Joint Cache-Channel Coding over Erasure Broadcast Channels

Roy Timo[†] and Michèle Wigger[‡]

[†]Technische Universität München, roy.timo@tum.de
[‡]Telecom ParisTech, michele.wigger@telecom-paristech.fr

Abstract—We consider a cache-aided communications system in which a transmitter communicates with many receivers over an erasure broadcast channel. The system serves as a basic model for communicating on-demand content during periods of high network congestion, where some content can be pre-placed in local caches near the receivers. We formulate the cache-aided communications problem as a joint cache-channel coding problem, and characterise some information-theoretic tradeoffs between reliable communications rates and cache sizes. We show that if the receivers experience different channel qualities, then using unequal cache sizes and joint cache-channel coding improves system efficiency.

I. INTRODUCTION

Consider a network with one transmitter and many receivers. Imagine that the transmitter has a library of *messages* (or, *data files*), and suppose that each receiver will request and download a message during a period of high network congestion. In such settings, it is advantageous to move traffic away from the congested period using *caching*. The basic idea of caching is that the transmitter sends and stores "parts" of the library in local *cache memories* near the receivers beforehand, during periods with low network traffic. The caches provide this data directly to the receivers, so that less data needs to be sent during the congested period.

The above problem is relevant to video-streaming services, where content providers pre-place data in clients' caches (or, on servers near the clients), with the goal of improving latency and rate performance in high demand periods. The content provider typically does not know in advance which specific movies the clients will request, and thus the cached data cannot depend on the clients' specific demands.

Let us call the pre-placement of data in caches the *caching phase*, and the remaining communications phase the *delivery phase*. Cache memories are typically much smaller than the library, and the caching phase occurs before the receivers demands are known. A key engineering challenge is, therefore, to carefully choose and cache only that data which is most useful during the delivery phase. That is, one should cache data that minimises the rate needed to complete the delivery-phase downloads for any feasible receiver demands.

Cache-aided communications systems have received significant attention in the information-theoretic literature in recent years, and those works most closely related to this paper are [1]–[11]. With the exception of [11], these works assume that the delivery phase takes place over a single rate-limited multicast noiseless channel (a bit-pipe) that connects the transmitter to every receiver. In practice, however, the communications medium is sometimes better modelled by a noisy broadcast channel (BC). This scenario is considered in [11], where the BC is essentially a set of parallel links with different qualities to the various receivers, which models a wireless fading BC.

Here we assume that the delivery phase takes place over a memoryless erasure BC. However, in contrast to [11], we assume that the caching phase takes place over error-free pipes. The motivation for this simplified assumption is that the caching phase typically occurs during periods of low network-congestion, where network resources are not a limiting factor.

Our main contribution in this paper is a joint cache-channel coding scheme for the described setup for general demands, and a characterization of the *capacity-memory* region when the receivers wish to learn the same message. Our results show that when the receivers experience different erasure probabilities (different channel qualities), then

- it is beneficial to employ unequal cache sizes at the receivers (larger cache memories at weaker receivers, and smaller cache memories at strong receivers); and
- joint cache-channel coding techniques can provide significant gains over separated cache and channel coding.

Allocating larger cache memories to the weaker receivers is quite natural because one then needs to communicate less data over noisier channels (see also [11]). Interestingly, there is an additional benefit to asymmetric caches that arises when joint cache-channel coding is used during the delivery phase. The basic idea is as follows: Consider a degraded BC communications scenario (such as the erasure BC) with separate cache and channel coding. Here a stronger receiver can decode all the data that is sent to a weaker receiver during the delivery phase. In fact, the strong receiver could decode even more data, but it is limited by the weaker receiver. Now suppose that part of the message intended for the stronger receiver is stored within the weaker receiver's cache: one can freely piggyback this part of the stronger receiver's message on the message intended for the weaker receiver. The weaker receiver is not penalised because it knows what data is being piggybacked on its desired message, and its channel decoder can still resolve its desired message. While, simultaneously, the stronger receiver has decoded something

about its desired message and therefore we have improved efficiently. Thus, thanks to the weaker receiver's cache and a simple joint cache-channel coding scheme, we can send additional data to stronger receivers without any extra cost, i.e., extra rate-constraints. This additional benefit of asymmetric cache memories was not observed in [11], because separate cache-channel coding was used for the delivery phase.

