Noisy Broadcast Networks with Receiver Caching

Shirin Saeedi Bidokhti, Michèle Wigger and Roy Timo

Abstract

We study noisy broadcast networks with local cache memories at the receivers, where the transmitter can pre-store information even before learning the receivers' requests. We mostly focus on packet-erasure broadcast networks with two disjoint sets of receivers: a set of weak receivers with all-equal erasure probabilities and equal cache sizes and a set of strong receivers with all-equal erasure probabilities and no cache memories. We present lower and upper bounds on the *capacity-memory tradeoff* of this network. The lower bound is achieved by a new *joint cache-channel coding* idea and significantly improves on schemes that are based on separate cache-channel coding. We discuss how this coding idea could be extended to more general discrete memoryless broadcast channels and to unequal cache sizes. Our upper bound holds for all stochastically degraded broadcast channels.

For the described packet-erasure broadcast network, our lower and upper bounds are tight when there is a single weak receiver (and any number of strong receivers) and the cache memory size does not exceed a given threshold. When there are a single weak receiver, a single strong receiver, and two files, then we can strengthen our upper and lower bounds so as they coincide over a wide regime of cache sizes.

Finally, we completely characterise the rate-memory tradeoff for general discrete-memoryless broadcast channels with arbitrary cache memory sizes and arbitrary (asymmetric) rates when all receivers always demand exactly the same file.

I. Introduction

We address a one-to-many broadcast communications problem where many users demand files from a single server during *peak-traffic* times — periods of high network congestion. To improve network performance, the server can pre-place information in local cache memories near users at the network edge. This pre-placement of information is called the *caching communications phase*, and it occurs during off-peak times when the communications rate is not a limiting network resource. The server typically does not know in advance which files the users will demand, so it can try to cache information that is likely to be useful for many users during the *delivery communications phase* (the peak-traffic time when the users demand files from the server). For example, researchers at Huawei Laboratories [4] recently used machine learning techniques to predict user behavior and proactively cache data to improve user request satisfaction ratios and reduce backhaul loads during the delivery phase.

The above caching problem is particularly relevant to video-streaming services in mobile networks. Here network operators pre-place information in clients' caches (or, on servers near the clients) to improve latency and throughput during peak-traffic times. The network operator does not know in advance which movies the clients will request, and thus the cached information cannot depend on the clients' specific demands. It is now widely expected that there will be a nine-fold increase in mobile data traffic by 2020, and around 60 percent of this traffic will be mobile video [5]. Smart data caching strategies, new bandwidth allocations, reduced cell sizes and new radio-access technologies will all be needed to meet these growing demands [6].

The information-theoretic aspects of cache-aided communications have received significant attention in recent years [7]–[36]. Maddah-Ali and Niesen [7] considered a one-to-many broadcast problem where the receivers have independent caches of equal sizes and the delivery phase takes place over a noiseless broadcast communications link. They showed that a smart design of the cache contents enables the server to send coded (XOR-ed) data during the delivery phase that can simultaneously meet the demands of multiple receivers with a single transmission. This *coded caching scheme*, by simultaneously satisfying user demands, allows the server to reduce the delivery rate beyond the obvious *local caching gain* (the data rate that each receiver can immediately retrieve from its cache without using coded caching). Intuitively, this performance improvement occurs because the receivers can profit from other receivers' caches, and was thus termed [7] *global caching gain*.

Improved caching and delivery strategies for the Maddah-Ali and Niesen model were presented in [8]–[10]. Fundamental converse (lower) bounds on the total required delivery rate were presented in [7], [11]–[16]. It was shown in [17] that the coded caching scheme of Maddah-Ali and Niesen is optimal among schemes that have uncoded caching placement when there are more files than users.

The Maddah-Ali and Niesen model considers a *worst-case scenario*, meaning that the goal is to satisfy all possible user demands. The caching problem has also been studied in *average-case scenarios* [14]–[16] where the receivers' demands follow a given probability distribution and the delivery rate is averaged over this demand distribution.

S. Saeedi Bidokhti is with the Department of Electrical Engineering at Stanford University, saeedi@stanford.edu. S. Saeedi Bidokhti is supported by the Swiss National Science Foundation fellowship no. 158487.

M. Wigger is with CNRS, LTCI, Telecom ParisTech, Université Paris-Saclay, 75013 Paris, michele.wigger@telecom-paristech.fr

R. Timo is with the Institute for Communications Engineering at the Technical University of Munich, roy.timo@tum.de. R. Timo is supported by the Alexander von Humboldt Foundation.

Parts of the material in this paper have been presented at the *IEEE International Symposium on Wireless Communications Systems (ISWCS)*, Bruxelles, Belgium, August 2015 [1], and will be presented at the *2016 IEEE International Symposium on Information Theory*, Barcelona, Spain [2] and at the *International Symposium on Turbo Codes & Iterative Information Processing*, Brest, France, September 2016 [3].

In contrast to [7], we will assume in this paper that the delivery phase takes place over a noisy broadcast channel (BC), and we will see that further global caching gains can be achieved by *joint cache-channel coding*. Intuitively, when the BC is noisy the cache content not only determines *what* to transmit but also *how* to transmit it.

We will focus on packet erasure broadcast channels that provide a first order model[1] of packet losses in congested networks. The importance of including a noisy channel model for the delivery phase was also observed in [1], [18]–[28]. For example, [22] and [23] illustrate interesting interplay between feedback or channel state information with caching, and [24] and [25] show that caches at the transmitter-side and receiver-side allow for load-balancing and interference mitigation in noisy interference networks. The works in [22]–[25] focus on the high signal-to-noise ratio regime.

Our main interest in this paper is to characterize some of the fundamental *rate-memory tradeoffs* for cache-aided broadcast networks; that is, we wish to determine the set of rates at which messages can be reliably communicated for given cache sizes. We focus on a worst-case (worst-demand) setup and consider two different communication scenarios.

Scenario 1: We assume that the receivers' demands are arbitrary[2] and the messages are all of equal rate. We focus on the packet-erasure BC illustrated in Figure 1, and divide the K receivers into two sets:

- A set of K_{w} *weak receivers* with equal "large" BC erasure probabilities $\delta_{\mathrm{w}} > 0$. These receivers are each equipped with an individual cache of equal memory M.
- A set of $K_{\mathrm{s}} = K - K_{\mathrm{w}}$ *strong receivers* with equal "small" BC erasure probabilities $\delta_{\mathrm{s}} \geq 0$ with $\delta_{\mathrm{s}} \leq \delta_{\mathrm{w}}$. These receivers are not provided with caches.

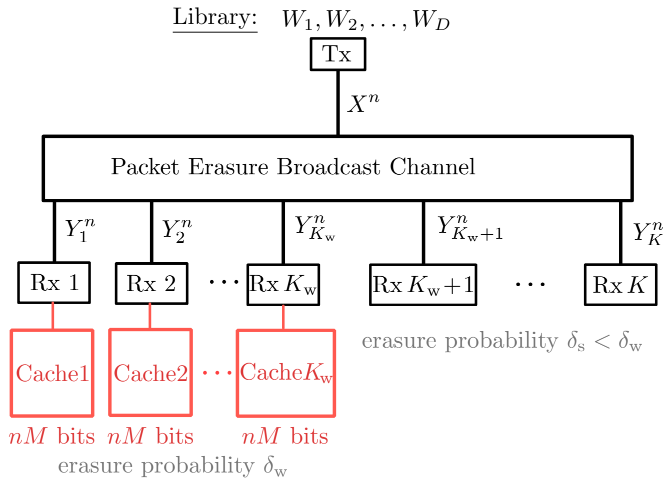

Fig. 1: K user packet-erasure BC with K_{w} weak and K_{s} strong receivers and where the weak receivers have cache memories.

This scenario is motivated by previous studies [1], [19] that showed the benefit of prioritizing cache placements near weaker receivers. In practical systems, this means that telecommunications operators with a limited number of caches might first place caches at houses that are further away from an optical fiber access point. Or, they might place caches at pico or femto base stations in heterogenous networks that are located in areas with notoriously bad coverage. Scenario 1 also arises as part of a more complex system model in which every receiver is equipped with a cache. Suppose, for example, that the stronger receivers want to decode additional data that will never be demanded by the weak receivers (see also Section VI). This additional data might represent, for example, a higher resolution of a video. A practical solution in this case is to separate transmission of files from the two libraries [32]–[34]: A first transmission sends the files that are of interest to all receivers, and a second transmission sends only files from the additional library to the strong receivers. The question is now how to divide the cache memory between the two transmissions. Based on the results we obtain in this paper, we propose to assign all the cache memory at the strong receivers to the second transmission, because through a careful design of the first transmission scheme, the strong receivers can already benefit from the weak receivers' caches without accessing their own cache memories.

The fundamental rate versus cache memory tradeoff of interest in Scenario 1 is the largest rate R at which all messages can be reliably transmitted (in the usual Shannon sense) for *a given* cache size M at the weak receivers. This largest rate is called the *capacity-memory tradeoff* and will be denoted by $C(M)$.

We present general achievable (lower) and converse (upper) bounds on the capacity-memory tradeoff $C(M)$. Our achievable bound on $C(M)$ is based on a joint cache-channel coding scheme that builds on the 'piggyback' coding idea in [39] (see Subsection III-D). The basic idea of piggyback coding is to carry messages to strong receivers on the back of messages to the

[1]Here we can assume that bit-level errors within a packet are handled on a link-by-link basis using physical layer error-correction techniques, and packets arrive at the users promptly or are lost due to, for example, buffer overflows.

[2]Each user can choose any file from the server.

weak receivers. These messages can be carried for "free" if the server pre-places appropriate "message side information" in the weak receivers' caches. We will see that joint cache-channel coding provides substantial gains over separate cache-channel coding with Maddah-Ali and Niesen's coded caching scheme [7] and a capacity achieving scheme for the packet-erasure BC. For example, if the library has D messages of rate R and M is smaller than approximately $\frac{DR(\delta_{\text{w}}-\delta_{\text{s}})}{K_{\text{w}}+K_{\text{s}}}$ (i.e., a 'small-memory' regime), then

$$C(M) \geq R_0 + \frac{M}{D} \cdot \gamma_{\text{local}} \cdot \gamma_{\text{global,sep}} \cdot \gamma_{\text{global,joint}}. \tag{1}$$

Here R_0 represents the capacity of the network without caches, and $\gamma_{\text{local}} \leq 1$, $\gamma_{\text{global,sep}} \geq 1$ and $\gamma_{\text{global,joint}} \geq 1$ are constants (depending on the number of receivers and erasure probabilities). The exact expressions of the three constants are presented later in (34), but they have the following interpretations. If we employ a standard capacity-achieving coding scheme to transmit the parts of the demanded messages that are not already stored in the intended receivers' caches, then we achieve a rate equal to $R_0 + \frac{M}{D} \cdot \gamma_{\text{local}}$. This strategy achieves only a local caching gain, and hence the subscript "local" for this factor. When some receivers in the network have no caches ($K_{\text{s}} \geq 1$), then $\gamma_{\text{local}} < 1$. A scheme that combines Maddah-Ali and Niesen's coded caching algorithm with a capacity-achieving scheme for the packet-erasure BC attains the lower bound $R_0 + \frac{M}{D} \cdot \gamma_{\text{local}} \cdot \gamma_{\text{global,sep}}$. The factor $\gamma_{\text{global,sep}}$ thus describes the global caching gain obtained by a separate cache-channel coding scheme, and hence the subscript "global,sep". Whenever $K_{\text{w}} > 1$, we have $\gamma_{\text{global,sep}} > 1$. Finally, our joint cache-channel coding scheme achieves the lower bound (1), and the parameter $\gamma_{\text{global,joint}}$ describes this scheme's gain over the previous separate cache-channel coding. In other words, the factor

$$\gamma_{\text{global,joint}} = 1 + \frac{2K_{\text{w}}}{1+K_{\text{w}}} \cdot \frac{K_{\text{s}}(1-\delta_{\text{w}})}{K_{\text{w}}(1-\delta_{\text{s}})} \tag{2}$$

describes the further global caching gain that is possible using our joint cache-channel coding scheme (that was not achievable with the aforementioned separate cache-channel coding scheme). By (2), the improvement of our joint cache-channel coding scheme over the separate cache-channel coding scheme is not bounded for small cache sizes. In particular, it is strictly increasing in the number of strong receivers K_{s}.

Our general lower and upper bounds match for

$$K_{\text{w}} = 1 \qquad \text{and} \qquad M \leq FD\frac{(1-\delta_{\text{s}})(\delta_{\text{w}}-\delta_{\text{s}})}{K_{\text{s}}(1-\delta_{\text{w}})+(1-\delta_{\text{s}})}. \tag{3a}$$

For the special case $K_{\text{w}} = K_{\text{w}} = 1$ and $D = 2$, we present a refined lower bound on $C(M)$ as well as a refined upper bound. The idea is to cache also the XOR of a part of the two messages in the library, similarly to [7, Appendix]. Our refined bounds coincide when

$$K_{\text{w}} = K_{\text{s}} = 1; \qquad D = 2; \qquad \delta_{\text{w}} = \delta_{\text{s}} \tag{3b}$$

and

$$K_{\text{w}} = K_{\text{s}} = 1; \qquad D = 2; \qquad M \geq F((1-\delta_{\text{s}}) + (\delta_{\text{w}}-\delta_{\text{s}})). \tag{3c}$$

Scenario 2: In our second scenario (section VIII) we allow for general discrete memoryless broadcast channels (DMBCs), arbitrary cache sizes $\{M_k\}_{k=1}^K$, and non-equal rates of the various messages $R_1, \ldots, R_D$. However, we impose that each receiver demands exactly the same message. For this scenario we completely characterize the entire rate-memorry tradeoff $(R_, \ldots, R_D, M_1, \ldots, M_k)$.

The remainder of this paper is organised as follows. In Sections II and III, we state the problem setup and some auxiliary results that are helpful in the design of our joint cache-channel coding scheme. Section IV summarizes our main results for the first scenario. We describe and analyze our joint source-cache channel coding scheme in Section V and sketch how our scheme can be extended to more general scenarios with arbitrary cache sizes and arbitrary DMBCs in Section VI. In Section VII, we prove an upper bound on the capacity-memory tradeoff of general (stochastically) degraded BCs with caches at the receivers. Our second scenario is discussed in section VIII.

II. Problem Definition

A. Notation

Random variables are identified by uppercase letters, e.g. A, their alphabets by matching calligraphic font, e.g. $\mathcal{A}$, and elements of an alphabet by lowercase letters, e.g. $a \in \mathcal{A}$. We also use uppercase letters for deterministic quantities like rate R, capacity C, number of users K, memory size M, and number of files in the library D.

The Cartesian product of $\mathcal{A}$ and $\mathcal{A}'$ is $\mathcal{A} \times \mathcal{A}'$, and the n-fold Cartesian product of $\mathcal{A}$ is $\mathcal{A}^n$. Vectors are identified by bold font symbols, e.g., $\mathbf{a}$, and matrices by the font A. We use the shorthand notation A^n for the tuple $(A_1, \ldots, A_n)$. LHS and RHS stand for left-hand side and right-hand side.

Finally, we use the notation $W_1 \bigoplus W_2$ to denote the bitwise XOR over the binary strings corresponding to the messages W_1 and W_2, which are assumed to be of equal length.

B. Message and channel models

Consider a broadcast channel (BC) with a single transmitter and K receivers as depicted in Figure 1. We have two sets of receivers: K_{w} weak receivers that statistically have a bad channel and $K_{\mathrm{s}} = K - K_{\mathrm{w}}$ strong receivers that statistically have a good channel. (The meaning of good and bad channels will be explained shortly.) For convenience of notation, we assume that the first K_{w} receivers are weak and the subsequent K_{s} receivers are strong, and we define the sets

$$\mathcal{K}_{\mathrm{w}} := \{1, \ldots, K_{\mathrm{w}}\}$$

and

$$\mathcal{K}_{\mathrm{s}} := \{K_{\mathrm{w}} + 1, \ldots, K\}.$$

We model the channel from the transmitter to the receivers by a memoryles *packet-erasure BC* with input alphabet

$$\mathcal{X} := \{0, 1\}^F$$

and equal output alphabet at all receivers

$$\mathcal{Y} := \mathcal{X} \cup \{\Delta\}.$$

Here $F \geq 0$ is a fixed positive integer, and each input symbol $x \in \mathcal{X}$ is an F-bit packet. The output erasure symbol Δ models loss of a packet at a given receiver. Each receiver $k \in \mathcal{K} := \{1, \ldots, K\}$ observes the erasure symbol Δ with a given probability $\delta_k \geq 0$, and it observes an output y_k equal to the input, $y_k = x$, with probability $1 - \delta_k$. The marginal transition laws[3] of the memoryless BC are thus described by

$$\mathbb{P}[Y_k = y_k | X = x] = \begin{cases} 1 - \delta_k & \text{if } y_k = x \\ \delta_k & \text{if } y_k = \Delta \\ 0 & \text{otherwise} \end{cases} \qquad \forall\ k. \tag{4}$$

We will assume throughout that

$$\delta_i = \begin{cases} \delta_{\mathrm{w}} & \text{if } i \in \mathcal{K}_{\mathrm{w}} \\ \delta_{\mathrm{s}} & \text{if } i \in \mathcal{K}_{\mathrm{s}} \end{cases} \tag{5}$$

for fixed erasure probabilities[4] $0 < \delta_{\mathrm{s}} \leq \delta_w \leq 1$. Since $\delta_{\mathrm{s}} \leq \delta_w$, the weak receivers have statistically worse channels than the strong receivers, hence the distinction between good and bad channels. In the sequel, we will assume that each weak receiver is provided with a cache memory of size nM bits. The strong receivers are not provided with cache memories. We explain shortly how the cache memory at the weak receivers can be exploited.

