Wireless Coded Caching: A Topological Perspective

Jingjing Zhang and Petros Elia

***Abstract*—We explore the performance of coded caching in a SISO BC setting where some users have higher link capacities than others. Focusing on a binary and fixed topological model where strong links have a fixed normalized capacity 1, and where weak links have reduced normalized capacity $\tau < 1$, we identify — as a function of the cache size and τ — the optimal throughput performance, within a factor of at most 8. The transmission scheme that achieves this performance, employs a simple form of interference enhancement, and exploits the property that weak links attenuate interference, thus allowing for multicasting rates to remain high even when involving weak users. This approach ameliorates the negative effects of uneven topology in multicasting, now allowing all users to achieve the optimal performance associated to $\tau = 1$, even if τ is approximately as low as $\tau \geq 1-(1-w)^g$ where g is the coded-caching gain, and where w is the fraction of users that are weak. This leads to the interesting conclusion that for coded multicasting, the weak users need not bring down the performance of all users, but on the contrary to a certain extent, the strong users can lift the performance of the weak users without any penalties on their own performance. Furthermore for smaller ranges of τ, we also see that achieving the near-optimal performance comes with the advantage that the strong users do not suffer any additional delays compared to the case where $\tau = 1$.**

I. Introduction

Recently the seminal work in [1] introduced *coded caching* as a means of using caches at the receivers in order to induce multicasting opportunities that lead to substantial removal of interference. This breakthrough provided impressive throughput gains, and inspired a sequence of works such as [2]–[8].

Focusing on the single-stream BC, the work in [1] considered a single transmitter with access to a library of N files, serving a set of K users, each requesting a single file from the library. As is typical with caching techniques, the communication had two phases: the caching phase and the delivery phase. During the caching phase (off peak hours), each user could cache the equivalent of M files (corresponding to a fraction $\gamma \triangleq M/N$ of the library in each cache) without knowledge of future request. During the delivery phase (peak hours), which would commence upon notification of each user's requested file (one requested file per user), the transmitter would deliver (the remaining of) the single requested file to each user.

Emphasis in [1] was placed on the symmetric, error free, single-stream BC, where each link from the transmitter to any of the receivers was identical, with normalized capacity equal to 1 file per unit of time. For this topologically symmetric setting, it was shown that a delivery phase with delay $T(K) \triangleq \frac{K(1-\gamma)}{1+K\gamma}$ suffices to guarantee the delivery of any K requested files to the users. This was achieved by caching a fraction γ of each file at each cache, and then by using cache-aided multicasting to send the remaining information to $K\gamma+1$ users at a time. In this symmetric setting, the resulting coding gain $g_{max} \triangleq \frac{K(1-\gamma)}{T(K)} = 1+K\gamma$ far exceeded the local caching gains typically associated to receiver-side caching.

What was also noticed though is that, because of multicasting, the performance suffered when the links had unequal capacities. Such uneven topologies, where some users have weaker channels than others, introduce the problem that any multicast transmission that is meant for at least one weak user, could conceivably have to be sent at a lower rate, thus 'slowing down' the rest of the strong users as well. For example, if we were to naively apply the delivery scheme in [1] — which consisted of a sequential transmission of $\binom{K}{K\gamma+1}$ different XORs (one XOR for each subset of $K\gamma+1$ users) — we would have the case that even a single weak user would suffice for the performance to deteriorate such that $T(K,\tau) > T(K,\tau=1),\ \forall \tau < 1$. Such topological considerations[1] have motivated work such as that in [2] which — for the setting of the broadcast erasure channel — includes a 'balancing' solution where only weak users have access to caches, while strong users do not.

Our motivation is to mitigate the performance degradation that coded caching experiences when some link capacities are reduced. The key to mitigating this topology-induced degradation, is a simple form of interference enhancement which exploits the natural interference attenuation in the direction of the weak links, and which allows us to maintain — to a certain degree — a constant multicasting flow of normalized rate 1.