II. PROBLEM DEFINITION

A. Message library and feasible receiver demands

We have a transmitter, K receivers and a library with D messages $W_1, \ldots, W_D$. The d-th message in the library W_d is independent of all other messages and uniform on[1] $\{0, 1, \ldots, 2^{nR_d} - 1\}$, where $R_d \geq 0$ is its rate and n is the transmission blocklength. We represent a particular combination of receivers' demands by a tuple $\mathbf{d} = (d_1, \ldots, d_K) \in \{1, \ldots, D\}^K$. That is, $\mathbf{d}$ represents the situation where receiver 1 demands (i.e., requests and downloads) message W_{d_1}, receiver 2 demands W_{d_2}, and so on. Let

$$\mathcal{D} \subseteq \{1, \ldots, D\}^K.$$

denote the *feasible set* of all possible receiver demands. The feasible set $\mathcal{D}$ is known to the transmitter and receivers during the caching and delivery phases, but the specific demand tuple $\mathbf{d}$ chosen from $\mathcal{D}$ is only revealed for the delivery phase.

B. Caching phase

For each receiver $k \in \{1, \ldots, K\}$, the size of its cache is described by a nonnegative integer $\mathcal{M}_k$, see (1) below. The transmitter sends

$$\mathbb{Z}_k := g_k(W_1, \ldots, W_D),$$

to receiver k's cache, where $g_k : \prod_{d=1}^{D} \{0, 1, \ldots, 2^{nR_d} - 1\} \to \mathcal{Z}_k$ such that

$$\log |\mathcal{Z}_k| \leq 2^{n\mathcal{M}_k}. \tag{1}$$

The caching phase occurs during a low congestion period, and we assume that $\mathbb{Z}_k$ is reliably conveyed to receiver k's cache (for each $k \in \{1, \ldots, K\}$).

C. Erasure Broadcast Channel Model

The delivery phase occurs during a high congestion period, which we model by an *erasure BC* with input alphabet $\mathcal{X} := \{0, 1\}^F$. Here $F \geq 0$ is a fixed positive integer, and each $x \in \mathcal{X}$ is an F-bit packet. Due to congestion, some packets may be lost when, for example, router buffers overload. We denote the event of a lost packet with the *erasure symbol* Δ, and the BC's output alphabet by $\mathcal{Y} := \mathcal{X} \cup \{\Delta\}$ (the same alphabet is used for all receivers). Fix $1 \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_K \geq 0$. Let $Q(y_1, \ldots, y_K | x) := \mathbb{P}[(Y_1, \ldots, Y_K) = (y_1, \ldots, y_K) | X = x]$ be any probability law for the BC with marginals

$$\mathbb{P}[Y_k = y_k | X = x] = \begin{cases} 1 - \delta_k & \text{if } y_k = x \\ \delta_k & \text{if } y_k = \Delta \\ 0 & \text{otherwise} \end{cases}, \quad \forall\, k.$$

[1]To simplify notation and help elucidate our main ideas, we assume throughout the paper that 2^{nR_d} is an integer.

For our purpose only these marginal probabilities are relevant.

We discuss a caching system in the next section that is built on separate cache and channel codes, and, for this reason, it is useful to recall the degraded message set capacity region for Q. A channel-coding rate tuple $(R_{\{1,\ldots,K\}}, R_{\{2,\ldots,K\}}, \ldots, R_{\{K\}})$ is said to be achievable on Q if the following holds: For any $\epsilon > 0$ there exists an encoder and K-decoders such that, for all k, the transmitter can send $(R_{\{k,\ldots,K\}} - \epsilon)$ information bits per channel use to every receiver in the set $\{k, k+1, \ldots, K\}$ with an average probability of error less than ϵ. The set of all achievable rates — the *capacity region* $\mathcal{C}^\dagger$ — is given by the next proposition. The proposition can be distilled from [12], and we omit these details.

