Coded Caching Gains at Low SNR Over Nakagami Fading Channels

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

***Abstract*—The promising gains provided by coded caching in ideal settings were thought to vanish in more realistic scenarios. This drawback was due to two main reasons: First, the limitation on the maximum number of chunks in which a given file can be split. Second, the so-called worst-user bottleneck, which refers to the fact that the transmission is limited by the user with the worst channel quality at each time, and which is exacerbated at low SNR. Recent works have shown that the first problem can be exploited to reduce the worst-user bottleneck through the use of shared caches and by dropping the use of XOR-based transmissions. This work analyzes for first time how this promising result impacts the delivery time of coded caching at low SNR. Motivated by the fact that the delivery time under Rayleigh fading does not have an expectation, Nakagami-m fading is assumed to model the physical-layer channel. We derive the analytical expression of the delivery time at low SNR, and we also consider the regime of large number of users in order to draw some valuable insights. Finally, we validate the correctness of the derived expressions by means of numerical evaluations.**

***Index Terms*—Coded caching, low SNR, Nakagami fading, and delivery time.**

I. Introduction

CURRENT and future networks are witnessing an ever-increasing demand of multimedia content, which poses a significant challenge in terms of congestion and delay. One of the most promising solutions is Coded Caching, which was introduced in [1] by Maddah-Ali and Niesen. Coded caching capitalizes a resource that had not been previously exploited to its full potential: the local memory (or cache) available at the receivers. The scheme presented in [1], hereinafter referred to as the MN scheme, leverages the receiver-side cache to simultaneously deliver content to many users through a shared link.

Specifically, the seminal coded caching work [1] considers a shared error-free link scenario where a single-antenna transmitter, which has access to a library of N equal-sized files, serves K users. Each of these users has a local cache of size equal to M files, such that each user can store a fraction $\gamma \triangleq \frac{M}{N} \in [0,1]$ of the library. The authors of [1] showed that, compared to uncoded caching transmission, the MN scheme reduces the delivery time by a multiplicative factor of $K\gamma+1$, which we refer to as the *theoretical coded caching gain*.

Driven by such gains, a great variety of different research directions have emerged from this result, mostly to characterize the extent of such an outcome in more practical settings, e.g., by considering multi-antenna transmission and the impact of CSI availability [2]–[5], uneven link capacities [6], [7], or files with different popularities [8].

Two main bottlenecks have emerged from this extensive analysis [2]–[5]. First, the MN scheme and many subsequent clique-based schemes are based on transmitting a common symbol that is intended by a set of $K\gamma+1$ users, which implies that the delivery rate is limited by the user with the worst channel quality. This limitation is called the *worst-user bottleneck* and is exacerbated at low SNR [2]. Indeed, the gain of the MN scheme vanishes when the SNR is sufficiently small for quasi-static Rayleigh fading channels [9], [10] and in the single-cell scenario [11]. This is crucial because many practical systems operate in the low-SNR regime, such as IoT networks [12] or massive Machine-Type Communications [13].

The other bottleneck is referred to as the subpacketization bottleneck, and arises as a result of the finite size of the files. In particular, any known high-performance coded caching scheme requires the file size to grow exponentially or near-exponentially with K [14]. However, files have a finite size and also the chucks (or subfiles) in which it can be divided have a minimum size. This implies that there exists a maximum number of subfiles ($S_{\max}$) into which a file can be divided, i.e., a *subpacketization constraint*. As a result, when K is large enough, the subpacketization bottleneck lessens the coded caching gain from $K\gamma+1$ to $\Lambda\gamma+1$, for $\Lambda \leq K$ [3]. We refer to this $\Lambda\gamma+1$ gain as the *nominal coded caching gain*.

Recently, a novel coded caching scheme so-called Aggregated Coded Caching (ACC) has been developed in [9], [10], which exploits the fact that the subpacketization constraint induces several users to cache the same content to alleviate (and asymptotically remove) the worst-user effect. It turns out that increasing K for a fixed Λ yields an effect equivalent to time diversity. However, the works [9]–[11] did not consider the delivery time as a metric, which is the common metric considered in coded caching literature, because the delivery time does not converge over quasi-static Rayleigh fading (the channel model adopted in these works).

In order to analyze the impact of the novel ideas in [9], [10] on the delivery time, we consider Nakagami-m fading and analyze the delivery time of both the MN and the ACC schemes. Nakagami-m fading is a broader model than Rayleigh fading and fits practical wireless channels, such as land–mobile and indoor–mobile multi-path propagation, as well as scintillating ionospheric radio links [15, Ch. 2]. Moreover, when m is a positive integer, the statistics of the received SNR over Nakagami-m fading are the same as for the received SNR in an

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).

m-antenna receiver over symmetric Rayleigh fading after applying maximal-ratio combining (MRC). We consider the low-SNR regime, where the worst-user bottleneck is aggravated, and we derive the analytical expressions of the delivery time in this regime. We also consider an approximation based on the large number of users regime that precisely characterizes the performance at any SNR and for not-so-large number of users. We show how the fading parameter m and the ratio $\frac{K}{\Lambda}$ impact in a similar manner the performance of coded caching, since both offer an effect equivalent to time diversity. In this respect, a setting with a given Λ and K single-antenna receivers approximately achieves the same performance as a setting with K K/Λ-antenna receivers that apply MRC.

II. System Model and Problem Definition

We consider a single-antenna transmitter that has full access to a library $\mathcal{L}$ composed of N files, where each file consists of F bits. The transmitter serves K single-antenna users, and each user is endowed with a local memory (or cache) of MF bits, i.e., of size equal to the size of M files. Therefore, we define the normalized cache size $\gamma \in [0,1]$ as $\gamma \triangleq \frac{M}{N}$. The files can be split in different chunks or sub-files, and we consider that there exists a minimum size for these subfiles. This assumption arises in practice because the file size (F) is finite and because communication protocols normally have a minimum size for the transmitted packets [3]. Consequently, there exists a maximum number of subfiles ($S_{\max}$) into which a file can be divided, i.e., a *subpacketization constraint*.