C. Message library and receiver demands

The transmitter has access to a library with $D \geq K$ messages

$$W_1, \ldots, W_D. \tag{6}$$

These messages are all independent of each other and each of them is uniformly distributed over the message set $\{1, \ldots, \lfloor 2^{nR} \rfloor\}$, where $R \geq 0$ is the rate of each message and n the blocklength of transmission.

Each receiver will demand (i.e., request and download) exactly one of these messages. Let

$$\mathcal{D} := \{1, \ldots, D\}.$$

We denote the demand of receiver 1 by $d_1 \in \mathcal{D}$, the demand of receiver 2 by $d_2 \in \mathcal{D}$, etc., to indicate that receiver 1 desires message W_{d_1}, receiver 2 desires message W_{d_2}, and so on. For most of the time in this manuscript we assume that the demand vector

$$\mathbf{d} := (d_1, \ldots, d_K) \tag{7}$$

can take on any value in $\mathcal{D}^K$.

Communication takes place in two phases: a first *caching phase* where information is stored in the weak receivers' cache memories and a subsequent *delivery phase* where the demanded messages are delivered to all the receivers. The next two subsections detail these two communication phases.

[3]As will become clear in the following, for our problem setup only this marginal transition law is relevant, but not the joint transition law.

[4]Though, in principle, we allow $\delta_s = \delta_{\mathrm{w}}$, our main interest will be $\delta_{\mathrm{s}} < \delta_{\mathrm{w}}$.

D. Caching phase

During the first communication phase the transmitter sends caching information V_i to each weak receiver $i \in \mathcal{K}_\text{w}$, who then stores this information in its cache memory. The strong receivers do not take part in the caching phase.

The demand vector $\mathbf{d}$ is unknown to the transmitter and receivers during the caching phase, and, therefore, the cached information V_i cannot depend on the users' specific demands $\mathbf{d}$. Instead, V_i is a function of the entire library:

$$V_i := g_i(W_1, \ldots, W_D) \qquad i \in \mathcal{K}_\text{w}$$

for some function

$$g_i \colon \left\{1, \ldots, \lfloor 2^{nR} \rfloor\right\}^D \to \mathcal{V}, \qquad i \in \mathcal{K}_\text{w}, \tag{8}$$

where

$$\mathcal{V} := \{1, \ldots, \lfloor 2^{nM} \rfloor\}.$$

The caching phase occurs during a low-congestion period. We therefore assume that this phase incurs no erasures or other types of errors, and each weak receiver $i \in \mathcal{K}_\text{w}$ can store V_i in its cache memory.

E. Delivery phase

The transmitter is provided with the demand vector $\mathbf{d}$, and it communicates the corresponding messages $W_{d_1}, \ldots, W_{d_K}$ over the packet-erasure BC. The entire demand vector $\mathbf{d}$ is assumed to be known to the transmitter and all receivers[5].

Depending on the demand vector $\mathbf{d}$, the transmitter chooses an encoding function

$$f_\mathbf{d} \colon \{1, \ldots, \lfloor 2^{nR} \rfloor\}^D \to \mathcal{X}^n \tag{9}$$

and it sends

$$X^n = f_\mathbf{d}(W_1, \ldots, W_D), \tag{10}$$

over the packet-erasure BC.

Each receiver $k \in \{1, \ldots, K\}$ observes Y_k^n according to the memoryless transition law in (4). Each weak receiver attempts to reconstruct its desired message from it channel output, cache contents and demand vector $\mathbf{d}$. Similarly, each strong receiver attempts to reconstructs its desired message from its channel output and the demand vector $\mathbf{d}$. More formally,

$$\hat{W}_i := \begin{cases} \varphi_{i,\mathbf{d}}(Y_i^n, V_i) & \text{if } i \in \mathcal{K}_\text{w} \\ \varphi_{i,\mathbf{d}}(Y_j^n) & \text{if } i \in \mathcal{K}_\text{s} \end{cases} \tag{11a}$$

where

$$\varphi_{i,\mathbf{d}} \colon \mathcal{Y}^n \times \mathcal{V} \to \{1, \ldots, \lfloor 2^{nR} \rfloor\} \qquad i \in \mathcal{K}_\text{w} \tag{11b}$$

and

$$\varphi_{j,\mathbf{d}} \colon \mathcal{Y}^n \to \{1, \ldots, \lfloor 2^{nR} \rfloor\} \qquad j \in \mathcal{K}_\text{s}. \tag{11c}$$

F. Capacity-memory tradeoff

An error is said to occur whenever

$$\hat{W}_k \neq W_{d_k} \quad \text{for some} \quad k \in \{1, \ldots, K\}. \tag{12}$$

For a given demand vector $\mathbf{d}$ the probability of error is thus

$$\mathsf{P}_\text{e}(\mathbf{d}) := \mathbb{P}\left[\bigcup_{k=1}^{K} \hat{W}_k \neq W_{d_k} \right]. \tag{13}$$

We consider a worst-case probability of error over all feasible demand vectors:

$$\mathsf{P}_\text{e}^\text{worst} := \max_{\mathbf{d} \in \mathcal{D}^K} \mathsf{P}_\text{e}(\mathbf{d}). \tag{14}$$

In definitions (8)-(14), we sometimes add a superscript (n) to emphasise the dependency on the blocklength n.

We say that a rate-memory pair (R, M) is *achievable* if for every $\epsilon > 0$ there exists a sufficiently large blocklength n and caching, encoding and decoding functions as in (8), (9) and (11) such that $\mathsf{P}_\text{e}^\text{worst} < \epsilon$. The main problem of interest in this paper is to determine the following capacity versus cache memory tradeoff.

Definition 1: Given cache memory size M, we define the *capacity-memory tradeoff* $C(M)$ as the supremum of all rates R such that the rate-memory pair (R, M) is achievable.

[5]It takes only $\lceil \log(D) \rceil$ bits to describe the demand vector $\mathbf{d}$. The demand vector can thus be revealed to all terminals using zero transmission rate.

G. Trivial and non-trivial memory sizes

The capacity-memory tradeoff $C(M)$ is trivially upper bounded by the capacity of the packet-erasure BC to the strong receivers only (see Proposition 1 in Section III-A ahead):

$$C(M) \leq F\left(\frac{1-\delta_{\rm s}}{K_{\rm s}}\right) \qquad \forall\ M \geq 0. \tag{15}$$

For $M = DF(1-\delta_{\rm s})/K_{\rm s}$, this upper bound is also achievable because the weak receivers can store the entire library in their caches and the transmitter thus needs to only serve the strong receivers during the delivery phase. Therefore,

$$C(M) = F\left(\frac{1-\delta_{\rm s}}{K_{\rm s}}\right), \qquad \forall\ M \geq DF\left(\frac{1-\delta_{\rm s}}{K_{\rm s}}\right). \tag{16}$$

We will henceforth restrict attention to nontrivial cache memories

$$M \in \left[0,\ \ DF\frac{(1-\delta_{\rm s})}{K_{\rm s}}\right].$$

III. Previous Related Results

This section reviews capacity results and coding schemes for three related scenarios that form the basis for our new bounds on the capacity-memory tradeoff $C(M)$ and the joint cache-channel coding scheme that we present in Section V ahead.

A. Capacity of packet-erasure BCs

Temporarily consider the K receiver packet-erasure BC illustrated in Figure 2. The BC is characterised by (4) with arbitrary erasure probabilities $0 < \delta_1, \delta_2, \ldots, \delta_K \leq 1$. Suppose that the are no caches (i.e., $M = 0$), and that each receiver k wishes to learn an independent message W_k that is uniformly distributed over the set $\{1, \ldots, \lfloor 2^{nR_k} \rfloor\}$. Notice that, in contrast to previous sections, messages have different rates, and it is a priori known which message is intended for which receiver. Let $\hat{W}_k$ denote receiver k's reconstruction of message W_k.

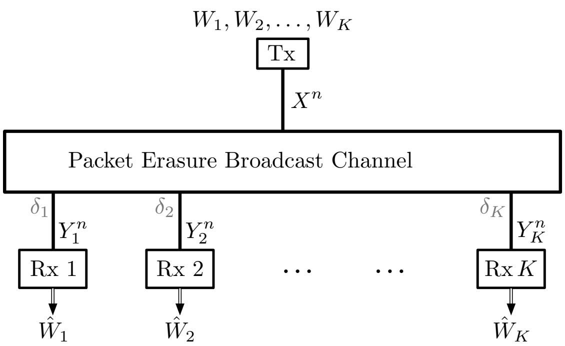

Fig. 2: Standard K-user packet-erasure BC with arbitrary erasure probabilities and no caches.

The capacity region of this standard packet erasure-BC is achieved by time-sharing capacity-achieving point-to-point codes [37]. The point-to-point capacity of the channel from the transmitter to receiver $k \in \{1, \ldots, K\}$ is

$$C_k = (1-\delta_k)F. \tag{17}$$

Proposition 1: The capacity region of the packet-erasure BC to K receivers with erasure probabilities $\delta_1, \delta_2, \ldots, \delta_K$ is the closure of the set of nonnegative rate-tuples $(R_1, \ldots, R_K)$ that satisfy [37]

$$\sum_{k=1}^{K} \frac{R_k}{(1-\delta_k)F} \leq 1.$$

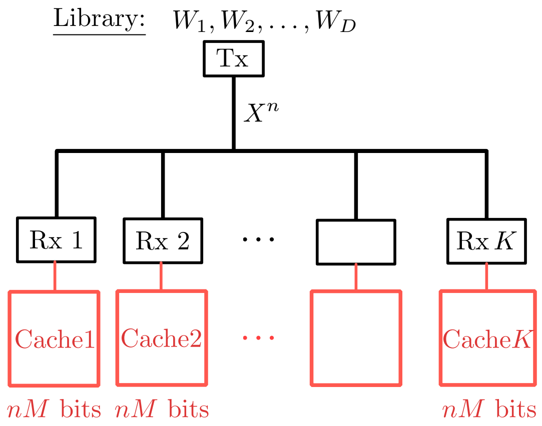

Fig. 3: BC with a common noise-free pipe of rate F to all $\tilde{K}$ receivers, which all have a cache memory of nM bits.

B. Coded caching over a BC with noise-free common bit-pipe

We briefly explain the Maddah-Ali and Niesen's *coded caching scheme* in [7]. The scheme has two parameters:

- a positive integer $\tilde{K}$ (representing the number of users with caches) and
- an index $\tilde{t} \in \{1, \ldots, \tilde{K}\}$.

Coded caching applies to the noiseless communication scenario illustrated in Figure 3. This scenario coincides with our original scenario in Section II in the special case where

$$K_{\mathrm{w}} = K = \tilde{K} \qquad \text{and} \qquad \delta_{\mathrm{w}} = 0.$$

As in (6), the messages $(W_1, W_2, \ldots, W_D)$ depicted in Figure 3 have a common rate R.

It will be convenient to describe coded caching using the following methods:

- *Method* `Ca` describing the caching phase for the setup in Figure 3.
- *Method* `En` describing the delivery-phase encoding.
- *Methods* $\{\mathtt{De}_k;\ k = 1, 2, \ldots, \tilde{K}\}$ describing the delivery-phase decoding at each user.

1) Preliminaries: Let

$$\mathcal{G}_1, \ldots, \mathcal{G}_{\binom{\tilde{K}}{\tilde{t}}}$$

denote the $\binom{\tilde{K}}{\tilde{t}}$ subsets of $\{1, \ldots, \tilde{K}\}$ of size $\tilde{t}$. Split each message W_d into $\tilde{K}$ choose $\tilde{t}$ independent submessages,

$$W_d = \left\{ W_{d,\mathcal{G}_\ell} \colon\ \ell = 1, \ldots, \binom{\tilde{K}}{\tilde{t}} \right\}.$$

Each of these submessages is of equal rate

$$R_{\text{sub}} := R \binom{\tilde{K}}{\tilde{t}}^{-1}. \tag{18}$$

2) Method `Ca`*:* This method takes the entire library $W_1, \ldots, W_D$ as an input, and it outputs the cache contents $V_1, V_2, \ldots, V_{\tilde{K}}$ where

$$V_k = \left\{ W_{d,\mathcal{G}_\ell} \colon\ d \in \{1, \ldots, D\} \text{ and } k \in \mathcal{G}_\ell \right\}, \quad k \in \{1, \ldots, \tilde{K}\}. \tag{19}$$

In other words, during the caching phase, the tuple

$$\left(W_{1,\mathcal{G}_\ell},\ W_{2,\mathcal{G}_\ell},\ \ldots,\ W_{D,\mathcal{G}_\ell} \right)$$

is stored in the cache memory of every receiver in $\mathcal{G}_\ell$.

3) Method `En`*:* This method takes the entire library $W_1, \ldots, W_D$ and the demand vector $\mathbf{d}$ as inputs, and it outputs

$$\left\{ W_{\text{XOR},\mathcal{S}} \colon\quad \mathcal{S} \subseteq \{1, \ldots, \tilde{K}\} \text{ and } |\mathcal{S}| = \tilde{t} + 1 \right\}, \tag{20}$$

where

$$W_{\text{XOR},\mathcal{S}} := \bigoplus_{s \in \mathcal{S}} W_{d_s, \mathcal{S} \setminus \{s\}}. \tag{21}$$

4) Methods $\mathtt{De}_k$ *(for* $k = 1, \ldots, \tilde{K}$*):* This method takes as inputs the demand vector $\mathbf{d}$; the XOR-messages $\{W_{\mathcal{S}} : k \in \mathcal{S}\}$ produced by method $\mathtt{En}$; and the cache content V_k produced by method $\mathtt{Ca}$. It outputs the $\binom{\tilde{K}}{\tilde{t}}$-tuple reconstruction

$$\hat{W}_{d_k} := \left(\hat{W}_{d_k,\mathcal{G}_1}, \ldots, \hat{W}_{d_k,\mathcal{G}_{\binom{\tilde{K}}{\tilde{t}}}}\right), \tag{22}$$

where

$$\hat{W}_{d_k,\mathcal{G}_\ell} = \begin{cases} W_{d_k,\mathcal{G}_\ell} & \text{if } k \in \mathcal{G}_\ell \\ \left(\bigoplus_{s\in\mathcal{G}_\ell} W_{d_s,\mathcal{G}_\ell\cup\{k\}\setminus\{s\}}\right) \bigoplus W_{\text{XOR},\mathcal{G}_\ell\cup\{k\}} & \text{if } k \notin \mathcal{G}_\ell. \end{cases} \tag{23}$$

Notice that all the (XOR) messages on the right are inputs of this method, because they are either part of the cache content V_k produced by method $\mathtt{Ca}$ or part of the XOR messages $\{W_{\text{XOR},\mathcal{S}} : k \in \mathcal{S}\}$ produced by method $\mathtt{En}$.

5) Analysis: We now analyse the three methods above for Figure 3.

Lemma 2: Consider the scenario in Figure 3. The XOR messages $\{W_{\text{XOR},\mathcal{S}}\}$ produced by method $\mathtt{En}$ can be sent over the common noise-free pipe if and only if the rate of the pipe satisfies

$$F \geq \binom{\tilde{K}}{\tilde{t}+1} R_{\text{sub}} = R\,\frac{\tilde{K}-\tilde{t}}{\tilde{t}+1}. \tag{24}$$

Moreover, each receiver $k \in \{1, \ldots, \tilde{K}\}$ can store the cache content V_k if, and only if, the cache memory size satisfies

$$M \geq D\binom{\tilde{K}-1}{\tilde{t}-1} R_{\text{sub}} = D\frac{\tilde{t}}{\tilde{K}}R. \tag{25}$$

Proof: Inequality (24) follows because there are $\binom{\tilde{K}}{\tilde{t}+1}$ XOR messages $W_{\mathcal{S}}$ of rate R_{sub} and by (18). Inequality (25) follows because each V_k produced by method $\mathtt{Ca}$, consists of $D\binom{\tilde{K}-1}{\tilde{t}-1}$ messages of rate R_{sub}, see (19). ■

C. Separate cache and channel coding for packet-erasure BCs (and the proof of Proposition 6)

Starting from Maddah-Ali and Niesen's coded-caching scheme one readily obtains a separation-based coding scheme for the packet-erasure BC with caches described in Section II. Details are as follows.

Choose $\tilde{K} = K_{\text{w}}$ and an arbitrary $\tilde{t} \in \{1, \ldots, K_{\text{w}}\}$. For the caching phase: Apply Method $\mathtt{Ca}$ to the library $W_1, \ldots, W_D$; take the resulting $V_1, \ldots, V_{\text{w}}$; and store each V_k the cache memories of receiver k. The delivery-phase transmitter proceeds in two steps:

T1: The transmitter applies method $\mathtt{En}$ to the library $W_1, W_2, \ldots, W_D$ and demand vector $\mathbf{d}_{\text{w}} := (d_1, d_2, \ldots, d_{K_{\text{w}}})$.

T2: The transmitter sends the XOR-messages produced in step T1 together with the messages that are demanded by receivers in $\mathcal{K}_{\text{s}}$ using a capacity-achieving scheme for the packet-erasure BC.

The strong receivers decode their intended messages using an optimal decoding method for the packet-erasure BC. The weak receivers decode in two steps:

R1: Each weak receiver uses an optimal decoder for the packet-erasure BC to recover all XOR-messages generated by method $\mathtt{En}$.

R2: It applies method $\mathtt{De}$ to the XOR messages decoded in step R1.

This separate cache and channel coding scheme can be easily analysed using Lemma 2 and Proposition 1, and from this analysis one obtains Proposition 6 in Section IV.

D. "Piggyback" coding for BCs with message side information

Consider the two-user BC with message side-information [40], [41] illustrated in Figure 4.

The transmitter observes two independent messages W_1 and W_2 of rates R_1 and R_2. Message W_1 is intended for receiver 1 and message W_2 for receiver 2. Suppose that receiver 1 is given W_2 prior to communications and the BC has arbitrary transition probabilities $P_{Y_1Y_2|X}$. The capacity region of this BC with message side-information at receiver 1 can be derived[6] from [41, Thm. 3].

