# Resolving the Worst-User Bottleneck of Coded Caching: Exploiting Finite File Sizes

Hui Zhao, Antonio Bazco-Nogueras, Petros Elia
Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France
Email: {hui.zhao; antonio.bazco-nogueras; petros.elia}@eurecom.fr

***Abstract*—In this work, we address the worst-user bottleneck of coded caching, which is known to diminish any caching gains due to the fundamental requirement that the multicast transmission rate should be limited by that of the worst channel among the served users. We consider the quasi-static Rayleigh fading Broadcast Channel, for which we first show that the coded caching gain of the XOR-based standard coded-caching scheme completely vanishes in the low-SNR regime. Yet, we show that this collapse is not intrinsic to coded caching by presenting a novel scheme that can completely recover the caching gains. The scheme exploits an aspect that has remained unexploited: the shared side information brought about by the file size constraint. The worst-user effect is dramatically ameliorated because it is replaced by the worst-group-of-users effect, where the users within a group have the same side information and the grouping is decided before the channel or the demands are known.**

## I. Introduction

Cache-aided communications is a promising approach aiming at reducing congestion in communication networks. Toward this aim, the seminal paper of Maddah-Ali and Niesen [1] proposed *coded caching* as a means of speeding up content delivery by exploiting receiver-side cached content to remove interference.

The work in [1] considers the error-free (or equivalently, high-SNR) shared-link Broadcast Channel (BC), where a transmitter with access to a library of $N$ content files serves $K$ users. Each such user enjoys a local (cache) memory of size equal to the size of $M$ files, i.e., equal to a fraction $\gamma \triangleq \frac{M}{N} \in [0,1]$ of the library size. The scheme of [1], henceforth referred to as the *MN scheme*, involves a cache placement phase and a subsequent delivery phase. During the first phase, each file is split into a generally large number of subfiles, which are selectively placed in various caches. During the second phase, the communication is split into an also large number of *transmission stages*, and, at each of them, a different subset of $K\gamma + 1$ users is simultaneously served, thus providing a theoretical speed-up factor (or gain) of $K\gamma + 1$ as compared to the uncoded case.

In recent years, a variety of works have investigated coded caching under more realistic wireless settings, considering for example uneven channel qualities [2]–[4] or the role of Channel State Information (CSI) availability [5]–[7]. Unfortunately, it is the case that coded caching suffers from two major constraints that severely limit its gains. The first is often referred to as the "file-size constraint", which stems from having to split each file into a number of subfiles that generally dwarfs any realistic file sizes that we may encounter [7], [8]. Hence, the MN scheme, as well as most known schemes [8], will inevitably require many users to share the same cache content. For example, for the scheme in [1], a file size constraint can effectively force the $K$ users to share a smaller number of $\Lambda < K$ distinct cache states, which would yield a (file-size) constrained caching gain of $\Lambda\gamma + 1$ in the error-free scenario.

On the other hand, there is a seemingly unrelated constraint which stems from the fact that the XOR multicast transmissions are fundamentally and inevitably limited by the rate of the worst user that they address [9]. This constraint, often referred to as the "worst-user bottleneck" of coded caching, arises when users experience different channel strengths, and it is a constraint that is exacerbated as the $\mathrm{SNR}$ becomes smaller. This bottleneck may indeed render coded caching unsuitable for many wireless scenarios that naturally operate at SNR regimes that are most often not much higher than $0$–$10$ dB [10].

This latter bottleneck has sparked considerable research interest. For example, for the well-established quasi-static fading setting, it has been shown by [11] that, in a single transmit antenna setting with finite power or SNR, the effective gain does not scale as $K$ becomes larger even in the absence of file-size constraint. Moreover, the work in [12] exploits channel heterogeneity to serve users at a different rate, mainly by employing superposition coding for opportunistic scheduling [12]. Another notable work is [13], which groups together users that experience similar SNR, and which, after neglecting users with the weakest channels, delivers to each group separately. To date, for a single transmit antenna setting without user selection, no scheme is known to overcome this bottleneck.