We consider that the physical-layer channel experiences quasi-static fading, which is a suitable model for low-mobility scenarios. In fact, low-mobility scenarios comprise coded-caching-suited use cases, mainly pedestrians or static users consuming video streaming. The quasi-static fading has been also considered in some existing works that analyze coded caching performance, such as [2], [10]. We consider Nakagami-m fading [15, Ch. 2] (note that the delivery time over Rayleigh fading does not have an expectation).

Coded Caching Schemes

In the coded caching framework, the transmission consists of two separated phases. The first one is called *placement phase*, and it happens during off-peak traffic hours and before the users requests are known. In this phase, the users fill their local caches with content from the library. The second phase is called *delivery phase*. In this phase, each user requests a different file from the library, and the transmitter serves the users aiming at minimizing the total delivery time, i.e., the time required to send one file to each user.

We consider two coded caching schemes. First, we consider the seminal MN scheme [1] introduced by Maddah-Ali and Niesen. We also consider the ACC scheme, which was presented in [9] and was shown to overcome the worst-user bottleneck by exploiting the fact that the subpacketization constraint motivates that different users cache the same content.

In the following, we briefly describe the ACC scheme. A detailed description of this scheme can be found in [10]. Due to space limitations, and because the MN scheme's structure can be inferred from the description of the ACC scheme, we omit the exposition of the MN scheme. We introduce the notation $[n] \triangleq \{1, 2, \cdots, n\}$ for a positive integer n, and $|\cdot|$ to denote the cardinality operator of a set.

1) Placement phase: Let us denote the files in the library $\mathcal{L}$ by W_n, $n \in [N]$. First, each file W_n is equally partitioned into $\binom{\Lambda}{\Lambda\gamma}$ subfiles, where Λ is defined as the maximum integer that satisfies that $\binom{\Lambda}{\Lambda\gamma} \leq S_{\max}$. Hence, Λ is given by the system parameters and it is considered fixed. Λ represents the number of different cache states, where *cache state* refers to the specific part of the library that users store in their cache.

The files are partitioned into $\binom{\Lambda}{\Lambda\gamma}$ subfiles such that: $W_n \rightarrow \left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\right\}$, for $n \in [N]$. Then, the different cache states are created as follows: The g-th cache state, $g \in [\Lambda]$, is composed of the following subfiles: $\mathcal{Z}_g = \left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, \mathcal{T} \ni g, \forall n \in [N]\right\}$. This placement procedure is analogous to the standard clique-based placement in coded caching, with the difference that normally there are K cache states and each of the K user stores one of them. Due to the subpacketization constraint, there are only $\Lambda \leq K$ cache states.[1] This implies that $B = \frac{K}{\Lambda}$ users store the same cache state (and thus the same content). The users that store the same cache state are said to belong to a *user group*. We label these user groups from 1 to Λ, and we use $\mathrm{U}_{g,b}$ to denote the b-th user in the g-th user-group for $g \in [\Lambda]$ and $b \in [B]$.

2) Delivery phase: The delivery phase consists of $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages. In each transmission stage, the transmitter serves a unique set of user groups $\mathcal{G}$ of $|\mathcal{G}| = \Lambda\gamma + 1$ groups (comprising $B(\Lambda\gamma+1)$ users). For a specific $\mathcal{G}$, $\Lambda\gamma+1$ users are simultaneously served, one from each group in $\mathcal{G}$, and the users within the same group are served in a round-robin manner. During such transmission, each user can receive its intended subfile at its single-link capacity. This is due to the fact that the user groups are served in a way that the subfiles delivered to the users of a given group are cached at all users belonging to the other $\Lambda\gamma$ groups.

When one of the $|\mathcal{G}|$ users successfully decodes its subfile, another user from the same group replaces it without impacting the transmission to the other groups. This is possible because users in the same group cache the same content and is the key factor of the ACC scheme. We refer to [9], [10] for a detailed description. Hence, the delivery time to a specific set $\mathcal{G}$ of user groups depends on the worst (slowest) group, and not on the specific worst user. Let us now present our metric of interest.

Delivery time: Consider the ℓ-th group set $\mathcal{G}_\ell$, $\ell \in \left[\binom{\Lambda}{\Lambda\gamma+1}\right]$. From the previous description of the ACC scheme, the delivery time for the set $\mathcal{G}_\ell$ is given by (cf. [10])

$$T_\ell = \frac{F}{B_w\binom{\Lambda}{\Lambda\gamma}} \max_{g \in \mathcal{G}_\ell} \left\{ \sum\nolimits_{b=1}^{B} \frac{1}{\log_2\left(1 + \mathrm{SNR}_{g,b}\right)} \right\}$$
$$= \rho^{-1} \frac{F \ln 2}{B_w\binom{\Lambda}{\Lambda\gamma}} \max_{g \in \mathcal{G}_\ell} \left\{ \sum\nolimits_{b=1}^{B} \frac{1}{|h_{g,b}|^2} \right\} + o(1), \quad (1)$$

[1] We consider that K is an integer multiple of Λ. The extension to arbitrary integers follows directly and with small impact on the performance [9].

where $\lim_{\rho\to 0} o(1) = 0$, $h_{g,b}$ denotes the channel coefficient between the transmitter and user $\mathrm{U}_{g,b}$, $\mathrm{SNR}_{g,b} = \rho|h_{g,b}|^2$ denotes the instantaneous SNR at user $\mathrm{U}_{g,b}$, B_w refers to the transmission bandwidth, and ρ denotes the normalized transmit power. Since we assume Nakagami-m fading, $|h_{g,b}|^2$ follows a Gamma distribution with shape parameter m and unitary scale parameter, which we denote by $|h_{g,b}|^2 \sim \mathrm{Gamma}(m,1)$. Since there are $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, the total delay averaged over the channel state is