A. Cache-aided SISO BC

We consider the topologically-uneven wireless SISO K-user BC, where $K-W$ users have strong links with unit-normalized capacity, while the remaining W users have weak links with normalized capacity $\tau \in [0,1]$. In the model, where a single-antenna transmitter communicates to K single-antenna users, at time t, the received signal at user k, takes the form

$$y_{k,t} = \sqrt{P^{\tau_k}} h_{k,t} x_t + z_{k,t} \quad k = 1,2,\cdots,K \qquad (1)$$

where the input x_t has bounded power $\mathbb{E}\{|x_t|^2\} \leq 1$, where the fading $h_{k,t}$ and the noise $z_{k,t}$ are assumed to be Gaussian with zero mean and unit variance, and where the link strength is $\tau_k = 1$ for strong users, and $\tau_k = \tau$ for weak users. In this setting, the average received signal to noise ratio (SNR) for the link to user k is given as [2]

$$\mathbb{E}\{|\sqrt{P^{\tau_k}} h_{k,t} x_t|^2\} = P^{\tau_k}.$$

The authors are with the Mobile Communications Department at EURECOM, Sophia Antipolis, 06410, France (email: jingjing.zhang@eurecom.fr, elia@eurecom.fr). The work of Petros Elia was supported by the ANR Jeunes Chercheurs project ECOLOGICAL-BITS-AND-FLOPS.

[1] In wireless communications, there is a variety of topological factors — including propagation path loss, shadow fading and inter-cell interference [9] — which lead to having some links that are much weaker or stronger than others; a reality that has motivated a variety of works (e.g. [10]–[12]) relating to *generalized* degrees of freedom (GDoF).

[2] Additionally in the high P regime of interest here, it is easy to see that $Pr(|\sqrt{P^{\tau_k}} h_{k,t}|^2 \doteq P^{\tau_k}) = 1$. We here use $\doteq$ to denote *exponential equality*, i.e., we write $g(P) \doteq P^B$ to denote $\lim_{P\to\infty} \frac{\log_2 g(P)}{\log_2 P} = B$.

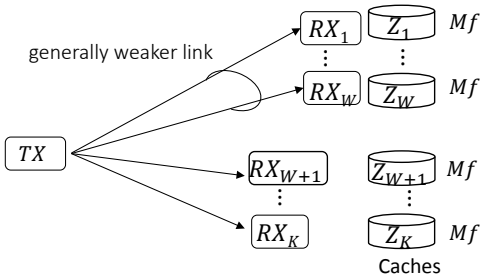

Fig. 1. Cache-aided K-user MISO BC.

For simplicity we assume that users $1,\dots,W$ are weak, and that users $W+1,\dots,K$ are strong.

We make the normalization, w.l.o.g., that each library file $W_n,\ n=1,\dots,N$, has size f (bits) which — in the high SNR setting of interest here — is set equal to $f=\log_2(P)$. Consequently the aforementioned capacity of a strong (interference free) link, is now *1 file per unit of time*, while the capacity of a weak link is τ files per unit of time. We consider that $N\geq K$. The cache Z_k of user k has size Mf bits, where M ($M\leq N$) defines the aforementioned normalized cache size

$$\gamma \triangleq \frac{M}{N}. \tag{2}$$

Our results consider the measure of performance $T(\tau)$ — in time slots, per file served per user — needed to complete the delivery process, *for any request*. After the aforementioned normalization $f=\log_2(P)$, this measure matches that in [1].

B. Notation and conventions

We will use $\mathcal{K}\triangleq\{1,2,\cdots,K\}$ to denote the (indices of the) set of all users, $\mathcal{W}\triangleq\{1,2,\cdots,W\}$ to denote the set of weak users, and $\mathcal{S}\triangleq\{W+1,\cdots,K\}$ to denote the set of strong users. We will also use $w\triangleq W/K$ to define the fraction of the users that are weak. We remind the reader that $\binom{n}{k}$ will be the n-choose-k operator, and $\oplus$ will be the bitwise XOR operation. If A and B are two sets, then $A\backslash B$ denotes the difference set. For a transmitted signal x, we will use $\text{dur}(x)$ to denote the transmission duration (in units of time) of that signal. We will use $\Gamma\triangleq\frac{KM}{N}=K\gamma$ to denote the cumulative (normalized) cache size, and for any integer L, we will use

$$T(L)\triangleq\frac{L(1-\gamma)}{1+L\gamma} \tag{3}$$

to denote the delay associated to the original coded caching solution in [1] with L strong users and no weak users ($\tau=1$).

