Wireless Coded Caching With Shared Caches Can Overcome the Near-Far Bottleneck

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

***Abstract*—We investigate the use of coded caching in the single-cell downlink scenario where the receiving users are randomly located inside the cell. We first show that, as a result of having users that experience very different path-loss, the real gain of the original coded caching scheme is severely reduced. We then prove that the use of shared caches — which, we stress, is a compulsory feature brought about by the subpacketization constraint in nearly every practical setting — introduces a spatial-averaging effect that allows us to recover most of the subpacketization-constrained gains that coded caching would have yielded in the error-free identical-link setting. For the ergodic-fading scenario with different pathloss, we derive tight approximations of the average (over the users) rate and of the coded caching gain by means of a basic high-SNR approximation on the point-to-point capacity. These derived expressions prove very accurate even for low SNR. We also provide a result based on the regime of large number of users which is nonetheless also valid for settings with few users. These results are extensively validated using Monte-Carlo simulations that adhere to 3GPP recommendations on system parameters for urban micro or macro cells.**

I. Introduction

Coded caching, first introduced in [1] by Maddah-Ali and Niesen, is an effective method to improve both spectral efficiency and network costs for content delivery. These improvements stem from the fact that coded caching leverages partially-shared information cached at the end users to create substantive multicasting opportunities. The scenario considered in [1] consisted of a transmitter with access to a library of N files serving a set of K users through an error-free shared-link Broadcast Channel. Each such receiving user has a local memory (cache) of size equal to the size of M files, thus being able to store a fraction $\gamma \triangleq \frac{M}{N}$ of the library content. The scheme in [1] (henceforth referred to as the MN scheme) allowed the transmitter to serve $K\gamma + 1$ users at a time, where $K\gamma + 1$ is referred to as the *coded caching gain*.

The work in [1] sparked interest in analyzing coded caching in wireless settings, where one naturally needs to account for feedback accuracy [2], [3], uneven channel capacities [4]–[6], multi-antenna transmissions [7]–[10], cell-free configurations [11], as well as a variety of other directions [12]–[16].

As of now, we know of two main constraints that severely limit the performance of coded caching. The first one is referred to as the finite file-size (or subpacketization) constraint, where, under realistic assumptions on γ and K, the file sizes (or subpacketization) required by caching schemes can dwarf any realistic file sizes that we encounter in wireless use-cases. Let us refer to the specific content stored at a user as the *cache state* of that user, and denote the number of *distinct* cache states by Λ. Under realistic assumptions, all known top-performing coded caching schemes require file sizes that grow exponentially or near-exponentially with Λ [17]–[19]. This constraint will realistically force a Λ that can be orders of magnitude smaller than K, thus forcing (when the MN scheme is applied over the original error-free scenario with identical link capacities) a *nominal* caching gain that falls from the original $K\gamma+1$ down to a much reduced gain of $\Lambda\gamma+1$.

The other main challenge, often referred to as the "worst-user bottleneck" [20], is brought about by the fact that multicast transmissions must always be limited by the rate of the worst-channel user [21]. The severity of this constraint was investigated in various works such as [22]–[26], where either outage probability or user selection are allowed to alleviate this worst-user effect, as well as in [5], where channel capacity knowledge is exploited during the cache-placement phase. The recent work in [27], [28] analyzed this worst-user effect under the assumption of finite file sizes and $\Lambda < K$. The authors developed the so-called Aggregated Coded Caching (ACC) scheme, which recovered most of the gains of coded caching without any user selection technique, effectively removing the worst-user constraint for large enough K. The scheme exploits that the same cache state is stored at several users [29] and uses a multi-rate scheme that replaces XOR-based transmissions.

A similar effect can be found in the realistic ergodic fading scenario considered here, where each user experiences a different long-term path-loss. This near-far effect, which indeed persists even in the presence of fast-fading, has been nicely analyzed in [11], [30], which focused on cell-free scenarios with multi-antenna receiving users, as well as in [31], where inner and outer bounds were obtained for distributed networks in which cache-aided nodes serve arbitrary demands.

Motivated by modern paradigms that consider dense networks of multimedia-demanding users [32], [33], our work here investigates the setting where file-size constrained coded caching is applied for cellular downlink. For this, we consider the fast-fading single-cell scenario where the users are randomly located within a ring around the transmitter. In this context, we derive the affine approximation [34] of the average (over the different users) rate, both for the standard MN scheme and for the aforementioned ACC variant. This simple approximation, while drawing from high-SNR simplifications, is here shown to tightly characterize the performance even when considering

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).

low-to-moderate SNR. We also provide additional large-K analysis, which is nicely tailored to settings such as dense urban cells. In these realistic dense settings, our analysis shows how the ACC scheme—which considerably outperforms the original MN scheme over symmetric Rayleigh channels [28]—manages to maintain most of the nominal coded caching gains that were originally associated to the error-free, identical-link scenario. This is partly done by exploiting the unavoidable use of shared cache states to neutralize the near-far effect.

Notations: For a positive integer n, we use the notation $[n] \triangleq \{1, 2, \cdots, n\}$. $|\cdot|$ denotes the cardinality for a set and the magnitude for a complex number. The notation $X \sim \mathcal{Y}$ denotes the random variable X following the distribution $\mathcal{Y}$.