$$\begin{aligned}\mathbb{E}\{T_{\mathrm{ACC}}\} &= \mathbb{E}\left\{\sum\nolimits_{\ell=1}^{\binom{\Lambda}{\Lambda\gamma+1}} T_\ell\right\} = \sum\nolimits_{\ell=1}^{\binom{\Lambda}{\Lambda\gamma+1}} \mathbb{E}\{T_\ell\} \\ &= \rho^{-1}\frac{\Lambda(1-\gamma)F\ln 2}{B_w(1+\Lambda\gamma)}\mathbb{E}\{T_{\mathcal{G}}\} + o(1), \qquad (2)\end{aligned}$$

where $T_{\mathcal{G}} \triangleq \max_{g\in\mathcal{G}}\left\{\sum_{b=1}^{B}\frac{1}{|h_{g,b}|^2}\right\}$ for a specific user-group set $\mathcal{G}$ during the ACC delivery phase. In this paper, we focus on the delivery time at low SNR, and thus we define the following approximation by omitting the term $o(1)$ in (2),

$$\mathring{T}_{\mathrm{ACC}} \triangleq \rho^{-1}\frac{\Lambda(1-\gamma)F\ln 2}{B_w(1+\Lambda\gamma)}\mathbb{E}\{T_{\mathcal{G}}\}. \qquad (3)$$

To show the extent to which the delivery performance is enhanced at reasonable values of the SNR, we define the *effective coded caching gain*. First, we will consider the performance of simple Time-Division Multiplexing (TDM) transmission without coded caching, such that the local memory at the users is only useful to provide *local caching gain*, which refers to the fact that γF bits of the requested file are already stored at the user. This scheme allows us to characterize the net gain generated by coded caching. Henceforth, we will refer to this scheme as the *uncoded TDM scheme*.

Definition 1 (Effective coded caching gain)**.** *The effective coded caching gain of a particular coded caching scheme is the ratio of the average delivery time of the uncoded TDM scheme over the average delivery time of the said particular scheme (e.g., MN or ACC schemes), where the average is with respect to the channel state over quasi-static fading.*

III. Delivery Time of the MN Scheme at Low SNR

In this section, we analyze the average delivery time of the MN scheme in order to present some insights about its performance at practical SNR values and to provide a benchmark for the ACC scheme.

In order to serve the K users by means of the MN scheme in the scenario where the subpacketization constraint induces a maximum $\Lambda < K$, we need to consider the MN transmission only over Λ users (one from each group), and then repeat the same process $B = K/\Lambda$ times to serve all the K users [10]. This modification was first presented in [16, Section V-A].

The delivery time required to serve Λ users with the MN scheme when the subpacketization is $\binom{\Lambda}{\Lambda\gamma}$ is equivalent to the delivery time required by the ACC scheme with $B = 1$ [10]. Then, the approximation of the average delivery time of the MN scheme at low SNR follows from (3) and writes as

$$\mathring{T}_{\mathrm{MN}} \triangleq \rho^{-1}B\frac{\Lambda(1-\gamma)F\ln 2}{B_w(1+\Lambda\gamma)}\mathbb{E}\{T_{\mathcal{G}}\}, \qquad (4)$$

where $T_{\mathcal{G}} \triangleq \max_{g\in\mathcal{G}}\left\{\frac{1}{|h_g|^2}\right\}$, and h_g denotes the channel coefficient between the transmitter and user U_g (we omit the subscript b because $B = 1$). In (4), the addition of the factor B is due to the $B = K/\Lambda$ times that we need to repeat the MN scheme to serve all K users. In the next lemma, we provide the analytical expression of $\mathring{T}_{\mathrm{MN}}$. We recall that $|\mathcal{G}| = 1+\Lambda\gamma$.

Lemma 1. *The approximate average delivery time of the MN scheme at low SNR over quasi-static Nakagami-m fading is given by*

$$\mathring{T}_{\mathrm{MN}} = \frac{K(1-\gamma)F\ln 2}{\rho B_w|\mathcal{G}|}\int_0^\infty 1-\left(\frac{\Gamma(m,1/x)}{\Gamma(m)}\right)^{|\mathcal{G}|}\mathrm{d}x, \quad (5)$$

where $\Gamma(\cdot)$ and $\Gamma(\cdot,\cdot)$ denote the Gamma function and the upper incomplete Gamma function [17], respectively.

Proof. Under the assumption of Nakagami-m fading, it follows that $|h_g|^2 \sim \mathrm{Gamma}(m,1)$. Then, we have that $\frac{1}{|h_g|^2}$ follows an inverse Gamma distribution with shape parameter m and scale parameter equal to 1. Therefore, the cumulative distribution function (CDF) of $\frac{1}{|h_g|^2}$ is (cf. [18])

$$F_{1/|h_g|^2}(x) = \frac{1}{\Gamma(m)}\Gamma(m,1/x), \quad x \geq 0. \qquad (6)$$

Hence, the CDF of $T_{\mathcal{G}} \triangleq \max_{g\in\mathcal{G}}\left\{\frac{1}{|h_g|^2}\right\}$ in (4) writes as

$$F_{T_{\mathcal{G}}}(x) = \left(F_{1/|h_g|^2}(x)\right)^{|\mathcal{G}|} = \left(\frac{1}{\Gamma(m)}\Gamma(m,1/x)\right)^{|\mathcal{G}|}. \quad (7)$$

Since $T_{\mathcal{G}}$ has a non-negative support, $\mathbb{E}\{T_{\mathcal{G}}\}$ can be obtained by integrating $1-F_{\mathcal{G}}(x)$ from 0 to infinity. Substituting $\mathbb{E}\{T_{\mathcal{G}}\}$ by this integral in (4) yields (5). □

The integral in Lemma 1 does not have a closed-form solution. Yet, we can obtain a closed-form expression for $\mathring{T}_{\mathrm{MN}}$ when m is a positive integer bigger than 1, as stated in the following corollary.

Before presenting this result, let us introduce some useful notations. Let $\mathbf{k} \triangleq [k_1, k_2, \cdots, k_m]$ be a non-negative integer vector and $\binom{|\mathcal{G}|-1}{\mathbf{k}} \triangleq \frac{(|\mathcal{G}|-1)!}{k_1!k_2!\cdots k_m!}$ be the multinomial coefficient. We also use $||\mathbf{k}||_{\mathbf{1}}$ to denote the norm-1 operator of any vector $\mathbf{k}$. We can present now the following result.

