# Wireless Coded Caching Can Overcome the Worst-User Bottleneck by Exploiting Finite File Sizes

Hui Zhao, *Graduate Student Member, IEEE,* Antonio Bazco-Nogueras, *Member, IEEE,* Petros Elia, *Member, IEEE*

***Abstract*—We address the worst-user bottleneck of wireless coded caching, which is known to severely diminish cache-aided multicasting gains due to the fundamental worst-channel limitation of multicasting. We consider the quasi-static Rayleigh fading Broadcast Channel, for which we first show that the effective coded caching gain of the standard XOR-based coded-caching scheme completely vanishes in the low signal-to-noise ratio (SNR) regime. Then, we reveal that this collapse is not intrinsic to coded caching. We do so by presenting a novel scheme that can fully recover the coded caching gains by capitalizing on one aspect that has remained unexploited to date: the shared side information brought about by the effectively unavoidable file-size constraint. As a consequence, the worst-user effect is dramatically ameliorated, as it is substituted by a much more subtle worst-group-of-users effect, where the suggested grouping is fixed, and it is decided before the channel or the demands are known. Furthermore, the theoretical gains are completely recovered as the number of users increases, and this is done without any user selection technique. We analyze the rate performance of the proposed scheme and derive approximations which prove to be very precise. Importantly, this novel approach can be translated to other coded caching schemes and scenarios, including decentralized scenarios.**

***Index Terms*—Coded-caching, finite SNR, shared caches, worst-user bottleneck, effective coded caching gain.**

## I. Introduction

CACHE-aided communication is a promising approach toward reducing congestion in modern communication networks [2], [3]. The promise of this approach has been recently fostered by the seminal paper of Maddah-Ali and Niesen [2], who proposed *coded caching* as a means to speed up content delivery by exploiting receiver-side stored content to remove interference.

The work in [2] 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 $K$ cache-aided 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 so-called *MN scheme* of [2] involves a cache placement phase and a subsequent delivery phase. During the first phase, each file is typically split into a very large number of subfiles, which are selectively placed in various different caches. During the second phase, the communication process is split into a generally large number of *transmission stages*, and, at each such stage, a different subset of $K\gamma+1$ users is simultaneously served via an XOR multicast transmission, thus allowing for a theoretical speed-up factor of $K\gamma+1$ as compared to the uncoded case. This speed-up factor of $K\gamma+1$ is also referred to as the *coded caching gain* achieved by this scheme.

The above algorithm was originally developed for the scenario where the channel is error-free and the capacity to each user is identical. In recent years, a variety of works have investigated coded caching under more realistic wireless settings, considering for example uneven channel qualities [4]–[7], the role of Channel State Information (CSI) availability [8]–[10], statistically diverse channels [11], [12], and a variety of other scenarios [13]–[18].

Unfortunately, it is the case that coded caching suffers from two major constraints. The first is often referred to as the "file-size constraint" of coded caching, which, as we will recall later, effectively forces different users to fill up their caches with identical content [19]–[21]. This constraint, which has been extensively analyzed in the literature [13], [20]–[23], essentially foregoes the freedom to endow users with their own dedicated caches, and rather forces these users to share a very limited number of cache states[1] that is considerably smaller than $K$. 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 [24]. 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 severely exacerbated as the $\mathrm{SNR}$ becomes smaller.

Both these realities, of bounded file sizes and limited SNR, are naturally inherent to any practical wireless content-delivery system. Let us look at these bottlenecks in greater detail.

### A. Subpacketization Bottleneck and the Need for Shared Caches

Our work builds on the premise that almost any realistic single-stream coded caching scenario will involve the use of shared, rather than dedicated, cache states. As we will see right below, this has to do with the simple fact that, under realistic assumptions on $\gamma$ and $K$, the file sizes required by caching schemes dwarf any realistic file sizes that we encounter in wireless downlink applications. The evidence for this is

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). Part of this work has been published in the 2020 IEEE Information Theory Workshop [1].

Hui Zhao and Petros Elia are with the Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France (email: hui.zhao@eurecom.fr; petros.elia@eurecom.fr).

Antonio Bazco-Nogueras was with the Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France. He is now with the IMDEA Networks Institute, 28918 Madrid, Spain (email: antonio.bazco@imdea.org).

[1]Hereinafter, *cache state* refers to the content stored at the cache of a certain user. Thus, two users storing the exact same content in their local cache are said to have the same *cache state*.

overwhelming, and, to date, under realistic assumptions, any high-performance coded caching scheme requires files to be split into a number of sub-files that grows exponentially or near-exponentially with $K$. For example, the MN algorithm requires files to be split into at least $\binom{K}{K\gamma}$ subfiles, and it is known from [22, Thm. 3] that this same subpacketization is indeed necessary for any algorithm to achieve this same gain under some basic symmetry conditions. Similarly, it was shown in [19] that decentralized schemes (cf. [25]) require exponential (in $K$) subpacketization in order to achieve linear caching gains, and, along similar lines, [26, Thm. 12] proved that, under basic assumptions, there exists no coded caching scheme that enjoys both linear caching gains and linear subpacketization.

From these previous results, we are in a position to say that such schemes will inevitably require many users to store the same cache content. While there is not a fundamental limitation that forces users to cache the same content, an extensive literature overview indicates that there are only two possible solutions to keep the subpacketization low while maintaining the gains: either to repeat the same cache state at several users, or to importantly increase $K$ [22], [23], [26]. Let us consider for instance the original MN scheme. Under the constraint that the subpacketization (number of subfiles) cannot exceed a realistic value $S_{max}$, we know that the best course of action is to encode over a limited number of $\Lambda < K$ users at a time [19], creating $\Lambda$ different cache states. This $\Lambda$ is indeed limited by the file size constraint that asks that $\binom{\Lambda}{\Lambda\gamma} \leq S_{max}$. This approach naturally limits the aforementioned (error-free) optimal gain to $\Lambda\gamma + 1$ [27], and it entails cache replication simply because now there are only $\Lambda$ cache states to be shared[2] or replicated among the $K$ users. As we will show later on, this forced replication can be exploited to circumvent another major problem: the worst-user bottleneck.

### *B. Worst-User Bottleneck: Motivation, Nature of the Problem, and Prior Work*

As we have mentioned, the worst-user limitation induced by the nature of the multicast transmission [24] is exacerbated when the SNR becomes smaller and when the channel strengths are different. Consequently, this dependence on multicasting can severely affect the applicability of coded caching in many wireless scenarios that possess such characteristics. These scenarios prevail in cellular or satellite communications settings [29] that suffer from heavy path-loss and/or shadowing and in IoT networks or massive Machine-Type Communication (mMTC) settings [30]. Similarly, we know that in 4G LTE networks the range of users' signal-to-interference-plus-noise ratio (SINR) is typically 0–20 dB [31], while the SINR of cell-edge users can be closer to 0–5 dB. The worst-user bottleneck is also exacerbated when considering the well-established setting of quasi-static fading that we will consider in the following, and which generally comes about in the presence of longer coherence periods and shorter latency constraints. This quasi-static setting applies to low-mobility scenarios, which nicely capture coded-caching use-cases where pedestrians or static users are consuming video streaming.

This bottleneck has sparked considerable research interest that resulted in a variety of notable results [4], [5], [32]–[34]. For example, the work in [32] shows that, in a single transmit-antenna setting with finite power and quasi-static fading, the effective gain does not scale as $K$ becomes larger *even in the absence of a file-size constraint*; moreover, the power must scale linearly with $K$ in order to preclude the collapse of the multicast rate (cf. [32, Table I]). Taking a different approach, the work in [33] employs superposition coding for opportunistic scheduling. Another notable work can be found in [34], which groups together users that experience similar SNR and delivers to each group in a separate way after neglecting users with the weakest channels.

However, to date, no scheme is known to overcome the worst-user bottleneck without user selection for the single transmit-antenna setting. In this context, we analyze in this work the worst-user bottleneck when no user selection techniques are applied, in order to expressly show that these techniques are not needed to overcome the worst-user bottleneck. This is an interesting result because user selection increases the complexity of the transmission in several aspects: First, because of the CSI required to take the decision of which users are going to be served in the next transmission; this requirement entails a trade-off: In order to better exploit the benefits of user-selection techniques, the transmitter would require CSI from many users (ideally, all) at every time, which in turn may consume a lot of resources. Second, the transmitter would need to add an extra step to select the suitable user subset. Both CSI acquisition and selection algorithm can become challenging when the number of users become large, as it is assumed in our analysis.

In all these previous scenarios, this bottleneck essentially diminishes the aforementioned coded caching gain[3]. Had the SNR been infinite, or the instantaneous link strengths identical, this hypothetical gain would have taken the form $\Lambda\gamma + 1$ for any allowable $\Lambda$ up to $K$ (where, we recall, this allowable $\Lambda$ is generally much less than $K$ due to the bounded file sizes). Yet, as the SNR decreases, the effect of the worst-user bottleneck becomes more accentuated[4], and the effective gain eventually collapses. This collapse will be rigorously described in Proposition 2, and it is illustrated in Fig. 1.

### *C. Contributions and Organization*

In this work, we consider coded caching with centralized placement in the standard single-antenna BC and in the context of finite SNR and quasi-static fading. The analysis holds for any SNR, whereas some of the subsequent approximations imply either many users or low SNR values. These asymptotic results are shown to precisely characterize the performance also for realistic values found in current wireless networks. Our contributions are outlined as follows.

[2]It is worth noting that the shared cache setting not only captures the effect of the file-size constraint, but also reflects promising heterogeneous scenarios where a main station serves users with the help of smaller cache-endowed helper nodes [27], [28].

[3]We remind the reader that the gain describes the cache-aided speed-up factor over the approach which employs the basic Time Division Multiplexing (TDM) method that serves one user at a time.

[4]To see this, simply recall that for smaller values of SNR and for $z < 1$, it follows that $\ln(1 + z\text{SNR}) \approx z \ln(1 + \text{SNR})$.

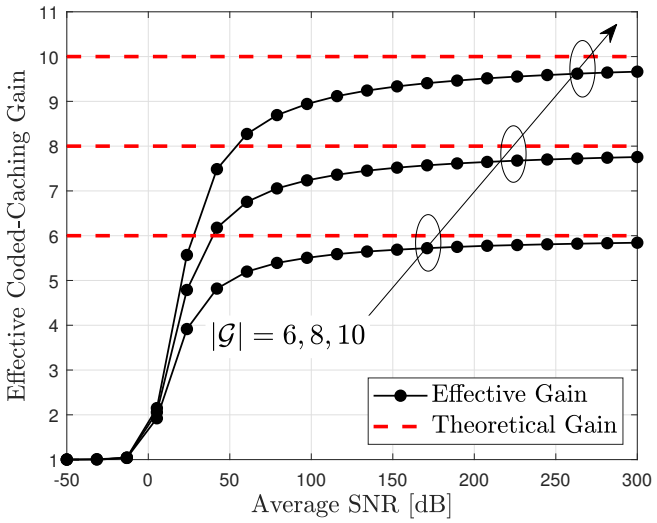



Fig. 1: Ratio between the average rates of the MN scheme and TDM (i.e., the *effective coded-cahing gain*) over quasi-static Rayleigh fading channel for different values of $|\mathcal{G}| = \Lambda\gamma + 1$.

- We first show that the coded caching gain of the MN scheme with respect to simple uncoded TDM (either with or without file size constraints) considerably deteriorates for any reasonable range of SNR values, and, in fact, completely vanishes in the low-SNR regime.
- Then, focusing on the file-size constrained scenario (which corresponds to having a limited number $\Lambda$ of different cache states), we present a novel transmission scheme that substantially improves the effective gain, and which manages to recover — without any user selection technique — the entire theoretical coded-caching gain $\Lambda\gamma + 1$ in the presence of sufficiently many users. The proposed scheme, which will be referred to as the *Aggregated Coded-Caching scheme*, builds on the practical 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. In fact, it turns out that having $B$ users per cache state is as efficient as having a time diversity of $B$ coherence times.
- We analyze the average rate (which we rigorously define later) and derive its exact analytical expression. To offer insight, we apply low-SNR and large-$K$ approximations to derive clear closed-form expressions for the average rate and the gain. These approximations are shown to retain a robust accuracy even for a very modest user count.
- As a consequence of these results, we now know a simple way to exploit the unavoidable nature of the file-size constraint in order to almost entirely remove the worst-user bottleneck. In essence, we show that, given the file-size constraint, the worst-user effect can be made negligible. Importantly, this dual idea of combining cache replication and multi-rate transmission can be applied to a variety of different coded caching algorithms in order to allow these algorithms to provide gains in fading scenarios. This ability is highlighted in Section V-C, where we apply our ideas in the context of decentralized coded caching.

The remainder of this paper is organized as follows: Section II defines the system model and the problem considered. In Section III, the proposed scheme is formally described. The average rate is investigated in Section IV, where we derive several tight approximations. Some numerical results and comparisons are presented in Section V, and Section VI concludes the paper.