II. System Model

A. System Settings

We analyze a single-cell setting in which a single-antenna transmitter has full access to a library $\mathcal{F} = \{W_n\}_{n=1}^N$ of N files $W_1, \ldots, W_N$, each of size F bits. The transmitter serves K cache-aided single-antenna users who are uniformly distributed throughout a ring with inner radius D_1 and outer radius D_2 surrounding the transmitter. Each user benefits from a local cache that can store a fraction $\gamma \triangleq \frac{M}{N} \in [0, 1]$ of the library content. We consider the instantaneous SNR at an arbitrary user U_i to be given by[1]

$$\mathrm{SNR}_i = \frac{P_t}{N_0 B_w N_f \beta_f L_c} |h_i|^2 r_i^{-\eta}, \tag{1}$$

where P_t, N_0, B_w, and η are the transmit power, the noise density, the bandwidth, and the path-loss exponent, respectively. In the above, h_i corresponds to the fast-fading channel coefficient, drawn from a zero-mean unit-variance complex Gaussian distribution, and r_i is the distance in meters from user i to the transmitter. In (1), β_f is a path-loss component that depends only on the carrier frequency (f_{GHz}), while N_f denotes the noise figure that measures the practical imperfections of the receiver, and L_c represents a constant loss term accounting for slow fading and other practical factors (rain, foliage, etc.). To facilitate notation, we define $\rho \triangleq \frac{P_t}{N_0 B_w N_f \beta_f L_c}$ to incorporate all the terms in SNR_i other than the distance and the effect of fast fading. As the users are uniformly distributed within a ring, the probability density function (PDF) of r_i is (cf. [36])

$$f_{r_i}(r) = \frac{2r}{D_2^2 - D_1^2}, \quad D_1 \le r \le D_2. \tag{2}$$

Under basic Gaussian signalling assumptions, we consider that a single-user transmission through a channel with instantaneous SNR denoted by X can attain an ergodic rate of $\mathbb{E}_h\{\ln(1+X)\}$ when averaged over the fast-fading channel coefficients. We will consider the average of such ergodic rate, averaged over the users, or, more precisely, over the random user locations. In this context, we will use $\mathbb{E}_{h|r}$ to denote the expectation over fast-fading channels for a given location realization r which remains fixed for the entire transmission, and we will use $\mathbb{E}_r$ to denote averaging over the user locations r. Hence, our *average rate* metric corresponds to averaging $\mathbb{E}_{h,r}\{\cdot\} \triangleq \mathbb{E}_r\{\mathbb{E}_{h|r}\{\cdot\}\}$, i.e., averaging the ergodic capacities across user locations.

We consider the basic linear (with respect to $\ln \rho$) approximation $\ln(1+x) \approx \ln x$ on the capacity, obtained from the fact that $\ln(1+x) = \ln x + o(1)$, where $\lim_{x\to\infty} o(1) = 0$.

Definition 1 (Affine approximation of average rate): The average-rate affine approximation $\tilde{R}$ is defined as the maximum long-term average achievable rate, averaged first over the fast-fading coefficients and then over user locations, after applying the linear[2] approximation (with respect to $\ln x$) $\ln(1+x) \approx \ln x$. Next, we define the effective coded caching gain.

Definition 2 (Effective coded caching gain): The effective coded caching gain of a particular scheme is the ratio of the average rate achieved by the said scheme over the average rate attained by simple Time-Division Multiplexing (TDM).

B. Aggregated Coded Caching Scheme

Let Λ denote the number of cache states allowed by the file-size constraint, where $\Lambda \le K$. For simplicity, we assume that K is an integer multiple of Λ. In order to apply the MN scheme in this scenario with finite file-size constraint, the same cache state is stored in $B \triangleq K/\Lambda$ users. Then, we apply the MN scheme only over Λ users, each endowed with a different cache state, and this must be repeated $\frac{K}{\Lambda}$ times.[3] Next, we briefly describe the ACC scheme first introduced in [27], [28].

1) Cache Placement: The subpacketization follows as for the MN scheme with Λ users: Each file W_n is split into $\binom{\Lambda}{\Lambda\gamma}$ equal-size subfiles as $W_n \to \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\}$. Then, Λ cache states are created such that the g-th cache state, $g \in [\Lambda]$, is given by $\mathcal{Z}_g = \{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, \mathcal{T} \ni g, \forall n \in [N]\}$. Users are evenly split into Λ groups, and all the $B = \frac{K}{\Lambda}$ users in group $g \in [\Lambda]$ cache the cache state $\mathcal{Z}_g$. We denote the b-th user in the g-th group as $\mathrm{U}_{g,b}$, where $g \in [\Lambda]$ and $b \in [B]$.

2) Delivery Phase: There are $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages. At each transmission stage, a unique user-group set $\mathcal{G}$ of $|\mathcal{G}| = \Lambda\gamma + 1$ user groups (thus of $B(\Lambda\gamma+1)$ users) is served. Throughout the transmission to a specific set $\mathcal{G}$, the ACC scheme serves $\Lambda\gamma + 1$ users at a time, but it has two main differences with respect to the MN scheme: i) The simultaneous transmission to a given set of $\Lambda\gamma + 1$ users does not depend on the worst-channel user, and instead each user receives its information at its single-link capacity,[4] and ii) once one of the $\Lambda\gamma + 1$ users has finished decoding its subfile, the transmitter can start transmitting to another user endowed with the same cache state (i.e., from the same user-group $g \in \mathcal{G}$) with no delay, while continuing the transmission to the other $\Lambda\gamma$ users.