Proposition 1:

$$\mathcal{C}^\dagger = \left\{ \left(R_{\{1,\ldots,K\}}, R_{\{2,\ldots,K\}}, \ldots, R_{\{K\}} \right) : \right.$$
$$\left. \sum_{k=1}^{K} \frac{R_{\{k,\ldots,K\}}}{F(1 - \delta_k)} \leq 1, \ R_{\{k,\ldots,K\}} \geq 0, \ \forall\, k \right\}.$$

D. Delivery phase

For each feasible demand $\mathbf{d} \in \mathcal{D}$, let

$$f_{\mathbf{d}} : \prod_{d'=1}^{D} \{0, 1, \ldots, 2^{nR_{d'}} - 1\} \to \mathcal{X}^n$$

denote the corresponding encoder at the transmitter. Given $\mathbf{d} \in \mathcal{D}$ and the library $(W_1, \ldots, W_D)$, the transmitter sends

$$X^n := f_{\mathbf{d}}(W_1, \ldots, W_D),$$

where $X^n = (X_1, \ldots, X_n)$. Receiver k observes $Y_k^n = (Y_{k,1}, \ldots, Y_{k,n})$ according to the memoryless law Q. Let

$$\varphi_{k,\mathbf{d}} : \mathcal{Y}^n \times \mathcal{Z}_k \to \{0, 1, \ldots, 2^{nR_{d_k}} - 1\}$$

denote the decoder at receiver k. Given demands $\mathbf{d} \in \mathcal{D}$, cache content $\mathbb{Z}_k$ and channel outputs Y_k^n, receiver k outputs

$$\hat{W}_k := \varphi_{k,\mathbf{d}}(Y_k^n, \mathbb{Z}_k)$$

as its reconstruction of the d_k-th message W_{d_k}.

E. Achievable rate-memory tuples

Let

$$\mathsf{P_e} := \mathbb{P}\left[\bigcup_{\mathbf{d} \in \mathcal{D}} \bigcup_{k=1}^{K} \{\hat{W}_k \neq W_{d_k}\} \right]$$

denote the probability of error at any receiver for any feasible demand. We call the collection of all encoders and decoders,

$$\{g_1, g_2, \ldots, g_K\} \text{ and } \{f_{\mathbf{d}}, \varphi_{1,\mathbf{d}}, \varphi_{2,\mathbf{d}}, \ldots, \varphi_{K,\mathbf{d}}\}_{\mathbf{d} \in \mathcal{D}},$$

an $(n, R_1, \ldots, R_D, \mathcal{M}_1, \ldots, \mathcal{M}_K)$-code.

We say that a rate-memory tuple $(R_1, \ldots, R_D, \mathcal{M}_1, \ldots, \mathcal{M}_K)$ is *achievable* if for any $\epsilon > 0$ there exists a sufficiently large blocklength n and an $(n, R_1, \ldots, R_D, \mathcal{M}_1, \ldots, \mathcal{M}_K)$-code with $\mathsf{P_e} \leq \epsilon$.

Definition 1: We define the *capacity-memory* region $\mathcal{C}$ to be the closure of the set of all achievable rate-memory tuples.

The main problem of interest in this paper is to determine the capacity-memory region $\mathcal{C}$ for a given erasure BC Q.

III. MOTIVATING EXAMPLES

We now demonstrate the potential of unequal cache memories and joint cache-channel coding with three examples. Fix $K = 2$; $\mathcal{D} = \{1, \ldots, D\}^2$; $R_d = R$ for all d; and

$$\delta_1 = 4/5 \quad \text{and} \quad \delta_2 = 1/5. \tag{2}$$

A. Coded caching with symmetric caches

Suppose that $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}$, and

$$\alpha := \mathcal{M}/R \in [0, D/2]. \tag{3}$$

Split each message W_d in the library into three sub-messages,

$$W_d = \left(W_d^{(c1)}, W_d^{(c2)}, W_d^{(u)}\right),$$

of rates $\mathcal{M}/D$, $\mathcal{M}/D$, and $R - 2\mathcal{M}/D$.