*Notation:* We use the notation $X \sim \mathcal{Y}$ to state that a random variable $X$ follows a distribution $\mathcal{Y}$. Given a real-valued function $f(x)$ over a variable $x$, $f(x) = o(x)$ stands for $\lim_{x\to 0} \frac{f(x)}{x} = 0$. $\mathbb{E}\{\cdot\}$ denotes the expectation operator. We use the short-hand notation $[n] \triangleq \{1, 2, \ldots, n\}$ for a positive integer $n$. $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution of mean $\mu$ and variance $\sigma^2$, and $\mathrm{Gamma}(m, \rho)$ denotes the Gamma distribution with shape and scale parameters $m$ and $\rho$ respectively. $|\cdot|$ denotes the cardinality operator of a set. All sets are assumed to be ordered.

## II. System Model and Problem Definition

We consider the quasi-static 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 $\{W_n\}_{n=1}^N$ of $N$ files, and each user is assisted by a cache of normalized size $\gamma \in [0, 1]$. We consider an arbitrary number $\Lambda$ of allowable cache states, and we assume that $K$ is an integer multiple of $\Lambda$.

The received signal at user $k \in [K]$ 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. [32]), 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 it is common in works that study coded caching under quasi-static fading [4], [32], we adopt the *average rate*[5] as the metric of interest. Toward this, we define the instantaneous rate $r$ as the maximal sum-rate that can be transmitted to the simultaneously served users for a instantaneous channel realization. Then, the *average rate* $\mathbb{E}_H\{r\}$ is defined as the average (over fading statistics) of the above instantaneous rate. It is important to not confuse this long-term average $\mathbb{E}_H\{r\}$ with the ergodic rate, since the latter implies an ability to encode over several fading realizations (cf. [32]). Henceforth, all the values for rate (bits/s) and time (s) are normalized to one Hz of bandwidth.

In this context, a coded caching scheme seeks to provide an *effective coded-caching gain*, which 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

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

file-size constrained coded caching in the error-free scenario with fixed and identical link capacities.

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

## III. Aggregated Coded-Caching Scheme

We now introduce a novel scheme, coined as the *Aggregated Coded Caching* (ACC) scheme, which will be shown to overcome the previous collapse of the effective gains. The scheme is based on the scheme presented in [19, Section V-A], which is henceforth denoted as "the $\Lambda$-MN" scheme. In turn, this $\Lambda$-MN scheme consists on the MN scheme but with cache replication. Yet, the main contribution of this work is to provide tools that can be applied to different coded caching schemes, namely multi-rate transmission combined with cache replication, and the base scheme is chosen to ease the readability.

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 (i.e., they share the same cache state). As we have seen, this is essentially inevitable under realistic file-size constraints. The scheme also follows a standard clique-based approach [2], such that the transmission is divided into *transmission stages* that experience a clique-side information pattern. As for [2], this implies that any desired subfile of some served user can be found in the cache of every other user involved in that same transmission stage. Thus, this approach defines a side-information structure that was addressed in the following well-known result from [35].

**Proposition 1** ( [35, 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 $\overline{\mathcal{W}}_i = \{W_j'\}_{j\neq i, j\in[t]}$, is given by*

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

*Proof.* Proposition 1 is known as a special case of [35, Thm. 6], and this particular form has been considered in [36], [37]. More details on this, as well as on the association to our setting, are described in Appendix I. □

Proposition 1 implies that, under this particular configuration of side information, each user can achieve its own point-to-point capacity, as if no other user was being served at the same time. There are various optimal *multi-rate transmission* schemes for this setting [36], [37], and the proposed ACC scheme can remain oblivious to the encoding choice.[6]

**Remark 1.** *We state in advance that the aforementioned multi-rate transmission must indeed be combined with the method of shared caches in order to yield the desired gains. While multi-rate transmission performs better than MN-based XORs, this rate improvement appears only when we focus our attention on a single isolated delivery stage that serves some fixed set of users $\mathcal{G}$. However, when considering the entire delivery problem over all sets $\mathcal{G}$, we would see no gain, because the MN placement and multicast group generation without shared caches would not allow for an additional subfile to be sent to a potentially 'fast' user in $\mathcal{G}$ without generating interference to the remaining (slower) users. This latter point, which is that the MN placement does not allow exploitation of fast users, is presented below in the original context of XORs.*

**Example 1.** *Consider the delivery of XOR $A_{2,3} \oplus B_{1,3} \oplus C_{1,2}$ meant for users $\mathcal{G} = \{1,2,3\}$ who respectively ask for files $W_1 = A, W_2 = B, W_3 = C$. Even if user 1 decodes $A_{2,3}$ very quickly, she must wait for $B_{1,3}$ and $C_{1,2}$ to be decoded, because (by definition of the MN placement) there exists only one subfile that is desired by user 1 and can be decoded by users 2 and 3. An illustrative example is shown in Fig. 2a.*

### 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 [19], [27], and thus it simply applies the MN placement of the $\Lambda$-user problem, and then each user of the same group stores the same cache content. In particular, each file $W_n$, $n \in [N]$, is partitioned into $\binom{\Lambda}{\Lambda\gamma}$ segments as $W_n \to \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\}$, and then each user in group $g \in [\Lambda]$ stores all the subfiles belonging to the set $\mathcal{Z}_g \triangleq \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, \mathcal{T} \ni g, \forall n \in [N]\}$.

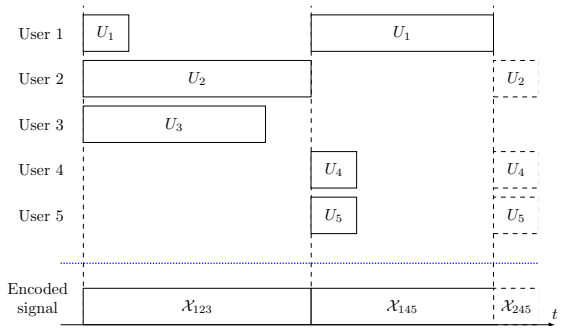
*2) Delivery Phase:* The delivery phase is split into $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, where each stage involves a set $\mathcal{G} \subseteq [\Lambda]$ of $|\mathcal{G}| = \Lambda\gamma + 1$ groups. During each stage, the transmitter *simultaneously* delivers to as many as $\Lambda\gamma + 1$ users, each from a different group in set $\mathcal{G}$. The users within each group are served one after the other in a round-robin manner. For a given set $\mathcal{G}$, the transmitter employs a multi-rate code that achieves the capacity in Proposition 1, which implies that the channel state information should be available at the transmitter *only* for the $|\mathcal{G}|$ served users. We emphasize that the multi-rate transmission in the ACC delivery does not require power-splitting and guarantees the successful information decoding in each served user due to the rate adaptation.

Let $\mathcal{G}(i)$ denote the $i$-th group in $\mathcal{G}$, $i \in [|\mathcal{G}|]$ (recall the group-set $\mathcal{G}$ is ordered). We represent the set of users that are being served at a particular time by the vector $\mathbf{v} \in \mathbb{Z}^{|\mathcal{G}|}$.[7] Consistently, $\mathbf{v}(i) \in [B]$ tells us which user of the $i$-th group in $\mathcal{G}$ is currently being served, and $d_{\mathbf{v}(i)} \in [N]$ denotes the file index requested by user $\mathbf{v}(i)$. Hence, the transmitter serves the users $\mathbf{v}$ of the groups in $\mathcal{G}$ by transmitting

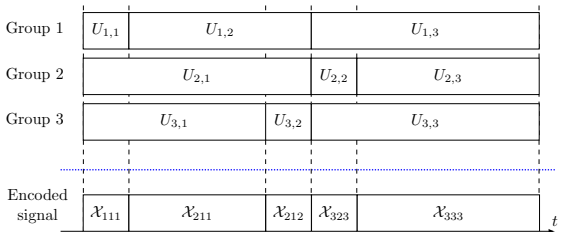
$$X_{\mathcal{G},\mathbf{v}} = \mathcal{X}\big(\{W_{d_{\mathbf{v}(i)}}^{\mathcal{G}\setminus\{\mathcal{G}(i)\}}\}_{i\in[|\mathcal{G}|]}\big), \tag{1}$$

[6] In terms of practicality, we know that very simplified schemes, such as nesting BPSK into M-QAM constellations (cf. [4]), come extremely close to achieving the above capacity region, and in fact achieve the single-user capacity when we restrict ourselves to QAM modulations [38]. Such practical codes can be directly applied in our cache-aided setting with minor performance losses.

[7] Please note here that the dependence of $\mathbf{v}$ on the time index and on $\mathcal{G}$ is assumed but omitted for simplicity.




(a) Dedicated caches: Delay depends on the worst-user capacity at each transmission stage. $\mathcal{X}_{abc}$ denotes the signal encoded for users $a$, $b$, and $c$.




(b) ACC scheme: Delay depends on the worst group sum rate. $\mathcal{X}_{abc}$ denotes the encoded signal for users $a$, $b$, and $c$ of groups 1, 2, and 3, respectively.

Fig. 2: Comparison of MN and ACC for a nominal coded-caching gain of 3.

where, for any set of messages $\Phi$, $\mathcal{X}(\Phi)$ denotes the transmitted signal obtained from encoding the messages in $\Phi$ with a coding scheme achieving the capacity region in Proposition 1. We recall that the ACC scheme is oblivious to the selected coding scheme, as long as it achieves the capacity region. As usual in coded caching schemes, $W^{\mathcal{G}\setminus\{\mathcal{G}(i)\}}_{d_{\mathbf{v}(i)}}$ represents the subfile intended by user $\mathbf{v}(i)$ that is stored in the cache of all groups in $\mathcal{G}$ except group $\mathcal{G}(i)$.

Algorithm 1 presents the transmission for a specific group set $\mathcal{G}$. Every time the user of some group $\mathcal{G}(i')$ obtains its subfile, $\mathbf{v}(i')$ is updated[8] as $\mathbf{v}(i') \leftarrow \mathbf{v}(i')+1$. This process is repeated until all users in all groups in $\mathcal{G}$ are served. If every user of a group has obtained its subfile, the transmission is 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 obtain their requested files. We reemphasize that the ACC scheme does not apply user selection. Let us proceed with a simple clarifying example.

**Example 2.** *Consider a transmission stage serving 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 group $g$ as $U_{g,b}$ and the subfile intended for this user as $W'_{g,b}$. Let us further assume that the normalized capacity of each user (expressed in transmitted subfiles per time slot) is as follows:*

| | 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 |

[8]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 remains only a part of such subfiles to be transmitted. This is intuitive from Fig. 2b.

**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** Number of 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 $\mathbf{v}(i)$, $i \in [|\mathcal{G}|]$, fully obtains its subfile*
  Set $i^\star$ as the index of the group $\mathcal{G}(i^\star)$ whose user has decoded its subfile
  **if** $\mathbf{v}(i^\star) = B$ **then**
    Number of finished groups $\longleftarrow$ Number of finished groups $+1$
  $\mathbf{v}(i^\star) \longleftarrow \mathbf{v}(i^\star)+1$

*which simply implies 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. 2b. 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 ($\log_2(1+\mathrm{SNR}_{g,b})$) because each user knows the subfiles of the other two served users.*

*After the first slot, user $U_{1,1}$ has successfully decoded its subfile. Hence, $U_{1,1}$ is substituted 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 serve any of the users storing the same cache state 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 group substitutes this user, while the other served users can continue decoding their subfile. In the same way, $U_{3,1}$ obtains its subfile after the fourth time slot, it is replaced by $U_{3,2}$, and the transmitter then sends $\mathcal{X}\big(W'_{1,2}, W'_{2,1}, W'_{3,2}\big)$. After the fifth slot, the three users obtain their subfile and the transmitter starts sending $\mathcal{X}\big(W'_{1,3}, W'_{2,2}, W'_{3,3}\big)$, and so on.*

## IV. Average Rate Analysis

In this section, we analyze the long-term average rate of the $\Lambda$-MN and ACC schemes. First, we will derive the exact expression of the average rate for both schemes. Afterward, we will approximate this rate at low SNR, and we will also derive the limit in the regime of many users. It will turn out, as we will see in the following, that these two approximations are very robust in realistic scenarios. Furthermore, we obtain the effective gain of this scheme with respect to TDM as well as its improvement with respect to the $\Lambda$-MN scheme, and we show that while the effective gain of the $\Lambda$-MN scheme vanishes at low SNR, the ACC scheme recovers — at any SNR value — the nominal (high-SNR) gain as the number of users per cache increases.

We recall that, under Rayleigh fading, the SNR follows an exponential distribution. Hence, for user $k \in [K]$, the

probability density function (PDF) and cumulative distribution function (CDF) of $\mathrm{SNR}_k$ are given respectively by $f_{\mathrm{SNR}_k}(x) = \frac{1}{\rho}\exp\left(-\frac{x}{\rho}\right)$ and $F_{\mathrm{SNR}_k}(x) = 1-\exp\left(-\frac{x}{\rho}\right)$, for any $x \geq 0$, where $\rho = \mathbb{E}_H\{\mathrm{SNR}_k\}$ denotes the average SNR with respect to channel states. We recall that the user channels are statistically symmetric. As for the ACC scheme, we will use $\mathrm{SNR}_{g,b}$, $f_{\mathrm{SNR}_{g,b}}(x)$, and $F_{\mathrm{SNR}_{g,b}}(x)$ 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 $\Lambda$-MN and ACC Schemes

*1) Average Rate of the $\Lambda$-MN Scheme:* We first note that the $\Lambda$-MN is an adaptation for finite-file sizes settings from [19] of the standard MN scheme of [2]. Placement is analogous to the one of the ACC scheme, and the transmission consists of repeating $B$ times the transmission of the dedicated caches setting. Consequently, the $\Lambda$-MN scheme consists of $B\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, each of them employed to deliver an XOR to a group of users of size $|\mathcal{G}| = \Lambda\gamma+1$.