Consequently, for a given transmission stage, the average rate achieved by group $g \in \mathcal{G}$ is given by $\frac{1}{B}\sum_{b=1}^{B} \ln(1 + \mathrm{SNR}_{g,b})$,

[1] The channel model follows the propagation models defined by the Third Generation Partnership Project (3GPP) [35].

[2] As we will see below, this linear approximation will provide affine expressions of the average rate, such that it can be written as $A \ln \rho + B$, where A and B are independent of ρ. Such affine approximations have been shown to accurately represent the rate in several scenarios [34], [37].

[3] In the original single-stream error-free scenario, this approach is indeed optimal under the assumption of uncoded cache placement [29].

[4] This follows from the result by Tuncel in [38] on the transmission with side information. We refer to [28] for more details. Some previous works on coded caching also considered Tuncel's coding idea, cf. [25], [26], [39], [40].

since the users are served sequentially within a group. Then, the transmission stage ends when all user groups have decoded their intended subfiles. Thus, the (affine-approximated) achievable rate in a given transmission stage takes the form (cf. [27], [28])

$$R_{\mathcal{G}} \triangleq |\mathcal{G}| \min_{g\in\mathcal{G}} \left\{ \frac{1}{B} \sum\nolimits_{b=1}^{B} \ln\left(\mathrm{SNR}_{g,b}\right) \right\} \text{ nats/s/Hz}, \quad (3)$$

where the factor $|\mathcal{G}| = \Lambda\gamma + 1$ is due to the fact that the transmitter simultaneously serves a set of $|\mathcal{G}|$ users.

III. MN Scheme in the Single-Cell Scenario

We first derive the affine-approximated average rate of TDM. We then provide a tight lower bound on the affine-approximated MN rate and a corresponding bound on the approximated MN effective gain.[5] We quickly recall that the affine-approximated rate is simply the sought average rate after substituting $\ln(1+\mathrm{SNR}_k)$ by $\ln \mathrm{SNR}_k$.

Lemma 1: The affine-approximated rate for TDM is

$$\tilde{R}_{\mathrm{TDM}} = \ln\rho - \xi + \frac{\eta}{2} - \eta \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2} \quad (4)$$

where $\xi = 0.5772\ldots$ is the Euler-Mascheroni constant.

Proof: The affine-approximated average rate can be written as

$$\tilde{R}_{\mathrm{TDM}} = \mathbb{E}_r\left\{ \mathbb{E}_{h|r}\left\{ \ln\left(\rho r_k^{-\eta} |h_k|^2\right)\right\}\right\}, \quad (5)$$

where $\rho r_k^{-\eta}$ is naturally constant with respect to the inner expectation, and where $|h_k|^2$ follows a unit-mean exponential distribution. It then follows that

$$\tilde{R}_{\mathrm{TDM}} \overset{(a)}{=} \mathbb{E}_r\left\{\ln(\rho r_k^{-\eta}) - \xi\right\} = \ln\rho - \xi - \eta \mathbb{E}_r\{\ln r_k\}, \quad (6)$$

where (a) holds because, for any X drawn from a unit-mean exponential distribution, $-\ln(X)$ follows a standard Gumbel distribution, whose mean equals ξ [34]. We then obtain (4) from (6) by deriving $\mathbb{E}_r\{\ln r_k\}$ from the PDF of r_k in (2). □

We present a lower bound on the affine approximation of the MN scheme in the following lemma, which is derived by means of Jensen's inequality. As we will see in Section V, this simple bound offers a tight approximation of the actual rate.

Lemma 2: The affine-approximated average rate of the MN scheme is lower bounded by

$$\tilde{R}_{\mathrm{MN}} \geq |\mathcal{G}| \ln\rho - |\mathcal{G}| \left[\ln\left(\frac{2|\mathcal{G}|(D_2^{\eta+2} - D_1^{\eta+2})}{(\eta+2)(D_2^2 - D_1^2)} \right) + \xi \right]. \quad (7)$$

Proof: Since the expression of the average rate of the MN scheme coincides with the one of the ACC scheme for $B = 1$ (i.e., with dedicated caches) [27], it follows from (3) that

$$\tilde{R}_{\mathrm{MN}} = |\mathcal{G}| \ln\rho + |\mathcal{G}| \, \mathbb{E}_{h,r}\left\{ \ln\left(\min\nolimits_{k\in\mathcal{G}} \left\{ |h_k|^2 r_k^{-\eta} \right\} \right) \right\}, \quad (8)$$

where we have applied the approximation $\ln(1+x) \approx \ln x$. Since $|h_k|^2 r_k^{-\eta}$ follows an exponential distribution with mean $r_k^{-\eta}$ for a fixed location realization, $\min_{k\in\mathcal{G}} \left\{ |h_k|^2 r_k^{-\eta} \right\}$ follows an exponential distribution with mean $1/\left(\sum_{k\in\mathcal{G}} r_k^{\eta}\right)$. Hence, applying the same step as for (a) in (6), the expectation of $\ln\left(\min_{k\in\mathcal{G}} \left\{ |h_k|^2 r_k^{-\eta} \right\}\right)$ conditioned on the user locations is

$$\mathbb{E}_{h|r}\left\{ \ln\left(\min_{k\in\mathcal{G}} \left\{ |h_k|^2 r_k^{-\eta} \right\} \right) \right\} = -\ln\left(\sum\nolimits_{k\in\mathcal{G}} r_k^{\eta} \right) - \xi. \quad (9)$$