### *Contributions and Organization*

In this work, we consider coded caching with centralized placement in the standard single-antenna BC, in the context of finite SNR and quasi-static fading. For this setting, we will show that the theoretical global caching gain of the file-size constrained MN scheme with respect to uncoded Time Division Multiplexing (TDM), which corresponds to $\Lambda\gamma+1$, deteriorates considerably at moderate SNR, and, in fact, it completely vanishes in the low-SNR regime. This low-SNR regime remains of interest because many wireless scenarios operate at below-moderate SNR values. As it turns out, this regime allows us to crisply and very reliably capture the worst-user effect of coded caching in the aforementioned realistic SNR ranges.

Then, we show that this collapse is not inherent to coded caching. We do so by presenting a novel scheme that recovers

This work is supported by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929 (ERC project DUALITY).

(without any user selection) the theoretical coded-caching gain in the presence of sufficiently many users. The proposed scheme builds on the inevitability of having users with identical cache content, and it employs multi-rate encoding that avoids XOR transmissions, thus allowing each user to receive at a rate that matches its single-link capacity. As we will see, this implies that, in the presence of finite file-sizes and sufficiently many users, the worst-user effect can be made negligible even in the absence of any user-selection or time-diversity technique.

*Notations:* $\mathbb{E}\{\cdot\}$ denotes the expectation operator, $|\cdot|$ denotes the cardinality operator of a set, and, for any $K \in \mathbb{N}$, $[K] \triangleq \{1, \ldots, K\}$. $X \sim \mathcal{Y}$ means that the random variable $X$ follows the distribution $\mathcal{Y}$. We assume that all the sets are ordered.

## II. System Model and Problem Definition

We consider the Rayleigh fading BC in which a single-antenna transmitter serves a set of $K$ users. As mentioned before, each user requests a file from a library $\mathcal{F} = \{W_n\}_{n=1}^N$ of $N$ files, and each user is assisted by a cache of normalized size $\gamma \in [0, 1]$. We will consider an arbitrary number $\Lambda$ of different allowable cache states, and we will assume for simplicity that $K$ is an integer multiple of $\Lambda$.

The received signal at user $k$ is given by $Y_k = H_k X + Z_k$, where $H_k$ denotes the channel coefficient for user $k$, $X$ denotes the transmit signal satisfying an average power constraint $\mathbb{E}[|X|^2] \leq P$, and $Z_k$ denotes the zero-mean, unit-power, additive white Gaussian noise at user $k$. Each user $k$ experiences an instantaneous SNR of $\mathrm{SNR}_k = P|H_k|^2$, and an average SNR of $\rho \triangleq \mathbb{E}_H\{\mathrm{SNR}_k\}$. As is common in the coded caching literature (cf. [11]), we will assume that $H_k$ remains fixed during a transmission stage, but may change between different transmission stages. We will further assume that the users experience statistically symmetric Rayleigh fading.

As with various other works that study coded caching under quasi-static fading [2], [11], we will adopt the transmission rate[1] as the metric of interest. Toward this, we define the instantaneous rate $r_k$ as the maximal rate that can be transmitted to user $k$ for the instantaneous channel realization, and similarly we will consider the *average rate* $\mathbb{E}_H\{r_k\}$ to be the above rate, averaged over the fading statistics.[2]

In this context, a coded caching scheme seeks to provide an *effective coded-caching gain*, where this effective gain represents the true (multiplicative) speed up factor, at finite SNR, that the said scheme offers over the average rate obtained by TDM. This effective gain is contrasted to the (ideal, or high-SNR) *nominal coded-caching gain*, which is the gain $\Lambda\gamma + 1$ provided by the file-size constrained coded caching in the error-free scenario with fixed and identical link capacities.

The proposed scheme and the analysis here derived are motivated by the fact that the effective gain of the MN scheme collapses at low SNR, which will be proven in Section IV. Furthermore, this collapse is irrespective of $\Lambda$ and $K$, i.e., it happens even in the absence of file-size constraints.

[1]We recall that, in the quasi-static Rayleigh fading scenario, the typical metric of the worst-case delivery time does not have an expectation.

[2]The long-term average $\mathbb{E}_H\{r_k\}$ should not be confused with the ergodic rate, which implies an ability to encode over several fading realizations [11].

## III. Aggregated Coded-Caching Scheme

We now introduce a novel scheme that we denote as the *Aggregated Coded-Caching* scheme (or ACC).

The ACC scheme clusters the users into $\Lambda$ *groups* of $B = K/\Lambda$ users per group, such that every member of the same group is assigned identical cache content. As we have seen, this is essentially inevitable under realistic file-size constraints.