Consider the delivery to a particular set $\mathcal{G}$ of $\Lambda\gamma+1$ users. We know from the multicast capacity theorem in [24] that the maximum instantaneous rate for any user $i \in \mathcal{G}$ takes the form

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

where the minimum operator guarantees the successful information decoding at all the users in $\mathcal{G}$. Note that the delay (or delivery time) required to transmit one sub-file to every user in $\mathcal{G}$ at this transmission stage is given by

$$T_{\mathrm{MN},\mathcal{G}} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}}\left[\log_2\left(1+\min_{k\in\mathcal{G}}\mathrm{SNR}_k\right)\right]^{-1} \quad \text{s}, \tag{3}$$

where $F$ is the total information bits of a file, and $F/\binom{\Lambda}{\Lambda\gamma}$ is the size of the subfiles generated from subpacketization. Since $\min_{k\in\mathcal{G}}\mathrm{SNR}_k$ follows an exponential distribution with rate $|\mathcal{G}|/\rho$, the expectation of $T_{\mathrm{MN},\mathcal{G}}$ diverges. For this reason, we consider the average rate as a main metric of interest, which crisply reflects the worst-user effect.

The instantaneous sum rate is given by $\sum_{i\in\mathcal{G}} r_{i,\mathcal{G}}^{(\mathrm{MN})}$, since we are simultaneously serving all the $|\mathcal{G}|$ users. Consequently, the average (sum) rate for that specific set $\mathcal{G}$ takes the form

$$\begin{aligned}\bar{R}_{\mathcal{G}}^{(\mathrm{MN})} &\triangleq \mathbb{E}_H\Big\{\sum_{i\in\mathcal{G}} r_{i,\mathcal{G}}^{(\mathrm{MN})}\Big\} \\ &= \frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\Big\{\ln(1+\min_{k\in\mathcal{G}}\mathrm{SNR}_k)\Big\}\end{aligned} \tag{4}$$

which follows because the users are statistically equivalent, which in turn also implies that the average sum rate $\bar{R}^{(\mathrm{MN})}$ remains the same for any set $\mathcal{G}$, i.e., it implies that $\bar{R}^{(\mathrm{MN})} = \bar{R}_{\mathcal{G}'}^{(\mathrm{MN})}\ \forall \mathcal{G}' \subseteq [\Lambda],\ |\mathcal{G}'| = \Lambda\gamma+1$.

Naturally, the average rate under the TDM scheme, which we denote as $\bar{R}^{(\mathrm{TDM})}$, is a special case of $\bar{R}^{(\mathrm{MN})}$ obtained by setting $|\mathcal{G}| = 1$. The variable $\min_{k\in\mathcal{G}}\{\mathrm{SNR}_k\}$ is the minimum of $|\mathcal{G}|$ i.i.d. exponential variables of rate $\frac{1}{\rho}$ (i.e., mean $\rho$), and, consequently, it follows an exponential distribution with rate $|\mathcal{G}|/\rho$ (or mean $\rho/|\mathcal{G}|$). Thus, it follows from [39, Eq. (15.26)] that

$$\bar{R}^{(\mathrm{MN})} = -\frac{|\mathcal{G}|}{\ln 2}\exp\left(\frac{|\mathcal{G}|}{\rho}\right)\cdot\mathrm{Ei}\left(-\frac{|\mathcal{G}|}{\rho}\right), \tag{5}$$

where $\mathrm{Ei}(\cdot)$ represents the exponential integral function [40]. Note that $|\mathcal{G}| = 1$ in (5) yields the closed-form expression for $\bar{R}^{(\mathrm{TDM})}$.

*2) Average Rate of the ACC Scheme:* Due to the symmetry of the ACC scheme and the statistical symmetry of the channel, we now focus on a particular 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 said user $b$ has completed the decoding of its subfile. Furthermore, in the ACC scheme, the delivery to a group-set $\mathcal{G}$ is completed when every user belonging to one of these groups has obtained its subfile. Consequently, the resulted delay (or delivery time) to serve all user-groups in $\mathcal{G}$ (which include $|\mathcal{G}|\cdot B$ users) is given by

$$T_{\mathrm{ACC},\mathcal{G}} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}}\max_{g\in\mathcal{G}}\sum_{b=1}^{B}\Big[\log_2\left(1+\mathrm{SNR}_{g,b}\right)\Big]^{-1} \quad \text{s}, \tag{6}$$

which will become (3) for $B = 1$. As for (3), the expectation of (6) diverges. Then, as explained for the $\Lambda$-MN scheme, we consider the average rate. The (per-user average) rate with which any group $j$ in the set $\mathcal{G}$ is served is here captured by

$$r_{j,\mathcal{G}}^{(\mathrm{ACC})} = \min_{g\in\mathcal{G}}\frac{1}{B}\sum\nolimits_{b=1}^{B}\log_2(1+\mathrm{SNR}_{g,b}) \quad \text{bits/s} \tag{7}$$

for all $j \in \mathcal{G}$. By applying the same reasoning as in (2)–(4), we obtain that the average rate with which the transmitter delivers data across the users is given by

$$\bar{R}^{(\mathrm{ACC})} = \frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\Big\{\min_{g\in\mathcal{G}}\frac{1}{B}\sum_{b=1}^{B}\ln(1+\mathrm{SNR}_{g,b})\Big\} \quad \text{bits/s}. \tag{8}$$

We quickly note that, by comparing (8) with (4), we can see how the worst-user effect is essentially averaged out into a cumulative "worst-group" effect. By considering dedicated caches (i.e., $B = 1$), we obtain the same average rate as that of the $\Lambda$-MN scheme in (4) despite having a different (not XOR-based) coding scheme, which is consistent with Remark 1.

In the following, $\jmath \triangleq \sqrt{-1}$ denotes the imaginary unit, $\mathrm{Im}\{\cdot\}$ denotes the imaginary part of a complex number, and $\mathrm{E}_{-\jmath t}(\cdot)$ denotes the exponential integral function of the $(-\jmath t)$-th order [40]. Next, we present our first main result.

**Lemma 1.** *The exact average rate of the ACC scheme over symmetric quasi-static Rayleigh fading can be derived in a double-integral form, which takes the form*

$$\bar{R}^{(\mathrm{ACC})} = \frac{|\mathcal{G}|}{B\ln 2}\times$$
$$\int_0^\infty\left(\frac{1}{2}+\frac{1}{\pi}\int_0^\infty\frac{\mathrm{Im}\left\{\exp(-\jmath xt)\frac{\exp(B/\rho)}{\rho^B}\mathrm{E}_{-\jmath t}^{B}\left(\frac{1}{\rho}\right)\right\}}{t}\mathrm{d}t\right)^{|\mathcal{G}|}\mathrm{d}y.$$

*Proof.* The proof is relegated to Appendix II. □

The numerical implementation of the above expression is very complex and it provides little insight. In the following, we obtain the effective gains in both the low-SNR limit and the large-$B$ limit for the $\Lambda$-MN scheme and the ACC scheme, and we derive approximations of their rates, from which some meaningful insights can be easily drawn.

### B. Rate Approximations and Effective Gains at Low SNR

*1) $\Lambda$-MN Scheme:* First, we present a low-SNR approximation for the average rate of the $\Lambda$-MN scheme, which is in fact a special case of the ACC scheme with $B = 1$. Although the exact form has been derived in (5), we can provide a simple but tight approximation which allows us to remove the special function $\mathrm{Ei}(\cdot)$ from the expression.

**Lemma 2.** *In the low-SNR region, the average rate of the $\Lambda$-MN scheme can be approximated by*

$$\bar{R}^{(\mathrm{MN})} \approx \frac{|\mathcal{G}|}{\ln 2}\left(\ln\left(1+\frac{\rho}{|\mathcal{G}|}\right) - \frac{\rho^2}{2|\mathcal{G}|^2\left(1+\rho/|\mathcal{G}|\right)^2}\right). \quad (9)$$

*Proof.* See Appendix III-A. □

In the numerical evaluation section (cf. Fig. 6 in Section V), it will be shown that this computationally efficient second-order approximation can in fact provide us with an extremely reliable estimation of the performance even in the medium-SNR region.

Let us now consider the exact effective gain of the $\Lambda$-MN scheme, which, directly from (5), takes the form

$$\frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}} = \frac{|\mathcal{G}|\exp\left(\frac{|\mathcal{G}|}{\rho}\right)\cdot \mathrm{Ei}\left(-\frac{|\mathcal{G}|}{\rho}\right)}{\exp\left(\frac{1}{\rho}\right)\cdot \mathrm{Ei}\left(-\frac{1}{\rho}\right)}. \quad (10)$$

As expected, the effective gain converges to the nominal gain $|\mathcal{G}|$ at high SNR, since the limit of (10) as $\rho \to \infty$ is $|\mathcal{G}|$. On the other hand, in the low-SNR region, this effective gain entirely vanishes, as stated in the following proposition.

**Proposition 2.** *For any value of $K$ and $\Lambda$, the effective gain of the $\Lambda$-MN scheme converges to*

$$\lim_{\rho\to 0}\frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}} = 1 \quad (11)$$

*meaning that this effective coded-caching gain entirely vanishes at low SNR.*

*Proof.* See Appendix III-B. □

As noted before, Proposition 2 holds for any scheme which requires decoding of single XORs.

*2) ACC Scheme:* After presenting the previous result for the $\Lambda$-MN scheme, let us now consider the ACC scheme. In the following, for any integer vector $\mathbf{b} \triangleq [b_1, b_2, \cdots, b_B] \in \mathbb{Z}^B$ composed of $B$ non-negative elements, we will use

$$\binom{n}{\mathbf{b}} \triangleq \frac{n!}{b_1! b_2! \cdots b_B!} \quad (12)$$

to denote the multinomial coefficient. We can now state our following result, which presents an expression of the rate of the ACC scheme for the low-SNR regime.

**Lemma 3.** *In the low-SNR region, the average rate of the ACC scheme can be approximated by $\bar{R}^{(\mathrm{ACC})} \approx \frac{\rho|\mathcal{G}|}{B\ln 2}\Psi_{|\mathcal{G}|}$, since it holds that*

$$\bar{R}^{(\mathrm{ACC})} = \frac{\rho|\mathcal{G}|}{B\ln 2}\Psi_{|\mathcal{G}|} + o(\rho), \quad (13)$$

*where $\Psi_{|\mathcal{G}|}$ is defined as*

$$\Psi_{|\mathcal{G}|} \triangleq \sum_{||\mathbf{b}||_1=|\mathcal{G}|}\binom{|\mathcal{G}|}{\mathbf{b}}\frac{|\mathcal{G}|^{-1-\sum_{t=1}^{B}(t-1)b_t}}{\prod_{t=1}^{B}((t-1)!)^{b_t}}\Big(\sum_{t=1}^{B}(t-1)b_t\Big)!\,,$$

*and where the sum is over all the vectors composed of $B$ non-negative integer elements and whose norm-1 equals $|\mathcal{G}|$.*

*Proof.* The proof is relegated to Appendix III-C. □

From Lemma 3 and Proposition 2, we obtain a corollary on the gain of the ACC scheme over the $\Lambda$-MN scheme.

**Corollary 1.** *In the limit of low SNR, the ratio of $\bar{R}^{(\mathrm{ACC})}$ over $\bar{R}^{(\mathrm{MN})}$ converges to the constant*

$$\lim_{\rho\to 0}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}} = \frac{|\mathcal{G}|}{B}\Psi_{|\mathcal{G}|} \quad (14)$$

*where we recall that $|\mathcal{G}| = \frac{K}{B}\gamma + 1$.*

*Proof.* The proof is relegated to Appendix III-D. □

The expression in Corollary 1 is illustrated in Fig. 3 for different values of $B$ and $|\mathcal{G}|$.

**Remark 2.** *In Fig. 3, we can see that $\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}}$ is concave with respect to $B$, and that this concavity increases with $|\mathcal{G}|$. This signals that, for large $|\mathcal{G}|$, most of the gain from having $B > 1$ is obtained quickly, at relatively small values of $B$. For example, when $|\mathcal{G}| = 100$ (which is unrealistic), we see that the ACC rate for $B = 2$ is up to 20 times higher than the $\Lambda$-MN rate ($B = 1$).*

### C. Effective Gain in the Large-$B$ Region

We now move away from the low-SNR regime, and we consider instead the limit of many users. This regime is nicely motivated by the ever increasing density of users in wireless networks. Therefore, we consider that $\Lambda$ remains fixed and $K$ can grow unboundedly, which also implies that $B \to \infty$ since $B = K/\Lambda$. The following shows that, in the limit of many users, the effective gain of the ACC scheme matches — for any SNR value — the nominal gain.