[5] In a small deviation of notation, and only for this section, we will use SNR_k, $\forall k \in [K]$, to denote the SNR of User k in both TDM and MN schemes.

Since $-\ln x$ is a convex function over $(0,\infty)$, we can apply Jensen's inequality to obtain

$$\begin{aligned} \mathbb{E}_{h,r}\left\{ \ln\left(\min_{k\in\mathcal{G}} \left\{ |h_k|^2 r_k^{-\eta} \right\} \right) \right\} &= \mathbb{E}_r\left\{ -\ln\left(\sum_{k\in\mathcal{G}} r_k^{\eta} \right) \right\} - \xi \\ &\geq -\ln\left(|\mathcal{G}| \mathbb{E}_r\left\{ r_k^{\eta} \right\} \right) - \xi. \end{aligned} \quad (10)$$

Finally, we obtain (7) by using the PDF of r_k to derive the closed-form expression for $\mathbb{E}_r\{r_k^{\eta}\}$ in (10). □

Let us now approximate the effective coded caching gain of the MN scheme (which was introduced in Definition 2) by the ratio $\tilde{G}_{\mathrm{MN}}$ of the corresponding average-rate affine approximations. The following is direct from Lemmas 1 and 2.

Corollary 1: The approximate effective coded caching gain for the MN scheme is bounded as

$$\tilde{G}_{\mathrm{MN}} \geq \frac{|\mathcal{G}| \ln\rho - |\mathcal{G}| \left[\ln\left(\frac{2|\mathcal{G}|(D_2^{\eta+2} - D_1^{\eta+2})}{(\eta+2)(D_2^2 - D_1^2)} \right) + \xi \right]}{\ln\rho - \xi + \frac{\eta}{2} - \eta \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2}}. \quad (11)$$

In the above, the approximation steps only include the use of Jensen's inequality and the use of the basic approximation on capacity. Consequently, as we will show, the above results offer a very accurate evaluation of the actual performance.

IV. ACC Scheme in the Single-Cell Scenario

In this section, we first derive the analytical expression for the affine-approximated average rate of the ACC scheme. Then, we provide a large-B approximation that precisely characterizes the actual performance even if B is as low as 2.

A. Affine-Approximated Average Rate

Before presenting the lemma describing the performance of the ACC scheme, let us introduce the notation $\jmath \triangleq \sqrt{-1}$ and let $\Im\{\cdot\}$ denote the imaginary part of a complex number.

Lemma 3: The affine-approximated average rate of the ACC scheme is given by the expression in (12) at the top of next page, where $\Gamma(\cdot)$ denotes the Gamma function [41] and $|\mathcal{G}| = \Lambda\gamma + 1$.

Proof: From (3), we can write that

$$\tilde{R}_{\mathrm{ACC}} = |\mathcal{G}| \ln\rho + \frac{|\mathcal{G}|}{B} \mathbb{E}_{h,r}\left\{ \min_{g\in\mathcal{G}} \left\{ \sum_{b=1}^{B} \ln\left(|h_{g,b}|^2 r_{g,b}^{-\eta} \right) \right\} \right\}. \quad (13)$$

For $X_{g,b} = \ln\left(|h_{g,b}|^2 r_{g,b}^{-\eta} \right)$, the characteristic function (CF) $\mathrm{CF}_{X_{g,b}}(t) = \mathbb{E}_{h,r}\{\exp(\jmath t X_{g,b})\}$ can be shown to be

$$\mathrm{CF}_{X_{g,b}}(t) = \mathbb{E}_h\left\{ \left(|h_{g,b}|^2 \right)^{\jmath t} \right\} \mathbb{E}_r\left\{ r_{g,b}^{-\jmath\eta t} \right\}. \quad (14)$$

By substituting the PDFs of $|h_{g,b}|^2$ and $r_{g,b}$ into (14), we have

$$\mathrm{CF}_{X_{g,b}}(t) = \Gamma(1+\jmath t) \frac{\jmath 2 (D_2^{2-\jmath\eta t} - D_1^{2-\jmath\eta t})}{(\eta t + \jmath 2)(D_2^2 - D_1^2)}. \quad (15)$$