Consequently we will use $T(K)\triangleq\frac{K(1-\gamma)}{1+K\gamma}$ to describe the performance for the case of $L=K$ users, as this was derived in [1] for integer $K\gamma\in\{0,1,\cdots,K\}$ (for the general $K\gamma$, the lower convex envelope of the integer points is achievable). Similarly $T(K-W)=\frac{(K-W)(1-\gamma)}{1+(K-W)\gamma}$ will simply correspond to the case of $L=K-W$, and $T(W)=\frac{W(1-\gamma)}{1+W\gamma}$ to the case of $L=W$, and *we stress that* $T(K),T(K-W),T(W)$ *all correspond to the case of* $\tau=1$. We here note that for clarity of exposition, we allow for an integer relaxation on $(K-W)\gamma$ and $W\gamma$. This relaxation, which allows for crisp expressions, will be lifted in Section V-C which, for completeness, presents the extension of the algorithm in [1] for any γ, using memory-sharing between files (see also [13]).

II. Throughput of topological cache-aided BC

The following describes, within a factor of 8, the optimal $T^*(\tau)$ as a function of K,W,γ,τ. It applies to the case of centralized placement[3]. The results use the expression

$$\bar{\tau}_{thr}=\frac{T(W)}{T(W)+T(K-W)}$$

and

$$\tau_{thr}=\begin{cases}1-\frac{\binom{K-W}{K\gamma+1}}{\binom{K}{K\gamma+1}}, & \text{for } W<K(1-\gamma)\\ 1, & \text{otherwise.}\end{cases} \tag{4}$$

Theorem 1: In the K-user topological cache-aided SISO BC with W weak users,

$$T(\tau)=\begin{cases}\frac{T(W)}{\tau}, & 0\leq\tau<\bar{\tau}_{thr}\\ \min\{T(K-W)+T(W),\frac{\tau_{thr}T(K)}{\tau}\}, & \bar{\tau}_{thr}\leq\tau\leq\tau_{thr}\\ T(K), & \tau_{thr}<\tau\leq 1\end{cases}$$

is achievable, and has a gap from optimal

$$\frac{T(\tau)}{T^*(\tau)}\leq 8 \tag{5}$$

that is always less than 8.

Proof: The achievable scheme is presented in Section III, while the corresponding gap is bounded in Appendix V-A. ■

What the above shows is that there are three regions of interest. In the first region where $\tau\geq\tau_{thr}$, despite the degradation in the link strengths, the performance of all users remains as if all links were uniformly strong (as if $\tau=1$). In this setting, instead of experiencing the phenomenon that the weak users 'pull down' the performance of all users, we observe the interesting effect of strong users bringing up the performance of the weak users, to the optimal $T(K)$ associated to $\tau=1$. It is easy to see that a naive sequential transmissions of the (scaled) XORs from [1] would result in τ_{thr} strictly less than one. The conclusion is that in this first region, the reduction in the capacity of the weak links τ, does not translate into a performance degradation. This is because, even when multicasting involves weak users, the employed superposition scheme allows for an overall multicasting rate of 1. Then, there is an intermediate region where there is a degradation in the overall performance by a factor $\frac{\tau_{thr}}{\tau}$ (rather than by a factor $\frac{1}{\tau}$). Finally there is the third region $\tau\leq\bar{\tau}_{thr}$, where due to the substantially limited capacity of the weak links, the transmission to the weak users becomes the bottleneck and the performance is dominated by the delay of serving the weak users, and it deteriorates by a factor $\frac{1}{\tau}$.

Example 1: ($K=500,W=50,\gamma=\frac{1}{50}$) Directly from the above we see that

$$T(\tau)=\begin{cases}\frac{24.5}{\tau}, & 0\leq\tau<0.36\\ \min\{68.6,\frac{30.7}{\tau}\}, & 0.36\leq\tau\leq 0.69\\ T(K)=44.5, & 0.69<\tau\leq 1\end{cases} \tag{6}$$

which means that, with a tenth of the users being weak, as long as $\tau\geq 0.69$, there is no performance degradation due to reduced-capacity links, and every user receives their file with delay $T(K)=\frac{K(1-\gamma)}{1+K\gamma}=44.5$ associated to $\tau=1$.

Regarding the region $\tau\in[0.69,1]$, the following quantifies the intuition that the topology threshold τ_{thr} (until which,

[3]The longer version [14] includes similar results for the decentralized case.

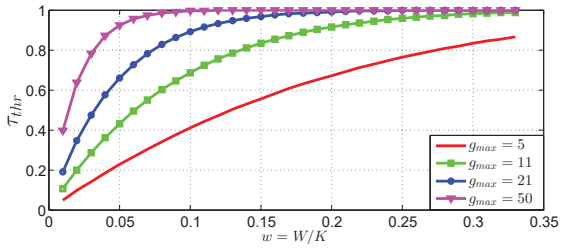

Fig. 2. τ_{thr} corresponding to distinct values for gains g_{max}. For example, for $g_{max} = 5$ and $w = 0.1$ then $\tau_{thr} \approx 0.4$.