**Lemma 4.** *For any average SNR $\rho$, the ACC scheme guarantees*

$$\lim_{B\to\infty}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{TDM})}} = \Lambda\gamma + 1, \quad (15)$$

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

*Proof.* The proof is relegated to Appendix III-E. □

We now proceed to compare the ACC scheme with the $\Lambda$-MN scheme, again in the limit of large $B$. We will also obtain the low-SNR approximation of this comparison, which nicely captures scenarios such as cell-free or satellite networks, where

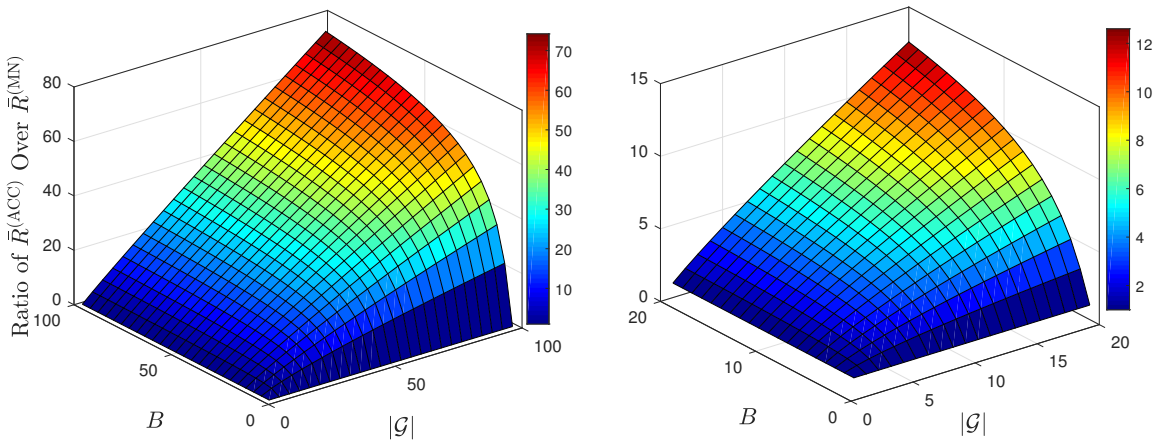



Fig. 3: The ACC improvement $\left(\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}}\right)$ over the MN scheme in Corollary 1 for different $B$ and $|\mathcal{G}|$.

the majority of the users is distributed in the edge area and/or suffers from heavy path-loss or heavy shadowing.

**Lemma 5.** *In a setting with $\Lambda$ cache states and $K = \Lambda B$ users, and for any average SNR $\rho$, the ratio $\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}}$ satisfies*

$$\lim_{B\to\infty} \frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}} = \exp\left(\frac{1-|\mathcal{G}|}{\rho}\right) \frac{\mathrm{Ei}\left(-\frac{1}{\rho}\right)}{\mathrm{Ei}\left(-\frac{|\mathcal{G}|}{\rho}\right)}. \tag{16}$$

*Furthermore, it holds that*

$$\lim_{\rho\to 0} \lim_{B\to\infty} \frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}} = \Lambda\gamma + 1. \tag{17}$$

*Proof.* The proof is relegated to Appendix III-F. Note that (17) follows from (16), but the same conclusion can be seen directly by combining Proposition 2 and Lemma 4. □

**Remark 3.** *The key for recovering the nominal gain is that a larger $B$ implies a smaller fluctuation around the average transmission rate within a user group, which inherently reduces the impact of the worst-user (or worst-group) bottleneck.*

**Remark 4.** *As previously mentioned, the preservation of the nominal gains in Lemma 5 would also hold for other coded caching schemes. Indeed, works that focus on the finite file-size constraint are normally based on XOR transmissions, and thus they are not robust to the worst-user effect. Consequently, we can improve the low-SNR performance in scenarios such as those found in [21], [22], [27] by incorporating our approach of multi-rate transmission and cache replication into these schemes. This is shown in Section V-C for the setting of [21].*

### *D. High-Fidelity Approximation of $\bar{R}^{(\mathrm{ACC})}$ for Any SNR Value*

The previous subsections offered crisp and insightful approximations of the performance of the ACC scheme. We now take a step back and seek to provide high-accuracy approximations that can be evaluated very easily.

Indeed, both the exact value of $\bar{R}^{(\mathrm{ACC})}$ in Lemma 1 and the approximation at low SNR in Lemma 3 have time-consuming implementations when $B$ is large. To counter this, we now provide a simple but very precise large-$B$ approximation of $\bar{R}^{(\mathrm{ACC})}$, which accurately approximates the average rate even if $B$ is relatively small. This expression involves the well-known Q-function $Q(\cdot)$, i.e., the tail distribution function of the standard normal distribution, and the Meijer's G-function $\mathrm{G}^{\cdot,\cdot}_{\cdot,\cdot}(\cdot)$ defined in [40, Eq. (9.301)].

Before presenting the new approximation, let us denote the expectation of the maximum of $|\mathcal{G}|$ i.i.d. standard normal random variables by $H_{|\mathcal{G}|}$. Consequently, the expectation of the minimum of such set of variables is given by $-H_{|\mathcal{G}|}$. We can now present our next result.

**Lemma 6.** *In the large-$B$ regime, the average rate of the ACC scheme can be approximated by*

$$\bar{R}^{(\mathrm{ACC})} \approx \frac{|\mathcal{G}|}{\ln 2}\left(\mu - \frac{\sigma}{\sqrt{B}} \times H_{|\mathcal{G}|}\right), \tag{18}$$

*where $\mu$ and $\sigma$ respectively represent the average and the standard deviation of $\ln(1+\mathrm{SNR}_{g,b})$ for $g \in [\Lambda]$ and $b \in [B]$, which are given by*

$$\mu = -\exp\left(\frac{1}{\rho}\right) \cdot \mathrm{Ei}\left(-\frac{1}{\rho}\right), \tag{19}$$

$$\sigma = \sqrt{2\exp\left(\frac{1}{\rho}\right) \mathrm{G}^{3,0}_{2,3}\left(\frac{1}{\rho}\,\middle|\,{}^{1,1}_{0,0,0}\right) - \mu^2}. \tag{20}$$

*Proof.* See Appendix IV. □

The term $H_{|\mathcal{G}|}$ is given by the following integral form,

$$H_{|\mathcal{G}|} = \frac{-|\mathcal{G}|}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y\Big(Q(y)\Big)^{|\mathcal{G}|-1} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y, \tag{21}$$

and the proof of (21) is relegated to Appendix IV.

At this point, we note that the value of $H_{|\mathcal{G}|}$ for $|\mathcal{G}| = 1, 2, 3, 4, 5$ is known (cf. [41, Sec. 5.16]) and it is given in the following table.

| $\lvert\mathcal{G}\rvert$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $H_{\lvert\mathcal{G}\rvert}$ | 0 | $\pi^{\frac{-1}{2}}$ | $\frac{3}{2}\pi^{\frac{-1}{2}}$ | $3\pi^{\frac{-3}{2}}\cos^{-1}\left(\frac{-1}{3}\right)$ | $\frac{5}{2}\pi^{\frac{-3}{2}}\cos^{-1}\left(\frac{-23}{27}\right)$ |

For larger values of $|\mathcal{G}|$, there are not known closed-form expressions, but it is known (cf. [42]) that one can have

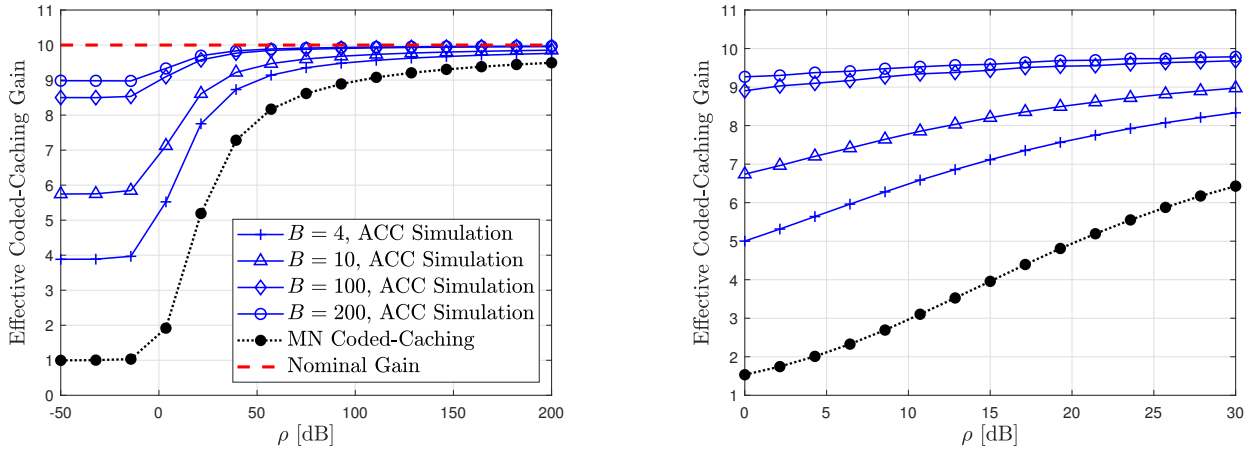



Fig. 4: Effective gain versus $\rho$ for $|\mathcal{G}| = 10$. Right-side plot focuses on realistic SNR values.

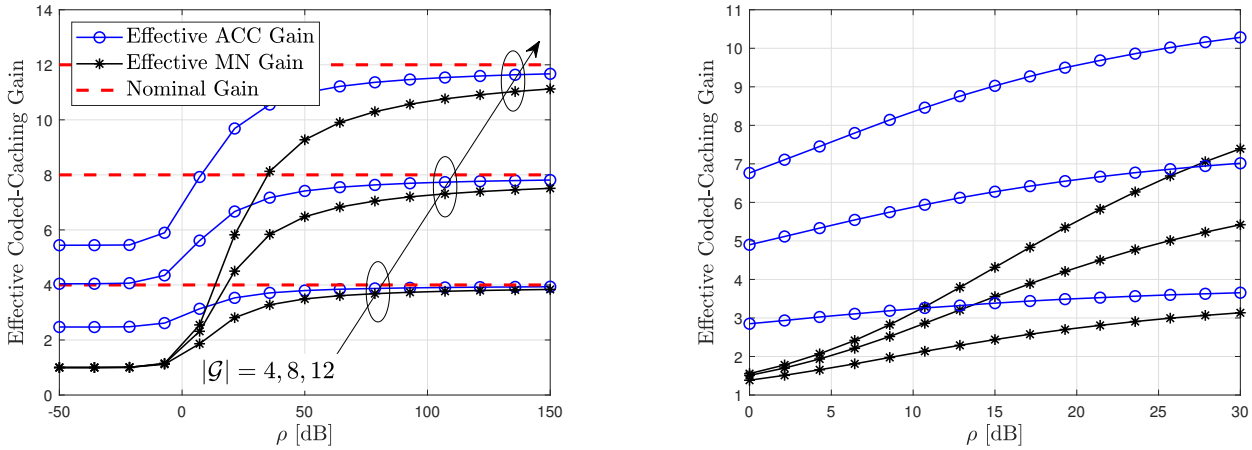



Fig. 5: Effective gain versus $\rho$ for $B = 6$. Right-side plot focuses on realistic SNR values.

a simple approximation by substituting $H_{|\mathcal{G}|}$ by $\sqrt{2\ln(|\mathcal{G}|)}$. This approximation is based on the fact that $H_{|\mathcal{G}|}$ is bounded as $\frac{1}{\sqrt{\pi \ln 2}}\sqrt{\ln(|\mathcal{G}|)} \leq H_{|\mathcal{G}|} \leq \sqrt{2\ln(|\mathcal{G}|)}$, and the fact that $\lim_{|\mathcal{G}|\to\infty} \frac{H_{|\mathcal{G}|}}{\sqrt{\ln(|\mathcal{G}|)}} = \sqrt{2}$ (cf. [42]).

In order to obtain a better approximation of $H_{|\mathcal{G}|}$ than $\sqrt{2\ln(|\mathcal{G}|)}$, which is simple but only accurate for large values of $|\mathcal{G}|$, a very interesting approximation is to adopt the Gauss-Hermite quadrature (GHQ) [43, Ch. 9], which nicely balances high accuracy and low complexity. Applying this method to the specific integral form in (21) yields

$$H_{|\mathcal{G}|} \approx \frac{-\sqrt{2}|\mathcal{G}|}{\sqrt{\pi}} \sum\nolimits_{v=1}^{V} \omega_v x_v \left(Q(\sqrt{2}x_v)\right)^{|\mathcal{G}|-1}, \qquad (22)$$

where $V$, $x_v$, and $\omega_v$ are the summation terms, sample points and weights in the GHQ, respectively. Generally speaking, we can get an approximate result with high accuracy by summing up several terms in the GHQ.

## V. Numerical Results

In the following, we illustrate through numerical analysis both the exact results and the previously obtained approximations. The derived approximations on the average rate are computationally efficient, can handle large-dimensional problems, and, as we will show via Monte-Carlo simulations, tightly approximate the true performance of the algorithms. In the following, we characterize the different considered scenarios in the simulations by the parameters $B$ and $|\mathcal{G}|$. Note that the use of these two parameters can apply to various $K, \Lambda, \gamma$ scenarios, where the relation follows from the fact that $|\mathcal{G}| = \Lambda\gamma + 1 = \frac{K}{B}\gamma + 1$.

