NOMA-Aided Aggregated Coded Caching

Hui Zhao, Dirk Slock, and Petros Elia
Communication Systems Department, EURECOM, Sophia Antipolis, France
Email: hui.zhao@eurecom.fr; dirk.slock@eurecom.fr; petros.elia@eurecom.fr

***Abstract*—We propose a new class of coded caching schemes for wireless networks, which we term as non-orthogonal multiple access (NOMA) aided aggregated coded caching (NACC), which manages to alleviate both the uneven channel bottleneck as well as the shared-cache limitation. In particular, NACC uses an aggregation principle to efficiently serve users with unequal channel strengths, as well as uses for the first time the NOMA principle to simultaneously transmit to multiple users that share the same exact cache state. This transmission strategy represents an efficient utilization of NOMA within the coded caching framework. We analyze the high-SNR transmission performance of the proposed scheme and derive analytical expressions for the achievable rate and the effective gain, providing a qualitative understanding of its spectral efficiency. Our numerical results for urban Micro-cell environments abiding by 5G standards, demonstrate that NACC achieves – with a single transmit antenna – the same spectral efficiency as a multi-user (MU) multicasting system with 32 transmit antennas, while it also matches the spectral efficiency of traditional MU unicasting system with 8 transmit antennas.**

I. Introduction

Coded caching has emerged as a promising technique to improve spectral efficiency by creating *new multiplexing space* through the use of cached content at receivers [1]–[6]. Originally proposed by Maddah-Ali and Niesen in [1], the coded caching scheme divides files into multiple subfiles and delivers coded multicast messages generated from these subfiles to users. This approach, known as the MN scheme, enables interference-free transmissions to multiple users, each with different content requests. Coded caching offers a theoretical coded caching gain (or multiplexing gain) of $K\gamma+1$, where K is the total number of users and $\gamma \in [0,1]$ is the cache size normalized to the library size. Such gains can be particularly beneficial in reducing peak traffic loads during busy periods in content delivery networks.

However, when coded caching is applied to wireless environments, where users experience varying channel qualities, the advantages of coded caching will be severely constrained by the so-called *worst-user bottleneck*. In wireless broadcast channels, the multicast rate is fundamentally limited by the user with the worst channel conditions, leading to a significant reduction in the effective multicast gain, particularly in low-to-medium SNR regimes common in practical deployments. Another critical bottleneck is the *finite file-size constraint*. This occurs because coded caching schemes require files to be divided into numerous subfiles, which typically grow exponentially with K. To address this, coded caching often repeats its operation over several rounds, forcing the theoretical gain to reduce to $\Lambda\gamma+1$, where Λ is determined by the subpacketization level, and which remains clearly within the *single-digit range* [3]. This also inevitably forces some users to cache the same content. These two bottlenecks significantly impact the performance of coded caching in wireless networks.

Our previous works in [7]–[10] addressed the worst-user bottleneck under the finite file-size constraint through the design of the so-called *aggregated coded caching (ACC)*. Specifically, in the presence of finite subpacketization, we developed ACC that effectively mitigated the worst-user bottleneck by fully exploiting the shared side information among the users with the same cached content. This approach replaced the worst-user limitation with a worst-group-of-users effect. This ACC scheme, however, relied on time-division multiplexing (TDM), which, while mitigating the impact of the worst-user bottleneck, is suboptimal for wireless environments characterized by the asymmetry in channel conditions.

The primary drawback of TDM lies in its sequential nature, which limits spectrum utilization, particularly in urban Micro-cell environments that are typically governed by the high signal-to-noise ratio (SNR) regime [11], [12]. In such settings, non-orthogonal multiple access (NOMA) offers a more efficient alternative [13]–[15], as it can serve multiple users simultaneously by leveraging power-domain multiplexing and successive interference cancellation (SIC). NOMA is particularly advantageous in high-SNR regimes, where it has been shown to address the asymmetry of wireless channels and significantly enhance spectral efficiency compared to TDM. We note that there have been studies integrating caching with NOMA, but the majority of these works focus on applying NOMA in uncoded caching scenarios (cf. [16], [17]). Existing coded caching studies, on the other hand, mostly consider ideal channel conditions and do not place significant emphasis on the physical-layer transmission design [18]. To the best of our knowledge, very few works have explored incorporating NOMA into coded caching to improve its delivery performance in realistic wireless environments. This combination holds significant research value, as NOMA, being a physical-layer multiple access technique, and coded caching, fundamentally a content-delivery strategy, operate at different layers. Effectively integrating these two

This work is supported in part by the projects EEMW4FIX (French ANR), CONVERGE (EU H2030), and ERC-PoC Project LIGHT (Grant 101101031).

technologies to maximize their complementary benefits poses an interesting and worthwhile challenge.