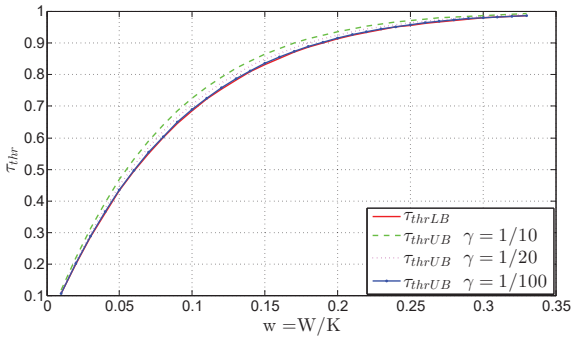

Fig. 3. $\tau_{thrLB} = 1-(1-w)^{g_{max}}$ denotes the lower bound of τ_{thr}, while $\tau_{thrUB} = 1-(1-w-\frac{w\gamma}{1-\gamma})^{g_{max}}$ denotes the upper bound.

capacity reductions do not degrade performance), is a function of the degree of multicasting (coding gain) $g_{max} \triangleq K\gamma + 1 = K(1-\gamma)/T(K)$ (see Fig.2 and Fig.4).

Corollary 1a: The threshold τ_{thr} which guarantees full-capacity performance $T(K)$, lies inside the region $\tau_{thr} \in [1-(1-w)^{g_{max}}, 1-(1-w-\frac{w\gamma}{1-\gamma})^{g_{max}}]$, which also means that

$$T(\tau) = T(K), \ \forall \tau \geq 1-(1-w)^{g_{max}} + \gamma^{g_{max}}.$$

As γ decreases, this threshold approaches (see Fig.3)

$$\tau_{thr} \approx 1-(1-w)^{g_{max}}.$$

Proof: The proof can be found in Appendix V-B in the longer version [14]. ■

We extend the above to the link-capacity threshold

$$\tau_{thr,G} \triangleq \arg\min\{\tau : T(\tau) \leq G \cdot T(K), \ G \geq 2\} \quad (7)$$

until which, the performance loss is restricted to a factor of $G \geq 2$. For example, for any $\tau \geq \tau_{thr,2}$, the scheme guarantees that $T(\tau) \leq 2T(K)$.

Corollary 1b: For any $\tau \geq \tau_{thr,G} = \frac{w}{1+w(g_{max}-1)}\frac{g_{max}}{G}$ ($G \geq 2$), the performance degradation is bounded as $T(\tau) \leq G \cdot T(K)$.

Proof: The proof is presented in Appendix V-B. ■

Example 2: ($w = \frac{1}{10}, g_{max} = 11$) Here, as we have seen, $\tau_{thr} = 0.686$, whereas

$$\tau_{thr,G} = \frac{0.55}{G}, \ G \geq 2 \quad (8)$$

which means that any link-capacity reduction down to, e.g., $\tau \geq \tau_{thr,2} = \frac{0.55}{2} = 0.275$, only comes with a performance deterioration of at most 2 ($T(\tau) \leq 2T(K), \ \forall \tau \geq 0.275$).

III. Coded Caching with Simple Interference Enhancement

We now focus on the scheme, for the cases in Theorem 1.

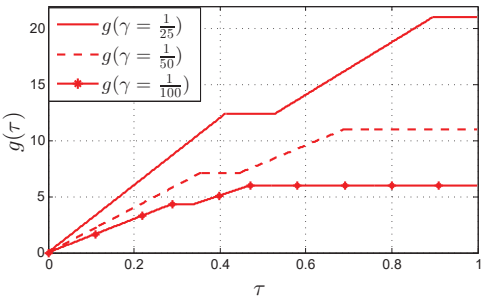

Fig. 4. The plot shows the gain as a function of τ when $K = 500, W = 50$. The horizontal lines represent the maximum gain g_{max} corresponding to $\tau = 1$, and reveal how these can be achieved even with lesser link capacities.

A. Scheme for $\tau > \tau_{thr}$

The following applies to the case where $W < K(1-\gamma)$. Note that when $W \geq K(1-\gamma)$, $\tau_{thr} = 1$.