Let us define $X_g \triangleq \sum_{b=1}^{B} X_{g,b}$. The CF of X_g takes the form $\mathrm{CF}_{X_g}(t) = \prod_{b=1}^{B} \mathrm{CF}_{X_{g,b}}(t)$. By using the Gil-Pelaez Theorem [42], the CDF of X_g can be obtained as

$$\begin{aligned} F_{X_g}(x) &= \frac{1}{2} - \frac{1}{\pi} \int_0^{\infty} \frac{\Im\{\exp(-\jmath t x) \mathrm{CF}_{X_g}(t)\}}{t} \mathrm{d}t \\ &= \frac{1}{2} - \frac{1}{\pi} \int_0^{\infty} \Im\left\{ \frac{\exp(-\jmath t x)}{t} \right. \\ &\qquad \left. \times \left[\Gamma(1+\jmath t) \frac{\jmath 2 (D_2^{2-\jmath\eta t} - D_1^{2-\jmath\eta t})}{(\eta t + \jmath 2)(D_2^2 - D_1^2)} \right]^B \right\} \mathrm{d}t. \end{aligned} \quad (16)$$

$$\tilde{R}_{\rm ACC} = |\mathcal{G}|\ln\rho - \frac{|\mathcal{G}|}{B}\int_0^\infty 1 - \left(\frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Im\left\{\frac{\exp(\jmath t x)}{t}\left[\Gamma(1+\jmath t)\frac{\jmath 2(D_2^{2-\jmath\eta t} - D_1^{2-\jmath\eta t})}{(\eta t + \jmath 2)(D_2^2 - D_1^2)}\right]^B\right\}\mathrm{d}t\right)^{|\mathcal{G}|}\mathrm{d}x. \quad (12)$$

Let $X \triangleq \min_{g\in\mathcal{G}}\{X_g\}$. The CDF of X is

$$F_X(x) = 1 - \mathbb{P}\big(\min_{g\in\mathcal{G}}\{X_g\} > x\big) = 1 - \big(1 - F_{X_g}(x)\big)^{|\mathcal{G}|}. \quad (17)$$

As $D_1^\eta \gg 1$ and $|h_{g,b}|^2 \sim \mathrm{Exp}(1)$, the probability that $\ln(|h_{g,b}|^2 r_{g,b}^{-\eta})$ is bigger than zero is negligible,[6] so we can consider the variable $X = \min_{g\in\mathcal{G}}\left\{\sum_{b=1}^B \ln\left(|h_{g,b}|^2 r_{g,b}^{-\eta}\right)\right\}$ to be a non-positive random variable. Hence $X' = -X$ satisfies

$$\mathbb{E}\{X'\} = \int_0^\infty F_X(-x)\mathrm{d}x = \int_0^\infty 1 - \big(1 - F_{X_g}(-x)\big)^{|\mathcal{G}|}\,\mathrm{d}x. \quad (18)$$

Combining (13), (16) and (18), we obtain (12). □

Remark 1: (12) provides an affine approximation of the rate, where $|\mathcal{G}| = \Lambda\gamma + 1$ in the first term defines the Degrees-of-Freedom and it is equal for both MN and ACC schemes. The second term, which does not depend on the transmit power, defines the rate-offset in the affine approximation, and it explains the improved performance of the ACC scheme. This type of affine approximations have been considered at high-SNR [34], but it will be shown later that, for this setting, it also offers an accurate characterization at practical SNR ranges.

B. Large B Approximation

We derive now an approximation based on assuming a large number of users, which is meant to simplify the previous results. This assumption is well justified by current trends in practical dense scenarios, where K is expected to be as large as 800 in a dense urban Micro-cell setting, and up to 4000 in a dense urban Macro-cell setting [32], [33]. Then, in addition to the linear approximation on the capacity, we will also apply the Central Limit Theory (CLT) after assuming that B is large.

Let us define $S_{g,b} \triangleq \rho|h_{g,b}|^2 r_{g,b}^{-\eta}$, and let μ_S and σ_S^2 denote respectively the mean and the variance of $\ln(S_{g,b})$. Furthermore, let $-H_{|\mathcal{G}|}$ denote the expectation of the minimum of $|\mathcal{G}|$ i.i.d. Gaussian random variables with zero-mean and unit-variance. From [28], we know that $H_{|\mathcal{G}|}$ can be expressed as

$$H_{|\mathcal{G}|} = \frac{-|\mathcal{G}|}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x\big(Q(x)\big)^{|\mathcal{G}|-1}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}y, \quad (19)$$

where $Q(\cdot)$ denotes the well-known complementary CDF (or tail distribution) of the standard Gaussian distribution.[7]

Lemma 4: For $B \to \infty$, the affine-approximated average rate of the ACC scheme satisfies that

$$\tilde{R}_{\rm ACC} \approx |\mathcal{G}|\Big(\ \mu_S - H_{|\mathcal{G}|}\sqrt{\sigma_S^2/B}\ \Big), \quad (20)$$

[6]The minimum distance between the base station and a user is generally considered to be at least 10 meters [32], [43]. Even considering that all the users were located in the inner border of the ring, the probability $\mathbb{P}\big(|h_{g,b}|^2 r_{g,b}^{-\eta} > 1\big)$ is as small as $\exp(-100) = 3.72 * 10^{-44}$ (for $\eta = 2$).