Our scheme follows a standard clique-based approach [1], such that the transmission is divided into *transmission stages* that experience a clique-side information pattern. In particular, as in [1], for each such stage, any desired subfile $W_i'$ of some served user $i$ can be found in the cache of every other user involved in that same transmission stage. This approach defines a side-information structure which was addressed in the following well known result from [14].

**Proposition 1** ([14, Thm. 6])**.** *The capacity region of a $t$-user Gaussian BC, where each user $i \in [t]$ is endowed with SNR equal to $\mathrm{SNR}_i$ and requests message $W_i'$ while having access to side information $\overline{\mathcal{W}}_i = \{W_j'\}_{j \neq i}$, is given by*

$$\mathcal{C} = \big\{(R_1, \cdots, R_t) : 0 \leq R_i \leq \log_2(1 + \mathrm{SNR}_i),\ i \in [t]\big\}.$$

Proposition 1 implies that, under this particular configuration of side information, each user can achieve its own point-to-point capacity. There are various optimal schemes for this setting [15], [16], and the proposed ACC scheme can remain oblivious to the encoding scheme employed.[3]

### A. Aggregated Coded-Caching Design

We proceed with the description of the placement and delivery phases of the ACC scheme. At the end, we will also present a small clarifying example.

*1) Placement phase:* This phase begins by arbitrarily splitting the $K$ users into $\Lambda$ ordered groups of $B = \frac{K}{\Lambda}$ users each. Placement is exactly as in [18], and thus it simply applies the MN placement of the $\Lambda$-user problem, such that each user of the same group shares the same cache content. In particular, each file $W_n$, $n \in [N]$, is partitioned into $\binom{\Lambda}{\Lambda\gamma}$ segments as $W_n \to \big\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\big\}$. Then, each user in group $g \in [\Lambda]$ stores all the subfiles in the following set $\mathcal{Z}_g = \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, \mathcal{T} \ni g, \forall n \in [N]\}$.

*2) Delivery phase:* The delivery phase is split into $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, where each stage involves a set $\mathcal{G} \subseteq [\Lambda]$ of $|\mathcal{G}| = \Lambda\gamma + 1$ groups. During each stage, the transmitter *simultaneously* delivers to $|\mathcal{G}| = \Lambda\gamma + 1$ users, each from a different group in set $\mathcal{G}$. The users within each group are served in a TDM round-robin manner. For a given set of $\Lambda\gamma + 1$ users simultaneously served, the transmitter employs a multi-rate scheme that achieves the capacity in Proposition 1.

[3]In terms of practicality, it is known that schemes such as nesting BPSK into M-QAM constellations [2] can approach this optimal capacity and can in fact achieve the single-user capacity insofar as we restrict ourselves to QAM modulations [17]. If simplicity necessitates, such simple encoding schemes can be directly applied in our cache-aided setting, with only minor losses.

**Algorithm 1:** Transmission stage for a set of groups $\mathcal{G}$

**Initialize** $\mathbf{v} \in \mathbb{Z}^{|\mathcal{G}|}$ as $\mathbf{v}(i) \longleftarrow 1$ for any $i \in [|\mathcal{G}|]$
**Initialize** finished groups $\longleftarrow 0$
**while** Number of finished groups $\neq |\mathcal{G}|$ **do**
  **Transmit**
    $X_{\mathcal{G},\mathbf{v}} \longleftarrow \mathcal{X}\Big(\big\{W^{\mathcal{G}\setminus\{\mathcal{G}(i)\}}_{d_{\mathbf{v}(i)}} \,\big|\, i \in [|\mathcal{G}|] \text{ and } \mathbf{v}(i) \leq B\big\}\Big)$
  **until** *A served user* $\mathbf{v}(i), i \in [|\mathcal{G}|]$, *decodes its subfile*
  Set $i^\star$ as the index of the group $\mathcal{G}(i^\star)$ whose user $\mathbf{v}(i^\star)$ has decoded its subfile
  **if** $\mathbf{v}(i^\star) = B$ **then**
    finished groups $\longleftarrow$ finished groups $+ 1$
  $\mathbf{v}(i^\star) \longleftarrow \mathbf{v}(i^\star) + 1$

Once a user decodes its desired subfile, the transmitter starts serving another user from the same group without delay, while continuing the transmission to the other users, which is possible because the users in the same group cache the same content.