In this paper, we first design the *NOMA-aided ACC (NACC)* scheme. Building on the original ACC framework, the NACC scheme effectively incorporates the NOMA technology and its advantages, *tailoring a transmission strategy specifically for coded caching* in realistic wireless environments with fading and asymmetric channels. Notably, under practical high-SNR conditions, this scheme substantially improves the spectral efficiency of the original ACC scheme. We then focus on analyzing its performance in the high-SNR regime. Specifically, we derive analytical expressions for the achievable rate and the corresponding effective gain over uncoded caching, providing valuable insights into the scheme's spectral efficiency. These expressions also offer an efficient tool for quick evaluation of the delivery performance under various conditions.

In our numerical results, which are based on 5G standards in an urban Micro-cell setting, we demonstrate that the proposed NACC scheme achieves a remarkable improvement in spectral efficiency compared to the original ACC. Moreover, the proposed scheme with *a single transmit antenna* achieves the same spectral efficiency as the system with *32 transmit antennas* in the multi-user multicasting scenario which can be treated as classical XOR-based multi-antenna coded caching (e.g., [2]). It is worth noting that such a compared system employs the optimal full-digital (FD) beamformer based on NP-hard and non-convex optimization. Furthermore, compared to the *multi-user unicasting scenario*, which corresponds to the uncoded caching scenario where the multi-antenna transmitter must simultaneously send multiple signal symbols to serve multiple users, our scheme again with *a single transmit antenna* matches the spectral efficiency of a multi-user unicasting system with *8 transmit antennas* using minimum mean squared error (MMSE) precoding optimized over the operational multiplexing gain.[1]

II. System Model and NACC Design

In a cache-aided system, a single-antenna transmitter serves a total of K users, each requesting *different* files from a shared library, denoted as $\mathcal{F}$. The library contains N equal-sized files with F bits[2], where $N \geq K$. While the transmitter has full access to the entire library, each user has limited storage, capable of caching only a fraction $\gamma \in [0,1]$ of the total library content. The communication system is designed to handle time-varying user requests, with peak and off-peak periods. To efficiently manage these variations, the process is divided into two phases: the cache placement phase, which occurs during off-peak times when network traffic is lower, and the content delivery phase, which happens during peak times when user requests are served by intelligently combining cached content and the transmitted data. This strategic use of cached content minimizes the load during high-demand periods and improves the delivery performance over peak times.

A. NACC Design

Before introducing the NACC design, we first give the following result for the multi-user Gaussian broadcast channel (BC) in Proposition 1.

Proposition 1: The capacity region of a t-user Gaussian BC, where each user $i \in [t]$ has a SNR of SNR_i and requests message W_i, while having access to side information $\overline{W}_i = \{W_j\}_{j \neq i, j \in [t]}$, is given by the set $\mathcal{C} = \{(R_1, R_2, \ldots, R_t) : \ 0 \leq R_i \leq \log_2(1 + \mathrm{SNR}_i),\ i \in [t]\}$.

Proof: We refer to [20, Thm. 6] for the proof. ■

Let Λ denote the number of distinct cache states (they will be explained in the cache placement design) imposed by the file-size constraint, where $\Lambda \leq K$, and assume for simplicity that K is an integer multiple[3] of Λ. In this setting, the same cache content is stored across $B \triangleq K/\Lambda$ users per cache state. In the following, we will elaborate on the NACC during the cache placement and content delivery phases.

1) Cache Placement: The cache placement process follows the standard approach of the MN scheme [1]. Specifically, each file W_n in the library $\mathcal{F}$ is divided into $\binom{\Lambda}{\Lambda\gamma}$ non-overlapping and equal-sized subfiles, $W_n \rightarrow \left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\right\}$, where $\mathcal{T}$ represents subsets of cache states. Then, Λ distinct cache states are created, with the g-th cache state, $\mathcal{Z}_g$, defined as $\mathcal{Z}_g = \left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, g \in \mathcal{T}, \forall n \in [N]\right\}$. The users are evenly distributed into Λ user groups, and all $B = \frac{K}{\Lambda}$ users in each user group $g \in [\Lambda]$ have the same cache state $\mathcal{Z}_g$. The b-th user in the g-th group is denoted by $\mathrm{U}_{g,b}$, where $g \in [\Lambda]$ and $b \in [B]$, and let $d_{g,b} \in [N]$ denote the file index requested by user $\mathrm{U}_{g,b}$.