[7]As discussed in [28], $H_{|\mathcal{G}|}$ has a known and simple closed-form expression for $|\mathcal{G}| \leq 5$, whereas we can closely approximate it for all the possible values of $|\mathcal{G}|$ by using the Gauss-Hermite quadrature (GHQ), which provides an accurate approximation even with a few summation terms.

where $X \approx Y$ denotes that X can be approximated by Y with vanishing error as $B \to \infty$, and where μ_S and σ_S^2 are given by

$$\mu_S = \ln\rho - \xi + \frac{\eta}{2} - \eta\frac{D_2^2\ln D_2 - D_1^2\ln D_1}{D_2^2 - D_1^2}, \quad (21)$$

$$\sigma_S^2 = \frac{\pi^2}{6} + \frac{\eta^2}{4} - \eta^2\frac{D_1^2 D_2^2\left(\ln D_2 - \ln D_1\right)^2}{(D_2^2 - D_1^2)^2}. \quad (22)$$

Proof: Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian distribution of mean μ and variance σ^2. It is clear that the variables $\{S_{g,b}\}_{g\in\mathcal{G}, b\in[B]}$ are i.i.d. variables, and thus we can apply the CLT to get

$$\frac{1}{B}\sum_{b=1}^B \ln\left(\rho|h_{g,b}|^2 r_{g,b}^{-\eta}\right) \xrightarrow{d.} \mathcal{N}\left(\mu_S, \frac{\sigma_S^2}{B}\right), \text{ as } B \to \infty, \quad (23)$$

where $d.$ stands for the convergence in distribution. Let Y_g, $g \in \mathcal{G}$, be i.i.d. Gaussian random variables with zero mean and unit variance. It then follows from (23) that (cf. [28, App. IV])

$$\min_{g\in\mathcal{G}}\frac{1}{B}\sum_{b=1}^B \ln\left(\rho|h_{g,b}|^2 r_{g,b}^{-\eta}\right) \xrightarrow{d.} \mu_S + \sqrt{\frac{\sigma_S^2}{B}}\min_{g\in\mathcal{G}}\{Y_g\}. \quad (24)$$

By considering the definition of $\tilde{R}_{\rm ACC}$, (24) implies that

$$\tilde{R}_{\rm ACC} = \mathbb{E}_{h,r}\left\{|\mathcal{G}|\min_{g\in\mathcal{G}}\left\{\frac{1}{B}\sum\nolimits_{b=1}^B \ln\left(\mathrm{SNR}_{g,b}\right)\right\}\right\} \quad (25)$$

$$\to |\mathcal{G}|\Big(\mu_S + \sqrt{\sigma_S^2/B}\ \mathbb{E}_{h,r}\big\{\min_{g\in\mathcal{G}}\{Y_g\}\big\}\Big) \quad (26)$$

as $B \to \infty$, which together with the definition of $-H_{|\mathcal{G}|}$ yields (20). Note that $\mu_S \triangleq \mathbb{E}_{h,r}\{\ln(\rho|h_{g,b}|^2 r_{g,b}^{-\eta})\}$ is equivalent to $\tilde{R}_{\rm TDM}$ (cf. (5)) and thus (21) follows from (4). For σ_S^2, it holds that

$$\sigma_S^2 = \mathrm{Var}\left\{\ln\rho + \ln\left(|h_{g,b}|^2 r_{g,b}^{-\eta}\right)\right\} = \mathrm{Var}\left\{\ln\left(|h_{g,b}|^2 r_{g,b}^{-\eta}\right)\right\}$$
$$= \mathrm{Var}\left\{\ln\left(|h_{g,b}|^2\right)\right\} + \eta^2\mathrm{Var}\left\{\ln r_{g,b}\right\}. \quad (27)$$

We can then derive $\mathrm{Var}\left\{\ln\left(|h_{g,b}|^2\right)\right\}$ and $\mathrm{Var}\left\{\ln r_{g,b}\right\}$ in (27) from the PDFs of $|h_{g,b}|^2$ and $r_{g,b}$, which yields (22). □

Lemma 4 and the definition of μ_S directly yield the next corollary, which approximates the effective coded caching gain of the ACC scheme by the ratio $\tilde{G}_{\rm ACC}$ of the affine-approximated average rates of the ACC scheme and TDM.

Corollary 2: The approximate effective coded caching gain of the ACC scheme takes the form

$$\tilde{G}_{\rm ACC} \approx \frac{|\mathcal{G}|}{\mu_S}\left(\ \mu_S - H_{|\mathcal{G}|}\sqrt{\sigma_S^2/B}\ \right), \text{ for } B \to \infty, \quad (28)$$

where μ_S is given by (21) and σ_S^2 by (22).

Note that the ACC scheme can be extended to the case where Λ does not divide K with minor impact in the performance, and such performance will converge to the result in (28) as $\frac{K}{\Lambda} \to \infty$. Due to the page limitation, we omit this extension.

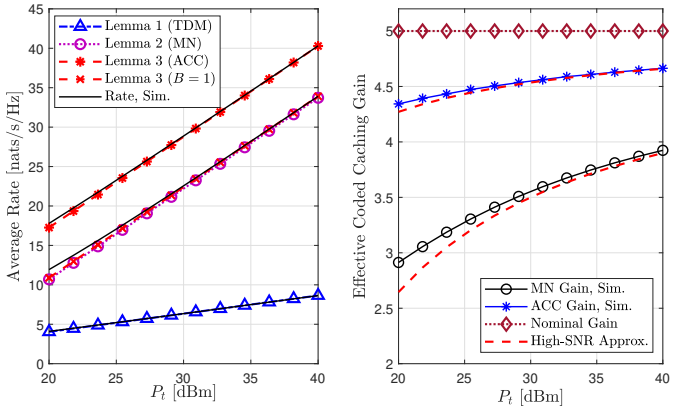

Fig. 1: Performance in a dense urban Micro-Cell, for the case $K = 800$, $\gamma = 5\%$, and $\Lambda = 80$ (edge SNR: $15 \to 35$dB).