The transmitted signal when the transmitter delivers the subfiles $A_1, \ldots, A_{|\mathcal{G}|}$ is denoted by $\mathcal{X}\big(A_1, \ldots, A_{|\mathcal{G}|}\big)$. Let us recall that $\mathcal{G}$ is ordered and let $\mathcal{G}(i)$ denote the $i$-th group in $\mathcal{G}$, $i \in [|\mathcal{G}|]$; consistently, the vector $\mathbf{v} \in \mathbb{Z}^{|\mathcal{G}|}$ represents[4] the set of currently served users, such that $\mathbf{v}(i)$ tells us which user of the group $\mathcal{G}(i)$ is currently being served, where $\mathbf{v}(i) \in [B]$. The transmit signal for serving users $\mathbf{v}$ of the groups in $\mathcal{G}$ is

$$X_{\mathcal{G},\mathbf{v}} = \mathcal{X}\big(\{W^{\mathcal{G}\setminus\{\mathcal{G}(i)\}}_{d_{\mathbf{v}(i)}}\}_{i\in[|\mathcal{G}|]}\big), \tag{1}$$

where $d_{\mathbf{v}(i)} \in [N]$ denotes the file index requested by user $\mathbf{v}(i)$.

Algorithm 1 presents the transmission for a specific set of groups $\mathcal{G}$. Every time the user of some group $\mathcal{G}(i')$ obtains its subfile, $\mathbf{v}(i')$ is updated[5] as $\mathbf{v}(i') \leftarrow \mathbf{v}(i') + 1$. This process is repeated until all the users in all the groups in $\mathcal{G}$ are served. If every user of a group has obtained its intended subfile, the transmission can be composed only of the remaining groups. Algorithm 1 is iterated over all possible $\binom{\Lambda}{\Lambda\gamma+1}$ sets $\mathcal{G}$. After this, the $K$ users can obtain their requested files. We reemphasize that the ACC scheme does not apply user selection. Let us proceed with a simple clarifying example.

**Example 1.** *Consider a transmission stage that serves groups* $\{1,2,3\} =: \mathcal{G}$*, where each group is composed of* $B = 3$ *users. To simplify the explanation of this example, let us denote the* $b$*-th user of the (ordered) group* $g$ *as* $U_{g,b}$*, and let* $W'_{g,b}$ *denote the subfile intended for this user. Let us further assume that the normalized capacity of each user is as in the next table:*

| | User 1 | User 2 | User 3 |
|---|---|---|---|
| Group 1 | 1 | 0.25 | 0.2 |
| Group 2 | 0.2 | 1 | 0.25 |
| Group 3 | 0.25 | 1 | 0.2 |

*where the capacity is expressed in subfiles per unit of time.*

[4]The dependence of $\mathbf{v}$ on the time index and on $\mathcal{G}$ is omitted for simplicity.

[5]We are actually incurring an abuse of notation in (1) and Algorithm 1. Specifically, when a group updates its served user, the transmitter continues encoding the partially-decoded subfiles taking into account that there only remains a part of such subfiles to be transmitted. This is intuitive from Fig. 1.

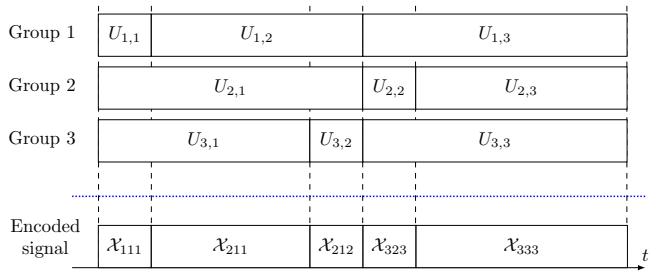


Fig. 1: ACC scheme for a nominal coded-caching gain of 3: Delay depends on the average per-user rate within the group. $\mathcal{X}_{abc}$ denotes the encoded signal for users $a$, $b$, and $c$, of groups 1, 2, and 3, respectively.