Corollary 1. *For any positive integer $m \geq 2$, the approximate delivery time of the MN scheme becomes*

$$\begin{aligned}\mathring{T}_{\mathrm{MN}} = \rho^{-1}\frac{K(1-\gamma)F\ln 2}{B_w(m-1)!}\sum\nolimits_{||\mathbf{k}||_1=|\mathcal{G}|-1}\binom{|\mathcal{G}|-1}{\mathbf{k}} \\ \times \frac{|\mathcal{G}|^{1-m-\sum_{j=1}^m(j-1)k_j}}{\prod_{j=1}^m((j-1)!)^{k_j}}\left(m-2+\sum_{j=1}^m(j-1)k_j\right)!. \quad (8)\end{aligned}$$

Proof. We first obtain the probability density function (PDF) of $T_{\mathcal{G}}$ by differentiating the CDF of $T_{\mathcal{G}}$ given in (7).

$$f_{T_{\mathcal{G}}}(x) = \frac{|\mathcal{G}|e^{-1/x}}{(\Gamma(m))^{|\mathcal{G}|}x^{m+1}}\left(\Gamma(m,1/x)\right)^{|\mathcal{G}|-1}. \qquad (9)$$

By applying [17, Eq. (8.352.2)], we can rewrite $f_{T_{\mathcal{G}}}(x)$ for any positive integer m as (10), shown at the top of next page. Since $\mathbb{E}\{T_{\mathcal{G}}\} = \int_0^\infty x f_{T_{\mathcal{G}}}(x)\mathrm{d}x$, we obtain $\mathbb{E}\{T_{\mathcal{G}}\}$ by considering

$$f_{T_{\mathcal{G}}}(x)=\frac{|\mathcal{G}|e^{-1/x}}{(\Gamma(m))^{|\mathcal{G}|}x^{m+1}}\left(\Gamma(m)e^{-1/x}\sum_{j=0}^{m-1}\frac{x^{-j}}{j!}\right)^{|\mathcal{G}|-1}=\frac{|\mathcal{G}|e^{-|\mathcal{G}|/x}}{\Gamma(m)}\sum_{||\mathbf{k}||_1=|\mathcal{G}|-1}\binom{|\mathcal{G}|-1}{\mathbf{k}}\frac{x^{-m-1-\sum_{j=1}^{m}(j-1)k_j}}{\prod_{j=1}^{m}((j-1)!)^{k_j}}. \quad (10)$$

the expression of $f_{T_{\mathcal{G}}}(x)$ in (10). Finally, substituting $\mathbb{E}\{T_{\mathcal{G}}\}$ in (4) yields (8). □

Next, we also consider the performance of simple TDM transmission without coded caching.

Corollary 2. *The approximate delivery time at low SNR of the uncoded TDM scheme is*

$$\mathring{T}_{\mathrm{TDM}}=\rho^{-1}\frac{K(1-\gamma)F\ln 2}{B_w(m-1)} \quad \forall m>1. \quad (11)$$

Proof. This result follows directly from Lemma 1 after fixing the number of simultaneously served users to $|\mathcal{G}|=1$. Then, it follows that $\mathbb{E}\{T_{\mathcal{G}}\}=\mathbb{E}\{\frac{1}{|h_i|^2}\}$ for $i\in K$, which yields (11) after substituting $\mathbb{E}\{T_{\mathcal{G}}\}=1/(m-1)$ in (4). The factor $1-\gamma$ in (11) is due to the fact that every user has stored γF bits of each file in $\mathcal{L}$, and the server only needs to deliver the remaining $(1-\gamma)F$ bits of each requested file. □

Remark 1. *The expectation of the delivery time does not exist for $m\leq 1$, where $m=1$ corresponds to Rayleigh fading. This can also be inferred from Corollaries 1 and 2. Accordingly, hereon we only consider $m>1$ unless otherwise stated.*

From the previous results, we can analyze the effective coded caching gain of the MN scheme in the considered scenario, which leads to the following lemma.

Lemma 2. *The effective coded caching gain of the MN scheme in the low-SNR limit, i.e., when $\rho\to 0$, is given by*

$$G_{\mathrm{MN}}=\frac{|\mathcal{G}|}{(m-1)}\left(\int_0^{\infty}1-\left(\frac{\Gamma(m,1/x)}{\Gamma(m)}\right)^{|\mathcal{G}|}\mathrm{d}x\right)^{-1}.$$

Proof. This result is directly obtained by applying Definition 1 as the ratio of $\mathring{T}_{\mathrm{TDM}}$ in (11) over $\mathring{T}_{\mathrm{MN}}$ in (5). □

When m is a positive integer, we can treat $|h_g|^2$ (Gamma distributed) as the summation over m independently and identically distributed (i.i.d.) exponential random variables with unit-mean. As $m\to\infty$, the Strong Law of Large Numbers implies that $\frac{1}{m}|h_g|^2\xrightarrow{a.s.}1$, where $\xrightarrow{a.s.}$ stands for almost sure convergence. Substituting $\frac{1}{m}|h_g|^2\xrightarrow{a.s.}1$ into (4) yields

$$\mathring{T}_{\mathrm{MN}}\xrightarrow{a.s.}\rho^{-1}B\frac{\Lambda(1-\gamma)F\ln 2}{B_w(1+\Lambda\gamma)m},\ \text{as } m\to\infty, \quad (12)$$

and thus the effective gain at low SNR satisfies that

$$G_{\mathrm{MN}}\xrightarrow{a.s.}|\mathcal{G}|=\Lambda\gamma+1\ \ \text{as } m\to\infty, \quad (13)$$

i.e., the worst-user effect in the MN scheme can be effectively mitigated (and asymptotically resolved) for $m\gg 1$.