We now present a specific random coding scheme for the BC in Figure 4, which we call *piggyback coding*.

1) Code Construction: Fix a distribution P_X on the input alphabet of the BC, a small $\epsilon > 0$ and a large blocklength n. Randomly generate a codebook $\mathcal{C}$ with $\lfloor 2^{nR_1} \rfloor \times \lfloor 2^{nR_2} \rfloor$ codewords of length n by independently picking each entry of each codeword using P_X. We place the codewords into a matrix with $\lfloor 2^{nR_1} \rfloor$ columns and $\lfloor 2^{nR_2} \rfloor$ rows, and denote the codeword in column w_1 and row w_2 by $x^n(w_1, w_2)$. Figure 5 sketches the codebook: Each dot represents a codeword; message W_1 determines the column of the codeword to pick and W_2 determines the row. The codebook is given to the transmitter and both receivers.

[6]Kramer and Shamai assume in [41, Thm. 3] that receiver 1 not only needs to decode message W_1 but also W_2. However, since receiver 1 has message W_2 as side information, this additional requirement is not a limitation and the two setups have identical capacity regions.

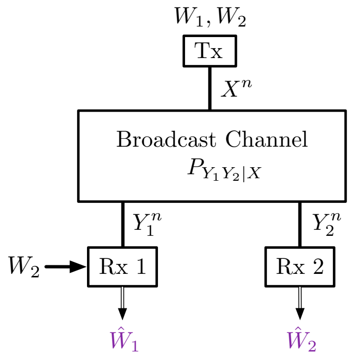

Fig. 4: Standard two-user BC with message side-information at receiver 2.

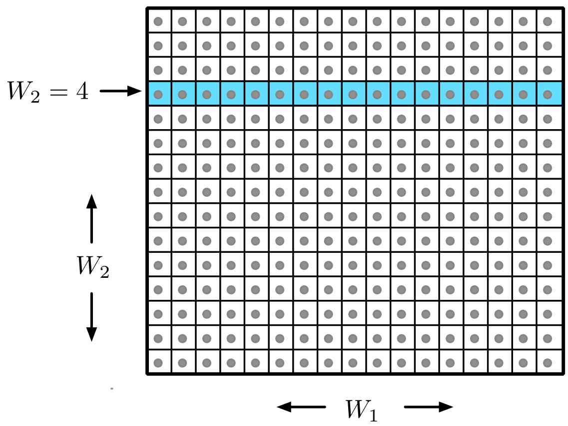

Fig. 5: Codebook $\mathcal{C}$ where the dots represent the codewords $x^n(w_1, w_2)$ arranged in a matrix array. Columns encode message W_1 and rows encode message W_2. Receiver 1 knows the value of message W_2 and thus can restrict its decoding to a single row of the codebook.

2) Encoding: Given that the transmitter wishes to send messages $W_1 = w_1$ and $W_2 = w_2$, it transmits the codeword $x^n(w_1, w_2)$ over the channel.

3) Decoding at receiver 1 (the receiver with side information): Since receiver 1 knows $W_2 = w_2$, it can restrict attention to the w_2-th row of the codeword. For example if $W_2 = 4$, then it needs only to consider the codewords (dots) that lie in the highlighted row of Figure 5.

Given that receiver 1 observes the channel outputs $Y_1^n = y_1^n$, it looks for a unique index $\hat{w}_1 \in \{1, \dots, \lfloor 2^{nR_1} \rfloor\}$ satisfying

$$(x^n(\hat{w}_1, w_2), y_1^n) \in \mathcal{T}_\epsilon^n(P_{XY_1}),$$

where $\mathcal{T}_\epsilon^n(P_{XY_1})$ denotes the typical set as defined in [42]. If there is no such index $\hat{w}_1$, then receiver 1 declares an error.

4) Decoding at receiver 2 (the receiver without side information): Receiver 2 will attempt to decode both messages W_1 and W_2. Given that it observes channel outputs $Y_2^n = y_2^n$, it looks for the unique pair of indices $(\hat{w}_1, \hat{w}_2) \in \{0, \dots, 2^{nR_1} - 1\} \times \{0, \dots, 2^{nR_2} - 1\}$ satisfying

$$(x^n(\hat{w}_1, \hat{w}_2), y_2^n) \in \mathcal{T}_\epsilon^n(P_{XY_2}).$$

If there is no such pair of indices, then receiver 2 declares an error.

5) Analysis & discussion: By the *covering lemma* [42] and because receiver 1 restricts attention to a single row in the codebook, the probability of decoding error at receiver 1, $\mathbb{P}[\hat{W}_1 \neq W_1]$, tends to 0 as $n \to \infty$ whenever

$$R_1 < I(X; Y_1). \tag{26a}$$

Moreover, by this covering lemma and because receiver 2 decodes both messages, the probability of error at receiver 2, $\mathbb{P}[\hat{W}_2 \neq W_2]$, tends to 0 as $n \to \infty$ whenever

$$R_1 + R_2 < I(X; Y_2). \tag{26b}$$

We have the following proposition.

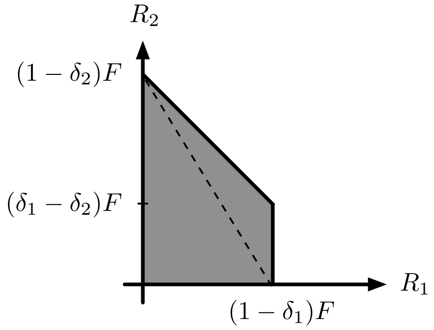

Fig. 6: The gray area depicts the rate-region achieved with piggyback coding over a two-user packet-erasure BC with message side-information at receiver 1 when receiver 2 has smaller erasure probability than receiver 1, i.e., $\delta_2 < \delta_1$. The dashed line indicates the border of the capacity region without message side-information. The figure shows that the message side-information allows to *piggyback* information to receiver 2 without reducing the achievable rate to receiver 1.

Proposition 3: For a DMBC $P_{Y_1Y_2|X}$ with message side-information at receiver 1 shown in Figure 4, piggyback coding achieves all nonnegative rate-pairs (R_1, R_2) satisfying (26) for some channel input distribution P_X.

Specialising this proposition to the packet-erasure BC, we obtain the following.

Proposition 4: For the two-user F bit packet-erasure BC, with erasure probabilities δ_1, δ_2 and message side information at receiver 1, piggyback coding achieves all rate-pairs (R_1, R_2) that satisfy

$$\max\left\{\frac{R_1}{(1-\delta_1)F}, \frac{R_1+R_2}{(1-\delta_2)F}\right\} \leq 1. \tag{27}$$

The achievable region in Proposition 4 is illustrated in Figure 6. This region covers the entire capacity region of the two-user F-bit packet-erasure BC with message side-information at receiver 1 whenever $\delta_2 < \delta_1$.

Remark 1: The piggyback coding scheme does not achieve the entire capacity region of an arbitrary DMBC with message side-information at receiver 1. Consider, for example, a degenerate DMBC in which the channel to receiver 2 is useless (i.e., $P_{Y_2|X}$ has 0 point-to-point capacity) and receiver 1's channel $P_{Y_1|X}$ has positive capacity. Over such a channel the piggyback coding achieves no positive rates because (26b) constrains the sum-rate to 0. It is clear that with a different scheme positive rates $R_1 > 0$ can be achieved. The optimal coding scheme splits message W_1 into $(W_{1,p}, W_{1,c})$ and combines piggyback coding with superposition coding: in the cloud center it sends $(W_{1,c}, W_2)$ as in piggyback coding, and in the satellite it sends $W_{1,p}$.

IV. New Results for Arbitrary Demands

In this section we assume that the demand vector

$$\mathbf{d} \text{ can take on every value in } \mathcal{D}^K.$$

Our main results are a general lower bound and a general upper bound on the capacity-memory tradeoff $C(M)$. The bounds are tight in certain regimes of M when $K_\mathrm{w} = 1$. We further present improved lower and upper bounds for the special case $K_\mathrm{s} = K_\mathrm{w} = 1$ and $D = 2$. These bounds match except for a small regime of M's.

A. General lower bounds

Define $K_\mathrm{w} + 2$ rate-memory pairs $\{(R_t, M_t);\ t = 0, 1, \ldots, K_\mathrm{w}+1\}$ as follows:

(i)

$$R_0 := F\left(\frac{K_\mathrm{w}}{1-\delta_\mathrm{w}} + \frac{K_\mathrm{s}}{1-\delta_\mathrm{s}}\right)^{-1}, \qquad M_0 := 0; \tag{28}$$

(ii) For each $t \in \{1, \ldots, K_\mathrm{w}\}$:

$$\begin{aligned} R_t &:= \frac{F(1-\delta_\mathrm{w})\left(1 + \dfrac{K_\mathrm{w}-t+1}{tK_\mathrm{s}}\ \dfrac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}\right)}{\dfrac{K_\mathrm{w}-t+1}{t}\left(1 + \dfrac{K_\mathrm{w}-t}{(t+1)K_\mathrm{s}}\ \dfrac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}\right) + K_\mathrm{s}\dfrac{1-\delta_w}{1-\delta_\mathrm{s}}}, \\ M_t &:= R_t\ \frac{D}{K_\mathrm{w}}\left(t - \left(1 + \frac{K_\mathrm{w}-t+1}{tK_\mathrm{s}}\ \frac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}\right)^{-1}\right); \end{aligned} \tag{29}$$

(iii)

$$R_{K_\text{w}+1} := F\frac{1-\delta_\text{s}}{K_\text{s}}, \qquad M_{K_\text{w}+1} := DF\ \frac{1-\delta_\text{s}}{K_\text{s}}. \tag{30}$$

Theorem 5 (Lower bound): The upper convex hull of the $K_\text{w}+2$ rate-memory pairs $\{(R_t, M_t);\ t = 0, 1, \ldots, K_\text{w}+1\}$ in (28)–(30) forms a lower bound on the capacity-memory tradeoff:

$$C(M) \geq \text{upper hull}\big\{(R_t, M_t):\ t = 0, \ldots, K_\text{w}+1\big\}. \tag{31}$$

Proof outline: The pair $(R_0,\ M_0 = 0)$ corresponds to the case without caches, and the achievability of R_0 follows from the usual capacity result for packet-erasure BCs, see Proposition 1 in the previous section III-A. Achievability of the pair $(R_{K_\text{w}+1},\ M_{K_\text{w}+1})$ follows from (16).

The remaining pairs (R_t, M_t), $t = 1, \ldots, K_\text{w}$, are more interesting and are achieved by the joint cache-channel coding scheme in section V ahead. The upper convex hull of $\{(R_t, M_t); t = 0, 1, \ldots, K_\text{w}+1\}$, finally, is achieved by time-sharing. ■

To better illustrate the strength of our lower bound, and thus of our joint cache-channel coding scheme in section V, consider the following separation-based lower bound:

Proposition 6: The upper convex hull of the rate-memory pairs $\big\{(R_{t,\text{sep}}, M_{t,\text{sep}}); t = \{0, \ldots, K_\text{w}\}\big\}$ is achievable, where

$$R_{t,\text{sep}} := F\left(\frac{K_\text{w}-t}{(t+1)(1-\delta_\text{w})} + \frac{K_\text{s}}{(1-\delta_\text{s})}\right)^{-1}, \tag{32a}$$

$$M_{t,\text{sep}} := D\frac{t}{K_\text{w}}R_{t,\text{sep}}. \tag{32b}$$

Proof: For $t = 0$, no cache memory is used and the achievability of $R_{0,\text{sep}}$ follows simply by the standard capacity region of the packet-erasure BC, Proposition 1 in section III-A. For $t \in \{1, \ldots, K_\text{w}\}$ achievability of the pair $(R_{t,\text{sep}}, M_{t,\text{sep}})$ can be shown by trivially combining the Maddah-Ali & Niesen coded caching [7, Algorithm 1] with a capacity-achieving scheme for the packet-erasure BC without caching, see the previous section III-C. ■

Of special interest is the regime of small cache size M. In this regime, Theorem 5 specializes as follows:

Corollary 6.1: For small cache memory sizes, $M \leq M_1$, the capacity-memory tradeoff is lower bounded as

$$C(M) \geq R_0 + \frac{M}{D}\cdot\gamma_\text{local}\cdot\gamma_\text{global,sep}\cdot\gamma_\text{global,joint}, \qquad M \leq M_1, \tag{33}$$

where R_0 is defined in (28) and

$$\gamma_\text{local} := \frac{K_\text{w}(1-\delta_\text{s})}{K_\text{w}(1-\delta_\text{s}) + K_\text{s}(1-\delta_\text{w})}, \tag{34a}$$

$$\gamma_\text{global,sep} := \frac{1+K_\text{w}}{2}, \tag{34b}$$

$$\gamma_\text{global,joint} := 1 + \frac{2K_\text{w}}{1+K_\text{w}}\cdot\frac{K_\text{s}(1-\delta_\text{w})}{K_\text{w}(1-\delta_\text{s})}. \tag{34c}$$

If in the above lower bound one replaces the product $\gamma_\text{global,sep}\cdot\gamma_\text{global,joint}$ by 1, then one obtains the lower bound that corresponds to a coding scheme with only local caching gain. If only the factor γ_joint is replaced by 1, then one obtains the lower bound implied by Proposition 6. The factor $\gamma_\text{global,sep}$ is thus due to the separation-based Maddah-Ali & Niesen coded caching idea. In contrast, the last factor $\gamma_\text{global,joint}$ is due to our joint cache-channel coding scheme. Notice that this factor $\gamma_\text{global,joint}$ is unbounded when one increases the number of strong receivers K_s or more generally the ratio $\frac{K_\text{s}(1-\delta_\text{w})}{K_\text{w}(1-\delta_\text{s})}$.

B. General upper bound

We now present our upper bound. Define for each $k_\text{w} \in \{0, \ldots, K_\text{w}\}$

$$R_{k_\text{w}}(M) := F\left(\frac{k_\text{w}}{1-\delta_\text{w}} + \frac{K_\text{s}}{1-\delta_\text{s}}\right)^{-1} + \frac{k_\text{w}M}{D}.$$

Theorem 7 (Upper bound): The capacity-memory tradeoff $C(M)$ is upper bounded as

$$C(M) \leq \min_{k_\text{w}\in\{0,\ldots,K_\text{w}\}} R_{k_\text{w}}(M). \tag{35}$$

Proof: In section VII we derive an upper bound on the capacity-memory tradeoff for a general degraded BC with arbitrary cache sizes at the receivers, see Theorem 9. We then specialize this upper bound to packet-erasure BCs with arbitrary cache sizes and erasure probabilities in Corollary 9.1, and we show how the upper bound in (35) is obtained from this corollary. ■

We numerically compare our upper and lower bounds on the capacity-memory tradeoff in Figures 7 and 8.

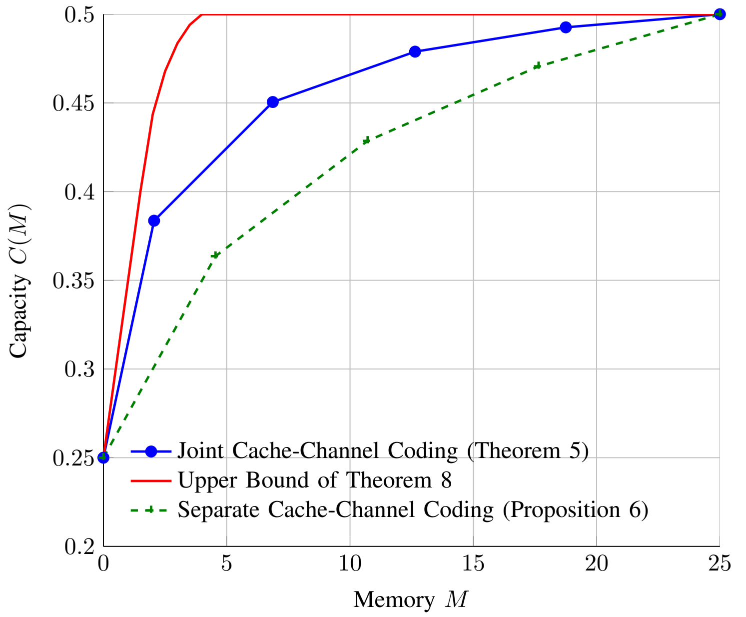

Fig. 7: Bounds on the capacity-memory tradeoff $C(M)$ for $K_\mathrm{w} = 4$, $K_\mathrm{s} = 16$, $D = 50$, $\delta_\mathrm{w} = 0.8$, $\delta_\mathrm{s} = 0.2$, and $F = 10$.

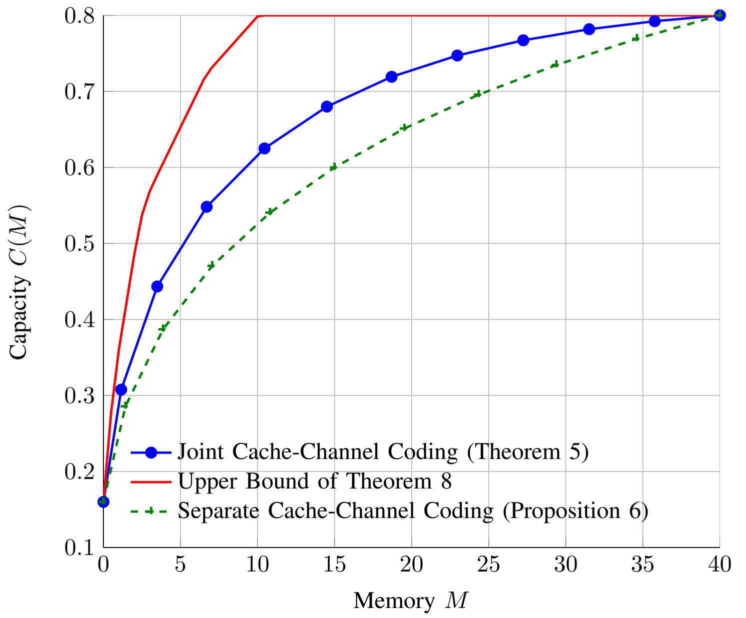

Fig. 8: Bounds on capacity-memory tradeoff $C(M)$ for $K_\mathrm{w} = K_\mathrm{s} = 10$, $D = 50$, $\delta_\mathrm{w} = 0.8$, $\delta_\mathrm{s} = 0.2$, and $F = 50$.