*This means that the point-to-point capacity of users* $U_{1,1}$*,* $U_{2,2}$*, and* $U_{3,2}$ *is four times the capacity of users* $U_{1,2}$*,* $U_{2,3}$*, and* $U_{3,1}$*, and five times the capacity of* $U_{1,3}$*,* $U_{2,1}$*, and* $U_{3,3}$*. The encoded signal for this example is illustrated in Fig. 1. Initially, the first user of each group is selected to be served, and the transmitter sends* $\mathcal{X}\big(W'_{1,1}, W'_{2,1}, W'_{3,1}\big)$*. Following the result of Proposition 1, each user can decode its own subfile at a rate matching its single-user capacity (*$\log_2(1 + \mathrm{SNR}_{g,b})$*).*

*After a single unit of time, user* $U_{1,1}$ *has successfully decoded its subfile. Hence,* $U_{1,1}$ *is replaced by* $U_{1,2}$*, and the transmitter sends* $\mathcal{X}\big(W'_{1,2}, W'_{2,1}, W'_{3,1}\big)$*. The key is that we can replace any of the users sharing the cache, because all of them can cache out the subfiles intended by the users of the other groups in* $\mathcal{G}$*, and vice versa. Thus, every time a user obtains its subfile, a new member of the same cache group replaces this user: After the fourth unit of time,* $U_{3,1}$ *obtains its subfile, and is replaced by* $U_{3,2}$*. Then, the transmitter sends* $\mathcal{X}\big(W'_{1,2}, W'_{2,1}, W'_{3,2}\big)$*. After the fifth unit of time, all the three served users obtain their desired subfiles, the transmitter begins to send* $\mathcal{X}\big(W'_{1,3}, W'_{2,2}, W'_{3,3}\big)$*, and so forth.*

## IV. Average Rate Analysis

We analyze the average rate of the MN and ACC schemes. We recall that, under Rayleigh fading, the probability density function (PDF) and cumulative distribution function (CDF) of $\mathrm{SNR}_k$ are respectively $f_{\mathrm{SNR}_k}(x) = {}^1\!/_\rho \, e^{-\frac{x}{\rho}}$, and $F_{\mathrm{SNR}_k}(x) = 1 - e^{-x/\rho}$, for $x \geq 0$. We will use $\mathrm{SNR}_{g,b}$, $f_{\mathrm{SNR}_{g,b}}(x)$, $F_{\mathrm{SNR}_{g,b}}(x)$ to refer to the SNR, PDF, and CDF corresponding to the $b$-th user of the group $g$, where $b \in [B]$ and $g \in [\Lambda]$.

### A. Average Rate of the MN and the ACC Schemes

*1) MN scheme:* Since the MN scheme serves $\Lambda\gamma + 1$ users at a time, let us consider the delivery to a particular group $\mathcal{G}$ of $\Lambda\gamma + 1$ users. From the multicast capacity theorem in [9], the maximum instantaneous rate for any user $i \in \mathcal{G}$ is given by

$$r^{(\mathrm{MN})}_{i,\mathcal{G}} = \log_2\big(1 + \min_{k\in\mathcal{G}} \mathrm{SNR}_k\big) \quad \text{bits/s/Hz}, \tag{2}$$

and thus the instantaneous sum rate is simply $\sum_{i\in\mathcal{G}} r^{(\mathrm{MN})}_{i,\mathcal{G}}$, since we are serving $|\mathcal{G}|$ users simultaneously. Consequently, the average sum rate for that specific group $\mathcal{G}$ takes the form

$$\bar{R}^{(\mathrm{MN})}_{\mathcal{G}} \triangleq \mathbb{E}_H\Big\{\sum_{i\in\mathcal{G}} r^{(\mathrm{MN})}_{i,\mathcal{G}}\Big\} = \frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\big\{\ln(1 + \min_{k\in\mathcal{G}} \mathrm{SNR}_k)\big\} \tag{3}$$

which follows because the users are statistically equivalent, which in turn also implies that, for any group $\mathcal{G}' \subseteq [\Lambda]$ of size $|\mathcal{G}'| = \Lambda\gamma + 1$, the average sum rate $\bar{R}^{(\mathrm{MN})} = \bar{R}^{(\mathrm{MN})}_{\mathcal{G}'}$ remains the same. Naturally, the average rate under the TDM scheme, which we denote as $\bar{R}^{(\mathrm{TDM})}$, is a special case of $\bar{R}^{(\mathrm{MN})}$ which is obtained by setting $|\mathcal{G}| = 1$.

Note that, even if we allowed multi-rate transmission in the MN scheme, we cannot start transmitting another XOR until the slowest user decodes the previous XOR, because it will create interference (cf. [19]).