To motivate the values of $B$ that we use, let us consider a scenario with $\gamma = 10\%$ and a realistic subpacketization limit of $10^5$. For a file size of $10^8$ bytes, this implies an atomic sub-file size of about 1000 bytes. This gives $\Lambda = \arg\max_{x\in\mathbb{Z}} \left\{ \binom{x}{0.1x} < 10^5 \right\} \approx 40$, which means that having $K = 800$ users reasonably allows for $B$ up to 20. Such (or even higher) values of $K$ are motivated by several different scenarios [44], [45]. In order to obtain the simulation results with high accuracy, $10^6$ channel states are generated and averaged over Rayleigh fading.

### A. Effective Gains With Respect to TDM

In Figs. 4–5, we present the effective coded-caching gains of the ACC and MN schemes versus $\rho$, for different values of $B$ and different nominal gains ($|\mathcal{G}|$). As expected, the effective gains of both the ACC scheme and the MN scheme converge to the nominal gain as $\rho$ increases. However, the convergence

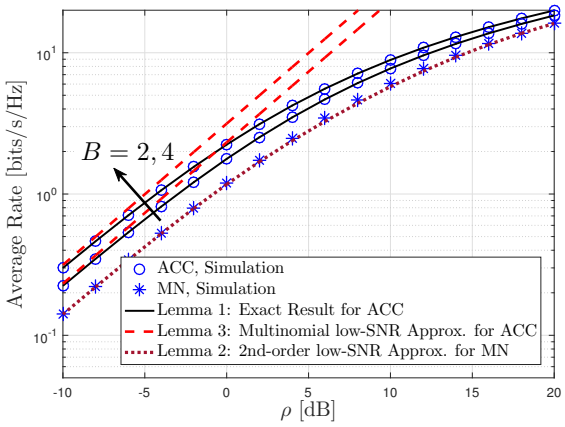



Fig. 6: $\bar{R}^{(\mathrm{ACC})}$ versus $\rho$ for $|\mathcal{G}| = 4$.

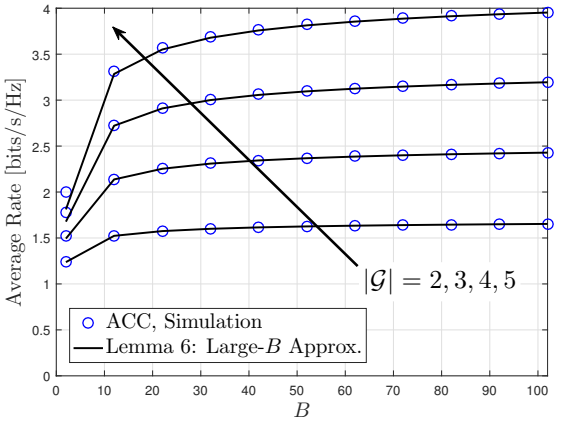



Fig. 8: $\bar{R}^{(\mathrm{ACC})}$ versus $B$ for $\rho = 0$ dB.

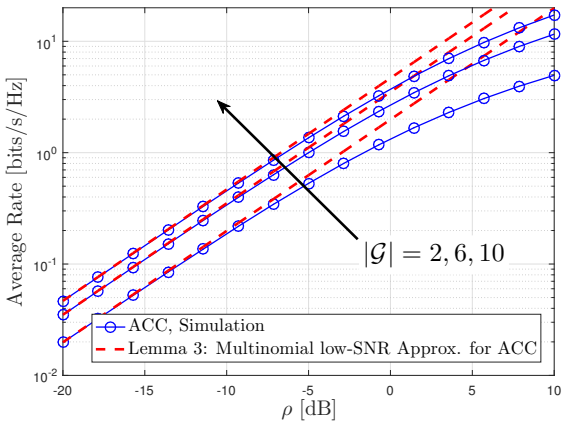



Fig. 7: $\bar{R}^{(\mathrm{ACC})}$ versus $\rho$ for $B = 3$.

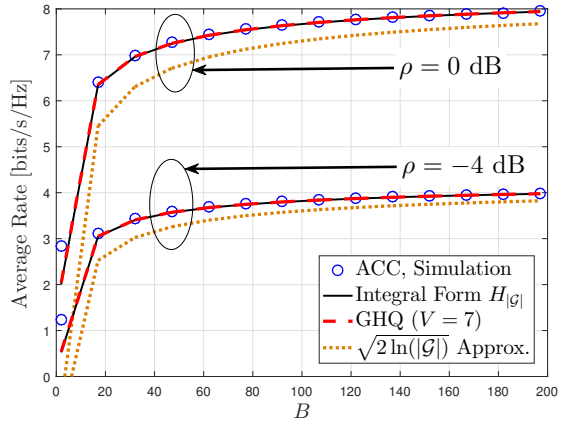



Fig. 9: $\bar{R}^{(\mathrm{ACC})}$ versus $B$ for $|\mathcal{G}| = 10$.

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.

From the same figures, it is also worth noting that, when $\rho$ is relatively small, the effective coded-caching gains of both schemes arrive to a flat lower bound. The lower bound for the ACC scheme is notably greater and improves as either $B$ or $|\mathcal{G}|$ become bigger. However, this behavior does not extend to the MN scheme, which is consistent with the result of Proposition 2 stating that the effective gain of the MN scheme collapses at low SNR regardless of the value of the high-SNR caching gain $|\mathcal{G}|$. Moreover, in Fig. 5, we can see that for the MN scheme the worst-user effect is amplified as $|\mathcal{G}|$ increases. Therefore, Figs. 4–5 show that the advantages of the ACC scheme in terms of average rate are still significant even for a small group size ($B = 4, 6$).

### *B. Approximations on the Average Rate $\bar{R}^{(\mathrm{ACC})}$*

In Figs. 6–9, we validate the derived analytical approximations and highlight some interesting trends and comparisons. First, Fig. 6 shows the average rate $\bar{R}^{(\mathrm{ACC})}$ versus $\rho$ for different values of $B$. Note that, for $B = 1$, $\bar{R}^{(\mathrm{ACC})} = \bar{R}^{(\mathrm{MN})}$. For comparison, Fig. 6 displays the simulated result (circle and asterisk symbols), the exact derived average rate $\bar{R}^{(\mathrm{ACC})}$ in Lemma 1 (solid line), the low-SNR multinomial approximation in Lemma 3 (dashed line), and the low-SNR second-order approximation for $\bar{R}^{(\mathrm{MN})}$ in Lemma 2 (dotted line). The rate enhancement due to the ACC scheme is exhibited by comparing the results of Lemma 1 and Lemma 2 (solid and dotted lines, respectively). Fig. 6 shows that the accuracy of the approximation for $\bar{R}^{(\mathrm{MN})}$ in Lemma 2 is better than the approximation for $\bar{R}^{(\mathrm{ACC})}$ in Lemma 3, mainly because Lemma 3 considers a first-order approximation. Fig. 7 reveals that the approximation derived in Lemma 3 becomes more accurate as $|\mathcal{G}|$ increases, which indicates that the value of $\rho$ at which the nonlinear part of the average rate becomes significant increases as $|\mathcal{G}|$ increases.

The large-$B$ approximation of $\bar{R}^{(\mathrm{ACC})}$ from Lemma 6 is validated in Fig. 8, where the average rate is plotted for different $|\mathcal{G}|$. The values of $H_{|\mathcal{G}|}$ are taken from the table presented in Section IV-D. This large-$B$ approximation tightly approximates the simulation results, even for a small $B$. In fact, this approximation is extremely tight for any value of $B$ bigger than 1. To further demonstrate the accuracy of Lemma 6, we show in Fig. 9 the results derived by using $i$) the integral calculation in (21), $ii$) the GHQ method in (22), and $iii$) the $\sqrt{2\ln(|\mathcal{G}|)}$ approximation of $H_{|\mathcal{G}|}$ for $|\mathcal{G}| > 5$.

After verifying the high accuracy of the approximation in Lemma 6, we exploit it to present some interesting comparisons

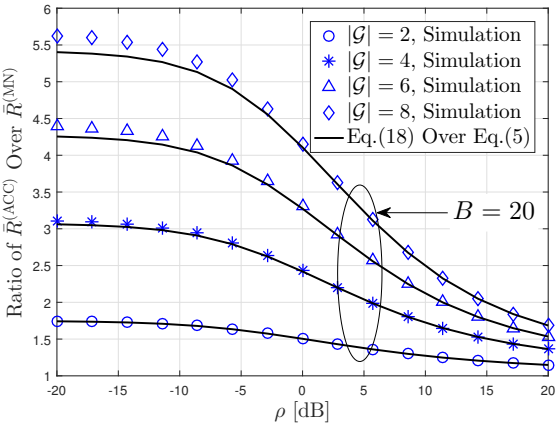



Fig. 10: $\bar{R}^{(\mathrm{ACC})}/\bar{R}^{(\mathrm{MN})}$ versus $\rho$ for $V = 7$ in GHQ.

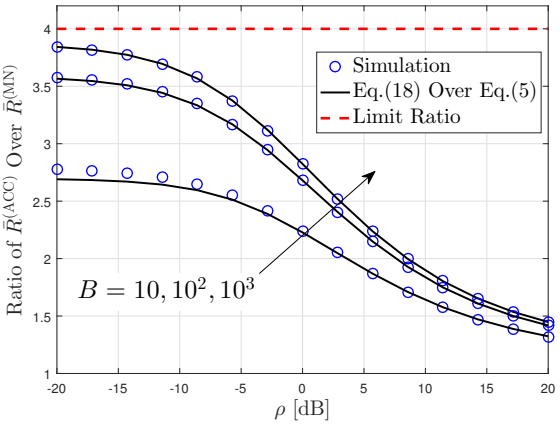



Fig. 11: $\bar{R}^{(\mathrm{ACC})}/\bar{R}^{(\mathrm{MN})}$ versus $\rho$ for $|\mathcal{G}| = 4$.

between the ACC scheme and the MN scheme in Figs. 10–11. In Fig. 10, we can see through the ratio $\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}}$ that $\bar{R}^{(\mathrm{ACC})}$ provides significant boost for realistic SNR values. In order to illustrate the extent to which this ratio approaches the theoretical gain in the low-SNR regime, we show in Fig. 11 the different ratios/improvements achieved by varying $B$.

### *C. Exploiting the Approach in other Settings: Delivery Time on the Decentralized-Placement Scenario*

In order to show the generality of the key ideas underlying the ACC scheme, we provide an example of its application in a decentralized coded caching setting with finite file-size constraints. We then compare our new decentralized scheme with the state-of-the-art scheme from [21].

In the decentralized scenario of [21], the subpacketization constraint induces a certain number $\Lambda$ of cache states. The main difference with our previous setting is that, during placement, each user *independently* selects one of the $\Lambda$ cache states *uniformly at random* and stores it in its cache, such that each of the $\Lambda$ cache states will be stored at a different number of users. Let $B_g$ denote the number of users storing the $g$-th cache state, and note that $\sum_{g=1}^{\Lambda} B_g = K$.

During the delivery phase, the scheme in [21] first serves one user from each cache state by implementing sequential XOR transmissions as if we applied the standard MN scheme for $\Lambda$ users. After that, the transmission procedure is repeated for the next user of each cache state for the cache states still including some not-served users. We refer to [21] for more details.

We recall that, in contrast to the scheme from [21], the proposed ACC scheme sequentially serves all users in a set of cache states $\mathcal{G}$ of $|\mathcal{G}| = \Lambda\gamma + 1$ cache states (i.e., $\sum_{g\in\mathcal{G}} B_g$ users). Once all these users have received their subfile, the transmitter starts to serve another cache-state set $\mathcal{G}'$. As mentioned in Remark 4, the ACC scheme can be directly applied to the case in which each cache state is stored at a different number of users.

Let us now analyze the benefits of using the ideas from the ACC scheme in this setting. Since now the number of users per cache state may (and probably will) differ, we need to consider the average delivery time instead of the average rate. Note, however, that the delivery time over Rayleigh fading channels does not converge, as previously mentioned. Hence, for comparative purposes, we consider Nakagami-$m$ fading to model the wireless propagation [39]. The delivery time of the centralized ACC scheme over Nakagami-$m$ fading channels has been recently analyzed in [46].

We can obtain from (6) the total delivery time of the decentralized ACC scheme as

$$T_{\mathrm{ACC}} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}} \sum_{\substack{\mathcal{G}\subseteq[\Lambda]\\ |\mathcal{G}|=\Lambda\gamma+1}} \max_{g\in\mathcal{G}} \left\{ \sum\nolimits_{b=1}^{B_g} \left[ \ln\left(1+\mathrm{SNR}_{g,b}\right)\right]^{-1} \right\} \text{ s.}$$

Upon defining $B_{\max} \triangleq \max_{g\in[\Lambda]}\{B_g\}$ and considering the same assumption of quasi-static fading as for the ACC scheme, the total delivery time in the coded caching scheme of [21] is

$$T_{\mathrm{Dec}} = \frac{F}{\binom{\Lambda}{\Lambda\gamma}} \sum_{b=1}^{B_{\max}} \sum_{\substack{\mathcal{G}\subseteq[\Lambda]\\ |\mathcal{G}|=\Lambda\gamma+1}} \max_{\substack{g\in\mathcal{G}\\ B_g\geq b}} \left\{ \left[ \ln\left(1+\mathrm{SNR}_{g,b}\right)\right]^{-1} \right\} \text{ s,}$$

where $\mathrm{SNR}_{g,b}$ is the SNR of the $b$-th ($b \in [B_g]$) user of the $g$-th cache state.