C. *Special case of* $K_\mathrm{w} = 1$

We first evaluate our bounds for a setup with a single weak receiver and any number of strong receivers. Let

$$\Gamma_1 := F\frac{(1-\delta_\mathrm{s})}{K_\mathrm{s}}\frac{(\delta_w - \delta_\mathrm{s})}{(K_\mathrm{s}(1-\delta_w) + (1-\delta_\mathrm{s}))}, \tag{36}$$

$$\Gamma_2 := \frac{(1-\delta_\mathrm{s})}{K_\mathrm{s}}F, \tag{37}$$

$$\Gamma_3 := F\frac{(1-\delta_\mathrm{s})}{K_\mathrm{s}}\frac{(1-\delta_\mathrm{s})}{(K_\mathrm{s}(1-\delta_w) + (1-\delta_\mathrm{s}))}. \tag{38}$$

Notice that $0 \leq \Gamma_1 \leq \Gamma_3 \leq \Gamma_2$. From Theorems 5 and 7 we obtain the following corollary.

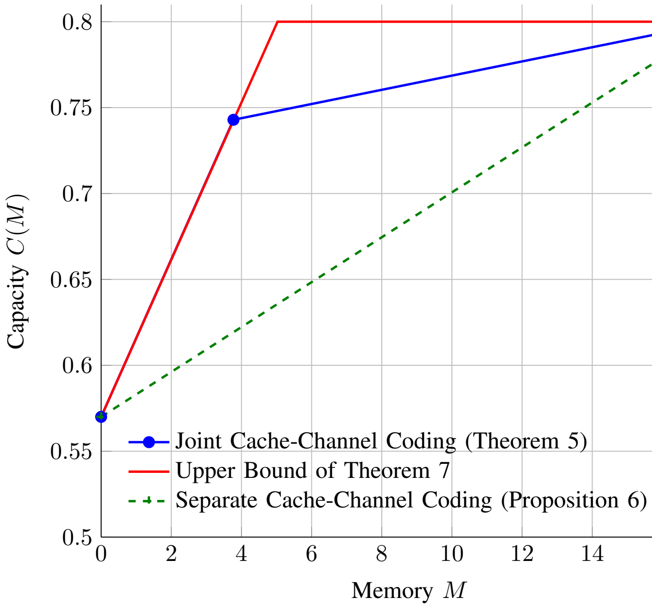

Fig. 9: Bounds on the capacity-memory tradeoff for $K_{\mathrm{w}}=1$, $K_{\mathrm{s}}=10$, $D=22$, $\delta_{\mathrm{w}}=0.8$, $\delta_{\mathrm{s}}=0.2$, and $F=10$.

Corollary 7.1: If $K_{\mathrm{w}}=1$ the capacity-memory tradeoff is lower bounded by:

$$C(M) \geq \begin{cases} F\frac{(1-\delta_{\mathrm{w}})(1-\delta_{\mathrm{s}})}{K_{\mathrm{s}}(1-\delta_{\mathrm{w}})+(1-\delta_{\mathrm{s}})}+\frac{M}{D}, & \text{if } \frac{M}{D}\in[0,\Gamma_1] \\ F\frac{(1-\delta_{\mathrm{s}})}{1+K_{\mathrm{s}}}+\frac{M}{(1+K_{\mathrm{s}})D}, & \text{if } \frac{M}{D}\in(\Gamma_1,\Gamma_2], \end{cases} \tag{39}$$

and upper bounded by:

$$C(M) \leq \begin{cases} F\frac{(1-\delta_{\mathrm{w}})(1-\delta_{\mathrm{s}})}{K_{\mathrm{s}}(1-\delta_{\mathrm{w}})+(1-\delta_{\mathrm{s}})}+\frac{M}{D}, & \text{if } \frac{M}{D}\in[0,\Gamma_3] \\ F\frac{(1-\delta_{\mathrm{s}})}{K_{\mathrm{s}}} & \text{if } \frac{M}{D}\in(\Gamma_3,\Gamma_2]. \end{cases} \tag{40}$$

Figure 9 shows these two bounds and the bound in Proposition 6 for $K_{\mathrm{w}}=1$, $K_{\mathrm{s}}=10$, $D=22$, $D=10$, $\delta_{\mathrm{w}}=0.8$, $\delta_{\mathrm{s}}=0.2$, and $F=10$.

We identify two regimes. In the first regime $0\leq\frac{M}{D}\leq\Gamma_1$, the cache memory allows reducing the rate R to each receiver by $\frac{M}{D}$. This is the same performance as when a naive uncoded caching strategy is used in a setup where *all* $K_{\mathrm{s}}+1$ *receivers* have cache memories of rate M. In the first regime, our joint cache-channel coding scheme thus enables all receivers to profit from the single cache memory and provides the best possible global caching gain. In the second regime $\Gamma_1<\frac{M}{D}\leq\Gamma_2$ the gains are not as significant as in the first regime, but increasing the cache size still results in an improved performance. This won't be the case for $\frac{M}{D}>\Gamma_2$.

In the first regime, $0\leq\frac{M}{D}\leq\Gamma_1$, our joint cache-channel coding scheme of section V achieves the capacity-memory tradeoff $C(M)$.

D. Special case $K_{\mathrm{w}}=K_{\mathrm{s}}=1$ and $D=2$

For this special case we present tighter upper and lower bounds on $C(M)$. These new bounds meet for a large range of cache memory sizes M. Let

$$\tilde{\Gamma}_1 := F\frac{(1-\delta_{\mathrm{w}})^2+(1-\delta_{\mathrm{s}})^2-(1-\delta_{\mathrm{w}})(1-\delta_{\mathrm{s}})}{(1-\delta_{\mathrm{w}})+(1-\delta_{\mathrm{s}})}, \tag{41}$$

$$\tilde{\Gamma}_2 := \frac{1}{2}F\left((1-\delta_{\mathrm{s}})+(\delta_{\mathrm{w}}-\delta_{\mathrm{s}})\right). \tag{42}$$

Theorem 8: If $K_{\mathrm{w}}=K_{\mathrm{s}}=1$ and $D=2$, the capacity-memory tradeoff is upper bounded as:

$$C(M) \leq \begin{cases} F\frac{(1-\delta_{\mathrm{w}})(1-\delta_{\mathrm{s}})}{(1-\delta_{\mathrm{w}})+(1-\delta_{\mathrm{s}})}+\frac{M}{2}, & \text{if } \frac{M}{2}\in[0,\tilde{\Gamma}_1] \\ F\frac{1}{3}(2-\delta_{\mathrm{s}}-\delta_{\mathrm{w}})+\frac{M}{3}, & \text{if } \frac{M}{2}\in(\tilde{\Gamma}_1,\tilde{\Gamma}_2] \\ F(1-\delta_{\mathrm{s}}) & \text{if } \frac{M}{2}\in(\tilde{\Gamma}_2,\Gamma_2]. \end{cases} \tag{43}$$

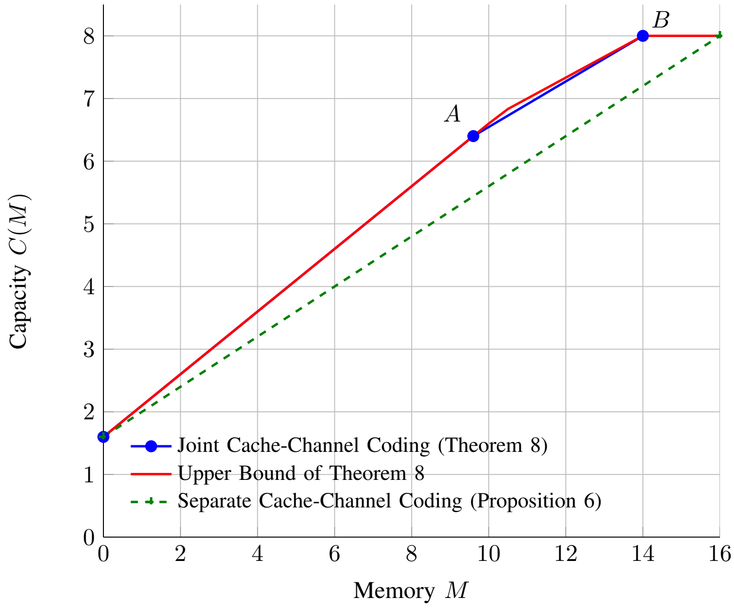

Fig. 10: Bounds on the capacity-memory tradeoff for $K_{\mathrm{w}}=1$, $K_{\mathrm{s}}=1$, $D=2$, $\delta_{\mathrm{w}}=0.8$, $\delta_{\mathrm{s}}=0.2$, and $F=10$.

and lower bounded as:

$$C(M) \geq \begin{cases} F\dfrac{(1-\delta_{\mathrm{w}})(1-\delta_{\mathrm{s}})}{(1-\delta_{\mathrm{w}})+(1-\delta_{\mathrm{s}})}+\dfrac{M}{2}, & \dfrac{M}{2}\in[0,\Gamma_1] \\ \dfrac{(1-\delta_s)}{3(1-\delta_{\mathrm{s}})-(1-\delta_{\mathrm{w}})}\left(F(1-\delta_{\mathrm{s}})+M\right), & \dfrac{M}{2}\in(\Gamma_1,\tilde{\Gamma}_2] \\ F(1-\delta_{\mathrm{s}}) & \dfrac{M}{2}\in(\tilde{\Gamma}_2,\Gamma_2]. \end{cases} \tag{44}$$

Proof: Lower bound (44) coincides with the upper convex hull of the three rate-memory pairs: $(R_0,\ M_0)$ in (28); $(R_1,\ M_1)$ in (29); and $(F(1-\delta_{\mathrm{s}}),\ 2\tilde{\Gamma}_2)$. Achievability of the former two pairs follows from Theorem 5. Achievability of the last pair is proved in appendix D. The upper bound is proved in appendix E. ■

Figure 10 shows the bounds of Theorem 8 for $\delta_{\mathrm{w}}=0.8$, $\delta_{\mathrm{s}}=0.2$, and $F=10$.

Corollary 8.1: The minimum cache size M for which communication is possible at the maximum rate $F(1-\delta_{\mathrm{s}})$ is $2\tilde{\Gamma}_2$.

Upper and lower bounds of Theorem 8 coincide in the case of equal erasure probabilities $\delta_{\mathrm{w}}=\delta_{\mathrm{s}}$:

Corollary 8.2: If $K_{\mathrm{w}}=K_{\mathrm{s}}=1$, $D=2$ and $\delta_{\mathrm{w}}=\delta_{\mathrm{s}}=\delta$:

$$C(M)=\begin{cases} F\frac{1}{2}(1-\delta)+\frac{M}{2}, & \text{if } \frac{M}{2}\in\left[0,\ \frac{1}{2}F(1-\delta)\right] \\ F(1-\delta) & \text{if } \frac{M}{2}\in\left(\frac{1}{2}F(1-\delta),\ \Gamma_2\right]. \end{cases} \tag{45}$$

Proof: It follows from Theorem 8 because for $\delta_{\mathrm{w}}=\delta_{\mathrm{s}}$: $\tilde{\Gamma}_1=\tilde{\Gamma}_2=\frac{1}{2}F(1-\delta)$, and in the regime $\frac{M}{2}\in(\Gamma_1,\tilde{\Gamma}_2]$ lower bound (44) specialises to $C(M)\geq F\frac{1}{2}(1-\delta)+\frac{M}{2}$. ■

V. A Joint Cache-Channel Scheme for Arbitrary Demands

We describe a joint cache-channel coding scheme parameterised by

$$t\in\{1,\ldots,K_{\mathrm{w}}\}. \tag{46}$$

We show in subsection V-D that, for parameter t, this scheme achieves the rate-memory pair (R_t, M_t) in (29).

A. Preparations

For each $d\in\{1,\ldots,D\}$, split message W_d into two parts:

$$W_d=\left(W_d^{(t-1)},\ W_d^{(t)}\right) \tag{47}$$

of rates

$$R^{(t-1)}=R\left(1+\frac{K_{\mathrm{w}}-t+1}{tK_{\mathrm{s}}}\cdot\frac{\delta_{\mathrm{w}}-\delta_{\mathrm{s}}}{1-\delta_{\mathrm{w}}}\right)^{-1} \tag{48a}$$

$$R^{(t)} = R\left(1 + \frac{tK_{\text{s}}}{K_{\text{w}} - t + 1} \cdot \frac{1 - \delta_{\text{w}}}{\delta_{\text{w}} - \delta_{\text{s}}}\right)^{-1} \tag{48b}$$

where $R^{(t-1)} + R^{(t)} = R$.

B. Caching phase

The first step is to cache information about both parts of each message in (47) at the weak receivers. Specifically, we first apply Method Ca with $\tilde{K} = K_{\text{w}}$ and $\tilde{t} = t$ to messages $W_1^{(t)}, \dots, W_D^{(t)}$. We then apply Method Ca with $\tilde{K} = K_{\text{w}}$ and $\tilde{t} = t-1$ to messages $W_1^{(t-1)}, \dots, W_D^{(t-1)}$.

In the following, we will use the superscript (t) to identify the outputs of Methods Ca, En, and De$_i$ with $\tilde{t} = t$. Similarly, we will use the superscript $(t-1)$ to identify these outputs for $\tilde{t} = t-1$. We use the notation that we introduced in Section III-B.

Consider an arbitrary weak receiver $i \in \mathcal{K}_{\text{w}}$. The total cache content at this receiver is

$$\begin{aligned} V_i &= V_i^{(t)} \cup V_i^{(t-1)} \\ &= \left\{ W^{(t)}_{d, \mathcal{G}_\ell^{(t)}} :\ d \in \{1, \dots, D\} \text{ and } k \in \mathcal{G}_\ell^{(t)} \right\} \\ &\quad \bigcup \left\{ W^{(t-1)}_{d, \mathcal{G}_\ell^{(t-1)}} :\ d \in \{1, \dots, D\} \text{ and } k \in \mathcal{G}_\ell^{(t-1)} \right\}. \end{aligned} \tag{49}$$

C. Delivery phase

The delivery phase takes place in three subphases consisting of $\beta_1 n$, $\beta_2 n$, and $\beta_3 n$ channel uses, where $\beta_1, \beta_2, \beta_3 \geq 0$ and

$$\beta_1 + \beta_2 + \beta_3 = 1. \tag{50}$$

1) Delivery subphase 1: Here we only consider the weak receivers, and we communicate the "t parts" of their demanded messages (see (47)) using separate source and channel coding. (The strong receivers will not participate in this subphase.)

The transmitter proceeds in two steps:

T1: The transmitter applies Method En with $\tilde{K} = K_{\text{w}}$ and $\tilde{t} = t$ to demand vector $(d_1, \dots, d_{K_{\text{w}}})$ and to messages

$$\left\{ W_{d_i}^{(t)} :\ i \in \mathcal{K}_{\text{w}} \right\}.$$

Let

$$\left\{ W^{(t)}_{\text{XOR}, \mathcal{S}} : \mathcal{S} \subseteq \{1, \dots, K_{\text{w}}\},\ |\mathcal{S}| = t+1 \right\} \tag{51}$$

denote the output of Method En.

T2: The transmitter uses a capacity-achieving code for the packet-erasure BC to send the XORs in (51) to the weak receivers.

Each weak receiver $i \in \mathcal{K}_{\text{w}}$ decodes in two steps:

R1: Receiver i recovers all transmitted XOR messages in (51) using an appropriate channel decoder.

R2: Receiver i applies method De$_i$ with $\tilde{K} = K_{\text{w}}$ and $\tilde{t} = t$ to demand vector $(d_1, \dots, d_{K_{\text{w}}})$ and to the XOR messages produced in step R1. For $i \in \mathcal{K}_{\text{w}}$, let

$$\hat{W}_{d_i}^{(t)} = \left(\hat{W}_{d_i, \mathcal{G}_1^{(t)}}, \hat{W}_{d_i, \mathcal{G}_2^{(t)}}, \dots, \hat{W}_{d_i, \mathcal{G}^{(t)}_{\binom{K_{\text{w}}}{t}}} \right) \tag{52}$$

denote the output produced by De$_i$.

2) Delivery subphase 2: Here we consider all receivers. To the strong receivers: We communicate the "t parts" of their demanded messages. To the weak receivers: We communicate the "$(t-1)$ parts" of their demanded messages. Both communications will be done simultaneously using joint cache-channel coding (via piggyback coding).

The transmitter proceeds in two steps:

T1: The transmitter applies Method En with $\tilde{K} = K_{\text{w}}$ and $\tilde{t} = t-1$ to demand vector $(d_1, \dots, d_{\text{w}})$ and messages

$$\left\{ W_{d_i}^{(t-1)} :\ i \in \mathcal{K}_{\text{w}} \right\}.$$

Method En outputs an XOR message for each size-t subset of $\{1, \dots, K_{\text{w}}\}$, for example, see (21). We denote these XOR messages by[7]

$$\left\{ W^{(t-1)}_{\text{XOR}, \mathcal{G}_\ell^{(t)}} :\ \ell = 1, \dots, \binom{K_{\text{w}}}{t} \right\}. \tag{53}$$

[7]The messages in (53) have the superscript $(t-1)$, because they correspond to the output of Method En with the parameter $\tilde{t} = t-1$. In contrast, $\{\mathcal{G}_\ell^{(t)}\}$ have superscript (t) because they correspond to subsets of size $\tilde{t} = t$.