*2) ACC scheme:* Due to the symmetry of the ACC scheme and the statistical symmetry of the channel, we will here focus on a set $\mathcal{G}$ of $|\mathcal{G}| = \Lambda\gamma + 1$ user groups, where we recall that each group is composed of $B$ users.

As explained in Section III, the ACC scheme allows us to serve some user $b$ of group $g$ at its own point-to-point capacity, and it allows us to immediately start serving another user of the same group as soon as the delivery to that user $b$ is completed. Furthermore, in the ACC scheme, the delivery to a set $\mathcal{G}$ of groups is completed when all the groups of that set are finished. Consequently, the (per-user average) instantaneous rate with which any group $j$ in the set $\mathcal{G}$ is served is here captured by

$$r^{(\mathrm{ACC})}_{j,\mathcal{G}} = \min_{g\in\mathcal{G}} \frac{1}{B}\sum\nolimits_{b=1}^{B} \log_2(1+\mathrm{SNR}_{g,b}) \quad \text{bits/s/Hz}, \tag{4}$$

and thus, by applying the same reasoning as in (2)–(3), we obtain that the average ACC rate takes the form

$$\bar{R}^{(\mathrm{ACC})} = \frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\Big\{ \min_{g\in\mathcal{G}} \frac{1}{B}\sum\nolimits_{b=1}^{B} \ln\big(1+\mathrm{SNR}_{g,b}\big)\Big\}. \tag{5}$$

Before analyzing the gain associated to the above rate, we note that, by comparing (5) with (3), we can see how the worst-user effect is essentially averaged out into a cumulative "worst-group" effect. Furthermore, the ACC scheme for $B = 1$ corresponds to the MN scheme.

### *B. Effective Gain of the MN and the ACC Schemes*

The complete exposition of the finite-SNR performance of the MN and ACC schemes can be found in the journal version of this work [19]. Due to lack of space, we choose here to highlight two important results. The first result concerns the MN scheme, and it reveals that its effective gain vanishes entirely at low SNR, irrespective of $K$ and $\Lambda$. The second result concerns the ACC scheme and it demonstrates that, for sufficiently large $B$, the ACC scheme entirely recovers the nominal high-SNR gain of $\Lambda\gamma + 1$ for any SNR value.

**Lemma 1.** *For any value of $K$ and $\Lambda$, the MN scheme satisfies*

$$\lim_{\rho\to 0} \frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}} = 1. \tag{6}$$

*Thus, its effective gain entirely vanishes at low SNR.*

*Proof.* The proof first considers the fact that $\min_{g\in\mathcal{G}}\{\mathrm{SNR}_g\}$ follows an exponential distribution with rate $|\mathcal{G}|/\rho$. Consequently, directly from [20, Eq. (15.26)], the above average rate takes the form $\bar{R}^{(\mathrm{MN})} = -\frac{|\mathcal{G}|}{\ln 2} e^{|\mathcal{G}|/\rho}\cdot \mathrm{Ei}\left(-|\mathcal{G}|/\rho\right)$, where $\mathrm{Ei}(\cdot)$ represents the exponential integral function [21]. Setting $|\mathcal{G}| = 1$ directly yields the average rate of TDM. Consequently, the ratio can be written as

$$\frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}} = |\mathcal{G}| e^{\frac{|\mathcal{G}|-1}{\rho}} \cdot \frac{\mathrm{Ei}\left(-|\mathcal{G}|/\rho\right)}{\mathrm{Ei}\left(-1/\rho\right)}. \tag{7}$$

Given that $-e^{-x}\ln(1+\frac{1}{x}) < \mathrm{Ei}(-x) < \frac{-e^{-x}}{2}\ln(1+\frac{2}{x})$ [21], we can upper bound the numerator and lower bound the denominator to obtain that

$$\lim_{\rho\to 0} \frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}} \le \lim_{\rho\to 0} \frac{|\mathcal{G}|}{2}\frac{\ln(1+\frac{2\rho}{|\mathcal{G}|})}{\ln(1+\rho)} = 1. \tag{8}$$

By interchanging the bounds to lower bound the ratio, we obtain that the limit is also lower bounded by 1, which concludes the proof of Lemma 1. □

It is worth noting that the collapse of the gain in Lemma 1 holds for any scheme which requires decoding of single XORs.

We proceed to prove that, for $K/\Lambda \gg 1$, the worst-user effect can be entirely eradicated.