1) Placement phase: The placement phase is identical to that in [1], where we recall that each file W_n, $n = 1, \ldots, N$ is equally split into $\binom{K}{\Gamma}$ subfiles $\{W_{n,\tau}\}_{\tau \in \Psi_\Gamma}$ where $\Psi_\Gamma \triangleq \{\tau \subset \mathcal{K} : |\tau| = \Gamma\}$, such that each cache Z_k is then filled according to $Z_k = \{W_{n,\tau}\}_{n \in [N], \tau \in \Psi_\Gamma, k \in \tau}$.

2) Delivery phase: Upon the requests $\{W_{R_k}\}_{k=1}^K$, the transmitter must deliver the remaining (uncached) subfiles $\{W_{R_k,\tau}\}_{k \notin \tau}$ for each user k.

We first recall from [1] that for any $\psi \in \Psi_{\Gamma+1} \triangleq \{\psi \in \mathcal{K} : |\psi| = \Gamma + 1\}$, then

$$X_\psi \triangleq \oplus_{k \in \psi} W_{R_k, \psi \setminus \{k\}} \quad (9)$$

suffices to deliver to each user $k \in \psi$, their requested file $W_{R_k,\psi\setminus\{k\}}$. To satisfy all requests $\{W_{R_k} \setminus Z_k\}_{k=1}^K$, the entire set $\mathcal{X}_\Psi \triangleq \{X_\psi\}_{\psi \in \Psi_{\Gamma+1}}$ consisting of $|\mathcal{X}_\Psi| = \binom{K}{\Gamma+1}$ folded messages (XORs), must be delivered. Each XOR has size

$$|X_\psi| = |W_{R_k,\tau}| = \frac{f}{\binom{K}{\Gamma}} \text{ (bits)}. \quad (10)$$

We distinguish between the subset of XORs $\mathcal{X}_{\Psi,s} \triangleq \{X_\psi : \forall \psi, \ s.t. \ \psi \cap \mathcal{W} = \emptyset\} \subset \mathcal{X}_\Psi$ that are only intended for strong users, and the remaining subset $\mathcal{X}_{\Psi,w} \triangleq \mathcal{X}_\Psi \setminus \mathcal{X}_{\Psi,s}$ that have at least one weak user as an intended recipient.

Let T_1 be the duration required to deliver all of $\mathcal{X}_{\Psi,w}$, to all weak users $k \in \mathcal{W}$. Let the transmission first take the form

$$x_t = c_t + b_t, \quad t \in [0, T_1] \quad (11)$$

where the power and rate are allocated such that

$$\mathbb{E}\{|c_t|^2\} \doteq P^0, \mathbb{E}\{|b_t|^2\} \doteq P^{-\tau}, r_t^{(c)} = \tau, r_t^{(b)} = 1-\tau \quad (12)$$

where $r_t^{(c)}$ (resp. $r_t^{(b)}$) denotes the prelog factor of the number of bits $r_t^{(c)} f$ carried by symbol c_t (resp. $r_t^{(b)}$) at time t. In the above, c_t will carry information from $\mathcal{X}_{\Psi,w}$, while b_t will carry the information from $\mathcal{X}_{\Psi,s}$. As we see, the reduced power of b_t guarantees that it does not interfere with weak users (at least not above the noise level).

During this period, the received signals $y_{k,t}$ take the form

$$y_{k,t} = \underbrace{\sqrt{P}h_{k,t}c_t}_{P} + \underbrace{\sqrt{P}h_{k,t}b_t}_{P^{1-\tau}} + \underbrace{z_{k,t}}_{P^0}, \quad k \in \mathcal{S} \quad (13)$$

$$y_{k,t} = \underbrace{\sqrt{P^\tau}h_{k,t}c_t}_{P^\tau} + \underbrace{\sqrt{P^\tau}h_{k,t}b_t}_{P^0} + \underbrace{z_{k,t}}_{P^0}, \quad k \in \mathcal{W} \quad (14)$$

allowing each weak user $k \in \mathcal{W}$ to directly decode c_t, and allowing each strong user $k \in \mathcal{S}$ to first decode c_t by treating b_t as noise, and to then decode b_t by removing c_t. This is achieved because the interference to the strong users was enhanced (see [15] and [16]) in order for it to be removed.