T2: Time-sharing is performed over $\binom{K_\text{w}}{t}$ different periods, where each period is associated with a size-t subset of $\{1,2,\ldots,K_\text{w}\}$. Consider the ℓ-th subset $\mathcal{G}_\ell^{(t)}$. First recall that the subset $\mathcal{G}_\ell^{(t)}$ of weak receivers has

$$\left\{W^{(t)}_{d_j,\mathcal{G}_\ell^{(t)}}\colon\ j\in\mathcal{K}_\text{s}\right\}$$

stored as "side information" in their cache memories. The transmitter uses piggyback coding to send

$$\left\{W^{(t)}_{d_j,\mathcal{G}_\ell^{(t)}}\colon\ j\in\mathcal{K}_\text{s}\right\} \tag{54a}$$

to all strong receivers $\mathcal{K}_\text{s}$ and the XOR message

$$W^{(t-1)}_{\text{XOR},\mathcal{G}_\ell^{(t)}} \tag{54b}$$

to all weak receivers in $\mathcal{G}_\ell^{(t)}$.

Each strong receiver $j\in\mathcal{K}_\text{s}$ performs piggyback decoding (for the receiver without side-information) for all $\binom{K_\text{w}}{t}$ transmission periods. It forms the $\binom{K_\text{w}}{t}$-tuple estimate

$$\hat{W}^{(t)}_{d_j} := \left(\hat{W}^{(t)}_{d_j,\mathcal{G}_1^{(t)}},\ldots,\hat{W}^{(t)}_{d_j,\mathcal{G}^{(t)}_{\binom{K_\text{w}}{t}}}\right),\quad j\in\mathcal{K}_\text{s}. \tag{55}$$

Each weak receiver $i\in\mathcal{K}_\text{w}$ proceeds in two steps:

R1: Receiver i considers each subset $\mathcal{G}_\ell^{(t)}$ of size t to which it belongs (i.e., each $\mathcal{G}_\ell^{(t)}\subseteq\{1,\ldots,K_\text{w}\}$ such that $\mathcal{G}_\ell^{(t)}\ni i$), and it decodes the XOR message $W^{(t)}_{\text{XOR},\mathcal{G}_\ell^{(t)}}$ by applying piggyback decoding (for the receiver with side-information) to the channel outputs of the period associated with $\mathcal{G}_\ell^{(t)}$.

R2: Receiver i then applies Method $\texttt{De}_i$ with $\tilde{K}=K_\text{w}$ and $\tilde{t}=t-1$ to demand vector $(d_1,\ldots,d_\text{w})$, the XOR messages decoded in step R1 and its cache content V_i. Let

$$\hat{W}^{(t-1)}_{d_i},\quad i\in\mathcal{K}_\text{w} \tag{56}$$

denote the output of Method $\texttt{De}_i$.

3) Delivery subphase 3: Here we consider only the strong receivers, and we will communicate the remaining "$(t-1)$ parts" of their demanded messages. (The weak receivers will not participate in this subphase.) The transmitter communicates

$$\left\{W^{(t-1)}_{d_j}\colon\ j\in\mathcal{K}_\text{s}\right\}$$

to the strong receivers using a capacity-achieving code for the packet-erasure BC. Each receiver uses an optimal decoding method to produce the estimate

$$\hat{W}^{(t-1)}_{d_j},\ j\in\mathcal{K}_\text{s}. \tag{57}$$

4) Final decoding: Each receiver $k\in\{1,\ldots,K\}$ outputs

$$\hat{W}_{d_k}=\left(\hat{W}^{(t-1)}_{d_k},\hat{W}^{(t)}_{d_k}\right). \tag{58}$$

D. Analysis

Fix $t\in\{1,\ldots,K_\text{w}\}$. We show that the above scheme achieves the rate-memory pair (R_t,M_t) in (29).

1) Caching phase: By (25), our caching strategy requires a cache memory size of

$$\begin{aligned} M &= R^{(t)}\cdot D\frac{t}{K_\text{w}}+R^{(t-1)}\cdot D\frac{t-1}{K_\text{w}} \\ &= R\cdot\frac{D}{K_\text{w}}\left(t-\left(1+\frac{K_\text{w}-t+1}{tK_\text{s}}\cdot\frac{\delta_\text{w}-\delta_\text{s}}{1-\delta_\text{w}}\right)^{-1}\right). \end{aligned} \tag{59}$$

We now analyse the probability of decoding error. We present conditions under which the probability that the estimates produced in Subphases 1–3, (52), (55), (56), and (57) are not equal to the corresponding submessages in (47) tends to 0 as $n\to\infty$.

2) Delivery subphase 1: Proposition 1 combined with Lemma 2 and (48b), prove that the probability that the estimates in (52) are incorrect tends to 0 as $n\to\infty$, whenever

$$\frac{R\cdot\frac{K_\text{w}-t}{t+1}}{\left(1+\frac{tK_\text{s}}{K_\text{w}-t+1}\cdot\frac{1-\delta_\text{w}}{\delta_\text{w}-\delta_\text{s}}\right)\cdot F(1-\delta_\text{w})}<\beta_1. \tag{60}$$

3) Delivery subphase 2: Consider a single period with the transmission of message in (54). Since all weak receivers and all strong receivers are statistically equivalent, the probability that the estimates in (55) and (56) are incorrect is at most $t \cdot K_\mathrm{s}$ times larger than the probability of error in (55) and (56) for a single weak and a single strong receiver. By Corollary 4 and Lemma 2, this latter probability of error tends to 0 (and thus also the original probability of error tends to 0) as $n \to \infty$, whenever

$$\max\left\{\frac{R^{(t-1)} \cdot \frac{K_\mathrm{w}-t+1}{t}}{F(1-\delta_\mathrm{w})}, \frac{R^{(t-1)} \cdot \frac{K_\mathrm{w}-t+1}{t} + R^{(t)}K_\mathrm{s}}{F(1-\delta_\mathrm{s})}\right\} < \beta_2. \tag{61}$$

By our choice (48) the two terms in the maximization are equal, and thus by (48a) we conclude that the probability of producing an error in (55) or (56) tends to 0 as $n \to \infty$, whenever

$$\frac{R\frac{K_\mathrm{w}-t+1}{t}}{\left(1 + \frac{K_\mathrm{w}-t+1}{tK_\mathrm{s}} \cdot \frac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}\right) F(1-\delta_\mathrm{w})} < \beta_2. \tag{62}$$

4) Delivery subphase 3: Proposition 1 combined with (48a) prove that the probability of producing a wrong guess in (57) tends to 0 as $n \to \infty$, whenever

$$\frac{RK_\mathrm{s}}{\left(1 + \frac{K_\mathrm{w}-t+1}{tK_\mathrm{s}} \cdot \frac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}\right) F(1-\delta_\mathrm{s})} < \beta_3. \tag{63}$$

5) Overall scheme: Combining (60), (62), and (63) and using (50), after some algebraic manipulations, we see that the probability of decoding error tends to 0 as $n \to \infty$, whenever

$$R < F(1-\delta_\mathrm{w}) \cdot \frac{1 + \frac{K_\mathrm{w}-t+1}{tK_\mathrm{s}} \cdot \frac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}}{\frac{K_\mathrm{w}-t+1}{t}\left(1 + \frac{K_\mathrm{w}-t}{(t+1)K_\mathrm{s}} \cdot \frac{\delta_\mathrm{w}-\delta_\mathrm{s}}{1-\delta_\mathrm{w}}\right) + K_\mathrm{s}\frac{1-\delta_\mathrm{w}}{1-\delta_\mathrm{s}}}.$$

Together with (59), this proves achievability of the rate-memory pair $(R_t,\ M_t)$ in (29).

VI. Extensions of our Joint Cache-Channel Coding Scheme

The scenario in Section II allowed for a compact exposition of our new joint cache-channel coding idea. This idea however extends also to more general scenarios. In the following subsections we present some ideas.

A. Weak receivers have different erasure probabilities:

For simplicity, assume that the weak receivers are ordered so that $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_{K_\mathrm{w}}$ holds. The scheme of section V may be modified as follows: For each XOR message sent in delivery phase 1, set the rate of the codebook to be equal to the capacity of the weakest receiver to whom the XOR message is intended.

B. Strong receivers have different erasure probabilities:

For simplicity, assume that the strong receivers are ordered so that $\delta_{K_\mathrm{w}+1} \geq \delta_{K_\mathrm{w}+2} \geq \ldots \geq \delta_K$. We split the set of strong receivers into a set of moderately strong receivers $K_\mathrm{w}+1, \ldots, j^\star$ and a set of very strong receivers $j^\star, \ldots, K$, where $j^\star$ is chosen depending on the various erasure probabilities and cache sizes. We now time-share two schemes whose lengths need to be optimized: In the first period, a standard capacity-achieving coding scheme for the packet-erasure BC is used to serve only the moderately strong receivers $K_\mathrm{w}+1, \ldots, j^\star$. In the second period, our joint cache-channel coding scheme is used to serve all other receivers.

C. Some weaker receivers do not have caches:

We time-share two schemes whose lengths need to be optimized. In the first period, a standard capacity-achieving code for the packet-erasure BC is used to serve the weak receivers without caches. In the second period, our joint cache-channel coding scheme is used to serve all other receivers.

D. Weak receivers have different cache sizes:

For simplicity, assume that the weak receivers are ordered so that $M_1 \geq M_2 \geq \ldots \geq M_{\mathcal{K}_\mathrm{w}}$, where M_i denotes the cache memory size at receiver $i \in \mathcal{K}_\mathrm{w}$.

We time-share K_w schemes of equal length. In period $i \in \mathcal{K}_\mathrm{w}$, we treat the weak receivers $i+1, \ldots, K_\mathrm{w}$ assuming that they have no zero cache memories and we treat the weak receivers $1, \ldots, i$ assuming that they have cache memories of size $(M_i - M_{i+1})$. We thus apply the coding scheme that we sketched in the previous subsection.

E. Strong receivers have cache memories

For simplicity, assume that all strong receivers in $\mathcal{K}_\text{s}$ have cache memories of equal size $M > M_\text{s} > 0$. We time-share two schemes. In the first period of length $n\big(1 - \frac{M_\text{s}}{M}\big)$ we suppose that the strong receivers have no caches: we thus use our joint cache channel coding scheme. In the second period of length $n\frac{M_\text{s}}{M}$ we suppose that all receivers in the network have cache memory M: we apply separate cache-channel coding combining Maddah-Ali & Niesen coded caching over all K receivers with capacity-achieving code for the packet-erasure BC.

F. Strong receivers have cache memories and are served additional data

The strong receivers have cache memories of size M_s as in the previous subsection. There are two libraries now:

$$\text{library A:} \quad \text{files } W_1^{(\text{A})}, \ldots, W_D^{(\text{A})}$$

and

$$\text{library B:} \quad \text{files } W_1^{(\text{B})}, \ldots, W_D^{(\text{B})}.$$

Each weak receiver $i \in \mathcal{K}_\text{w}$ demands only file $W_{d_i}^{(\text{A})}$ from library A, whereas each strong receiver demands a file $W_{d_j}^{(\text{A})}$ from library A and a file $W_{d_j}^{(\text{B})}$ from library B.

We time-share two schemes whose lengths need to be optimized. In the first period, we suppose that the strong receivers have no cache memories and use our joint cache-channel coding scheme for library A. In the second period, we only serve the strong receivers from library B. To this end, we apply separate cache-channel coding combining Maddah-Ali & Niesen coded caching over all K receivers with capacity-achieving code for the packet-erasure BC.

In the scheme that we propose, the caches at the strong receivers are used only to cache messages from library B, but not from library A. The idea is that when the strong receivers are sufficiently strong, then our joint cache-channel coding scheme is as performant as if all receivers in the network had caches. One might therefore choose to dedicate the caches at the strong receivers entirely to the transmission of files from library B.

G. General DMBCs

Packet-erasure BCs are simpler than general BCs because time-sharing of optimal point-to-point codes for various receivers achieves capacity without caches. For our joint cache-channel coding scheme to be effective on more general DMBCs, we will have to partially replace time-sharing by superposition coding (and more generally Marton coding). More specifically, we use superposition coding and superpose the codewords sent in delivery subphase 3 on the piggyback codeword sent in delivery subphase 2 and on the codewords sent in delivery subphase 1. When there is no clear notion of "weaker" and "stronger", i.e., the channel is neither *degraded*, *less noisy*, *more capable*, or *essentially less noisy*, then we use Marton coding where the piggyback codewords serve as cloud centers and the other codewords as satellites.

VII. Upper Bound for General Degraded BCs under Arbitrary Demands

We consider a more general setup for the upper bound, where each receiver i has a cache of size M_i, and where the broadcast channel is a discrete memoryless *degraded BC* with input alphabet $\mathcal{X}$ and equal output alphabets $\mathcal{Y}_1, \ldots, \mathcal{Y}_K$. The joint transitional law of the memoryless BC is given by $P_{Y_1 Y_2 \cdots Y_K|X}(y_1, \ldots, y_K|x)$. We assume that the BC is *degraded*, i.e., the transition law satisfies the Markov chain

$$X - Y_K - Y_{K-1} - \cdots - Y_1. \tag{64}$$

For our problem setup, only the marginal transition law is relevant. Therefore our upper bound holds also for *stochastically degraded BC*, i.e., for transition laws $P_{Y_1,\ldots,Y_K|X}$ for which there exists a conditional probability distribution $\tilde{P}_{Y_2|Y_1}, \tilde{P}_{Y_3|Y_2}, \ldots, \tilde{P}_{Y_{K-1}|Y_K}$ that satisfies

$$\begin{aligned} &P_{Y_1 Y_2 \cdots Y_K|X}(y_1, \ldots, y_K|x) \\ &= P_{Y_K|X}(y_k|x)\tilde{P}_{Y_{K-1}|Y_K}(y_{k-1}|y_k)\ldots\tilde{P}_{Y_1|Y_2}(y_1|y_2) \end{aligned} \tag{65}$$

for all $(x, y_1, y_2, \ldots, y_K) \in \mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_K$. Note that the packet-erasure BC that we study falls in the class of (stochastically) degraded BCs.

The library and the probability of worst-case error $\mathsf{P}_\text{e}^\text{worst}$ is defined as before. A rate-memory tuple $(R,\ M_1, \ldots, M_K)$ is said *achievable* if for every $\epsilon > 0$ there exists a sufficiently large blocklength n and caching, encoding and decoding functions as in (8)–(11) such that $\mathsf{P}_\text{e}^\text{worst} < \epsilon$. The capacity-memory tradeoff $C(M_1, \ldots, M_K)$ is defined as the supremum over all rates R so that $(R, M_1, \ldots, M_K)$ are achievable.

For each $\mathcal{S} \in \mathcal{K}$, let $R_{\text{sym},\mathcal{S}}$ denote the largest equal-rate that is achievable over a BC with receivers in $\mathcal{S}$ when there are no cache memories. We prove the following upper bound on $C(M_1, \ldots, M_K)$:

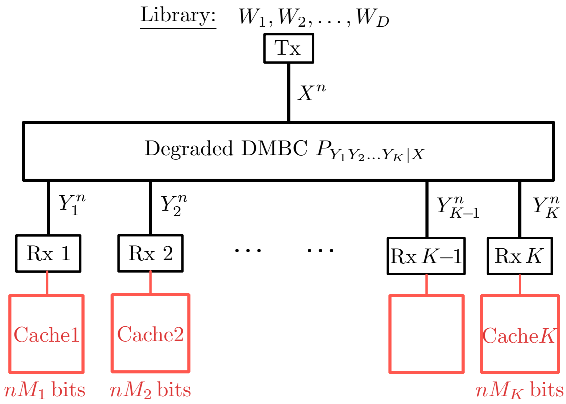

Fig. 11: Degraded K-user BC $P_{Y_1Y_2\cdots Y_K|X}$ where each Receiver $k \in \mathcal{K}$ has cache memory of size nM_k bits.

Theorem 9: The capacity-memory tradeoff $C(M_1, \ldots, M_K)$ of a degraded BC is upper bounded as follows:

$$C(M_1, \ldots, M_K) \leq \min_{\mathcal{S} \subseteq \{1,\ldots,K\}} \left(R_{\text{sym},\mathcal{S}} + \frac{M_{\mathcal{S}}}{D} \right),$$

where $M_{\mathcal{S}} = \sum_{k \in \mathcal{S}} M_k$ is the total cache size at receivers in $\mathcal{S}$.

Remark 2: Theorem 9 also holds for stochastically degraded BCs.

Before proving Theorem 9, let us specialize it to packet-erasure BCs by using Proposition 1 (to find $R_{\text{sym},\mathcal{S}}$):

Corollary 9.1: The capacity-memory tradeoff $C(M_1, \ldots, M_K)$ of the packet-erasure BC with packet size F, erasure probabilities $\delta_1, \ldots, \delta_K \geq 0$, and cache memory sizes $M_1, \ldots, M_K$, is upper bounded as follows:

$$C(M_1, \ldots, M_K) \leq \min_{\mathcal{S} \subseteq \{1,\ldots,K\}} \left(F \left(\sum_{k \in \mathcal{S}} \frac{1}{1-\delta_k} \right)^{-1} + \frac{M_{\mathcal{S}}}{D} \right). \tag{66}$$

We now consider our original setup of section II, where all weak receivers $i \in \mathcal{K}_\text{w}$ have the same erasure probability δ_w and the same cache memory $M_i = M$ and where all strong receivers $j \in \mathcal{K}_\text{s}$ have the same erasure probability $\delta_\text{s} \leq \delta_\text{w}$ and no cache memory $M_j = 0$. In this case, (66) simplifies to

$$C(M) \leq \min_{\substack{k_\text{w} \in \{0,\ldots,K_\text{w}\} \\ k_\text{s} \in \{0,\ldots,K_\text{s}\}}} \left(F \left(\frac{k_\text{w}}{1-\delta_\text{w}} + \frac{k_\text{s}}{1-\delta_\text{s}} \right)^{-1} + \frac{k_\text{w} M}{D} \right). \tag{67}$$

Since the right-hand side of inequality (67) is decreasing in k_s, the tightest upper bound is given by $k_\text{s} = K_\text{s}$. This also concludes the proof of Theorem 7.