For comparison, we also consider the performance of uncoded caching. When users request different files, the total delivery time is $T_{\mathrm{unCC}} = \sum_{k=1}^{K} \frac{(1-\gamma)F}{\ln(1+\mathrm{SNR}_k)}$ s.

Next, we numerically evaluate the ratios $T_{\mathrm{unCC}}/T_{\mathrm{ACC}}$ and $T_{\mathrm{unCC}}/T_{\mathrm{Dec}}$, averaged over channel states and cache-state allocations, to compare the delivery time boost of the proposed approach over Nakagami-$m$ fading channels. We consider that the distribution of users in cache states follows a Multinomial distribution with $\Lambda$ equally probable outcomes (cf. [21]).

We can observe in Fig. 12 how the decentralized ACC scheme considerably improves the performance of coded caching, and this enhancement is more acute in the low-to-moderate SNR region. As previously pointed out, the main reason for this improvement is the amelioration of the worst-user bottleneck, where this amelioration is again the result of using shared caches as a leverage to reduce delay variability. The fact that this reduction of the worst-user bottleneck is improved as the total number of users $K$ increases (cf. Lemma 4) is exemplified by comparing the $K = 600$ case in Fig. 12 with the $K = 300$ case.

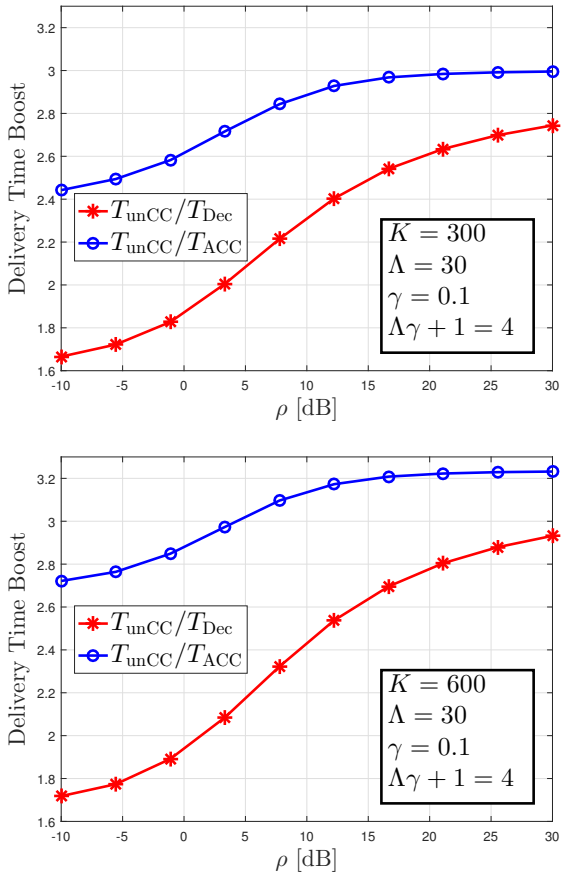



Fig. 12: Comparison of Delivery Time versus $\rho$ for $m = 2$ in the decentralized-placement scenario.

## VI. Conclusions

This work is motivated by the fact that any attempt to successfully adopt wireless coded caching in large-scale settings must account for the effects of low-to-moderate SNR fading channels. In this respect, we have first revealed that dedicated caches and XOR-based transmissions may no longer be suitable for various realistic SNR regimes. As we have seen, as the SNR becomes smaller, the effective gains of XOR-based schemes collapse, irrespective of either the nominal gain or the number of users. We have then proposed a novel dual idea that combines the use of cache replication and a multi-rate transmission scheme. This approach recovers a big fraction of the lost gains and does so for any SNR value. These gains are fully recovered in the regime of many users, again for any SNR value, thus essentially resolving the worst-user bottleneck.

The use of cache replication or shared cache states, which is enforced in practical coded caching settings due to the file-size constraint, turns out to be beneficial in the low SNR regime. These gains appear in practical values of SNR and for realistically many users. We have shown how having as few as 2 users per cache state allows the proposed scheme to approximately double the effective coded caching gain. As stated before, these gains do not involve user selection, and the corresponding user-grouping is done prior to cache-placement and is oblivious to the demands and of course oblivious to the channel. Finally, the derived expressions are simple but very precise. For example, the low SNR approximation for the MN scheme in Lemma 2 is essentially identical to the actual performance even for SNR values as high as 20 dB. Similarly, as Fig. 8 shows, the large-$B$ approximation is almost exact even for values of $B$ as low as 10.

In the end, the proposed scheme applies toward showing that properly designed coded caching has the ability to substantially speed up delivery of multimedia content even in the challenging environment of low-to-moderate SNR fading channels.

## Appendix I: Capacity Region of Proposition 1

This appendix is meant to orient the reader as to how the existing results in [35] on multicasting with side information[9] can be applied to our setting.

Using the notation of [35] and following the same derivation as in [36], we recover Proposition 1 from [35, Thm. 6] by choosing $X^n$ to be $(X_1,\ X_2,\ \cdots,\ X_t)^n$, selecting $m = n$ in [35, Thm. 6], setting the side information $Y_i$ to be $Y_i = \{X_\ell\}_{\ell\in[t]\setminus i}$, and applying invertible mappings between $X_i^n$ and $W_i'$ for any $i \in t$. From the maximum entropy theorem [48, Thm. 9.6.5], we obtain Proposition 1.

For the achievability part, we proceed as in [35] and consider a codebook of $2^{n(\sum_{\ell=1}^t R_\ell)}$ codewords. The codewords are denoted by $x^n(w_1, w_2, \cdots, w_t)$, with $w_\ell \in [2^{nR_\ell}]$ for any $\ell \in [t]$. The letters of the codewords, denoted by $x_j(w_1, w_2, \cdots, w_t)$, $j \in [n]$, are i.i.d. distributed as $\mathcal{N}(0, P)$. Each user can decode its intended message from the received signal and from the (cached) side information using typical set decoding. The intuition behind the successful decoding at a certain user $i$ is that, after receiving one of the $2^{n(\sum_{\ell=1}^t R_\ell)}$ codewords and thanks to the cached information, user $i$ applies typical decoding over only $2^{nR_i}$ possible codewords.

## Appendix II: Proof of Lemma 1

Let us start by introducing the notation $S_g \triangleq \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b})$, for any group $g \in [\Lambda]$ of users, such that we can write the average rate of the ACC scheme as

$$\bar{R}^{(\text{ACC})} = \frac{|\mathcal{G}|}{B \ln 2} \mathbb{E}_H \left\{ \min_{g\in\mathcal{G}} \{S_g\} \right\}. \tag{23}$$

For $t \in (-\infty, +\infty)$, the characteristic function (CF) in probability [49, Ch. 5] of $S_g$ is defined as

$$\begin{aligned} \text{CF}_{S_g}(t) = \mathbb{E}\left\{\exp(jtS_g)\right\} &= \mathbb{E}\left\{ \exp\left( jt \sum_{b=1}^B \ln(1 + \text{SNR}_{g,b}) \right) \right\} \\ &= \left[ \mathbb{E}\left\{ (1 + \text{SNR}_{g,b})^{jt} \right\} \right]^B. \end{aligned} \tag{24}$$

Substituting the PDF of $\text{SNR}_{g,b}$ into (24) yields

$$\begin{aligned} \text{CF}_{S_g}(t) &= \frac{1}{\rho^B} \left[ \int_0^\infty (1+x)^{jt} \exp\left(-\frac{x}{\rho}\right) \mathrm{d}x \right]^B \\ &\overset{(a)}{=} \frac{1}{\rho^B} \exp\left(\frac{B}{\rho}\right) \mathrm{E}_{-jt}^B\left(\frac{1}{\rho}\right), \end{aligned} \tag{25}$$

[9] Several works have considered this Gaussian setting after [35]. In [36], the capacity region was derived for the 2-user case, the 3-user case was studied in [37, Group 8, case $\mathcal{G}_{18} \bigcup \mathcal{G}_{28}$], and the converse of Prop. 1 can be also found in [47, Thm. 4].

where $(a)$ follows from [40, Eq. (3.382.4)]. By considering the Gil-Pelaez Theorem [50], the CDF of $S_g$ is obtained as

$$F_{S_g}(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\mathrm{Im}\left\{\exp(-\jmath x t)\frac{\exp(B/\rho)}{\rho^B}\mathrm{E}^B_{-\jmath t}\left(\frac{1}{\rho}\right)\right\}}{t}\,\mathrm{d}t.$$

Define $J \triangleq \min_{g\in\mathcal{G}}\{S_g\} = \min_{g\in\mathcal{G}}\left\{\sum_{b=1}^B \ln(1+\mathrm{SNR}_{g,b})\right\}$. The CDF of $J$ can be expressed by

$$\begin{aligned} F_J(y) &= \Pr\left\{\min_{g\in\mathcal{G}}\{S_g\} \le y\right\} \\ &= 1 - \Pr\left\{\min_{g\in\mathcal{G}}\{S_g\} > y\right\} = 1 - \left(\Pr\{S_g > y\}\right)^{|\mathcal{G}|} \\ &= 1 - \left(\frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\mathrm{Im}\left\{\exp(-\jmath x t)\frac{\exp(\frac{B}{\rho})}{\rho^B}\mathrm{E}^B_{-\jmath t}\left(\frac{1}{\rho}\right)\right\}}{t}\mathrm{d}t\right)^{|\mathcal{G}|}. \end{aligned} \quad (26)$$

As $J$ is a non-negative random variable, it holds that $\mathbb{E}\{J\} = \mathbb{E}\left\{\int_0^J \mathrm{d}x\right\}$, and furthermore,

$$\begin{aligned} \mathbb{E}\left\{\int_0^J \mathrm{d}x\right\} &= \mathbb{E}\left\{\int_0^\infty \mathbb{I}\{x \le J\}\,\mathrm{d}x\right\} \\ &= \int_0^\infty \mathbb{E}\{\mathbb{I}\{x \le J\}\}\,\mathrm{d}x = \int_0^\infty [1 - F_J(y)]\,\mathrm{d}y, \end{aligned} \quad (27)$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function, which, for claim $\mathcal{A}$, takes the value $\mathbb{I}\{\mathcal{A}\} = 1$ if $\mathcal{A}$ is true and $\mathbb{I}\{\mathcal{A}\} = 0$ otherwise. Combining (26) and (27) yields that the expectation of $J$ is

$$\mathbb{E}\{J\} = \int_0^\infty \left(\frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\mathrm{Im}\left\{\exp(-\jmath x t)\frac{\exp(B/\rho)}{\rho^B}\mathrm{E}^B_{-\jmath t}\left(\frac{1}{\rho}\right)\right\}}{t}\mathrm{d}t\right)^{|\mathcal{G}|}\mathrm{d}y.$$

It follows from (23) that $\bar{R}^{(\mathrm{ACC})} = \frac{|\mathcal{G}|}{B\ln 2}\mathbb{E}\{J\}$, which gives Lemma 1 by considering the integral form of $\mathbb{E}\{J\}$, and therefore Lemma 1 is proven. □

## Appendix III: Proofs for Section IV-B and Section IV-C

### A. Proof of Lemma 2

The fact that $\mathrm{SNR}_g$ is distributed as $\mathrm{Exp}(|\mathcal{G}|/\rho)$ implies that $\mathrm{Var}(\min_{g\in\mathcal{G}}\{\mathrm{SNR}_g\}) = \rho^2/|\mathcal{G}|^2 = o(\rho)$. Thus, in a similar way as in [51, Eq. (4)], in the low-SNR region we can approximate $\bar{R}^{(\mathrm{MN})}$ by its robust approximation based on the Taylor series: Let $P(X)$ be a real-valued function with respect to a random variable $X$ with mean $\mu_X$ and variance $\sigma_X^2$. The expectation of $P(X)$ can be tightly approximated in the low $\sigma_X^2$ region as

$$\mathbb{E}\{P(X)\} \approx P(\mu_X) + \frac{\sigma_X^2}{2}\left.\frac{\partial^2 P(X)}{\partial X^2}\right|_{X=\mu_X} \quad (28)$$

where $\frac{\partial^2 P(X)}{\partial X^2}$ stands for the second derivative of $P(X)$ with respect to $X$ (cf. [52]).