Caching phase: Store the sub-messages

$$\left(W_1^{(c1)}, \ldots, W_D^{(c1)}\right) \quad \text{and} \quad \left(W_1^{(c2)}, \ldots, W_D^{(c2)}\right)$$

in the caches of receiver 1 and 2 respectively.

Delivery phase: The transmitter sends

$$W_{d_1}^{(c2)} \oplus W_{d_2}^{(c1)}, \tag{4}$$

as a common message to both receivers, where the addition is modulo $2^{n(\mathcal{M}/D)}$. It then sends $W_{d_1}^{(u)}$ as a private message to receiver 1 and $W_{d_2}^{(u)}$ as a private message to receiver 2. Notice that receiver 1 can recover W_{d_1} from the common message and $W_{d_1}^{(u)}$, while receiver 2 can recover W_{d_2} from the common message and $W_{d_2}^{(u)}$. We use a good channel code to communicate common message $W_{d_1}^{(c2)} \oplus W_{d_2}^{(c1)}$ and private messages $W_{d_1}^{(u)}$ and $W_{d_2}^{(u)}$ over the BC.

Achievable rate-memory tuples: Proposition 1 asserts that the common message (4) and $W_{d_1}^{(u)}$ can be decoded by both receivers and $W_{d_2}^{(u)}$ can be decoded by receiver 2 whenever

$$\frac{R - \frac{\mathcal{M}}{D}}{F(1 - \delta_1)} + \frac{R - \frac{2\mathcal{M}}{D}}{F(1 - \delta_2)} \leq 1. \tag{5}$$

On substituting (2), the inequality (5) simplifies to

$$R \leq \frac{4}{5}F(1 - \delta_1) + \frac{6}{5}\frac{\mathcal{M}}{D}. \tag{6}$$

All rate-memory tuples $(R, \ldots, R, \mathcal{M}, \ldots, \mathcal{M})$, with R and M satisfying (3) and (6), are achievable.

B. Separate cache-channel coding and asymmetric caches

Now suppose that we have asymmetric caches $\mathcal{M}_1 = 2\mathcal{M}$ and $\mathcal{M}_2 = 0$ for some $\mathcal{M}$ satisfying (3). The total cache memory available at both receivers remains unchanged, only now the memory at receiver 2 has been reallocated to receiver 1.

Split each message W_d into two sub-messages,

$$W_d = \left(W_d^{(c1)}, W_d^{(u)}\right) \tag{7}$$

with rates $2\mathcal{M}/D$ and $R - (2\mathcal{M}/D)$ respectively.

Caching phase: Store $\left(W_1^{(c1)}, \ldots, W_D^{(c1)}\right)$ in receiver 1's cache.

Delivery phase: We use a good channel code for Proposition 1 to reliably communicate the above sub-messages. The transmitter sends $W_{d_1}^{(u)}$ as a common message to both receivers (although it is only used by receiver 1), and it sends $W_{d_2}^{(c1)}$ and $W_{d_2}^{(u)}$ as a private message to receiver 2.

Achievable rate-memory tuples: Proposition 1 asserts that reliable communication is possible if

$$\frac{R - \frac{2\mathcal{M}}{D}}{F(1 - \delta_1)} + \frac{R}{F(1 - \delta_2)} \leq 1. \tag{8}$$

On substituting (2), the inequality (8) simplifies to

$$R \leq \frac{4}{5}F(1 - \delta_1) + \frac{8}{5}\frac{\mathcal{M}}{D}. \tag{9}$$

All rate-memory tuples $(R, \ldots, R, \mathcal{M}, \ldots, \mathcal{M})$, with R and $\mathcal{M}$ satisfying (3) and (9) are achievable.

C. Joint cache-channel coding and asymmetric caches

As in Section III-B: Let $\mathcal{M}_2 = 0$ and $\mathcal{M}_1 = 2\mathcal{M}$, for some $\mathcal{M}$ satisfying (3), and split each message W_d into two sub-messages (7) with rates $2\mathcal{M}/D$ and $R - (2\mathcal{M}/D)$ respectively.