IV. Delivery Time of the ACC Scheme at Low SNR

Next, we analyze the delivery time of the ACC scheme in the low-SNR regime over quasi-static Nakagami-m fading. First, we present the exact expression of $\mathring{T}_{\mathrm{ACC}}$, which is obtained through the Characteristic Function (CF) method, and results in a complex integral. To simplify the derived result and obtain further insights, we will consider the regime of large number of users to derive approximations that are also robust for not-so-large number of users.

A. Low SNR Characterization

Before presenting our next result, let us introduce some useful notation. In the following, $\jmath\triangleq\sqrt{-1}$ denotes the imaginary unit, $\mathrm{Im}\{\cdot\}$ refers to the imaginary part of a complex number, and $K_n(\cdot)$ denotes the the modified Bessel function of the second kind [17].

We present now the exact expression of the approximated delivery time $\mathring{T}_{\mathrm{ACC}}$. We recall that $\mathring{T}_{\mathrm{ACC}}$ is the delivery time obtained after applying the low-SNR approximation $\ln(1+x)=x+o(1)\approx x$ when $x\to 0$.

Lemma 3. *Under quasi-static Nakagami-m fading (and $m>1$), the approximate delivery time at low SNR for the ACC scheme is given by* (14)*, shown at the top of next page.*

Proof. The proof is relegated to the appendix. □

Due to its double-integral form, the expression in (14) provides little insight and its numerical computation is challenging. For this reason, we consider in the following the regime of large number of users, i.e., the case in which Λ remains fixed but $B\to\infty$, and thus $K=B\Lambda\to\infty$. Thanks to this assumption, we will obtain simplified expressions that accurately characterize the performance also for practical values of B.

B. Large Number of Users at Low SNR

In this subsection, we consider that the number of users per group grows unboundedly (i.e., that $B\to\infty$) and the number of different cache states ($\Lambda=\frac{K}{B}$) remains fixed. A fixed Λ implies that the high-SNR coded-caching gain (under subpacketization constraint) $1+\Lambda\gamma$ also remains fixed.

Let us first introduce some useful notations. We use μ and σ^2 to denote the expectation and variance of $1/\mathrm{SNR}_{g,b}$, respectively. We also use $H_{|\mathcal{G}|}$ to denote the expectation of the maximum of $|\mathcal{G}|$ i.i.d. Gaussian random variables with zero-mean and unit-variance.

Armed with the previous notations, we can present the approximation of $\mathring{T}_{\mathrm{ACC}}$ for the regime of large number of users in the following lemma.

Lemma 4. *For a sufficiently large number of users, the expectation of the delivery time over quasi-static Nakagami-m fading channels with $m>2$ can be tightly approximated at low SNR as*

$$\mathring{T}_{\mathrm{ACC}}\approx\frac{\Lambda(1-\gamma)F\ln 2}{B_w|\mathcal{G}|}\left(B\mu+H_{|\mathcal{G}|}\sqrt{B\sigma^2}\right), \quad (15)$$

$$\mathring{T}_{\mathrm{ACC}} = \rho^{-1}\frac{\Lambda\left(1-\gamma\right)F\ln 2}{|\mathcal{G}|B_w}\int_0^{\infty} 1-\left(\frac{1}{2}-\frac{1}{\pi}\int_0^{\infty}\mathrm{Im}\left\{\frac{e^{-\jmath tz}}{t}\left(\frac{2(-\jmath t)^{\frac{m}{2}}}{\Gamma\left(m\right)}K_m\left(\sqrt{-4\jmath t}\right)\right)^B\right\}\mathrm{d}t\right)^{|\mathcal{G}|}\mathrm{d}z. \tag{14}$$

where μ and σ^2 are given respectively by

$$\mu = \frac{1}{\rho(m-1)}, \quad \sigma^2 = \frac{1}{\rho^2(m-1)^2(m-2)}. \tag{16}$$

Proof. The proof is based on the Central Limit Theorem (CLT), and the condition $m > 2$ guarantees that $1/\mathrm{SNR}_{g,b}$ has finite variance, such that we can apply the CLT.

As $1/\mathrm{SNR}_{g,b} = \rho^{-1}|h_{g,b}|^{-2}$ is drawn from an inverse Gamma distribution of mean and variance given respectively by μ and σ^2 in (16), we can apply the CLT to obtain that

$$\sum\nolimits_{b=1}^{B}\frac{1}{\mathrm{SNR}_{g,b}} \xrightarrow{d.} \mathcal{N}\left(B\mu,\ B\sigma^2\right), \tag{17}$$

where $d.$ denotes the convergence in distribution and $\mathcal{N}$ denotes the normal distribution. Therefore, $T_{\mathcal{G}}/\rho = \max_{g\in\mathcal{G}}\{\sum_{b=1}^{B}\rho^{-1}|h_{g,b}|^{-2}\}$ converges in distribution to the maximum of $|\mathcal{G}|$ i.i.d. normal random variables with mean $B\mu$ and variance $B\sigma^2$. Upon defining $\{X_g\}_{g\in\mathcal{G}}$ as a set of $|\mathcal{G}|$ i.i.d. variables distributed as $\mathcal{N}(0,1)$, we can write that $T_{\mathcal{G}}/\rho \xrightarrow{d.} B\mu + \sqrt{B}\sigma\max_{g\in\mathcal{G}}X_g$. Since we have defined $H_{\mathcal{G}}$ as the expectation of $\max_{g\in\mathcal{G}}X_g$, we obtain (15) from the definition of $\mathring{T}_{\mathrm{ACC}}$ in (3). □

Regarding $H_{|\mathcal{G}|}$ in (15), the exact expression is known for $|\mathcal{G}| \leq 5$ (cf. [10, Table I]). For general $|\mathcal{G}|$, there is no simple closed-form for $|\mathcal{G}| > 5$, but $H_{|\mathcal{G}|}$ can be expressed in an integral form as (cf. [10, Lemma 6])

$$\begin{aligned} H_{|\mathcal{G}|} &= \frac{-|\mathcal{G}|}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x\big(Q(x)\big)^{|\mathcal{G}|-1}e^{\frac{-x^2}{2}}\mathrm{d}x \\ &\overset{(a)}{\approx} \frac{-\sqrt{2}|\mathcal{G}|}{\sqrt{\pi}}\sum\nolimits_{v=1}^{V}\omega_v x_v\big(Q(\sqrt{2}x_v)\big)^{|\mathcal{G}|-1}, \end{aligned} \tag{18}$$

where $Q(\cdot)$ represents the complementary CDF (or tail distribution) of the standard Gaussian distribution, and (a) follows from the Gauss-Hermite quadrature (GHQ) [19]. In (18), V, x_v and ω_v are the summation terms, sample points and weights of GHQ, respectively. Summing up as less as 5 terms in the GHQ method provides a very accurate approximation of $H_{|\mathcal{G}|}$ [10].