Depending on the size of $\mathcal{X}_{\Psi,w}$ and $\mathcal{X}_{\Psi,s}$, we will have two cases. In the first case, all the information in $\mathcal{X}_{\Psi,s}$ is delivered by b_t within the aforementioned duration T_1, and thus $T = T_1$. In the second case though, the delivery of $\mathcal{X}_{\Psi,s}$ takes longer than the delivery of $\mathcal{X}_{\Psi,w}$ (longer than T_1), in which case the remaining information is transmitted during an additional period of duration T_2, during which the transmission (as it is intended only for strong users) takes the simpler form

$$x_t = c_t, \quad t \in [T_1, T_1 + T_2] \tag{15}$$

during which the power and rate are set as

$$\mathbb{E}\{|c_t|^2\} \doteq P^0, \quad r_t^{(c)} = 1 \tag{16}$$

which allows each strong user to directly decode c_t.

In both cases, each strong user can decode $\mathcal{X}_{\Psi,w}$ and $\mathcal{X}_{\Psi,s}$, while each weak user can decode $\mathcal{X}_{\Psi,w}$, and the delivery process is completed.

3) Calculation of T : First, let us use $Q_{\bar{w}} \triangleq |\mathcal{X}_{\Psi,s}||X_\psi| = \frac{\binom{K-W}{\Gamma+1} f}{\binom{K}{\Gamma}}$ (bits) to denote the size (in bits) of $\mathcal{X}_{\Psi,s}$, and let us use $Q_w = |\mathcal{X}_\Psi||X_\psi| - Q_{\bar{w}}$ (bits) to denote the size of $\mathcal{X}_{\Psi,w}$. We now treat the aforementioned two cases.

a) Case 1a: $T_1 > \frac{Q_{\bar{w}}}{(1-\tau)f}$ (corresponds to $\tau \in [0, \tau_{thr}]$): Here $T = T_1$ is directly calculated, and takes the form

$$T = T_1 = \frac{Q_w}{\tau f} = \frac{1}{\tau}\big(1 - \frac{\binom{K-W}{\Gamma+1}}{\binom{K}{\Gamma+1}}\big)\frac{K(1-\gamma)}{1+K\gamma} = \frac{\tau_{thr} T(K)}{\tau}.$$

b) Case 1b: $T_1 \leq \frac{Q_{\bar{w}}}{(1-\tau)f}$ (corresponds to $\tau \in (\tau_{thr}, 1]$): The transition to this new case, happens as soon as $T_1 < \frac{Q_{\bar{w}}}{(1-\tau)f}$, which happens as soon as $\tau > \tau_{thr}$ (i.e., $\tau = \tau_{thr}$ is derived by setting $T_1 = \frac{Q_{\bar{w}}}{(1-\tau)f}$). Recall that now $T = T_1 + T_2$. We can easily calculate that the second period (during which we multicast to strong users at full rate) has duration

$$T_2 = \frac{Q_{\bar{w}} - (1-\tau) f T_1}{f}$$

where $Q_{\bar{w}} - (1-\tau) f T_1$ is the amount of the remaining data of $\mathcal{X}_{\Psi,s}$ that had not been handled during the first period of duration T_1. Adding the two components gives us

$$T = T_1 + T_2 = \frac{K(1-\gamma)}{1+K\gamma} = T(K) \tag{17}$$

which matches the aforementioned performance $T(K)$ corresponding to uniformly strong topology ($\tau = 1$).

B. Scheme for the case of $\tau \leq \tau_{thr}$

The following applies for all $W \leq K$. Here the idea is that, we treat the weak users separately from the strong users. While we generally transmit to both strong and weak users simultaneously, caching at the strong users is independent of the caching at the weak users, and each XOR is meant either for strong users exclusively, or for weak users exclusively. Transmission again takes the form $x_t = c_t + b_t$, and c_t will deliver the group of XORs meant for weak users, while b_t will deliver the group of XORs for the strong users.

For the case of weak users, the total information that will be sent is $fT(W)\log(P)$ bits, while for the strong users, this will be $fT(K-W)\log(P)$ bits. There will be again two cases, where the split is again a function of the amount of information that needs to be sent to the weak vs. to the strong users. In the first case, the transmission and allocation of power and rate, are the same as in (11) and (12), while in the second case they will be the same as in (15) and (16).