Notice that the choice of k_w in (66) that leads to the tightest upper bound depends on the cache memory size M. For small values of M the choice $k_\text{w} = K_\text{w}$ leads to the tightest bound, and for increasing cache sizes smaller values of k_w lead to tighter bounds.

A. *Proof of Theorem 9*

For ease of exposition, we only prove the bound corresponding to $\mathcal{S} = \mathcal{K}$:

$$C(M_1, \ldots, M_K) \leq \left(R_{\text{sym},\mathcal{K}} + \frac{1}{D} \sum_{k=1}^{K} M_k \right), \tag{68}$$

where here $R_{\text{sym},\mathcal{K}}$ denotes the largest symmetric rate that is achievable over the BC $P_{Y_1Y_2\cdots Y_K|X}$ when there are no caches. The inequalities in the theorem that correspond to other subsets $\mathcal{S} \subseteq \{1, \ldots, K\}$ can be proved in an analogous way.

We start the proof of (68). Fix the rate of communication

$$R < C(M_1, \ldots, M_K).$$

Since R is achievable, for each sufficiently large blocklength n and for each demand vector $\mathbf{d}$, there exist K caching functions $\{g_i^{(n)}\}$, an encoding function $\{f_\mathbf{d}^{(n)}\}$, and K decoding functions $\{\varphi_{i,\mathbf{d}}^{(n)}\}$ so that the probability of worst-case error $\mathsf{P_e}^{(n)}(\mathbf{d})$ tends to 0 as $n \to \infty$. For each n let

$$V_k^{(n)} = g_k^{(n)}(W_1, \ldots, W_D), \qquad k \in \{1, \ldots, K\},$$

denote the cache contents for the chosen caching functions.

Lemma 10: For any $\epsilon > 0$, any demand vector $\mathbf{d} = (d_1, \ldots, d_K)$ with all different entries, and any blocklength n that is sufficiently large (depending on ϵ), there exist random variables $(U_{1,\mathbf{d}}, \ldots, U_{K,\mathbf{d}}, X_{\mathbf{d}}, Y_{1,\mathbf{d}}, \ldots, Y_{K,\mathbf{d}})$ such that

$$U_{1,\mathbf{d}} - U_{2,\mathbf{d}} - \cdots - U_{K,\mathbf{d}} - X_{\mathbf{d}} - Y_{K,\mathbf{d}} - Y_{K-1,\mathbf{d}} \cdots - Y_{1,\mathbf{d}} \tag{69}$$

forms a Markov chain, and given $X_{\mathbf{d}} = x \in \mathcal{X}$:

$$(Y_{1,\mathbf{d}}, Y_{2,\mathbf{d}}, \ldots, Y_{K,\mathbf{d}}) \sim P_{Y_1 \cdots Y_K|X}(\cdots|x),$$

and so that the following K inequalities hold:

$$R - \epsilon \leq \frac{1}{n} I\big(W_{d_1}; V_1^{(n)}, \ldots, V_K^{(n)}\big) + I\big(U_{1,\mathbf{d}}; Y_{1,\mathbf{d}}\big), \tag{70a}$$

$$\begin{aligned} R - \epsilon \leq & \frac{1}{n} I\big(W_{d_k}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{d_1}, \ldots, W_{d_{k-1}}\big) \\ & + I\big(U_{k,\mathbf{d}}; Y_{k,\mathbf{d}} | U_{k-1,\mathbf{d}}\big), \qquad \forall k \in \{2, \ldots, K\}. \end{aligned} \tag{70b}$$

Proof: The proof is similar to the converse proof of the capacity of degraded BCs without caching [42, Theorem 5.2]. It is deferred to Appendix A. ■

Fix $\epsilon > 0$ and a blocklength n (depending on this ϵ) so that Lemma 10 holds for all demand vectors $\mathbf{d}$ that have all different entries. We average the bound obtained in (76) over different demand vectors. Let $\mathcal{Q}$ be the set of all the $\binom{D}{K} K!$ demand vectors whose K entries are all different. Also, let Q be a uniform random variable over the elements of $\mathcal{Q}$ and independent of all other random variables. Define: $U_1 := (U_{1,Q}, Q)$; $U_k := U_{k,Q}$, for $k \in \{2, \ldots, K\}$; $X_k := X_Q$; and $Y_k := Y_{k,Q}$ for $k \in \{1, \ldots, K\}$. Notice that they form the Markov chain

$$U_1 \to U_2 \to \cdots \to U_K \to X \to (Y_1, \ldots, Y_K) \tag{71}$$

and given $X = x \in \mathcal{X}$ satisfy

$$(Y_1, Y_2, \ldots, Y_K) \sim P_{Y_1 \cdots Y_K|X}(\cdots|x). \tag{72}$$

Averaging inequalities (76) over the demand vectors in $\mathcal{Q}$ and using standard arguments to take care of the time-sharing random variable Q, we obtain:

$$R - \epsilon \leq \alpha_1 + I\big(U_1; Y_1\big), \tag{73a}$$

$$R - \epsilon \leq \alpha_k + I\big(U_k; Y_k | U_{k-1}\big), \quad \forall k \in \{2, \ldots, K\}, \tag{73b}$$

where we defined $\alpha_1, \ldots, \alpha_K$ as follows:

$$\alpha_1 := \frac{1}{\binom{D}{K} K!} \sum_{\mathbf{d} \in \mathcal{Q}} \frac{1}{n} I(W_{d_1}; V_1^{(n)}, \ldots, V_K^{(n)}), \tag{74a}$$

$$\alpha_k := \frac{1}{\binom{D}{K} K!} \sum_{\mathbf{d} \in \mathcal{Q}} \frac{1}{n} I(W_{d_k}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{d_1}, \ldots, W_{d_{k-1}}). \tag{74b}$$

Lemma 11: Parameters α_k, $k = 1, \ldots, K$, defined in (74), satisfy the following constraints:

$$\alpha_k \geq 0, \qquad k \in \{1, \ldots, K\} \tag{75a}$$

$$\alpha_{k'} \leq \alpha_k, \qquad k, k' \in \{1, \ldots, K\}, \; k' \leq k, \tag{75b}$$

$$\sum_{k \in \mathcal{K}} \alpha_k \leq \frac{K}{D} \sum_{k \in \mathcal{K}} M_k. \tag{75c}$$

Proof: See Appendix B. ■

Taking $\epsilon \to 0$, by (73) and (74) and by Lemma 11, we conclude that the capacity-memory tradeoff $C(M_1, \ldots, M_K)$ is upper bounded by the following K inequalities:

$$C(M_1, \ldots, M_K) \leq \alpha_1 + I\big(U_1; Y_1\big), \tag{76a}$$

$$C(M_1, \ldots, M_K) \leq \alpha_k + I\big(U_k; Y_k \big| U_{k-1}\big), \qquad \forall k \in \{2, \ldots, K\}, \tag{76b}$$

for some $\alpha_1, \ldots, \alpha_K$ satisfying (75) and some $U_1, \ldots, U_K, X, Y_1, \ldots, Y_K$ satisfying (71) and (72).

Lemma 12: Replacing each and every real number $\alpha_1, \ldots, \alpha_K$ in (76) by $\frac{1}{D} \sum_{k \in \{1, \ldots, K\}} M_k$ leads to a relaxed upper bound on $C(M_1, \ldots, M_K)$.

Proof: See Appendix C. ■

Thus,

$$C(M_1,\dots,M_K)-\frac{1}{D}\sum_{k\in\{1,\dots,K\}} M_k \le I\big(U_1;Y_1\big), \tag{77a}$$

$$C(M_1,\dots,M_K)-\frac{1}{D}\sum_{k\in\{1,\dots,K\}} M_k \le I\big(U_k;Y_k\big|U_{k-1}\big), \quad \forall k\in\{2,\dots,K\}, \tag{77b}$$

for some $U_1,\dots,U_K,X,Y_1,\dots,Y_K$ satisfying (71) and (72).

All K constraints in (77) have the same LHS, and their RHSs coincide with the rate-constraints of a degraded BC without caches. Therefore, the choice of the auxiliaries $(U_1,\dots,U_K)$ that leads to the most relaxed constraint on $C(M_1,\dots,M_K)$ coincides with the choice of auxiliaries that determines the largest symmetric rate-point of the degraded BC without caches. This establishes the equivalence of (77) with the desired bound in (68), and thus concludes the proof.

VIII. New Results for All-Equal Demands

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

$$d_1=d_2=\cdots=d_K\in\mathcal{D}.$$

Restricting to such "all-equal demands" allows us to treat arbitrary (unequal) message rates and arbitrary DMBCs. So, we consider the scenario in figure 12 where the transmitter communicates with K receivers over a DMBC $P_{Y_1\dots Y_K|X}$ and where each receiver $k\in\{1,\dots,K\}$ has a cache memory of size $M_k\ge 0$. The messages $W_1,\dots,W_D$ are independent of each other and each message W_d, for $d\in\{1,\dots,D\}$, is uniformly distributed over the set $\{1,\dots,\lfloor 2^{nR_d}\rfloor\}$ for some positive rate $R_d\ge 0$.

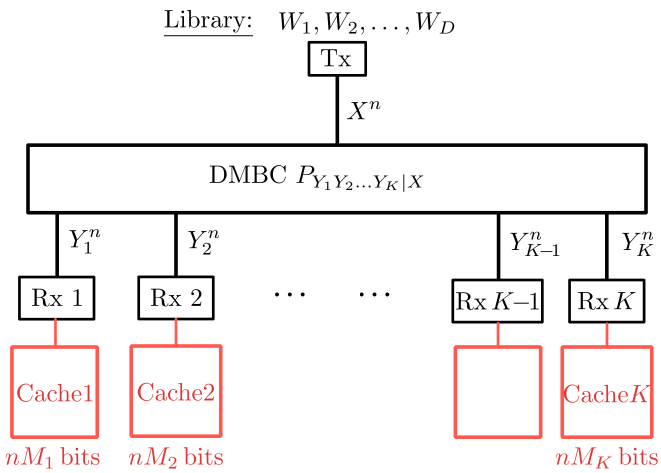

Fig. 12: K user DMBC where each Receiver $k\in\{1,\dots,K\}$ has cache memory of size nM_k bits.

The probability of worst-case error $\mathsf{P}_\mathrm{e}^{\text{worst}}$ is defined as before. A rate-memory tuple $(R_1,\dots,R_D,\ M_1,\dots,M_K)$ is said *achievable* if for every $\epsilon>0$ there exists a sufficiently large blocklength n and caching, encoding and decoding functions as in (8)–(11) (accounting for the different rates) such that $\mathsf{P}_\mathrm{e}^{\text{worst}}<\epsilon$.

Under the assumption that all receivers demand the same message, we can completely characterize the set of all achievable rates-memory tuples.

Theorem 13: A rate-memory tuple $(R_1,\dots,R_D,M_1,\dots,\ M_K)$ is achievable under all-equal demands if and only if

$$R_d\le \max_{P_X}\min_{k\in\{1,\dots,K\}}\big(I(X;Y_k)+M_{k,d}\big),\quad d\in\{1,\dots,D\}$$

for some nonnegative real numbers $\{M_{k,d}\}$ that satisfy

$$\sum_{d=1}^{D} M_{k,d}\le M_k,\quad k\in\{1,\dots,K\}. \tag{78}$$

Clearly, one wishes to allocate small cache sizes to strong receivers and large cache sizes to weak receivers.

If we used separate cache-channel codes, constraint (78) had to be replaced by

$$\max_{k\in\{1,\dots,K\}}(R_d-M_{k,d})\le \max_{P_X}\min_{k\in\{1,\dots,K\}} I(X;Y_k), \tag{79}$$

and the benefit of having unequal cache sizes $\{M_k\}$ at the different receivers would disappear.

A. Proof of Achievability of Theorem 13

We propose the following scheme. For each $d \in \{1, \ldots, D\}$ and each $k \in \{1, \ldots, K\}$, fix a positive integer number

$$M_{k,d} \leq R_k \tag{80}$$

so that for every $k \in \{1, \ldots, K\}$:

$$\sum_{d=1}^{N} M_{k,d} \leq M_k. \tag{81}$$

Fix a probability distribution P_X over the channel input alphabet $\mathcal{X}$. Construct a codebook $\mathcal{C}$ containing $\lfloor 2^{nR} \rfloor$ codewords of length n by drawing all symbols of all codewords i.i.d. according to the probability distribution P_X. Reveal codebook $\mathcal{C}$ to the transmitter and to all receivers.

Caching phase: For each receiver $k \in \{1, \ldots, K\}$, cache the first $\lfloor nM_{k,d} \rfloor$ bits of each Message W_d in receiver k's cache. By (81) this caching strategy satisfies the cache memory constraint.

Delivery phase: Let $d^\star = d_1 = \ldots = d_K$. The transmitter picks the codeword from codebook $\mathcal{C}$ that corresponds to message $W_{d^\star}$ and sends it over the channel.

Each receiver k decodes the same desired message $W_{d^\star}$. Since it knows the first $nM_{k,d^\star}$ bits of $W_{d^\star}$, in its decoding it restricts attention to the part of the codebook corresponding to messages starting with these bits.

Error analysis: The probability of error (averaged over codebooks and messages) at receiver k is the same as if only the last $n(R - M_{k,d^\star})$ bits of message $W_{d^\star}$ had been sent. Thus, by the packing lemma [42], the probability of decoding error at receiver k tends to 0 as $n \to \infty$, whenever

$$R_{d^\star} - M_{k,d^\star} \leq I(X; Y_k), \quad d^\star \in \{1, \ldots, D\},\ k \in \{1, \ldots, K\}.$$

This proves achievability of Theorem 13.

B. Proof of Converse to Theorem 13

Let $(R_1, \ldots, R_N, M_1, \ldots, M_K)$ be an achievable rate-memory tuple. For each sufficiently large n and demand $d \in \{1, \ldots, D\}$ fix K caching functions $\{g_i^{(n)}\}$, an encoding function $\{f_{\mathbf{d}}^{(n)}\}$, and K decoding functions $\{\varphi_{i,\mathbf{d}}^{(n)}\}$ so that the probability of worst-case error $\mathsf{P}_{\mathsf{e}}^{\text{worst}}$ tends to 0 as $n \to \infty$. Fix now a blocklength n, a demand $d^\star \in \{1, \ldots, D\}$, and a receiver $k \in \{1, \ldots, K\}$. We have

$$\begin{aligned}
R_{d^\star} &\leq \frac{1}{n} H(W_{d^\star}) \\
&= \frac{1}{n} I(W_{d^\star}; Y_k^n, V_k^{(n)}) + \frac{1}{n} H(W_{d^\star} | Y_k^n, V_k^{(n)}) \\
&\leq \frac{1}{n} I(W_{d^\star}; Y_k^n | V_k^{(n)}) + \frac{1}{n} I(W_{d^\star}; V_k^{(n)}) + \epsilon_n \\
&= \frac{1}{n} \sum_{t=1}^{n} \big(H(Y_{k,t} | V_k^{(n)}, Y_k^{t-1}) - H(Y_{k,t} | W_{d^\star}, Y_k^{t-1}, V_k^{(n)}) \big) \\
&\quad + M_{k,d^\star} + \epsilon_n \\
&\leq \frac{1}{n} \sum_{t=1}^{n} \big(H(Y_{k,t}) - H(Y_{k,t} | X_{k,t}) \big) + M_{k,d^\star} + \epsilon_n \\
&= \frac{1}{n} \sum_{t=1}^{n} I(Y_{k,t}; X_{k,t}) + M_{k,d^\star} + \epsilon_n,
\end{aligned} \tag{82}$$

where we defined

$$M_{k,d} \triangleq \frac{1}{n} I(W_d; V_k^{(n)}), \quad k \in \{1, \ldots, K\},\ d \in \{1, \ldots, D\}.$$

The second inequality above follows by Fano's inequality; the third inequality holds because conditioning does not increase entropy and because $(M_d, Y_k^{t-1}, V_k^{(n)}) \to X_t \to Y_{k,t}$ forms a Markov chain.

Moreover, for each $k \in \{1, \ldots, K\}$,

$$\begin{aligned}
\sum_{d=1}^{N} M_{k,d} &= \frac{1}{n} \sum_{d=1}^{D} I(W_d; V_k^{(n)}) \\
&\leq \frac{1}{n} \sum_{d=1}^{D} I(W_d; V_k^{(n)} | W_1, \ldots, W_{d-1})
\end{aligned}$$

$$= \frac{1}{n} I(W_1, \dots, W_D; V_k^{(n)})$$
$$\le \frac{1}{n} H(V_k^{(n)}) \le M_k, \tag{83}$$

where the first inequality follows because messages $W_1, \dots, W_D$ are independent.

Letting $n \to \infty$, and thus $\epsilon_n \to 0$, establishes the desired converse.

Appendix A
Proof of Lemma 10

Fix a small $\epsilon > 0$ and a demand vector $\mathbf{d}$ with all different entries. Then, let the blocklength n be sufficiently large as will be come clear in the following. Also, let

$$V_i^{(n)} = g_i^{(n)}(W_1, \dots, W_D), \qquad i \in \{1, \dots, K\}, \tag{84}$$
$$X_{\mathbf{d}}^n = f_{\mathbf{d}}^{(n)}(W_1, \dots, W_D) \tag{85}$$

denote cache contents and the input of the degraded BC for demand vector $\mathbf{d} \in \mathcal{D}^K$ and for above chosen caching and encoding functions. Also, let $Y_{k,\mathbf{d}}^n$ denote the corresponding channel outputs at Receiver k.