Lemma 4 allows us to characterize the effective coded caching gain of the ACC scheme.

Lemma 5. *In the regime of large number of users, and for $m > 2$, we can approximate the effective coded caching gain of the ACC scheme at low SNR by*

$$G_{\mathrm{ACC}} \approx \frac{|\mathcal{G}|}{1 + H_{|\mathcal{G}|}/\sqrt{B(m-2)}}. \tag{19}$$

Proof. The result follows by considering Definition 1 and the expressions of $\mathring{T}_{\mathrm{TDM}}$ in (11) and $\mathring{T}_{\mathrm{ACC}}$ in (15). □

It is direct to see that, when $B \to \infty$ or $m \to \infty$, G_{ACC} converges to $|\mathcal{G}|$, which is the optimal gain at high SNR. This means that we recover the nominal gains of coded caching even at low SNR provided that there are enough users or big enough m (e.g., with a massive number of receiving antennas).

C. Extension — Large Number of Users at any SNR

In the previous subsection, we have considered the low-SNR approximation $\ln(1+\mathrm{SNR}) \approx \mathrm{SNR}$ as $\mathrm{SNR} \to 0$. We can remove this approximation and directly consider $\ln(1+\mathrm{SNR})$ to obtain an approximation that is valid for any SNR value in the regime of large number of users.

In this case, (15) in Lemma 4 holds but with a different value of μ and σ^2. Specifically, μ and σ^2 will respectively denote the mean and variance of $1/\ln(1+\mathrm{SNR}_{g,b})$, instead of $1/\mathrm{SNR}_{g,b}$. From the PDF of $\mathrm{SNR}_{g,b}$ over Nakagami-m fading, we obtain the following integral forms for μ and σ^2,

$$\mu = \frac{1}{\rho^m\Gamma(m)}\int_0^{\infty}\frac{x^{m-1}e^{-\frac{x}{\rho}}}{\ln(1+x)}\mathrm{d}x, \tag{20}$$

$$\sigma^2 = \frac{1}{\rho^m\Gamma(m)}\int_0^{\infty}\left(\frac{1}{\ln(1+x)}-\mu\right)^2 x^{m-1}e^{-\frac{x}{\rho}}\mathrm{d}x. \tag{21}$$

Unfortunately, there are not closed-form solutions for these integrals. In view of the structure of these integrals, a valid method is to adopt the Gauss-Laguerre quadrature (GLQ) [19], from which the integrals are respectively approximated as

$$\mu \approx \frac{1}{\Gamma(m)}\sum\nolimits_{v=1}^{V}\varpi_v\frac{y_v^{m-1}}{\ln(1+\rho y_v)}, \tag{22}$$

$$\sigma^2 \approx \frac{1}{\Gamma(m)}\sum\nolimits_{v=1}^{V}\varpi_v y_v^{m-1}\left(\frac{1}{\ln(1+\rho y_v)}-\mu\right)^2, \tag{23}$$

where V, y_v and ϖ_v are the summation terms, sample points and weights of the GLQ, respectively. As for the GHQ, the accuracy of the GLQ is typically reasonable after summing up a few terms.

V. Numerical Results

In this section, we demonstrate the accuracy of the derived expressions through Monte-Carlo simulations. Hereinafter, we assume that the file size is $F = 8\times10^9$ bits (i.e., 1 Gigabyte), and that the bandwidth for each user is 20 MHz (as in 4G standard). In order to implement the Monte-Carlo simulations, 10^6 channel states are generated and averaged over Nakagami-m fading channels.

In Fig. 1, we plot the delivery time of the MN and ACC schemes versus ρ for different values of m. We can observe that the delivery time decreases as m increases. This happens because a bigger m is somehow equivalent to enjoying a richer multi-path environment, thereby enhancing the spatial diversity. Indeed, as it is known, if a m-antenna receiver applies MRC over Rayleigh fading, the magnitude of the equivalent channel coefficient after MRC follows a Nakagami-m fading distribution.

Fig. 1 also shows the high accuracy of the derived approximations at low SNR. We observe that the SNR regime in which we obtain high accuracy for the low-SNR approximations is reduced as m increases. This is due to the fact that the bigger m, the higher is the effective SNR received at the user for the

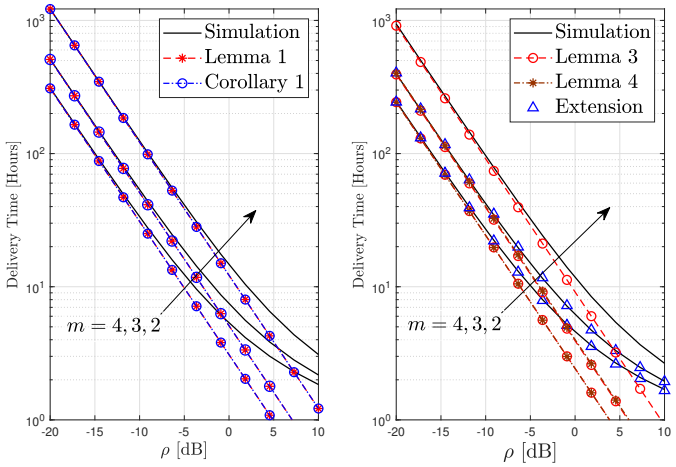

Fig. 1: Average delivery time of MN (left) and ACC (right) schemes versus ρ for $K = 300$, $B = 5$, $\gamma = 5\%$ and $V = 7$ in the GHQ and GLQ.

same ρ. However, it is worth noting that the approximation based on the regime of large number of users without low-SNR approximation (Subsection IV-C) tightly approximates the delivery time *even for very small values of* B, such as $B = 5$, for $m > 2$ and for any SNR.