Considering that $P(X) = \frac{|\mathcal{G}|}{\ln 2}\ln\left(1 + \min_{g\in\mathcal{G}}\{\mathrm{SNR}_g\}\right)$ and that $X = \min_{g\in\mathcal{G}}\{\mathrm{SNR}_g\}$ and adopting the robust approximation of (28) yields that $\bar{R}^{(\mathrm{MN})}$ can be tightly approximated at low SNR by (9). □

### B. Proof of Proposition 2

From the fact that $\mathrm{Ei}(-x)$ is bounded as (cf. [53])

$$-e^{-x}\ln\left(1+\frac{1}{x}\right) < \mathrm{Ei}(-x) < \frac{-e^{-x}}{2}\ln\left(1+\frac{2}{x}\right), \quad (29)$$

we can upper bound the numerator and lower bound the denominator of the exact expression of $\frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}}$ in (10) to obtain that

$$\lim_{\rho\to 0}\frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}} \le \lim_{\rho\to 0}\frac{|\mathcal{G}|}{2}\frac{\ln\left(1+\frac{2\rho}{|\mathcal{G}|}\right)}{\ln(1+\rho)} = 1. \quad (30)$$

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 Proposition 2. □

### C. Proof of Lemma 3

We start by proving that

$$\mathbb{E}\left\{\sum_{b=1}^B \ln(1+\mathrm{SNR}_{g,b})\right\} = \mathbb{E}\left\{\sum_{b=1}^B \mathrm{SNR}_{g,b}\right\} + o(\rho), \quad (31)$$

which is obtained from the fact that $\mathbb{E}\left\{\sum_{b=1}^B \ln(1+\mathrm{SNR}_{g,b})\right\} = \mathbb{E}\left\{\sum_{b=1}^B \left(\ln(1+\mathrm{SNR}_{g,b}) - \mathrm{SNR}_{g,b}\right)\right\} + \mathbb{E}\left\{\sum_{b=1}^B \mathrm{SNR}_{g,b}\right\}$. In the above, we obtain (31) from the Lebesgue's Dominated Convergence Theorem [54, Thm. 16.4] as follows: First, we know that $\lim_{x\to 0}(\ln(1+x)-x)/x = 0$, and hence $\ln(1+x) - x = o(x)$ as $x \to 0$. In order to prove that the expectation is also $o(\rho)$ as $\rho \to 0$, we need to prove that $|\ln(1+x) - x|$ is bounded by some integrable function. For that, since $\ln(1+x) \le x$ for any $x > 0$, it follows that $|\ln(1+\mathrm{SNR}_{g,b}) - \mathrm{SNR}_{g,b}| \le |\mathrm{SNR}_{g,b}|$, which satisfies that $\mathbb{E}\{|\mathrm{SNR}_{g,b}|\} = \rho < \infty$. Hence, we can apply the Dominated Convergence Theorem and obtain (31).

Since $\mathrm{SNR}_{g,b}$ is distributed as $\mathrm{Exp}(\frac{1}{\rho})$, $\sum_{b=1}^B \mathrm{SNR}_{g,b}$ follows a $\mathrm{Gamma}(B,\rho)$ distribution, with shape and scale parameters $B$ and $\rho$. Then, the CDF of $\Phi \triangleq \min_{g\in\mathcal{G}}\{\sum_{b=1}^B \mathrm{SNR}_{g,b}\}$ is given by

$$\begin{aligned} F_\Phi(y) &= 1 - \left(\frac{1}{\Gamma(B)}\Gamma\left(B, \frac{y}{\rho}\right)\right)^{|\mathcal{G}|} \\ &\overset{(a)}{=} 1 - \left(\exp\left(-\frac{y}{\rho}\right)\sum_{t=0}^{B-1}\frac{y^t}{t!\,\rho^t}\right)^{|\mathcal{G}|}, \end{aligned} \quad (32)$$

where $\Gamma(\cdot,\cdot)$ denotes the upper incomplete Gamma function [40], and $(a)$ follows from [40, Eq. (8.352.2)] since $B$ is a positive integer. For $\mathbf{b} \in \mathbb{Z}^B$, let $b_t \triangleq \mathbf{b}(t) \ge 0$, $t \in [B]$, denote its $t$-th element. Recalling that $\binom{n}{\mathbf{b}} \triangleq \frac{n!}{b_1!b_2!\cdots b_B!}$, we apply the Multinomial theorem [55] to get that

$$\begin{aligned} F_\Phi(y) = 1 &- \exp\left(-\frac{|\mathcal{G}|y}{\rho}\right) \\ &\times \sum_{||\mathbf{b}||_1=|\mathcal{G}|}\binom{|\mathcal{G}|}{\mathbf{b}}\frac{\rho^{-\sum_{t=1}^B (t-1)b_t}}{\prod_{t=1}^B((t-1)!)^{b_t}}y^{\sum_{t=1}^B(t-1)b_t}. \end{aligned}$$

In view of the relationship between the CDF and the expectation in (27), the average rate of the ACC scheme can be approximated in the low-SNR region by

$$\begin{aligned}\bar{R}^{(\mathrm{ACC})} &= \frac{|\mathcal{G}|}{B\ln 2}\big(\mathbb{E}\{\Phi\}+o(\rho)\big)\\ &= \frac{|\mathcal{G}|}{B\ln 2}\int_0^\infty \big[1-F_\Phi(y)\big]\mathrm{d}y + o(\rho)\\ &= \frac{|\mathcal{G}|}{B\ln 2}\sum_{||\mathbf{b}||_1=|\mathcal{G}|}\binom{|\mathcal{G}|}{\mathbf{b}}\frac{\rho^{-\sum_{t=1}^B (t-1)b_t}}{\prod_{t=1}^B((t-1)!)^{b_t}}\\ &\quad\times\int_0^\infty \exp\left(-\frac{|\mathcal{G}|y}{\rho}\right)y^{\sum_{t=1}^B(t-1)b_t}\mathrm{d}y+o(\rho), \quad (33)\end{aligned}$$

which can be solved by using the definition of Gamma function [40, Eq. (8.312.2)]. □

### D. Proof of Corollary 1

From Lemma 3 we have that $\bar{R}^{(\mathrm{ACC})} = \frac{\rho|\mathcal{G}|}{B\ln 2}\Psi_{|\mathcal{G}|}+o(\rho)$ and also that $\bar{R}^{(\mathrm{TDM})} = \bar{R}^{(\mathrm{ACC})}\big|_{B=|\mathcal{G}|=1} = \frac{\rho}{\ln 2}+o(\rho)$, whereas from Proposition 2 it follows that $\lim_{\rho\to 0}\frac{\bar{R}^{(\mathrm{MN})}}{\bar{R}^{(\mathrm{TDM})}}=1$. These results yield the desired $\bar{R}^{(\mathrm{MN})} = \frac{\rho}{\ln 2}+o(\rho)$ and $\lim_{\rho\to 0}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}} = \lim_{\rho\to 0}\frac{\frac{\rho|\mathcal{G}|}{B\ln 2}\Psi_{|\mathcal{G}|}+o(\rho)}{\frac{\rho}{\ln 2}+o(\rho)} = \frac{|\mathcal{G}|}{B}\Psi_{|\mathcal{G}|}$. □

### E. Proof of Lemma 4

We want to prove that $\lim_{B\to\infty}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{TDM})}} = \Lambda\gamma+1$ for a fixed number of caches $\Lambda$ and for any $\rho$. Since $\mathbb{E}\{|\ln(1+\mathrm{SNR}_{g,b})|\}<\infty$, the Strong Law of Large Numbers implies that

$$\frac{1}{B}\sum_{b=1}^B \ln(1+\mathrm{SNR}_{g,b}) \xrightarrow{a.s.} \mathbb{E}\{\ln(1+\mathrm{SNR}_{g,b})\} \quad (34)$$

as $B\to\infty$, which implies that

$$\lim_{B\to\infty}\frac{1}{B}\sum_{b=1}^B \ln(1+\mathrm{SNR}_{g,b}) = \mathbb{E}\{\ln(1+\mathrm{SNR}_{g,b})\}, \quad (35)$$

except for zero-probability events. Then since $\ln(1+x)\le x$ $\forall x>0$, we get that

$$\begin{aligned}&\mathbb{E}_H\Big\{\min_{g\in\mathcal{G}}\frac{1}{B}\sum_{b=1}^B\ln(1+\mathrm{SNR}_{g,b})\Big\}\\ &\qquad\le \mathbb{E}_H\Big\{\frac{1}{B}\sum_{b=1}^B \mathrm{SNR}_{g,b}\Big\}\overset{(a)}{=}\rho<\infty, \quad (36)\end{aligned}$$

where $(a)$ comes from the fact that $\mathrm{SNR}_{g,b}\sim\mathrm{Exp}(\frac{1}{\rho})$ and thus $\sum_{b=1}^B\mathrm{SNR}_{g,b}\sim\mathrm{Gamma}(B,\rho)$.

From (35) and (36), we can apply Lebesgue's Dominated Convergence Theorem [54, Thm. 16.4] to interchange the order of expectation and limit and show that

$$\begin{aligned}&\lim_{B\to\infty}\bar{R}^{(\mathrm{ACC})}/\bar{R}^{(\mathrm{TDM})}\\ &\overset{(a)}{=}\frac{\lim_{B\to\infty}\frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\left\{\min_{g\in\mathcal{G}}\frac{1}{B}\sum_{b=1}^B\ln(1+\mathrm{SNR}_{g,b})\right\}}{\frac{1}{\ln 2}\mathbb{E}_H\{\ln(1+\mathrm{SNR}_{g,b})\}} \quad (37)\\ &\overset{(b)}{=}|\mathcal{G}|\,\frac{\mathbb{E}_H\left\{\min_{g\in\mathcal{G}}\lim_{B\to\infty}\frac{1}{B}\sum_{b=1}^B\ln(1+\mathrm{SNR}_{g,b})\right\}}{\mathbb{E}_H\{\ln(1+\mathrm{SNR}_{g,b})\}}\\ &\overset{(c)}{=}|\mathcal{G}|=\Lambda\gamma+1, \quad (38)\end{aligned}$$

where $(a)$ follows from substituting $\bar{R}^{(\mathrm{ACC})}$ and $\bar{R}^{(\mathrm{TDM})}$ by their respective expressions, $(b)$ comes from the Dominated Convergence Theorem and the fact that the minimum of several continuous functions is a continuous function, and $(c)$ is due to (35). □

### F. Proof of Lemma 5

From (34), and by applying the same steps as in (37)–(38), we obtain (16) as

$$\begin{aligned}\lim_{B\to\infty}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}} &= \frac{\frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\{\ln(1+\mathrm{SNR}_{g,b})\}}{\frac{|\mathcal{G}|}{\ln 2}\mathbb{E}_H\{\ln(1+\min_{g\in\mathcal{G}}\{\mathrm{SNR}_{g,b}\})\}}\\ &\overset{(a)}{=}\exp\left(\frac{1}{\rho}-\frac{|\mathcal{G}|}{\rho}\right)\frac{\mathrm{Ei}\left(-\frac{1}{\rho}\right)}{\mathrm{Ei}\left(-\frac{|\mathcal{G}|}{\rho}\right)}, \quad (39)\end{aligned}$$

where $(a)$ follows from (5). To prove (17), we first obtain from (39) that

$$\lim_{\rho\to 0}\lim_{B\to\infty}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}} = \lim_{\rho\to 0}\exp\left(\frac{1}{\rho}-\frac{|\mathcal{G}|}{\rho}\right)\frac{\mathrm{Ei}\left(-\frac{1}{\rho}\right)}{\mathrm{Ei}\left(-\frac{|\mathcal{G}|}{\rho}\right)}. \quad (40)$$

Then, in a similar manner as for the proof of Proposition 2 in Appendix III-B, we can apply the relations $-e^{-x}\ln(1+\frac{1}{x})<\mathrm{Ei}(-x)<\frac{-e^{-x}}{2}\ln(1+\frac{2}{x})$ [53] in (40) to obtain that

$$\begin{aligned}&\lim_{\rho\to 0}\lim_{B\to\infty}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}}\\ &\le\lim_{\rho\to 0}\exp\left(\frac{1-|\mathcal{G}|}{\rho}\right)\frac{\frac{1}{2}\exp\left(\frac{-1}{\rho}\right)\ln(1+2\rho)}{\exp\left(\frac{-|\mathcal{G}|}{\rho}\right)\ln(1+\frac{\rho}{|\mathcal{G}|})}=|\mathcal{G}|, \quad (41)\\ &\lim_{\rho\to 0}\lim_{B\to\infty}\frac{\bar{R}^{(\mathrm{ACC})}}{\bar{R}^{(\mathrm{MN})}}\\ &\ge\lim_{\rho\to 0}\exp\left(\frac{1-|\mathcal{G}|}{\rho}\right)\frac{2\exp\left(\frac{-1}{\rho}\right)\ln(1+\rho)}{\exp\left(\frac{-|\mathcal{G}|}{\rho}\right)\ln(1+\frac{2\rho}{|\mathcal{G}|})}=|\mathcal{G}|, \quad (42)\end{aligned}$$

which concludes the proof of Lemma 5. □

## Appendix IV: Proof of Lemma 6

To prove Lemma 6, we first derive the approximation in (18). Afterward, we obtain the values of $\mu$ and $\sigma$ in (19) and (20), and finally we derive the integral expression of $H_{|\mathcal{G}|}$ in (21).