Caching phase: Store $\left(W_1^{(c1)}, \ldots, W_D^{(c1)}\right)$ at receiver 1.

Delivery phase: Transmission takes place in two phases. Let $\beta_1 \in [0, 1]$. First phase of length $\beta_1 n$: The transmitter sends

$$\left(W_{d_1}^{(u)}, W_{d_2}^{(c1)}\right)$$

as a common message to both receivers using a joint cache-channel code. Second phase of length $(1 - \beta_1)n$: The transmitter sends $W_{d_2}^{(u)}$ to receiver 2 using a point-to-point channel code. Receiver 1 tries to decode $W_{d_1}^{(u)}$ and receiver 2 tries to decode $\left(W_{d_1}^{(u)}, W_{d_2}^{(c1)}, W_{d_2}^{(u)}\right)$. A key observation here is that $W_{d_2}^{(c1)}$ is stored in receiver 1's cache. As we see in a moment, for $\alpha \in [0, \frac{3D}{8}]$, this allows us to freely piggyback receiver 2's message $W_{d_2}^{(c1)}$ on receiver 1's message $W_{d_1}^{(u)}$ without compromising the rate to receiver 1.

Achievable rate-memory tuples: By Tuncel's seminal *broadcasting with side-information* result [13], communication in phase 1 (to both receivers) is very likely to be successful if the following two conditions hold:

$$R - \frac{2\mathcal{M}}{D} \leq F(1 - \delta_1)\beta_1 \tag{10a}$$

$$R \leq F(1 - \delta_2)\beta_1; \tag{10b}$$

communication in phase 2 is very likely to be successful if

$$R - \frac{2\mathcal{M}}{D} \leq F(1 - \delta_2)(1 - \beta_1). \tag{10c}$$

For $\mathcal{M}/R \in [0, 3D/8]$, constraint (10a) is tighter than (10b), and constraint (10b) is tighter when $\mathcal{M}/R \in \left(3D/8, D/2\right]$. Moreover, for $\mathcal{M}/R \in [0, 3D/8]$ the choice $\beta_1 = 4/5$ maximizes the set of rates R satisfying constraints (10), and for $\mathcal{M}/R \in (3D/8, D/2]$ the choice $\beta_1 = (1/2)(1 - \mathcal{M}/(DR))^{-1}$ maximizes the set of rates R satisfying constraints (10). Combining these observations with (10) proves achievability of all rate-memory tuples $(R, \ldots, R, \mathcal{M}, \ldots, \mathcal{M})$, where R and $\mathcal{M}$ satisfy (3) and

$$R \leq \begin{cases} \frac{4}{5}F(1 - \delta_1) + 2\frac{\mathcal{M}}{D}, & \text{if } \frac{\mathcal{M}}{R} \in \left[0, \frac{3D}{8}\right] \\ 2F(1 - \delta_1) + \frac{\mathcal{M}}{D} & \text{if } \frac{\mathcal{M}}{R} \in \left(\frac{3D}{8}, \frac{D}{2}\right]. \end{cases} \tag{11}$$

D. Discussion

Comparing the rate-memory tradeoffs in (6), (9) and (11), we see that it is advantageous to use unequal cache sizes and joint cache-channel coding. In particular, allowing larger caches at the weaker receivers (with higher packet erasure probabilities) both reduces the delivery-phase rates to the weaker receivers and increases rates to the stronger receivers.

IV. A JOINT CACHE-CHANNEL CODE FOR ARBITRARY DEMANDS

We now describe a joint cache-channel code. We consider any set of feasible demands $\mathcal{D}$, but equal message rates

$$R_d = R, \qquad \forall\, d \in \{1, \dots, D\}.$$

We first treat the case where the K_0 weakest receivers (receivers 1 to K_0) have equal cache sizes and the remaining receivers do not have caches:

$$\mathcal{M}_k = \begin{cases} \mathcal{M} & \text{if } k \le K_0 \\ 0 & \text{if } k > K_0 \end{cases}. \tag{12}$$

See Section IV-B for an extension to unequal cache sizes.