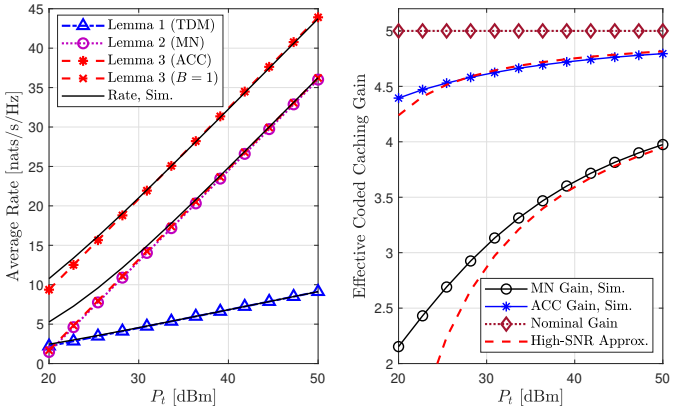

Fig. 2: Performance in a dense urban Macro-Cell for the case $K = 2000$, $\gamma = 5\%$, and $\Lambda = 80$ (edge SNR: $7 \to 35$dB).

V. Numerical Results

We validate our analytical results through Monte-Carlo simulations for two of the main scenarios proposed by 3GPP: the dense urban Micro-cell setting and the dense urban Macro-cell setting. In accordance to 5G standards [35], [43], we consider $f_{\text{GHz}} = 3.5\,\text{GHz}$, $B_w = 20\,\text{MHz}$, $N_0 = -174\,\frac{\text{dbm}}{\text{Hz}}$, and $N_f = L_c = 10\,\text{dB}$; following the 3GPP proposition in [35], we consider $\beta_f(\text{dB}) = 32.4 + 20\log_{10} f_{\text{GHz}}$ and $\eta = 2.1$ for the Micro-cell scenario, and $\beta_f(\text{dB}) = 28 + 20\log_{10} f_{\text{GHz}}$ and $\eta = 2.2$ for the Macro-cell scenario. Following the guidelines in [32], [43] for cell sizes, and recalling that users are located in a ring around the transmitter, we consider a distance range of $[D_1 = 10, D_2 = 100]$ meters in the Micro-cell setting and of $[D_1 = 35, D_2 = 300]$ meters in the Macro-cell setting.

Regarding the SNR operation range, we note the following: In the Macro-cell setting, a power level of $P_t = 20$ dBm entails an average SNR at the users on the cell edge ($D_2 = 300$) of 7 dB, while $P_t = 50$ dBm corresponds to 35 dB of SNR at the edge users. Typical values of transmit power are commonly considered to be $P_t = 33$ dBm and $P_t = 40$ dBm for Micro-cell and Macro-cell, respectively [32], [35]. We have broadened the range of power values to provide a wider perspective.

Following [32], [33], we consider $K = 800$ in the Micro-cell case and $K = 2000$ in the Macro-cell case. We plot

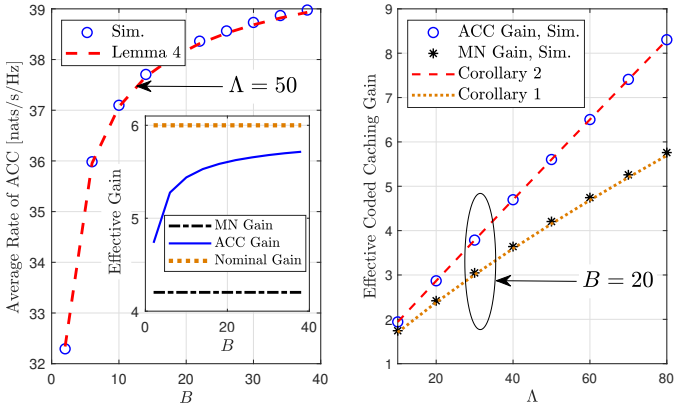

Fig. 3: Performance in Macro-Cell setting with $\gamma = 10\%$ and $P_t = 40\,\text{dBm}$. $H_{|\mathcal{G}|}$ is computed through GHQ with 10 terms.

in Figs. 1–2 the results from Lemmas 1–3 (dashed lines with markers), and validate these by also presenting the corresponding exact performance computed from Monte-Carlo simulations (solid lines) in Figs. 1–2. As we can see, the high-SNR analysis closely characterizes the realistic dense Micro-cell and Macro-cell settings for practical P_t values. The ACC scheme considerably outperforms the MN scheme, recovering most of the distance to the (ideal) nominal gain. Indeed, the robustness of the ACC scheme with respect to low P_t values is made clear as the performance boost over MN increases as P_t decreases. The ACC gain approaches the nominal gain $(\Lambda\gamma + 1)$ much earlier, in terms of SNR, than the MN scheme.