By Fano's inequality, by the independence of the messages $W_1, \dots, W_D$, and because the caching, encoding, and decoding functions have been chosen so that the worst case probability of error tends to 0 for increasing blocklengths, we obtain that that for all sufficiently large n the following K inequalities hold:

$$\begin{aligned} R - \epsilon &\le \frac{1}{n} I\big(W_{d_1}; Y_{k,\mathbf{d}}^n, V_1^{(n)}, \dots, V_K^{(n)}\big) \\ &= \frac{1}{n} I\big(W_{d_1}; V_1^{(n)}, \dots, V_K^{(n)}\big) + \frac{1}{n} I\big(W_{d_1}; Y_{k,\mathbf{d}}^n \big| V_1^{(n)}, \dots, V_K^{(n)}\big) \end{aligned} \tag{86a}$$

and for all $k \in \{2, \dots, K\}$:

$$\begin{aligned} R - \epsilon_n &\le \frac{1}{n} I\big(W_{d_k}; Y_{k,\mathbf{d}}^n, V_1^{(n)}, \dots, V_K^{(n)} \big| W_{d_1}, \dots, W_{d_{k-1}}\big) \\ &= \frac{1}{n} I\big(W_{d_k}; V_1^{(n)}, \dots, V_K^{(n)} \big| W_{d_1}, \dots, W_{d_{k-1}}\big) \\ &\quad + \frac{1}{n} I\big(W_{d_k}; Y_{k,\mathbf{d}}^n \big| V_1^{(n)}, \dots, V_K^{(n)}, W_{d_1}, \dots, W_{d_{k-1}}\big) \end{aligned} \tag{86b}$$

We further develop the second summands in (86a) and (86b). For the second summand in (86a) we obtain

$$\begin{aligned} &\frac{1}{n} I\big(W_{d_1}; Y_{k,\mathbf{d}}^n \big| V_1^{(n)}, \dots, V_K^{(n)}\big) \\ &\quad = \frac{1}{n} \sum_{t=1}^{n} I\big(W_{d_1}; Y_{k,\mathbf{d},t} \big| V_1^{(n)}, \dots, V_K^{(n)}, Y_{k,\mathbf{d}}^{t-1}\big) \\ &\quad = \frac{1}{n} \sum_{t=1}^{n} I\big(Y_{k,\mathbf{d}}^{t-1}, W_{d_1}; Y_{k,\mathbf{d},t} \big| V_1^{(n)}, \dots, V_K^{(n)}\big) \\ &\quad = I\big(U_{1,\mathbf{d},T}; Y_{k,\mathbf{d},T} \big| V_1^{(n)}, \dots, V_K^{(n)}, T\big) \\ &\quad \le I\big(U_{1,\mathbf{d}}; Y_{k,\mathbf{d}} \big| V_1^{(n)}, \dots, V_K^{(n)}\big), \end{aligned} \tag{87}$$

where T denotes a random variable that is uniformly distributed over $\{1, \dots, n\}$ and independent of all other random variables, and where we defined

$$\begin{aligned} U_{1,\mathbf{d},T} &:= (V_1^{(n)} \dots, V_K^{(n)}, W_{d_1}, Y_{1,\mathbf{d}}^{t-1}), \\ U_{1,\mathbf{d}} &:= (U_{1,\mathbf{d},T}, T), \\ Y_{k,\mathbf{d}} &:= Y_{k,\mathbf{d},T}, \qquad k \in \{1, \dots, K\}. \end{aligned}$$

We also define for $k \in \{2, \dots, K\}$:

$$\begin{aligned} U_{k,\mathbf{d},T} &:= (V_1^{(n)} \dots, V_K^{(n)}, W_{d_1}, W_{d_2}, \dots, W_{d_k}, Y_{1,\mathbf{d}}^{t-1}, \dots, Y_{k,\mathbf{d}}^{t-1}), \\ U_{k,\mathbf{d}} &:= U_{k,\mathbf{d},T}, \end{aligned}$$

in order to expand the second summand in (86b) as follows:

$$\frac{1}{n} I\big(W_{d_k}; Y_{k,\mathbf{d}}^n \big| V_1^{(n)}, \dots, V_K^{(n)}, W_{d_1}, \dots, W_{d_{k-1}}\big)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{t=1}^{n} I\big(W_{d_k}; Y_{k,\mathbf{d},t} \big| V_1^{(n)}, \ldots, V_K^{(n)}, W_{d_1}, \ldots, W_{d_{k-1}}, Y_{k,\mathbf{d}}^{t-1}\big) \\
&= \frac{1}{n}\sum_{t=1}^{n} I\big(W_{d_k}; Y_{k,\mathbf{d},t} \big| V_1^{(n)}, \ldots, V_K^{(n)}, W_{d_1}, \ldots, W_{d_{k-1}}, Y_{1,\mathbf{d}}^{t-1}, \ldots, Y_{k-1,\mathbf{d}}^{t-1}, Y_{k,\mathbf{d}}^{t-1}\big) \\
&\leq \frac{1}{n}\sum_{t=1}^{n} I\big(W_{d_k} Y_{k,\mathbf{d}}^{t-1} \ ; \ Y_{k,\mathbf{d},t} \big| V_1^{(n)}, \ldots, V_K^{(n)}, W_{d_1}, \ldots, W_{d_{k-1}}, Y_{1,\mathbf{d}}^{t-1}, \ldots, Y_{k-1,\mathbf{d}}^{t-1}\big) \\
&= I\big(U_{k,\mathbf{d},T}; Y_{k,\mathbf{d},T} \big| U_{k-1,\mathbf{d},T}, T\big) \\
&= I\big(U_{k,\mathbf{d}}; Y_{k,\mathbf{d}} \big| U_{k-1,\mathbf{d}}\big) \qquad (88)
\end{aligned}
$$

where the second equality follows from the degradedness of the outputs, see (64).

Notice that if we also define $X_{\mathbf{d}} := X_{\mathbf{d},T}$, then (69) and (76) hold. Combining this observation with (86)–(88) concludes the proof.

APPENDIX B
PROOF OF LEMMA 11

Constraint (75a) follows by the nonnegativity of mutual information. To prove Constraint (75b), we fix a demand vector $\mathbf{d} \in \mathcal{Q}$, and consider the cyclic shifts of this vector. For $\ell \in \{0, \ldots, K-1\}$, let $\overrightarrow{\mathbf{d}}^{(\ell)}$ be the vector obtained from $\overrightarrow{\mathbf{d}}$ when the elements are cyclically shifted ℓ positions to the right. (E.g., if $\mathbf{d} = (1,2,3)$ then $\overrightarrow{\mathbf{d}}^{(2)} = (2,3,1)$.) For each $\ell \in \{0, \ldots, K-1\}$ and $k \in \{1, \ldots, K\}$, let $\overrightarrow{d}_k^{(\ell)}$ denote the k-th index of demand vector $\overrightarrow{\mathbf{d}}^{(\ell)}$. So,

$$
\overrightarrow{d}_k^{(\ell)} = d_{(k-\ell) \bmod K} \qquad (89)
$$

where for each positive integer ξ the term $(\xi \mod K)$ takes value in $\{1, \ldots, K\}$ so that

$$
\xi \mod K = \xi - bK \quad \text{for some positive integer } b. \qquad (90)
$$

For each $\ell \in \{1, \ldots, K-1\}$ and $k, k' \in \{2, \ldots, K\}$, $k' \leq k$:

$$
\begin{aligned}
& I(W_{d_1}; V_1^{(n)}, \ldots, V_K^{(n)}) \stackrel{(a)}{=} I(W_{\overrightarrow{d}_{k'}^{(k'-1)}}; V_1^{(n)}, \ldots, V_K^{(n)}) \\
& \stackrel{(b)}{\leq} I(W_{\overrightarrow{d}_{k'}^{(k'-1)}}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{\overrightarrow{d}_1^{(k'-1)}}, \ldots, W_{\overrightarrow{d}_{k'-1}^{(k'-1)}}) \\
& \stackrel{(a)}{=} I(W_{\overrightarrow{d}_k^{(k-1)}}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{\overrightarrow{d}_2^{(k-1)}}, \ldots, W_{\overrightarrow{d}_{k-1}^{(k-1)}}) \\
& \stackrel{(b)}{\leq} I(W_{\overrightarrow{d}_k^{(k-1)}}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{\overrightarrow{d}_1^{(k-1)}}, \ldots, W_{\overrightarrow{d}_{k-1}^{(k-1)}}), \qquad (91)
\end{aligned}
$$

where (a) follows by (89) and (b) is by the independence of messages.

Fix a demand vector $\mathbf{d} \in \mathcal{Q}$ and sum up the above inequality (91) over all K cyclic shifts $\mathbf{d}^{(0)}, \mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(K-1)}$ of $\mathbf{d}$ to obtain:

$$
\begin{aligned}
& \sum_{\ell=0}^{K-1} I(W_{\overrightarrow{d}_1^{(\ell)}}; V_1^{(n)}, \ldots, V_K^{(n)}) \\
& \quad \leq \sum_{\ell=0}^{K-1} I(W_{\overrightarrow{d}_{k'}^{(\ell)}}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{\overrightarrow{d}_1^{(\ell)}}, \ldots, W_{\overrightarrow{d}_{k'-1}^{(\ell)}}) \\
& \quad \leq \sum_{\ell=0}^{K-1} I(W_{\overrightarrow{d}_k^{(\ell)}}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{\overrightarrow{d}_1^{(\ell)}}, \ldots, W_{\overrightarrow{d}_{k-1}^{(\ell)}}). \qquad (92)
\end{aligned}
$$

Since the set $\mathcal{Q}$ can be partitioned into subsets of demand vectors that are cyclic shifts of each others and all cyclic shifts of a demand vector in $\mathcal{Q}$ are also in $\mathcal{Q}$, we conclude from (92):

$$
\begin{aligned}
& \sum_{\mathbf{d} \in \mathcal{Q}} I(W_{d_1}; V_1^{(n)}, \ldots, V_K^{(n)}) \\
& \quad \leq \sum_{\mathbf{d} \in \mathcal{Q}} I(W_{d_{k'}}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{d_1}, \ldots, W_{d_{k'-1}}) \\
& \quad \leq \sum_{\mathbf{d} \in \mathcal{Q}} I(W_{d_k}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{d_1}, \ldots, W_{d_{k-1}}). \qquad (93)
\end{aligned}
$$

This proves (75b).

We proceed to prove Constraint (75c). For each $\mathbf{d} \in \mathcal{Q}$:

$$\begin{aligned}
&I(W_{d_1}; V_1^{(n)}, \ldots, V_K^{(n)}) \\
&\quad + \sum_{k=2}^{K} I(W_{d_k}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{d_1}, W_{d_2}, \ldots, W_{d_{k-1}}) \\
&\qquad = I(W_{d_1}, W_{d_2}, \ldots, W_{d_{K-1}}; V_1^{(n)}, \ldots, V_K^{(n)}).
\end{aligned} \tag{94}$$

So,

$$\begin{aligned}
&\sum_{\mathbf{d}\in\mathcal{Q}} \Big[I(W_{d_1}; V_1^{(n)}, \ldots, V_K^{(n)}) \\
&\qquad + \sum_{k=2}^{K} I(W_{d_k}; V_1^{(n)}, \ldots, V_K^{(n)} | W_{d_1}, W_{d_2}, \ldots, W_{d_{k-1}}) \Big] \\
&= \sum_{\mathbf{d}\in\mathcal{Q}} I(W_{d_1}, W_{d_2}, \ldots, W_{d_K}; V_1^{(n)}, \ldots, V_K^{(n)}) \\
&\overset{(a)}{=} \sum_{\mathbf{d}\in\mathcal{Q}} \Big[H(W_{d_1}) + H(W_{d_2}) + \ldots + H(W_{d_K}) \\
&\qquad\qquad - H(W_{d_1}, \ldots, W_{d_K} | V_1^{(n)}, \ldots, V_K^{(n)}) \Big] \\
&\overset{(b)}{=} \frac{K}{D} |\mathcal{Q}| H(W_1, \ldots, W_D) \\
&\quad - \sum_{\mathbf{d}\in\mathcal{Q}} H(W_{d_1}, \ldots, W_{d_K} | V_1^{(n)}, \ldots, V_K^{(n)}) \\
&\overset{(c)}{\leq} \frac{K}{D} K! \binom{D}{K} H(W_1, \ldots, W_D) \\
&\quad - \frac{K}{D} K! \binom{D}{K} H(W_1, \ldots, W_D | V_1^{(n)}, \ldots, V_K^{(n)}) \\
&\overset{(b)}{=} \frac{K}{D} K! \binom{D}{K} I(W_1, \ldots, W_D; V_1^{(n)}, \ldots, V_K^{(n)}) \\
&\leq \frac{K}{D} K! \binom{D}{K} n \sum_{k=1}^{K} M_k,
\end{aligned}$$

where (a) holds by the chain rule of mutual information, (b) by the independence and uniform rate of messages $W_1, \ldots, W_D$ and the definition of the set $\mathcal{Q}$, which is of size $\binom{D}{K} K!$, and (c) by the generalized Han-Inequality (the following Proposition 14).

Proposition 14: Let L be a positive integer and $A_1, \ldots, A_L$ be a finite random L-tuple. Denote by $A_{\mathcal{S}}$ the subset $\{A_\ell,\ \ell \in \mathcal{S}\}$. For every $\ell \in \{1, \ldots, L\}$:

$$\frac{1}{\binom{L}{\ell}} \sum_{\mathcal{S}\subseteq\{1,\ldots,L\}:|\mathcal{S}|=\ell} \frac{H(A_{\mathcal{S}})}{\ell} \geq \frac{1}{L} H(A_1, \ldots, A_L). \tag{95}$$

Proof: See [44, Theorem 17.6.1]. ■

Appendix C
Proof of Lemma 12

Fix random variables $U_1, U_2, \ldots, U_K, X$ satisfying the Markov chain (71) and real numbers $\alpha_1, \ldots, \alpha_K$ satisfying (75). We will show that if $\alpha_{\tilde{k}} \neq \alpha_{\tilde{k}+1}$ for some $\tilde{k} \in \mathcal{K}$, then we can find new random variables $\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_K, \bar{X}$ satisfying the Markov chain (71) and real numbers $\bar{\alpha}_1, \ldots, \bar{\alpha}_K$ satisfying (75) so that the upper bound on $C(M_1, \ldots, M_K)$ in (76) is relaxed if we replace

$$(U_1, U_2, \ldots, U_K, X) \qquad \text{and} \qquad (\alpha_1, \ldots, \alpha_K)$$

by

$$(\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_K, \bar{X}) \qquad \text{and} \qquad (\bar{\alpha}_1, \ldots, \bar{\alpha}_K).$$

This proves that we obtain a relaxed upper bound on $C(M_1,\ldots,M_K)$ if in (76) we replace all numbers $\alpha_1,\ldots,\alpha_K$ by the same number α. By (75c) this number $\alpha \leq \frac{1}{D}\sum_{k\in\{1,\ldots,K\}} M_k$, and by the monotonicity of the RHSs of (76) in $\alpha_1,\ldots,\alpha_K$ the choice $\alpha = \frac{1}{D}\sum_{k\in\{1,\ldots,K\}} M_k$ leads to the most relaxed upper bound. This will conclude the proof.

Assume that $\alpha_{\tilde{k}} \neq \alpha_{\tilde{k}+1}$ for some $\tilde{k} \in \{1,\ldots,K-1\}$. By (75b), the strict inequality

$$\alpha_{\tilde{k}} < \alpha_{\tilde{k}+1} \tag{96}$$

must hold. Choose

$$\bar{\alpha}_k = \alpha_k, \qquad k \in \mathcal{K},\ k \notin \{\tilde{k}, \tilde{k}+1\}, \tag{97}$$

$$\bar{\alpha}_{\tilde{k}} = \bar{\alpha}_{\tilde{k}+1} = \frac{1}{2}(\alpha_{\tilde{k}} + \alpha_{\tilde{k}+1}), \tag{98}$$

$$\bar{U}_k = U_k, \qquad k \in \mathcal{K},\ k \neq \tilde{k}. \tag{99}$$

The choice of $\bar{U}_{\tilde{k}}$ depends on whether

$$I(U_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1}) \leq I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}), \tag{100a}$$

or

$$I(U_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1}) > I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}). \tag{100b}$$

If (100a) holds, choose

$$\bar{U}_{\tilde{k}} = U_{\tilde{k}}. \tag{101}$$

If (100b) holds, let $E \in \{0,1\}$ be a Bernoulli-β random variable independent of everything else, where

$$\beta := \frac{1}{2} + \frac{1}{2}\frac{I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}})}{I(U_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1})}. \tag{102}$$

Choose

$$\bar{U}_{\tilde{k}} = \begin{cases} (U_{\tilde{k}}, E), & \text{if } E = 0 \\ (U_{\tilde{k}-1}, E), & \text{if } E = 1. \end{cases} \tag{103}$$

The proposed choice satisfies the Markov chain $\bar{U}_1 - \bar{U}_2 - \cdots \bar{U}_K - X$. Moreover, by (103) and (102):

$$\begin{aligned} &I(\bar{U}_{\tilde{k}}; Y_{\tilde{k}} | \bar{U}_{\tilde{k}-1}) \\ &= \frac{1}{2}\left(I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}) + I(U_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1})\right). \end{aligned} \tag{104}$$

Trivially, for $k \notin \{\tilde{k}, \tilde{k}+1\}$, constraint (76) is unchanged if we replace $(U_1, U_2, \ldots, U_K, X)$ by $(\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_K, \bar{X})$ and $(\alpha_1,\ldots,\alpha_K)$ by $(\bar{\alpha}_1,\ldots,\bar{\alpha}_K)$.