**Lemma 2.** *For any average SNR $\rho$, the ACC scheme guarantees*

$$\lim_{B\to\infty} \frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{TDM})}} = \Lambda\gamma + 1, \tag{9}$$

*and thus, its effective gain matches the nominal gain for any value of SNR.*

*Proof.* Since $\mathbb{E}\left\{|\ln\left(1+\mathrm{SNR}_{g,b}\right)|\right\} < \infty$, it follows from the Strong Law of Large Numbers that, as $B\to\infty$,

$$\lim_{B\to\infty} \frac{1}{B}\sum\nolimits_{b=1}^{B} \ln\left(1+\mathrm{SNR}_{g,b}\right) = \mathbb{E}\left\{\ln\left(1+\mathrm{SNR}_{g,b}\right)\right\}, \tag{10}$$

except for zero-probability events. Now, by taking into account that $\ln(1+x)\le x$ for any $x>0$, we obtain that

$$\mathbb{E}_H\Big\{\min_{g\in\mathcal{G}}\frac{1}{B}\sum_{b=1}^{B}\ln(1+\mathrm{SNR}_{g,b})\Big\} \le \mathbb{E}_H\Big\{\frac{1}{B}\sum_{b=1}^{B}\mathrm{SNR}_{g,b}\Big\} \overset{(a)}{=} \rho < \infty, \tag{11}$$

where $(a)$ follows because $\mathrm{SNR}_{g,b} \sim \mathrm{Exp}(1/\rho)$ and thus $\frac{1}{B}\sum_{b=1}^{B}\mathrm{SNR}_{g,b} \sim \mathrm{Gamma}(B, \frac{B}{\rho})$. From (10)–(11), we can apply Lebesgue's Dominated Convergence Theorem [22, Theo. 16.4] to interchange expectation and limit to show that

$$\begin{aligned}\lim_{B\to\infty} &\frac{\frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\left\{\min_{g\in\mathcal{G}}\frac{1}{B}\sum_{b=1}^{B}\ln(1+\mathrm{SNR}_{g,b})\right\}}{\frac{1}{\ln 2}\mathbb{E}_H\left\{\ln\left(1+\mathrm{SNR}_{g,b}\right)\right\}} \\ &\overset{(a)}{=} |\mathcal{G}|\ \mathbb{E}_H\Big\{\min_{g\in\mathcal{G}}\lim_{B\to\infty}\frac{\frac{1}{B}\sum_{b=1}^{B}\ln(1+\mathrm{SNR}_{g,b})}{\mathbb{E}_H\left\{\ln\left(1+\mathrm{SNR}_{g,b}\right)\right\}}\Big\} \qquad (12)\\ &\overset{(b)}{=} |\mathcal{G}| = \Lambda\gamma+1, \qquad (13)\end{aligned}$$

where $(b)$ is due to (10), and where $(a)$ derives from the Dominated Convergence Theorem and the fact that the minimum of several continuous functions is a continuous function. □

The above directly implies that, in the low SNR regime with large $B$, the ACC scheme improves upon the MN scheme (for any $\Lambda$) by a factor of $\Lambda\gamma + 1$.

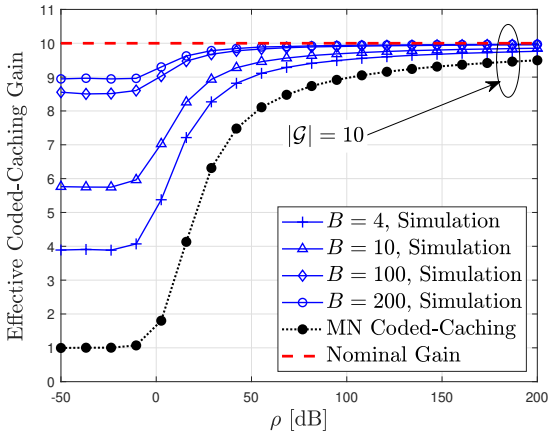


Fig. 2: Effective coded-caching gain versus $\rho$ for $|\mathcal{G}| = 10$.