We plot the effective coded caching gains of the ACC and MN schemes in Fig. 2 with the same system settings of Fig. 1. Besides validating the correctness of Lemmas 2 and 5, Fig. 2 also shows that the gains of both schemes arrive to a plateau at very low SNR, but that the infimum for the ACC scheme (which becomes bigger as B grows) is larger than the one for the MN scheme. Moreover, the effective gains of both schemes increase as m increases. Indeed, as stated before, both gains G_{ACC} and G_{MN} converge to $|\mathcal{G}|$ as $m \to \infty$.

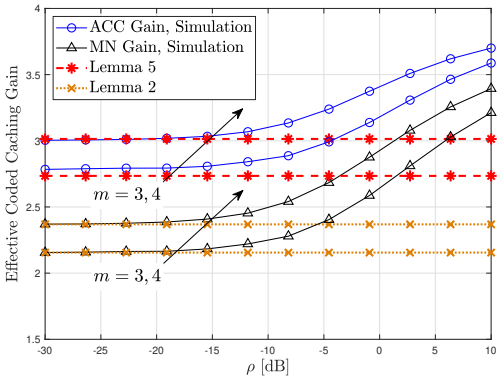

Fig. 2: Effective coded caching gains of ACC and MN schemes for $K = 300$, $B = 5$, $\gamma = 5\%$, and $m = 3, 4$.

Fig. 3 represents the delivery time improvement of the ACC scheme over the MN scheme versus ρ for different values of B. The delivery time of both schemes converge as ρ increases, while the boost effect of the ACC scheme becomes significant in the low SNR region. It is also obvious that the larger B is, the better performance we can achieve.

We illustrate the effective coded caching gains of the MN and ACC schemes for $B = 10$ and for different values of m and $|\mathcal{G}|$ as $\rho \to 0$ in Fig. 4. It is clear how the effective gain

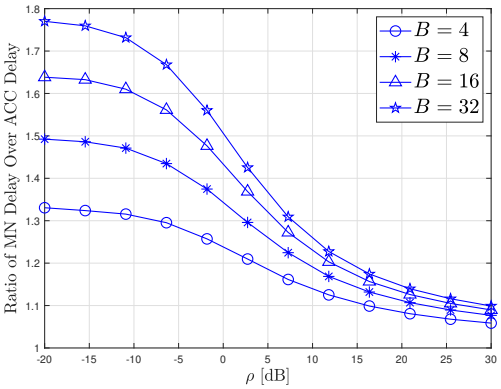

Fig. 3: Delay Ratio of $\mathring{T}_{\text{MN}}$ over $\mathring{T}_{\text{ACC}}$ versus ρ for $m = 3$, $\gamma = \frac{1}{12}$, and $|\mathcal{G}| = 6$.

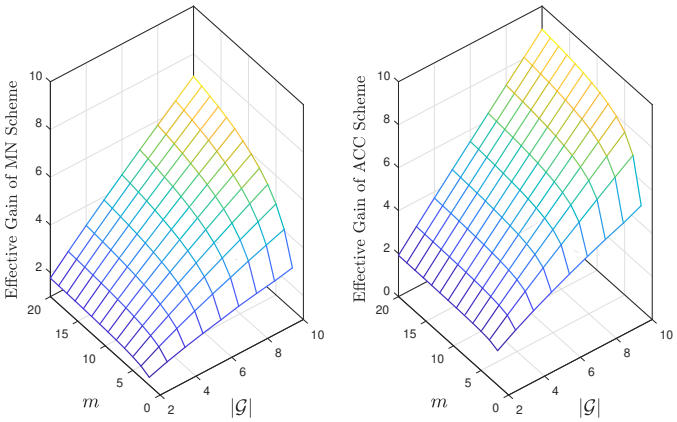

Fig. 4: Effective coded caching gains of MN (left) and ACC (right) schemes for $B = 10$ and $\gamma = 10\%$ as $\rho \to 0$.

is increased as m increases. This is due to the fact that both MN and ACC schemes are limited by some kind of worst-user effect, although this effect is strongly mitigated in the ACC scheme as it becomes a "worst-group" effect. A larger m implies more spatial diversity, which alleviates the worst-user effect. As shown in [10], the parameter B plays a similar role, as it provides the equivalent of spatial diversity when the scenario does not provide it. Because of that, we can see how the improvement of the ACC scheme over the MN scheme in terms of effective gain is bigger for small m. We can observe that the ACC scheme offers a considerable gain even for not-so-large values of B, and especially when $|\mathcal{G}|$ is large.

To further show the impact of B on the effective gain, we plot in Fig. 5 the effective gains of the MN (left) and ACC (right) schemes versus m and B for $|\mathcal{G}| = 5$ in the low-SNR limit. The effective gain of the MN scheme is independent of B. In contrast, the effective gain of the ACC scheme increases as B increases, and the increasing trend becomes more obvious in the small m region. Indeed, the symmetry of the plot illustrates that the number of users per cache state (B) and the fading parameter m are somehow equivalent. Thus, a setting with K single-antenna receivers and B users per cache group achieves approximately the same performance as a setting with K m-antenna receivers that apply MRC.

VI. Conclusions

We have analyzed the performance of coded caching under Nakagami-m fading with a focus on the low-SNR regime. We have shown how the gains offered by coded caching are preserved at low SNR, where the worst-user bottleneck was thought to erase these gains. In order to keep the coded caching gains, we exploit the inevitable subpacketization constraint and the fact that several users sharing the same cache content can alleviate the worst-user effect. We have derived approximations for the low-SNR regime and for the regime of large number of users (valid for any SNR value). These results show that coded caching gains on delivery time can be translated to practical settings.