c) Case 2a: $\frac{fT(K-W)}{(1-\tau)f} < \frac{fT(W)}{\tau f}$ (corresponds to $\tau \in [\bar{\tau}_{thr}, \tau_{thr}]$): For this case — corresponding to the scenario where the delivery to the strong users does not take longer than the delivery to the weak users — T is calculated to be

$$T = \frac{fT(W)}{\tau f} = \frac{T(W)}{\tau}.$$

d) Case 2b: $\frac{fT(K-W)}{(1-\tau)f} \geq \frac{fT(W)}{\tau f}$ (corresponds to $\tau \in [0, \bar{\tau}_{thr}]$): In addition to the above mentioned $T_1 = \frac{T(W)}{\tau}$, the second period duration T_2 is readily calculated to be

$$T_2 = \frac{fT(K-W) - (1-\tau) f T_1}{f}$$

which eventually gives

$$T = T_1 + T_2 = T(K-W) + T(W). \tag{18}$$

Combining this with the result corresponding to case 1a gives

$$T = \min\{T(K-W) + T(W), \frac{\tau_{thr} T(K)}{\tau}\}.$$

IV. Conclusion

In this work we explored the behavior of coded caching in the topological broadcast channel (BC), identifying the optimal cache-aided performance within a multiplicative factor of 8. Our proposed scheme uses a simple form of interference enhancement to alleviate the negative effect of having to multicast to both strong and weak links. By showing that the optimal performance can be achieved even in the presence of weaker links, the work reveals a new role of coded caching which is to partially balance the performance between weaker and stronger users, and to a certain degree without any penalty to the performance of the stronger users.

V. Appendix

A. Proving the gap to optimal

To prove the gap to optimal in Theorem 1, we first recall from [13] (which corresponds to the case of $\tau = 1$) that $\frac{T(K)}{T^*(\tau=1)} \leq 4$. There are three distinct regions of τ that must be treated separately. Due to lack of space, we here consider only one case, where $\tau \in [\bar{\tau}_{thr}, \tau_{thr}]$. The rest of the cases are handled in the extended version [14] of this work.

We first recall that $T(K)$ is increasing with K, since

$$T(K) = \frac{K(1-\gamma)}{1+K\gamma} = \frac{1-\gamma}{\gamma}(1 - \frac{1}{1+K\gamma}).$$

This means that $T(K-W) \leq T(K)$ and $T(W) \leq T(K)$, and consequently that

$$\begin{aligned} T(\tau) &= \min\{T(K-W) + T(W), \frac{\tau_{thr} T(K)}{\tau}\} \\ &\leq T(K-W) + T(W) \leq 2T(K) \end{aligned} \tag{19}$$

which yields the desired

$$\frac{T(\tau)}{T^*(\tau)} \leq \frac{2T(K)}{T^*(\tau)} \leq \frac{2T(K)}{T^*(\tau=1)} \leq 8.$$

B. Proof of Corollary 1b

Let us recall from (19) that when $\bar{\tau}_{thr} \leq \tau \leq \tau_{thr}$ then

$$T(\tau) \leq T(K-W) + T(W) \leq 2T(K) \tag{20}$$

which, together with the fact that $G \geq 2$, implies that such a performance degradation (beyond a factor of 2), requires that $\tau < \bar{\tau}_{thr}$, which in turn says that the achievable $T(\tau)$ takes the form $T(\tau) = \frac{T(W)}{\tau}$. Applying this in the definition in (7), yields the presented $\tau_{thr,G}$.

C. Removing the integer relaxation constraint

To remove the aforementioned integer relaxation, we consider the extension of the centralized MN algorithm in [1], to any γ (not just when $K\gamma$ is an integer). This has already been addressed in [13] which plots the intermediate values. For the sake of completeness we proceed to explicitly describe the corresponding performance, achieved here by the memory-sharing scheme described below, holding for any γ and $\tau = 1$.

Proposition 1: In the K-user cache-aided SISO BC, with $N \geq K$ files and cache size such that $K\gamma \in [t, t+1], t = 0, 1, \cdots, K-1$, then

$$T''(K) = \frac{K-t}{t+1} + \frac{(K\gamma - t)(K+1)}{(t+1)(t+2)} \tag{21}$$

is achievable and it has a gap from optimal

$$\frac{T''(K)}{T^*} \leq 4 \tag{22}$$

that is less than 4.

The above maintains the gap from optimal of 4, simply because the interpolation gives an improved performance over the case where $K\gamma \in \{1, 2, \ldots, K\}$ (see also [13]). The expression coincides with the original $T(K)$ for integer values of $K\gamma$. The purpose of this proposition is to allow for the applicability of Theorem 1 without the integer relaxation assumption. With $T''(L)$ in place, Theorem 1 can apply, simply now with slightly different values for $\bar{\tau}_{thr}$ and τ_{thr}, which though are more complicated and which do not offer any additional insight and are thus omitted.

Below we briefly describe the scheme.