### A. Approximation for the Rate of the ACC Scheme

Let $A_g \triangleq \frac{1}{B}\sum_{b=1}^B\ln(1+\mathrm{SNR}_{g,b})$, for any $g\in[\Lambda]$, represent the arithmetic mean of the user capacity over the set of $B$ users of group $g$, normalized by $\ln(2)$. Let us consider the Central Limit Theorem (CLT) in the large $B$ case. According to the Lindeberg-Lévy CLT [56], we have that $A_g \xrightarrow{d.} \mathcal{N}\left(\mu,\frac{\sigma^2}{B}\right)$ as $B\to\infty$, where $d.$ stands for *convergence in distribution*, and where $\mu=\mathbb{E}\{\ln(1+\mathrm{SNR}_{g,b})\}$ and $\sigma^2=\mathrm{Var}\{\ln(1+\mathrm{SNR}_{g,b})\}$[10]. We consider now the

[10]Note that, if we focused on the low-SNR region, we could apply the approximations $\mu\approx\mathbb{E}\{\mathrm{SNR}_{g,b}\}$ and $\sigma^2\approx\mathrm{Var}\{\mathrm{SNR}_{g,b}\}$. We do not consider them here for sake of generality, and our approximation holds for any value of SNR.

average rate for the ACC scheme when $B \to \infty$. Recall that $A_1, \cdots, A_{|\mathcal{G}|}$ are i.i.d. normal random variables with mean $\mu$ and variance $\sigma^2/B$. Although convergence in distribution does not generally imply convergence in mean, it was shown in [57] that this indeed holds in the specific case of extreme values of i.i.d. random variables. Consequently, $\bar{R}^{(\mathrm{ACC})}$ is given by

$$\lim_{B\to\infty} \bar{R}^{(\mathrm{ACC})} = \frac{|\mathcal{G}|}{\ln 2}\mathbb{E}\left\{\min\left\{A_1, \cdots, A_{|\mathcal{G}|}\right\}\right\}. \tag{43}$$

Deriving a simple closed-form expression for (43) is challenging. Consequently, we propose a simple method to obtain an approximation to this expectation. Since $B \to \infty$ and $A_1, \cdots, A_{|\mathcal{G}|}$ are i.i.d. normal random variables, we can write each $A_i$, $i \in [|\mathcal{G}|]$, as $A_i = \mu + \frac{\sigma}{\sqrt{B}}A'_i$, where $A'_i \sim \mathcal{N}(0,1)$. Then, the minimum of $A_1, \cdots, A_{|\mathcal{G}|}$ is re-written as

$$\min_{i\in[|\mathcal{G}|]}\{A_i\} = \mu + \frac{\sigma}{\sqrt{B}}\min_{i\in[|\mathcal{G}|]}\{A'_i\}. \tag{44}$$

Then (18) is obtained by taking the expectation of both sides, multiplying (44) by $\frac{|\mathcal{G}|}{\ln 2}$, and recalling that $H_{|\mathcal{G}|} \triangleq -\mathbb{E}\left\{\min_{i\in[|\mathcal{G}|]}\{A'_i\}\right\}$, as defined in Section IV-D. □

### B. Proof of (19)-(20): Mean and Variance of $\ln(1+\mathrm{SNR}_{g,b})$

We derive now the expressions for $\mu$ in (19) and $\sigma$ in (20). Note that $\frac{\mu}{\ln(2)}$=$\mathbb{E}\left\{\log_2(1+\mathrm{SNR}_{g,b})\right\}$ is exactly $\bar{R}^{(\mathrm{TDM})}$, so that we have (19) by considering (5) with $|\mathcal{G}| = 1$. Moreover, we have that

$$\mathbb{E}\left\{\left(\ln(1+\mathrm{SNR}_{g,b})\right)^2\right\} = \frac{1}{\rho}\int_0^\infty \left(\ln(1+x)\right)^2 \exp\left(-\frac{x}{\rho}\right)\mathrm{d}x.$$

To obtain a closed-form expression for the previous integral, we re-write both the logarithmic function and the exponential function into their Meijer's G-function forms [40, Eq. (9.301)], given by $\ln(1+x) = \mathrm{G}^{1,2}_{2,2}\left(x\left|{}^{1,1}_{1,0}\right.\right)$ and $\exp\left(-\frac{x}{\rho}\right) = \mathrm{G}^{1,0}_{0,1}\left(\frac{x}{\rho}\left|{}^{-}_{0}\right.\right)$, respectively. Then, the previous integral becomes

$$\begin{aligned}
&\mathbb{E}\left\{\left(\ln(1+\mathrm{SNR}_{g,b})\right)^2\right\} \\
&\quad= \frac{1}{\rho}\int_0^\infty \mathrm{G}^{1,2}_{2,2}\left(x\left|{}^{1,1}_{1,0}\right.\right)\mathrm{G}^{1,2}_{2,2}\left(x\left|{}^{1,1}_{1,0}\right.\right)\mathrm{G}^{1,0}_{0,1}\left(\frac{x}{\rho}\left|{}^{-}_{0}\right.\right)\mathrm{d}x \\
&\quad\overset{(a)}{=} 2\exp\left(\frac{1}{\rho}\right)\mathrm{G}^{3,0}_{2,3}\left(\frac{1}{\rho}\left|{}^{1,1}_{0,0,0}\right.\right)
\end{aligned}$$

where $(a)$ follows from [58, Eq. (07.34.21.0081.01)] after basic simplifications. By combining this expression with the relationship $\sigma^2 = \mathbb{E}\{\left(\ln(1+\mathrm{SNR}_{g,b})\right)^2\} - \left(\mathbb{E}\left\{\ln(1+\mathrm{SNR}_{g,b})\right\}\right)^2$, we obtain (20). □

### C. Proof of (21)

To derive the integral form of $H_{|\mathcal{G}|}$, we calculate the CDF of $\Omega \triangleq \min\{A'_1, \cdots, A'_{|\mathcal{G}|}\}$ to be

$$\begin{aligned}
F_\Omega(y) &= 1 - \Pr\left\{\min\left\{A'_1, \cdots, A'_{|\mathcal{G}|}\right\} > y\right\} \\
&= 1 - \left(\Pr\left\{A'_1 > y\right\}\right)^{|\mathcal{G}|} \overset{(a)}{=} 1 - \left(Q(y)\right)^{|\mathcal{G}|},
\end{aligned} \tag{45}$$

where $(a)$ holds because the CDF of the standard normal distribution is $F_{A'_i}(x) = 1 - Q(x)$.

The corresponding PDF is then derived by

$$\begin{aligned}
f_\Omega(y) &= \frac{\partial F_\Omega(y)}{\partial y} = -|\mathcal{G}|\left(Q(y)\right)^{|\mathcal{G}|-1}\frac{\partial Q(y)}{\partial y} \\
&\overset{(a)}{=} \frac{1}{\sqrt{2\pi}}|\mathcal{G}|\left(Q(y)\right)^{|\mathcal{G}|-1}\exp\left(-\frac{y^2}{2}\right),
\end{aligned}$$

where $(a)$ follows from the integral form of the Q-function and by applying the Leibniz's Rule for differentiation under the integral sign. The value of $H_{|\mathcal{G}|}$ in (21) is then obtained by writing the expectation of $\Omega$ as an integral form by using the above PDF of $\Omega$. □

## REFERENCES


[1] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Resolving the worst-user bottleneck of coded caching: Exploiting finite file sizes," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.

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

[3] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[4] 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.

[5] 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.

[6] 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.

[7] E. Lampiris *et al.*, "Fundamental limits of wireless caching under uneven-capacity channels," in *Int. Zurich Seminar*, Feb. 2020.

[8] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.

[9] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized degrees of freedom of the symmetric cache-aided MISO broadcast channel with partial CSIT," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5799–5815, Sep. 2019.

[10] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," *IEEE Trans. Inf. Theory*, pp. 1–1, 2021.

[11] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[12] M. Bayat, R. K. Mungara, and G. Caire, "Achieving spatial scalability for coded caching via coded multipoint multicasting," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 227–240, Jan. 2019.

[13] E. Lampiris and P. Elia, "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.

[14] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[15] S. Zhong and X. Wang, "Joint multicast and unicast beamforming for coded caching," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3354–3367, Aug. 2018.

[16] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[17] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.

[18] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.

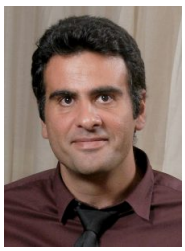
[19] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, Oct. 2016.

[20] K. Wan, D. Tuninetti, and P. Piantanida, "Novel delivery schemes for decentralized coded caching in the finite file size regime," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2017, pp. 1–6.

[21] S. Jin *et al.*, "A new order-optimal decentralized coded caching scheme with good performance in the finite file size regime," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5297–5310, Aug. 2019.
[22] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5821–5833, Sep. 2017.
[23] W. Song, K. Cai, and L. Shi, "Some new constructions of coded caching schemes with reduced subpacketization," 2019. [Online]. Available: https://arxiv.org/abs/1908.06570
[24] N. Jindal and Z. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2006, pp. 1841–1845.
[25] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.
[26] C. Shangguan, Y. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5755–5766, Aug. 2018.
[27] 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.
[28] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE Int. Conf. on Comput. Commun. (INFOCOM)*, 2012, pp. 1107–1115.
[29] M. Jia *et al.*, "Broadband hybrid satellite-terrestrial communication systems based on cognitive radio toward 5G," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 96–106, Dec. 2016.
[30] C. Bockelmann *et al.*, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
[31] Teltonika-Networks. Mobile signal strength recommendations. [Online]. Available: https://wiki.teltonika-networks.com/view/Mobile_Signal_Strength_Recommendations
[32] 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.
[33] A. M. Daniel and W. Yu, "Optimization of heterogeneous coded caching," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1893–1919, Mar. 2020.
[34] 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.
[35] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1469–1482, Apr. 2006.
[36] 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.
[37] 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.
[38] 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.
[39] M. K. Simon and M.-S. Alouini, *Digital communication over fading channels*. John Wiley & Sons, 2005.
[40] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Academic press, 2007.
[41] S. R. Finch, *Mathematical constants*. Cambridge university press, 2003.
[42] G. Kamath, "Bounds on the expectation of the maximum of samples from a Gaussian," 2015. [Online]. Available: http://www.gautamkamath.com/writings/gaussianmax.pdf
[43] S. Venkateshan and P. Swaminathan, *Computational Methods in Engineering*. Academic Press, 2014.
[44] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.
[45] M. Z. Shafiq *et al.*, "A first look at cellular network performance during crowded events," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 1, pp. 17–28, 2013.
[46] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Coded caching gains at low SNR over Nakagami fading channels," in *Proc. Asilomar Conf. Signals, Syst., and Comput. (ACSSC)*, Nov. 2021.
[47] J. W. Yoo, T. Liu, and F. Xue, "Gaussian broadcast channels with receiver message side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2009, pp. 2472–2476.
[48] T. Cover and A. Thomas, *Elements of information theory*. Wiley-Interscience, Jul. 1991.
[49] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*, 4th ed. Tata McGraw-Hill Ed., 2001.
[50] J. Gil-Pelaez, "Note on the inversion theorem," *Biometrika*, vol. 38, no. 3-4, pp. 481–482, 1951.
[51] H. Zhao *et al.*, "A simple evaluation for the secrecy outage probability over generalized-K fading channels," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1479–1483, Sep. 2019.
[52] J. Holtzman, "A simple, accurate method to calculate spread-spectrum multiple-access error probabilities," *IEEE Trans. Commun.*, vol. 40, no. 3, pp. 461–464, Mar. 1992.
[53] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1970, vol. 55.
[54] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. Wiley, 1995.
[55] K. Kataria, "A probabilistic proof of the multinomial theorem," *Amer. Math. Monthly*, vol. 123, no. 1, pp. 94–96, Jan. 2016.
[56] M. Inlow, "A moment generating function proof of the Lindeberg-Lévy central limit theorem," *Amer. Statist.*, vol. 64, no. 3, pp. 228–230, Aug. 2010.
[57] J. Pickands III, "Moment convergence of sample extremes," *Ann. of Math. Statist.*, vol. 39, no. 3, pp. 881–889, Jun. 1968.
[58] Wolfram Functions. [Online]. Available: http://functions.wolfram.com/07.34.21.0081.01

**Hui Zhao** (S'18) received the B.S. degree in telecommunications engineering from Southwest University, Chongqing, China, in 2016, and the M.S. degree in electrical engineering from the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, in 2019. He is currently pursuing the Ph.D. degree in communication systems with EURECOM, Sophia Antipolis, France. His current research interest includes the modeling, design, and performance analysis of wireless communication systems.

**Antonio Bazco-Nogueras** (M'20) received the B.S. and M.S. degrees in Telecommunications Engineering, both from University of Zaragoza, Spain, in 2014 and 2016, respectively. He obtained the Ph.D. degree from Sorbonne Université, Paris, France, in collaboration with the Mitsubishi Electric R&D Centre Europe, Rennes, France, in 2019. He was a post-doctoral researcher at EURECOM, Sophia-Antipolis, France, from 2020 to 2021. He is currently a post-doctoral researcher at IMDEA Networks Institute, Madrid, Spain. He is recipient of the Madrid Talent Attraction Grant 2021. His research interests include multi-user information theory, intelligent networks, decentralized systems, content delivery networks, and cooperative wireless networks.



**Petros Elia** received the B.Sc. degree from the Illinois Institute of Technology, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 2001 and 2006 respectively. He is now a professor with the Department of Communication Systems at EURECOM in Sophia Antipolis, France. His latest research deals with distributed computing as well as with the intersection of caching and communications in multiuser settings. He has also worked in the area of complexity-constrained communications, MIMO, queueing theory and cross-layer design, coding theory, information theoretic limits in cooperative communications, and surveillance networks. He is a Fulbright scholar, the co-recipient of the NEWCOM++ distinguished achievement award 2008-2011 for a sequence of publications on the topic of complexity in wireless communications, and the recipient of the ERC Consolidator Grant 2017-2022 on cache-aided wireless communications.