Appendix : Proof of Lemma 3

In the following, we prove Lemma 3, i.e., we obtain the exact expression of $\mathring{T}_{\mathrm{ACC}} \triangleq \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}\{T_{\mathcal{G}}\}$. For that, we need to obtain $\mathbb{E}\{T_{\mathcal{G}}\}$, and, since $T_{\mathcal{G}}$ has non-negative support, it follows that $\mathbb{E}\{T_{\mathcal{G}}\} = \int_0^\infty (1 - F_{T_{\mathcal{G}}}(z))\mathrm{d}z$. Consequently, we have to obtain $F_{T_{\mathcal{G}}}(z)$.

Let us define $\tau_g \triangleq \sum_{b=1}^{B} |h_{g,b}|^{-2}$ for the sake of readability, such that we can write $T_{\mathcal{G}}$ as $T_{\mathcal{G}} = \max_{g\in\mathcal{G}}\{\tau_g\}$. Then, the CDF of $T_{\mathcal{G}}$ can be obtained as

$$F_{T_{\mathcal{G}}}(z) = \Pr\left\{\max_{g\in\mathcal{G}}\{\tau_g\} \leq z\right\} = \left[F_{\tau_g}(y)\right]^{|\mathcal{G}|}. \tag{24}$$

Since $|h_{g,b}|^2 \sim \mathrm{Gamma}(m,1)$, $1/|h_{g,b}|^2$ follows an inverse Gamma distribution, it follows that (cf. [18])

$$\mathrm{CF}_{1/|h_{g,b}|^2}(t) = \frac{2(-\jmath t)^{\frac{m}{2}}}{\Gamma(m)} K_m\left(\sqrt{-4\jmath t}\right). \tag{25}$$

From (25), the CF of τ_g can be expressed as

$$\mathrm{CF}_{\tau_g}(t) = \mathbb{E}\left\{\exp\left(\jmath t \sum\nolimits_{b=1}^{B} \frac{1}{|h_{g,b}|^2}\right)\right\}$$

$$= \mathbb{E}\left\{\prod_{b=1}^{B} \exp\left(\frac{\jmath t}{|h_{g,b}|^2}\right)\right\} = \left(\frac{2(-\jmath t)^{\frac{m}{2}} K_m\left(\sqrt{-4\jmath t}\right)}{\Gamma(m)}\right)^B.$$

We can apply the Gil-Pelaez Theorem [20] to obtain the CDF of τ_g from its CF, which yields

$$F_{\tau_g}(y) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\mathrm{Im}\left\{\exp(-\jmath t y)\,\mathrm{CF}_{\tau_g}(t)\right\}}{t}\mathrm{d}t$$

$$= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \mathrm{Im}\left\{\frac{\exp(-\jmath t y)}{t}\left(\frac{2(-\jmath t)^{\frac{m}{2}} K_m\left(\sqrt{-4\jmath t}\right)}{\Gamma(m)}\right)^B\right\}\mathrm{d}t.$$

By plugging this expression in (24), we obtain $F_{T_{\mathcal{G}}}(z)$. Next, we apply the facts that $\mathbb{E}\{T_{\mathcal{G}}\} = \int_0^\infty (1 - F_{T_{\mathcal{G}}}(z))\mathrm{d}z$ and that $\mathring{T}_{\mathrm{ACC}} \triangleq \rho^{-1} \frac{\Lambda(1-\gamma)F \ln 2}{B_w(1+\Lambda\gamma)} \mathbb{E}\{T_{\mathcal{G}}\}$ to obtain (14).

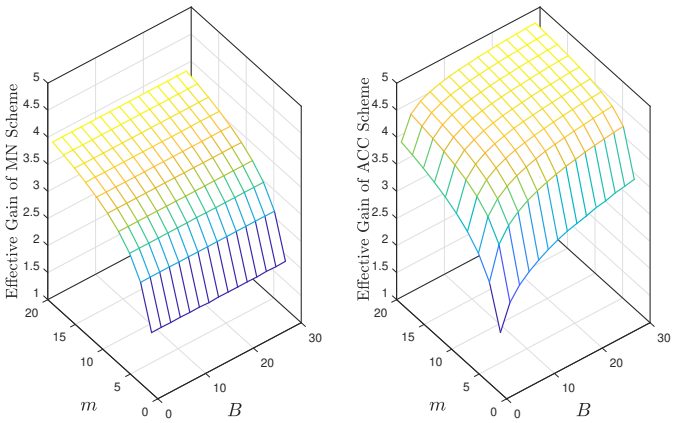

Fig. 5: Effective coded caching gains of MN (left) and ACC (right) schemes for $|\mathcal{G}| = 5$ and $\gamma = 5\%$ as $\rho \to 0$.

References

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

[2] 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.

[3] 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.

[4] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," 2020. [Online]. Available: https://arxiv.org/abs/1811.03935

[5] 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.

[6] H. Joudeh, E. Lampiris, P. Elia, and G. Caire, "Fundamental limits of wireless caching under mixed cacheable and uncacheable traffic," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1693–1698.

[7] 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.

[8] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 349–366, Jan. 2018.

[9] 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.

[10] ——, "Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes," 2020, submitted to IEEE Trans. Wireless Commun. [Online]. Available: https://arxiv.org/abs/2103.02967

[11] ——, "Wireless coded caching with shared caches can overcome the near-far bottleneck," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 350–355.

[12] K. E. Nolan, W. Guibene, and M. Y. Kelly, "An evaluation of low power wide area network technologies for the internet of things," in *Proc. Int. Wireless Commun. and Mobile Computing Conf. (IWCMC)*, Sep. 2016, pp. 439–444.

[13] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[14] 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.

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

[16] 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.

[17] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Academic press, 2007.

[18] F. J. Girón and C. del Castill, "A note on the convolution of inverted-gamma distributions with applications to the behrens-fisher distribution," *RACSAM*, vol. 95, no. 1, pp. 39–44, 2001.

[19] S. Venkateshan and P. Swaminathan, *Computational Methods in Engineering*. Academic Press, 2014.

[20] J. Gil-Pelaez, "Note on the inversion theorem," *Biometrika*, vol. 38, no. 3-4, pp. 481–482, 1951.