In Fig. 3, we plot the average rate and effective gain versus the number of users per cache state (B) for the dense urban Macro-cell scenario. Although the rate approximation in Lemma 4 is based on the assumption of large B, Fig. 3 (left) shows that the approximation is very accurate even when B is as low as $B = 2$. As expected, the ACC effective gain increases as B increases, since B offers a spatial averaging effect. Finally, Fig. 3 (right) shows that the performance gap between the ACC and MN schemes is proportional to Λ.

VI. Conclusions

This work analyzes the ergodic single-cell cache-aided setting under the assumption of limited number of cache states, which comes along with the unavoidable subpacketization constraint. We provide simple and accurate expressions for the performance of the MN and ACC schemes in realistic scenarios. Unlike previous analysis, which used low-SNR approximations, we consider here a high-SNR approach that nevertheless manages to tightly quantify a bottleneck that is associated to low or moderate SNR. With its substantial gains, the ACC scheme can overcome the very damaging near-far bottleneck that was known to severely diminish caching gains in cellular downlink transmissions. Indeed, in a scenario where the edge-cell users would normally diminish the general performance of XOR-based coded multicasting, the ACC scheme yields a spatial-averaging effect across users that share the same cache state. In the end, this result helps to further solidify the role that coded caching can play in various demanding wireless settings.

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

[3] A. Bazco-Nogueras and P. Elia, "Rate-memory trade-off for the cache-aided MISO Broadcast Channel with hybrid CSIT," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 121–125.

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

[5] 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.

[6] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, "Fundamental limits of wireless caching under uneven-capacity channels," in *Proc. Int. Zurich Seminar on Inf. and Commun. (IZS)*, Feb. 2020, pp. 120–124.

[7] I. Bergel and S. Mohajer, "Cache-aided communications with multiple antennas at finite SNR," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1682–1691, Aug. 2018.

[8] J. Zhang, F. Engelmann, and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," in *Proc. Allerton Conf. on Commun., Control, and Comput. (Allerton)*, 2015, pp. 1099–1105.

[9] 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.

[10] 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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] N. Mital, D. Gündüz, and C. Ling, "Coded caching in a multi-server system with random topology," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4620–4631, Aug. 2020.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] M. Ji and R.-R. Chen, "Caching and coded multicasting in slow fading environment," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[23] 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.

[24] 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.

[25] S. S. Bidokhti, M. Wigger, A. Yener, and A. E. Gamal, "State-adaptive coded caching for symmetric broadcast channels," in *Proc. 51st Asilomar Conf. on Signals, Syst., and Comput.*, 2017, pp. 646–650.

[26] M. M. Amiri and D. Gündüz, "Caching and coded delivery over Gaussian broadcast channels for energy efficiency," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1706–1720, Aug. 2018.

[27] 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. 236–240.

[28] ——, "Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes," 2021. [Online]. Available: https://arxiv.org/abs/2103.02967

[29] 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.

[30] 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.

[31] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[32] "5G Implementation Guidelines," GSMA, Tech. Rep. version 2.0, Jul. 2019, accessed on: 20/01/2021. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2019/03/5G-Implementation-Guideline-v2.0-July-2019.pdf

[33] P. Popovski *et al.*, "Scenarios, requirements and KPIs for 5G mobile and wireless system," Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), Tech. Rep. ICT-317669-METIS/D1.1, Apr. 2013, accessed on: 01/12/2020. [Online]. Available: https://cordis.europa.eu/docs/projects/cnect/9/317669/080/deliverables/001-METISD11v1pdf.pdf

[34] A. Lozano, A. Tulino, and S. Verdú, "High-SNR power offset in multiantenna communication," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4134–4151, Dec. 2005.

[35] "Study on channel model for frequencies from 0.5 to 100 ghz," 3GPP, Tech. Rep. 38.901, version 16.1.0, Release 16, Dec. 2019, accessed on: 22/12/2020. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/16.01.00_60/tr_138901v160100p.pdf

[36] C. Zhang, J. Ye, G. Pan, and Z. Ding, "Cooperative hybrid VLC-RF systems with spatially random terminals," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6396–6408, Dec. 2018.

[37] A. Bazco-Nogueras, P. de Kerret, D. Gesbert, and N. Gresset, "Asymptotically achieving centralized rate on the $M \times K$ decentralized network MISO," 2020. [Online]. Available: https://arxiv.org/abs/2004.11918

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

[39] S. S. Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6996–7016, Nov. 2018.

[40] M. M. Amiri and D. Gündüz, "Cache-aided content delivery over erasure broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 370–381, Jan. 2018.

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

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

[43] P. von Butovitsch, D. Astely, C. Friberg, A. Furuskar, B. Goransson, B. Hogan, J. Karlsson, and E. Larsson, "Advanced antenna systems for 5G networks," Ericsson, Tech. Rep. GFMC-18:000530, Nov. 2018, accessed on: 22/01/2021. [Online]. Available: https://www.ericsson.com/4a8a87/assets/local/reports-papers/white-papers/10201407_wp_advanced_antenna_system_nov18_181115.pdf