A. Scheme for cache sizes satisfying (12)

Preliminaries: Choose a positive integer $t < K_0$, and let

$$\tau := \binom{K_0}{t}.$$

Split each message W_d into $(\tau + 1)$-sub-messages,

$$W_d = \left(W_d^{(1)}, \dots, W_d^{(\tau+1)} \right),$$

where $W_d^{(i)} \in \left\{ 0, 1, \dots, 2^{nR^{(i)}} - 1 \right\}$ and

$$R^{(i)} := \begin{cases} \dfrac{\mathcal{M}}{D\binom{K_0-1}{t-1}}, & \text{for } i = 1, 2, \dots, \tau \\[2ex] R - \dfrac{\mathcal{M} K_0}{Dt}, & \text{for } i = \tau + 1. \end{cases}$$

Caching Phase: Consider the K_0 weakest receivers. Let

$$\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_\tau$$

denote the τ different subsets of $\{1, \dots, K_0\}$ with size t. For each $i = 1, 2, \dots, \tau$, take the tuple $(W_1^{(i)}, W_2^{(i)}, \dots, W_D^{(i)})$ and store it in the cache of each and every receiver in $\mathcal{R}_i$. Here we have stored $D\binom{K_0-1}{t-1}$ sub-messages in receiver k's cache (for each $k \in \{1, 2, \dots, K_0\}$) with a total memory requirement

$$\left(2^{n\frac{\mathcal{M}}{D}\binom{K_0-1}{t-1}^{-1}} \right)^{D\binom{K_0-1}{t-1}} = 2^{n\mathcal{M}}.$$

Delivery phase: The demand tuple $\mathbf{d} \in \mathcal{D}$ is given, and we are required to communicate message W_{d_1} to receiver 1, W_{d_2} to receiver 2, and so on.

We consider sets of $(t+1)$-receivers in $\{1, \dots, K_0\}$. Within these sets, each subset of t receivers shares a sub-message that is demanded (but unknown) by the remaining $(t+1)$-th receiver. For each set of $(t+1)$-receivers, we form the "XOR" of the $(t+1)$ sub-messages having the two above mentioned

properties, that is, being known at t of the receivers and demanded by the remaining $(t+1)$-th receiver. For example, for the subset of receivers $\{1, \dots, t+1\}$, we form

$$\bigoplus_{k=1}^{t+1} W_{d_k}^{(i_k)}, \tag{13}$$

where the addition is modulo $2^{n\mathcal{M}/(D\binom{K_0-1}{t-1})}$ (or, equivalently, a bitwise XOR operation); and for each $k \in \{1, \dots, t+1\}$, i_k is such that

$$\mathcal{R}_{i_k} \triangleq \{1, \dots, k-1, k+1, \dots, t+1\}. \tag{14}$$

Notice that (14) implies that $W_{d_k}^{(i_k)}$ is stored in the caches of receivers $1, \dots, k-1, k+1, \dots, t$, but not at receiver k.

We use a time-sharing scheme to send the XOR messages as well as all other messages to be transmitted. The time-sharing comprises K phases. Each phase $k \in \{1, \dots, K\}$ is constructed so that it can be decoded by Receivers $k, k+1, \dots, K$. Phase k is of length $\beta_k n$, where

$$\sum_{k=1}^{K} \beta_k = 1, \quad 0 \le \beta_k \le 1. \tag{15}$$

In phase $k \in \{1, \dots, K_0\}$, we send

- the XOR messages that are demanded by receiver k but not by receivers 1 to $k-1$;
- the uncached message $W_{d_k}^{(\tau+1)}$ demanded by Receiver k;
- the first $nC_{k,\tilde{k}}$ bits of sub-messages $W_{d_{\tilde{k}}}^{(i)}$, for every $\tilde{k} \in \{K_0+1, \dots, K\}$ and every $i \in \{1, \dots, \tau\}$ such that $k \in \mathcal{R}_i$. These messages are all known to receiver k and therefore do not limit the decoding at receiver k. The rates $\{C_{k,\tilde{k}}\}$ are parameters of a scheme. As we shall see, when they are chosen sufficiently small, but positive, and $\delta_{k+1} < \delta_k$, then sending these bits does not limit the decoding at receivers $k+1, k+2, \dots, K$. In fact, similarly to our motivating example, in this case, the transmitted bits of sub-messages $W_{d_{\tilde{k}}}^{(i)}$ can be freely piggybacked on the other messages transmitted in this phase k.