1) Proof of Proposition 1: Let $\Gamma = \frac{KM}{N} \in [t, t+1]$, for some $t = 0, 1, \cdots, K-1$. Let us start by splitting each file W_n into two parts $W_n^{(1)}$ and $W_n^{(2)}$, where $W_n^{(1)}$ has size $\big((t+1) - K\gamma\big)f$ and $W_n^{(2)}$ has size $(K\gamma - t)f$. Split each cache Z_k into two parts, $Z_{k,1}, Z_{k,2}$ such that $\frac{|Z_{k,1}|}{|Z_{k,2}|} = \frac{\big((t+1)-K\gamma\big)}{(K\gamma - t)}$. Focusing on the first part, apply the original MN algorithm, where now the library is $\{W_n^{(1)}\}_{n=1}^N$, the caches are $\{Z_{k,1}\}_{k=1}^K$, and caching is performed as though $K\gamma = t$, i.e., by splitting each half-file $W_n^{(1)}$ into $\binom{K}{t}$ equally-sized subfiles $W_{n,\tau}^{(1)}, \tau \in \Psi_t$ (each subfile now has size $((t+1) - K\gamma)f/\binom{K}{t}$), and by filling the caches according to $Z_{k,1} = \{W_{n,\tau}^{(1)}\}_{n\in[N], \tau \in \Psi_t, k \in \tau}$. Then simply create the sequence of $\binom{K}{t+1}$ XORs (where now each XOR is intended for $t+1$ users), the delivery of which requires

$$T^{(1)} = (t+1-K\gamma)\frac{\binom{K}{t+1}}{\binom{K}{t}} = \frac{K-t}{t+1}. \tag{23}$$

We then do the same for the second half of the files (second library $\{W_n^{(2)}\}_{n=1}^N$) except that now we substitute t with $t+1$, to get a corresponding duration of

$$T^{(2)} = (K\gamma - t)\frac{\binom{K}{t+2}}{\binom{K}{t+1}} = \frac{(K\gamma - t)(K+1)}{(t+1)(t+2)}. \tag{24}$$

Combining the two cases gives us

$$T = T^{(1)} + T^{(2)} = \frac{K-t}{t+1} + \frac{(K\gamma - t)(K+1)}{(t+1)(t+2)} \tag{25}$$

which completes the proof.

REFERENCES

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

[2] S. S. Bidokhti, M. A. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *CoRR*, vol. abs/1605.02317, 2016. [Online]. Available: http://arxiv.org/abs/1605.02317

[3] S. P. Shariatpanahi, A. S. Motahari, and B. H. Khalaj, "Multi-server coded caching," *CoRR*, vol. abs/1503.00265, 2015. [Online]. Available: http://arxiv.org/abs/1503.00265

[4] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," *CoRR*, vol. abs/1509.02074, 2015. [Online]. Available: http://arxiv.org/abs/1509.02074

[5] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *Proc. Allerton Conf. Communication, Control and Computing*, Monticello, Illinois, USA, Sep. 2015.

[6] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback," *CoRR*, vol. abs/1511.03961, 2015. [Online]. Available: http://arxiv.org/abs/1511.03961

[7] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *CoRR*, vol. abs/1602.04207, 2016. [Online]. Available: http://arxiv.org/abs/1602.04207

[8] M. A. Wigger, R. Timo, and S. Shamai, "Complete interference mitigation through receiver-caching in wyner's networks," *CoRR*, vol. abs/1605.03761, 2016. [Online]. Available: http://arxiv.org/abs/1605.03761

[9] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[10] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5534 – 5562, Dec. 2008.

[11] C. S. Vaze, S. Karmakar, and M. K. Varanasi, "On the generalized degrees of freedom region of the MIMO interference channel with no CSIT," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Aug. 2011.

[12] S. Karmakar and M. K. Varanasi, "The generalized degrees of freedom of the MIMO interference channel," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Aug. 2011.

[13] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *CoRR*, vol. abs/1501.06003, 2015. [Online]. Available: http://arxiv.org/abs/1501.06003

[14] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," *CoRR*, vol. abs/1606.08253, 2016. [Online]. Available: http://arxiv.org/abs/1606.08253

[15] A. G. Davoodi and S. A. Jafar, "Transmitter cooperation under finite precision csit: A GDoF perspective," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[16] I. Maric, R. Dabora, and A. J. Goldsmith, "Relaying in the presence of interference: Achievable rates, interference forwarding, and outer bounds," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4342–4354, July 2012.