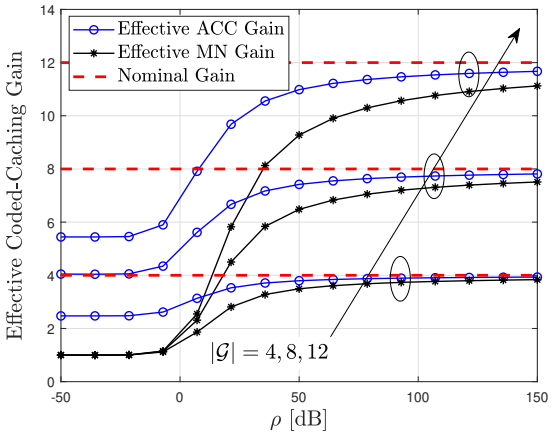


Fig. 3: Effective coded-caching gain versus $\rho$ for $B = 6$.

In Figs. 2–3, we present the effective coded-caching gains of the ACC and MN schemes versus $\rho$ for different values of the number of users per group ($B$) and nominal gain ($|\mathcal{G}|$). The gains are obtained from evaluating the average rates given in (3) and (5), whose exact expressions are derived in the journal version [19]. As expected, the effective gains of both schemes converge to the nominal gain as $\rho$ increases. However, the convergence of the ACC scheme is much faster than that of the MN scheme and, furthermore, the convergence of the ACC scheme becomes faster as $B$ grows. Figs. 2–3 also show the significant advantage of the ACC scheme in terms of average rate at low SNR, even for a small group size ($B = 4, 6$).

## V. Conclusions

We have shown that the key components of the coded-caching schemes achieving the optimal gains in the high-SNR regime, namely dedicated caches and XOR-based transmissions, are not suitable for the low-SNR regime inasmuch as the performance converges to that of simple TDM transmission without coded caching. To overcome this problem, we have designed a transmission scheme, so-called ACC scheme, which exploits the unavoidable nature of the subpacketization bottleneck in a simple way so as to reduce and asymptotically remove the worst-user bottleneck. Remarkably, the fact that users share the same cache content provides a time diversity effect even if there is no luxury for such an effect. The preservation of the asymptotic gains at low SNR under quasi-static fading, which were thought to vanish, evidences that coded caching has the ability to delivery multimedia content (generally involving a large volume of data) even in the presence of previously prohibitive values of low-to-moderate SNR. This result fosters the use of coded caching in settings such as satellite or cell-free networks, where most of the users are located in the edge area and suffer from heavy path-loss or heavy shadowing.

## References


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

[2] A. Tang, S. Roy, and X. Wang, "Coded caching for wireless backhaul networks with unequal link rates," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 1–13, Jan. 2018.

[3] S. Saeedi Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 6999–7019, Nov. 2019.

[4] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gündüz, "Coded caching with asymmetric cache sizes and link qualities: The two-user case," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6112–6126, Sep. 2019.

[5] E. Lampiris, J. Zhang, and P. Elia, "Cache-aided cooperation with no CSIT," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 2960–2964.

[6] E. Lampiris and P. Elia, "Achieving full multiplexing and unbounded caching gains with bounded feedback resources," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2018, pp. 1440–1444.

[7] ——, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[8] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5524 – 5537, Oct. 2016.

[9] N. Jindal and Z. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2006, pp. 1841–1845.

[10] Teltonika-Networks. Mobile signal strength recommendations. [Online]. Available: https://wiki.teltonika-networks.com/view/Mobile_Signal_Strength_Recommendations

[11] K.-H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.

[12] A. M. Daniel and W. Yu, "Optimization of heterogeneous coded caching," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1893–1919, Mar. 2020.

[13] B. Tegin and T. M. Duman, "Coded caching with user grouping over wireless channels," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 920–923, Jun. 2020.

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

[15] G. Kramer and S. Shamai, "Capacity for classes of Broadcast Channels with receiver side information," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2007, pp. 313–318.

[16] B. Asadi, L. Ong, and S. J. Johnson, "Optimal coding schemes for the three-receiver AWGN BC with receiver message side information," *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5490–5503, Oct. 2015.

[17] F. Xue and S. Sandhu, "PHY-layer network coding for broadcast channel with side information," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2007, pp. 108–113.

[18] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.

[19] H. Zhao, A. Bazco Nogueras, and P. Elia, "Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes," 2020, submitted to IEEE Trans. Wireless Commun. (TWC).

[20] M. K. Simon and M.-S. Alouini, *Digital communication over fading channels*. John Wiley & Sons, 2005.

[21] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1970, vol. 55.

[22] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. Wiley, 1995.