In phase $k \in \{K_0+1, \dots, K\}$ we send:

- the sub-messages of W_{d_k} that have not been sent in any previous phase.

Achievable Rate-Memory Tuples:

Proposition 2: A rate-memory tuple $(R, \dots, R, \mathcal{M}_1 = \mathcal{M}, \dots, \mathcal{M}_{K_0} = \mathcal{M}, 0, \dots, 0)$ is achievable if for some

- positive integer t;
- $(\beta_1, \dots, \beta_K)$ satisfying (15); and
- nonnegative real numbers $\{C_{k,\tilde{k}}\}$ with $k \in \{1, \dots, K_0\}$ and $\tilde{k} \in \{K_0+1, \dots, K\}$;

the following $(K + K_0)$-conditions in (16) hold.

1) For each $k \in \{1, \dots, K_0 - t - 1\}$, we have

$$R \le \beta_k F(1 - \delta_k) + \frac{\mathcal{M}}{D\binom{K_0-1}{t-1}} \left(\binom{K_0}{t} - \binom{K_0-k}{t} \right) \tag{16a}$$

and

$$R + \sum_{\tilde{k}=K_0+1}^{K} C_{k,\tilde{k}} \leq \beta_k F(1 - \delta_{k+1})$$
$$+ \frac{\mathcal{M}}{D\binom{K_0-1}{t-1}} \left(\binom{K_0}{t} - \binom{K_0 - k}{t} \right).$$
(16b)

2) For each $k \in \{K_0 - t, \ldots, K_0\}$, we have

$$R \leq \beta_k F(1 - \delta_k) + \frac{\mathcal{M}K_0}{Dt}$$
(16c)

and

$$R + \sum_{\tilde{k}=K_0+1}^{K} C_{k,\tilde{k}} \leq \beta_k F(1 - \delta_{k+1}) + \frac{\mathcal{M}K_0}{Dt}.$$
(16d)

3) Finally, for each $k \in \{K_0 + 1, \ldots, K\}$, we have

$$R - \sum_{k'=1}^{K_0} C_{k',k} \leq \beta_k F(1 - \delta_k).$$
(16e)

Proof outline: For each $k \in \{1, \ldots, K_0 - t\}$, Condition (16a) ensures that receiver k can reliably decode the submessages sent in phase k: $W_{d_k}^{(\tau+1)}$ of rate $R - \frac{\mathcal{M}K_0}{Dt} = R - \frac{\mathcal{M}}{D\binom{K_0-1}{t-1}}\binom{K_0}{t}$ and the $\binom{K-k}{t}$ x-or messages of the form (13) each having rate $\frac{\mathcal{M}}{D\binom{K_0-1}{t-1}}$. (Here, $\binom{K-k}{t}$ is the number of size-$(t + 1)$ subsets of $\{k, \ldots, K\}$ that include index k.) There is no constraint from the piggibacked sub-messages of rates $\{C_{k,\tilde{k}}\}_{\tilde{k}=K_0+1}^{K}$ because receiver k already knows them.

Condition (16b) ensures that all of the stronger receivers in $\{k + 1, \ldots, K\}$ can reliably decode all sub-messages sent in phase k, i.e., $W_{d_k}^{(\tau+1)}$, the x-or sub-messages, as well as the piggibacked sub-messages of rates $\{C_{k,\tilde{k}}\}_{\tilde{k}=K_0+1}^{K}$.

Similarly, Condition (16c) ensures that each receiver $k \in \{K_0 - t, \ldots, K_0\}$ can reliably decode the sub-messages sent in phase k, and Condition (16d) ensures that all of the stronger receivers in $\{k + 1, \ldots, K\}$ can also reliably decode these sub-messages. Finally, Condition (16e) ensures that each receiver $k \in \{K_0 + 1, \ldots, K\}$ can decode the sub-messages sent in phase k. ∎

Discussion: The parameters $\{C_{k,\tilde{k}}\}$ describe the gain that our scheme achieves over separate cache-channel coding schemes. If some of these rates are strictly larger than 0, then our scheme strictly outperforms separate cache-channel coding. It is possible to choose them strictly positive whenever the erasure probabilities $\delta_1, \ldots, \delta_{K_0+1}$ are not all equal.