If (100a) holds, then the proposed replacement relaxes constraint (76) for $k = \tilde{k}$ and it tightens it for $k = \tilde{k}+1$. However, the new constraint for $k = \tilde{k}+1$ is less stringent than the original constraint for $k = \tilde{k}$. We conclude that when (100a) holds, the upper bound on $C(M_1,\ldots,M_K)$ in (76) is relaxed if everywhere one replaces $(U_1, U_2, \ldots, U_K, X)$ and $(\alpha_1,\ldots,\alpha_K)$ by $(\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_K, \bar{X})$ and $(\bar{\alpha}_1,\ldots,\bar{\alpha}_K)$.

If (100b) holds, then the new constraint obtained for $k = \tilde{k}$ coincides with the average of the two original constraints for $k = \tilde{k}$ and for $k = \tilde{k}+1$, see (98) and (104). This average constraint cannot be more stringent than the most stringent of the two original constraints. The new constraint obtained for $k = \tilde{k}+1$ is more relaxed than the new constraint obtained for $k = \tilde{k}$, because of (98) and because

$$\begin{aligned} &I(\bar{U}_{\tilde{k}+1}; Y_{\tilde{k}+1} | \bar{U}_{\tilde{k}}) \\ &\overset{(a)}{=} \beta I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}) + (1-\beta) I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}-1}) \\ &\overset{(b)}{=} \beta I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}) + (1-\beta) I(U_{\tilde{k}+1}, U_{\tilde{k}}; Y_{\tilde{k}+1} | U_{\tilde{k}-1}) \\ &\overset{(c)}{=} I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}) + (1-\beta) I(U_{\tilde{k}}; Y_{\tilde{k}+1} | U_{\tilde{k}-1}) \\ &\overset{(d)}{\geq} I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}) + (1-\beta) I(U_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1}) \\ &\overset{(e)}{=} \frac{1}{2} I(U_{\tilde{k}+1}; Y_{\tilde{k}+1} | U_{\tilde{k}}) + \frac{1}{2} I(U_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1}) \\ &\overset{(f)}{=} I(\bar{U}_{\tilde{k}}; Y_{\tilde{k}} | U_{\tilde{k}-1}), \end{aligned} \tag{105}$$

where (a) follows by the definition of $\bar{U}_{\tilde{k}}$ and $\bar{U}_{\tilde{k}+1}$; (b) by the Markov chain (71); (c) by the chain rule of mutual information; (d) by the degradedness of the channel (71); (e) by the definition of β in (102); and (f) by (104).

We can thus conclude that also when (100b) holds, the upper bound on $C(M_1,\ldots,M_K)$ in (76) is relaxed if one replaces $(U_1, U_2, \ldots, U_K, X)$ and $(\alpha_1,\ldots,\alpha_K)$ by $(\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_K, \bar{X})$ and $(\bar{\alpha}_1,\ldots,\bar{\alpha}_K)$.

Appendix D
Achievability Proof for Rate-Memory Pair $(F(1-\delta_{\mathrm{s}}),\ 2\tilde{\Gamma}_2)$

The following scheme achieves the rate-memory pair

$$R = F(1-\delta_{\mathrm{s}}) \quad \text{and} \quad M = 2\tilde{\Gamma}_2. \tag{106}$$

Split messages W_1 and W_2 into two independent submessages

$$W_d = (W_d^{(1)}, W_d^{(2)}), \quad d \in \{1,\dots,D\}$$

of rates

$$R^{(1)} := F(\delta_{\mathrm{w}} - \delta_{\mathrm{s}}) \tag{107a}$$
$$R^{(2)} := F(1-\delta_{\mathrm{w}}) - \epsilon, \tag{107b}$$

for an arbitrarily small $\epsilon > 0$.
Caching Phase: Cache the pair

$$V_1 := \left(W_1^{(1)}, W_2^{(1)}, W_1^{(2)} \bigoplus W_2^{(2)}\right) \tag{108}$$

in the weak receiver's cache.
Delivery Phase: Use *piggyback coding*, see subsection III-D, to send $W_{d_2}^{(1)}$ to the strong receiver and $W_{d_2}^{(2)}$ to the weak receiver who already has side-information $W_{d_2}^{(1)}$.

The strong receiver applies piggyback decoding (for the receiver without side-information), where it in fact decodes both transmitted messages. This way it produces the estimate

$$\hat{W}_{d_2} := \left(\hat{W}_{d_2,\mathrm{Rx2}}^{(1)}, \hat{W}_{d_2,\mathrm{Rx2}}^{(2)}\right). \tag{109}$$

The weak receiver applies piggyback decoding (for the receiver with side-information), which produces $\hat{W}_{d_2,\mathrm{Rx1}}^{(2)}$. Using its cache content V_1, it produces the guess

$$\hat{W}_{d_1} := \begin{cases} \left(W_{d_1}^{(1)}, \hat{W}_{d_2,\mathrm{Rx1}}^{(2)}\right) & \text{if } d_1 = d_2 \\ \left(W_{d_1}^{(1)}, \hat{W}_{d_2,\mathrm{Rx1}}^{(2)} \bigoplus W_1^{(2)} \bigoplus W_2^{(2)}\right), & \text{if } d_1 \neq d_2. \end{cases} \tag{110}$$

Analysis: By Corollary 4 and due to the choice of rates $R^{(1)}$ and $R^{(2)}$ in (107), the probability of error tends to 0 as the blocklength n tends to infinity. Since $\epsilon > 0$ can be chosen arbitrarily close to 0, we have proved achievability of the rate-memory pair in (106).

Appendix E
Proof of Theorem 8

The first and last terms in (43) are special cases of (7) with $k_{\mathrm{w}} = 1$ and $k_{\mathrm{w}} = 0$, respectively. Here, we prove the second term by showing that for every achievable memory-rate pair $(R,\ M)$,

$$3R \leq M + (1-\delta_{\mathrm{w}})F + (1-\delta_{\mathrm{s}})F. \tag{111}$$

Since the capacity-memory tradeoff only depends on the conditional marginal distributions of the channel law (4), we will assume that the packet-erasure BC is physically degraded. So, for each $t \in \{1,\dots,n\}$,

$$X_t \to Y_{2,t} \to Y_{1,t}. \tag{112}$$

For all sufficiently large blocklengths n, choose caching functions $\{g_i^{(n)}\}$ as in (8), encoding functions $f_{\mathbf{d}}^{(n)}$ as in (9), and decoding functions $\{\varphi_{i,\mathbf{d}}^{(n)}\}$ as in (11) so that the probability of worst-case error $P_{\mathrm{e}}^{\mathrm{worst}}$ tends to 0 as the blocklength $n \to \infty$. Consider now a fixed blocklength n that is sufficiently large for the purposes that we describe in the following. Let

$$V_1^{(n)} = g_1^{(n)}(W_1,\dots,W_D), \tag{113}$$
$$X_{\mathbf{d}}^n = f_{\mathbf{d}}^{(n)}(W_1,\dots,W_D) \tag{114}$$

denote cache contents and the input of the packet-erasure BC for a given demand vector $\mathbf{d} \in \mathcal{D}^2$ and for above chosen caching and encoding functions. Also, let $Y_{1,\mathbf{d}}^n$ and $Y_{2,\mathbf{d}}^n$ denote the corresponding channel outputs at the weak and strong receivers.

We focus on the two demand vectors

$$\mathbf{d}_1 := (1,2) \qquad \text{and} \qquad \mathbf{d}_2 := (2,1).$$

So, W_1 should be decodable from $(Y^n_{1,\mathbf{d}_1}, V_1^{(n)})$ and from $Y^n_{2,\mathbf{d}_2}$, and W_2 should be decodable from $(Y^n_{1,\mathbf{d}_2}, V_1^{(n)})$. Thus, by Fano's inequality, for all $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and sufficiently large blocklength n, we have

$$nR \leq I(W_1; V_1^{(n)}, Y^n_{1,\mathbf{d}_1}) + n\epsilon_1 \tag{115a}$$

$$nR \leq I(W_1; Y^n_{2,\mathbf{d}_2}) + n\epsilon_2 \tag{115b}$$

$$nR \leq I(W_2; V_1^{(n)}, Y^n_{1,\mathbf{d}_1}, Y^n_{1,\mathbf{d}_2}|W_1) + n\epsilon_3, \tag{115c}$$

where for the last inequality we also used the independence of messages W_1 and W_2.

We first develop the second constraint using the chain rule of mutual information and [43, Lemma 1]:

$$\begin{aligned} nR &\leq \sum_{t=1}^{n} I(W_1; Y_{\mathbf{d}_2,t}|Y^{t-1}_{2,\mathbf{d}_2}) + n\epsilon_2 \\ &\leq (1-\delta_{\text{s}}) \sum_{t=1}^{n} I(W_1; X_{\mathbf{d}_2,t}|Y^{t-1}_{2,\mathbf{d}_2}) + n\epsilon_2. \end{aligned} \tag{116}$$

We then jointly develop the first and the third constraints, where we also define $\epsilon' := \epsilon_1 + \epsilon_3$:

$$\begin{aligned} &2nR \\ &\leq I(W_1, W_2; V_1^{(n)}, Y^n_{1,\mathbf{d}_1}) + I(W_2; Y^n_{1,\mathbf{d}_2}|W_1, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}) + n\epsilon' \\ &\overset{(a)}{\leq} I(W_1, W_2; V_1^{(n)}) + I(W_1, W_2; Y^n_{1,\mathbf{d}_1}|V_1^{(n)}) \\ &\quad + I(W_2; Y^n_{2,\mathbf{d}_2}|W_1, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}) + n\epsilon' \\ &\leq I(W_1, W_2; V_1^{(n)}) + \sum_{t=1}^{n} I(W_1, W_2; Y_{1,\mathbf{d}_1,t}|V_1^{(n)}, Y^{t-1}_{1,\mathbf{d}_1}) \\ &\quad + \sum_{t=1}^{n} I(W_2; Y_{2,\mathbf{d}_2,t}|W_1, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}, Y^{t-1}_{2,\mathbf{d}_2}) + n\epsilon' \\ &\leq I(W_1, W_2; V_1^{(n)}) + (1-\delta_{\text{w}}) \sum_{t=1}^{n} I(W_1, W_2; X_{\mathbf{d}_1,t}|V_1^{(n)}, Y^{t-1}_{1,\mathbf{d}_1}) \\ &\quad + (1-\delta_{\text{s}}) \sum_{t=1}^{n} I(W_2; X_{\mathbf{d}_2,t}|W_1, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}, Y^{t-1}_{2,\mathbf{d}_2}) + n\epsilon' \\ &\leq I(W_1, W_2; V_1^{(n)}) + (1-\delta_{\text{w}}) \sum_{t=1}^{n} I(W_1, W_2; X_{\mathbf{d}_1,i}|V_1^{(n)}, Y^{t-1}_{1,\mathbf{d}_1}) \\ &\quad + (1-\delta_{\text{s}}) \sum_{t=1}^{n} I(W_1, W_2, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}; X_{\mathbf{d}_2,t}|W_1, Y^{t-1}_{2,\mathbf{d}_2}) + n\epsilon' \\ &\leq nM + n(1-\delta_{\text{w}})F \\ &\quad + (1-\delta_{\text{s}}) \sum_{t=1}^{n} I(W_2, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}; X_{\mathbf{d}_2,t}|W_1, Y^{t-1}_{2,\mathbf{d}_2}) + n\epsilon'. \end{aligned} \tag{117}$$

In (a), we used that the physically degradedness of the channel in (112) implies the Markov chain

$$(W_1, W_2, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}) \to Y^n_{2,\mathbf{d}_2} \to Y^n_{1,\mathbf{d}_2}.$$

Adding up (116) and (117) and letting $\epsilon_1, \epsilon_2, \epsilon_3$ tend to 0, we obtain the missing converse bound in (111), because

$$\begin{aligned} &I(W_2, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}; X_{\mathbf{d}_2,t}|W_1, Y^{t-1}_{2,\mathbf{d}_2}) + I(W_1; X_{\mathbf{d}_2,t}|Y^{t-1}_{2,\mathbf{d}_2}) \\ &= I(W_1, W_2, V_1^{(n)}, Y^n_{1,\mathbf{d}_1}; X_{\mathbf{d}_2,t}|Y^{t-1}_{2,\mathbf{d}_2}) \\ &\leq H(X_{\mathbf{d}_2,t}) \leq F. \end{aligned} \tag{118}$$

REFERENCES

[1] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," in *IEEE International Symposium on Wireless Communications Systems (ISWCS)*, Bruxelles, Belgium, August, 2015.

[2] S. Saeedi Bidokhti, R. Timo and M. Wigger, "Erasure Broadcast Networks with Receiver Caching," in *IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, July 2016.

[3] S. Saeedi Bidokhti, M. Wigger and R. Timo, "An Upper Bound on the Capacity-Memory Tradeoff of Degraded Broadcast Channels," in *International Symposium on Turbo Codes & Iterative Information Processing*, Brest, France, September 2016.

[4] E. Baştuğ, M. Bennis, E. Zeydan, M. A. Kader, A. Karatepe, A. S. Er and M. Debbah, "Big data meets telcos: A proactive caching perspective," *arXiv*, 1602.06215, February, 2016.
[5] P. Cerwall, P. Jonsson, R. Möller, S. Bävertoft, S. Carson, I. Godor, P. Kersch, A. Kälvemark, G. Lemne, and P. Lindberg, Ericsson Mobility Report, June, 2015.
[6] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, June, 2014.
[7] M. A. Maddah-Ali, U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856 – 2867, May 2014.
[8] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for small buffer users," *arXiv*, 1407.1935, November, 2015.
[9] Mohammad Mohammadi Amiri and Deniz Gndz, "Fundamental limits of caching: improved delivery rate-cache capacity trade-off," *arXiv*, Jan. 2016.
[10] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," arXiv: 1601.063834v2, Jan. 2016.
[11] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," in *IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, July, 2015.
[12] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in Proc. *IEEE International Symposium on Information Theory (ISIT),* Hong Kong, Jun. 2015, pp. 1691–1695.
[13] C. Tian, "A note on the fundamental limits of coded caching," arXiv: 1503.00010v1, Feb. 2015.
[14] C.-Y. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," *arXiv.org*, 1601.05690, January, 2016.
[15] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv*, 1502.03124, February, 2015.
[16] C.-Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *arXiv.org*, 1504.00553, April, 2015.
[17] K. Wan, D. Tuninetti, and P. Piantanida, "On the Optimality of Uncoded Cache Placement," arXiv:1511.02256, Nov. 2015.
[18] S. Wang, X. Tian and H. Liu, "Exploiting the unexploited of coded caching for wireless content distribution," in *IEEE International Conference on Computing, Networking and Communications (ICNC)*, February, 2015.
[19] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv*, 1504.01452, April, 2015.
[20] Y. Ugur, Z. H. Awan and A. Sezgin, "Cloud radio access networks with coded caching," *arXiv*, 1512.02385, December, 2015.
[21] P. Hassanzadeh, E. Erkip, J. Llorca,and A. Tulino, "Distortion-memory tradeoffs in Cache-aided wireless video delivery," in *Allerton Conference on Communications, Control and Computing*, Monticello (IL), USA, October, 2015.
[22] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," in *Allerton Conference on Communications, Control and Computation*, Monticello (IL), USA, October, 2015.
[23] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback," *arXiv*, 1511.03961, November, 2015.
[24] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, June, 2015.
[25] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *arXiv*, 1602.04207, February 2016.
[26] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *arXiv.org*, 1601.02460, January, 2016.
[27] B. Azari, O. Simeone, U. Spagnolini and A. Tulino, "Hypergraph-based analysis of clustered cooperative beamforming with application to edge caching," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 84 – 87, February, 2016.
[28] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: opportunistic cooperation for video streaming," *IEEE Transactions on Signal Processing*, vol. 63, no. 1, pp. 57 – 69, January, 2015.
[29] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849 – 869, February, 2016.
[30] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, December, 2013.
[31] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *arXiv*, 1503.00265, March, 2015.
[32] S. Sahraei and M. Gastpar, "Multi-library coded caching," *arXiv*, 1601.06016, January, 2016.
[33] J. Hachem, N. Karamchandani, and S. Diggavi, "Content caching and delivery over heterogeneous wireless networks," *arXiv*, 1404.6560, March, 2015.
[34] J. Hachem, N. Karamchandani and S. Diggavi, "Multi-level coded caching," in *IEEE International Symposium on Information Theory*, July, 2014.
[35] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," *arXiv*, 1409.7626, September 2014.
[36] R. Timo, S. Saeedi Bidokhti and M. Wigger, "A rate-distortion approach to caching," in *International Zurich Seminar on Communications (IZS)*, Zurich, Switzerland, March, 2016.
[37] R. Urbanke and A. Wyner, "Packetizing for the erasure broadcast channel with an internet application," technical report, 1999.
[38] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1469 –1482, April, 2006.
[39] Y. Wu and M. Wigger, "Coding schemes with rate-limited feedback that improve over the nofeedback capacity for a large class of broadcast channels," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2009–2033, 2016.
[40] A. Bracher and M. Wigger, "Feedback and partial message side-information on the semideterministic broadcast channel," *arXiv*, 1508.01880, August, 2015.
[41] G. Kramer and S. Shamai, "Capacity for classes of broadcast channels with receiver side information," in *IEEE Information Theory Workshop (ITW)*, Lake Tahoe, California, September, 2007.
[42] A. El Gamal and Y. H. Kim, "Network information theory," Cambridge University Press, 2012.
[43] A. F. Dana and B. Hassibi "The capacity region of multiple input erasure broadcast channels," in *IEEE International Symposium on Information Theory (ISIT)*, September, 2005.
[44] T. M. Cover and J. Y. Thomas. *Elements of Information Theory*, Wiley-Interscience, New York, 2006.