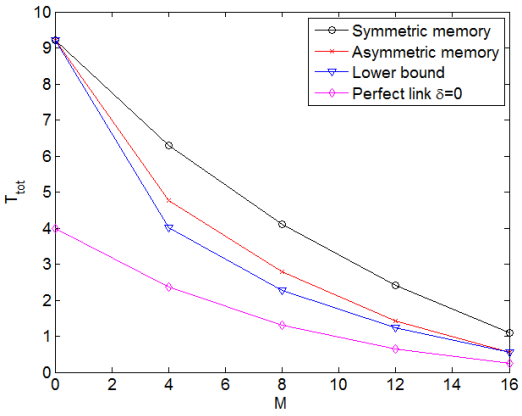

Fig. 3. T_{tot} vs. averaged memory size M for $N = 20$, $K = 4$.

References

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

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

[3] S. Wang, W. Li, X. Tian, H. Liu, "Coded Caching with Heterogenous Cache Sizes", arXiv preprint arXiv:1504.01123v3, 2015.

[4] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-Enabled Broadcast Packet Erasure Channels with State Feedback", proceeding of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015.

[5] J. Zhang, and P. Elia, "Fundamental Limits of Cache-Aided Wireless BC: Interplay of Coded-Caching and CSIT Feedback", arXiv preprint arXiv:1511.03961, 2015.

[6] C. C. Wang, "On the Capacity of 1-to-Broadcast Packet Erasure Channels with Channel Output feedback", *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 931–956, February 2012.

[7] M. Gatzianas, L. Georgiadis, and L. Tassiulas, "Multiuser Broadcast Erasure Channel With Feedback-Capacity and Algorithms", *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5779–5804, September 2013.

[8] A. Ghorbel, M. Kobayashi, and S. Yang, "Content delivery in Erasure Broadcast Channels with cache and Feedback", arXiv preprint arXiv:1602.04630, 2016.

[9] S. Yang and M. Kobayashi, "Secrecy Communications in K-user Multi-Antenna Broadcast Channel with State Feedback", in *Proceedings of the IEEE International Symposium on Information Theory (ISIT'2015)*, Hong-Kong, China, 2015.

[10] M. A. Maddah-Ali and D. N. C. Tse, "Completely Stale Transmitter Channel State Information is Still Very Useful," *IEEE Trans. Inf. Theory* vol. 58, no. 7, pp. 4418–4431, July 2012.

[11] S. Karthikeyan, M. Ji, A. Tulino, J. Llorca and A. Dimakis "Finite Length Analysis of Caching-Aided Coded Multicasting", in *Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IL, USA, 2014.