B. Scheme for unequal cache sizes

Assume now that $\mathcal{M}_1 \geq \mathcal{M}_2 \geq \cdots \mathcal{M}_K \geq 0$. Our scheme in the previous subsection is easily extended to this more general setup using time-sharing. Specifically: Let $\beta_1, \ldots, \beta_K$ be real numbers in the interval $[0, 1]$ that sum up to 1. Over a fraction of time β_i, $i \in \{1, \ldots, K\}$, we use our scheme in the previous subsection assuming that only the first

$K_0^{(i)} = K + 1 - i$ receivers have caches of equal cache size $\mathcal{M}^{(i)} = \beta_i^{-1}(\mathcal{M}_{K-i+1} - \mathcal{M}_{K-i+2})$. (Set $\mathcal{M}_{K+1} \triangleq 0$.)

V. Single Common Demand

In this section we consider the optimistic case where all receivers demand the same message. This corresponds to

$$\mathcal{D} = \left\{ (d_1, \ldots, d_K) \in \{1, \ldots, D\}^K : d_1 = d_2 = \cdots = d_K \right\}.$$

The rates $R_1, \ldots, R_D$ can be arbitrary, i.e., do not have to be equal as in the previous section.

It can be shown that in this setup, a joint cache-channel code based on Tuncel's virtual binning technique [13] achieves the capacity-memory region.

Theorem 3: A rate-memory tuple $(R_1, \ldots, R_D, \mathcal{M}_1, \ldots, \mathcal{M}_K)$ is achievable if and only if,

$$R_d \leq \min_{k \in \{1, \ldots, K\}} \left((1 - \delta_k)F + \mathcal{M}_{k,d} \right), \quad d \in \{1, \ldots, D\}, \quad (17)$$

for some nonnegative numbers $\{\mathcal{M}_{k,d}\}$ that satisfy

$$\sum_{d=1}^{D} \mathcal{M}_{k,d} \leq \mathcal{M}_k, \quad k \in \{1, \ldots, K\}. \quad (18)$$

We thus again wish to allocate small cache sizes to strong receivers and large cache sizes to weak receivers.

If we used separate cache-channel codes, Constraint (17) is replaced by

$$\max_{k \in \{1, \ldots, K\}} (R_d - \mathcal{M}_{k,d}) \leq \min_{k \in \{1, \ldots, K\}} (1 - \delta_k)F, \quad (19)$$

and the benefit of having unequal cache sizes $\{\mathcal{M}_k\}$ at the different receivers disappears.

References

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

[2] M. A. Maddah-Ali, U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *to appear IEEE/ACM Trans. Net.*

[3] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE INFOCOM Workshop*, 2014.

[4] S. Wang, X. Tian and H. Liu, "Exploiting the unexploited of coded caching for wireless content distribution," in *Proc. IEEE ICNC, 2015.*

[5] R. Pedarsani, M. A. Maddah-Ali and U. Niesen, "Online coded caching," *to appear IEEE/ACM Trans. Networking.*

[6] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, S. Diggavi, "Hierarchical coded caching." arxiv.org 1403.7007.

[7] J. Hachem, N. Karamchandani, and S. Diggavi, "Content caching and delivery over heterogeneous wireless networks." arxiv.org 1404.6560.

[8] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching" arxiv.org 1501.06003.

[9] C.-Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: sequential coding for computing." arxiv.org 1504.00553v1.

[10] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands." arxiv.org 1502.03124.

[11] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel." arxiv.org 1504.01452.

[12] R. Urbanke and A. Wyner "Packetizing for the erasure broadcast channel with an internet application," 1999. lthcwww.epfl.ch/~ruediger/papers

[13] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1469–1482, 2006.