2) Content Delivery: There are $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages. Each such stage involves serving a unique set $\mathcal{G}$ consisting of $\Lambda\gamma+1$ distinct cache states (or the collection of the indexes of user groups). In total, there are $B(\Lambda\gamma+1)$ users in the set $\mathcal{G}$ who are served during this transmission stage, but whether the users within a group are served simultaneously depends on the specific multiple access method. In this scheme, NOMA allows multiple users within the same group

[1] *Notations:* For a positive integer N, we use the notation $[N] \triangleq \{1, 2, \ldots, n\}$. For two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \setminus \mathcal{B}$ represents the set difference, i.e., the elements in $\mathcal{A}$ excluding those in $\mathcal{B}$. $|\cdot|$ denotes the cardinality of a set or the magnitude of a complex number. $X \sim \mathcal{Y}$ indicates that the random variable X follows the distribution $\mathcal{Y}$. Specifically, $\mathcal{CN}$ represents the complex Gaussian distribution. For a matrix $\mathbf{A}$, we use $\mathbf{A}^T$, $\mathbf{A}^*$, and $\mathbf{A}^H$ to denote its transpose, element-wise conjugate, and conjugate transpose, respectively. $\mathbf{I}_L$ denotes the $L \times L$ identity matrix, and $\mathbf{0}_L$ represents a $L \times 1$ zero vector.

[2] In practical applications like Netflix, each popular movie is divided into multiple equal-sized files, each typically several tens of megabytes. For example, the top 100 most popular movies can be split into such equally sized files, forming the library $\mathcal{F}$. We refer to [19] for more details.

[3] We note that the core concepts of the ACC scheme can be readily extended to various existing coded caching schemes to improve their delivery performance over realistic wireless channels, including decentralized ones where K is typically not an integer multiple of Λ (cf. [7]).

to be served simultaneously, unlike the original ACC where users in a group are served sequentially using TDM.

In a given transmission stage for serving the users in set $\mathcal{G}$, the transmitter sends a superposition of B distinct signal symbols, with each symbol carrying information for $\Lambda\gamma + 1$ users. The users associated with the same signal symbol belong to different user groups, meaning that each user has a distinct cache state. Specifically, the b-th symbol is generated using a multi-rate modulation on the requested subfiles $\big\{W_{d_{g,b}}^{\mathcal{G}\setminus\{g\}} : g \in \mathcal{G}\big\}$, where the adopted multi-rate scheme achieves the channel capacity region described in Proposition 1. As arranged in the cache placement phase and stated in Proposition 1, due to $g \in \mathcal{G}\setminus\{g'\}$ for $g' \neq g$, user $\mathrm{U}_{g,b}$ has cached the subfiles

$$\big\{W_n^{\mathcal{G}\setminus\{g'\}} :\ g' \in \mathcal{G},\ g' \neq g,\ n \in [N]\big\}, \tag{1}$$

which allows it to decode the intended subfile $W_{d_{g,b}}^{\mathcal{G}\setminus\{g\}}$ after successfully receiving the b-th signal symbol. By transmitting B such signal symbols, all $B(\Lambda\gamma+1)$ users in the set $\mathcal{G}$ are served. After $\binom{\Lambda}{\Lambda\gamma}$ transmission stages, all users will have obtained their requested files *in full*.

The users in each user group are ordered by their channel strengths, *from the weakest to the strongest*. In SIC, each user first decodes the signals intended for users with weaker channels, removes these from the received signal, and then decodes its own signal. NACC allows us to improve spectral efficiency by enabling simultaneous transmission within a user group, ensuring that the system effectively uses the available bandwidth and mitigates the limitations posed by variations in user channel conditions. This contrasts with the original (TDM-based) ACC which often leads to underutilization of resources, especially in high-SNR scenarios.

Remark 1: The total CSI overheads in the NACC and in the original ACC should be approximately the same; both them use CSI for rate adaptation to $|\mathcal{G}|B$ users. The only additional complexity brought about by NOMA is that each user has to perform SIC to decode its own message.

B. NACC Signal Processing

For a specific user-group set $\mathcal{G}$ with $|\mathcal{G}| = \Lambda\gamma+1$ selected user groups for service during the ACC delivery phase (cf. Section II-A), the transmitted signal is designed as

$$s = \sum\nolimits_{b=1}^{B} \sqrt{\rho\alpha_b}\, s_b, \tag{2}$$

where ρ is the transmit SNR, and s_b is the signal symbol that carries the messages intended by $|\mathcal{G}|$ users, each with the b-th weakest channel link quality in each user-group of $\mathcal{G}$, and where $\alpha_b \in [0,1]$ is the power-splitting factor for the symbol s_b which satisfies that $\sum_{b=1}^{B}\alpha_b = 1$. The received signal at user $\mathrm{U}_{g,b}$ for any $g \in \mathcal{G}$ and $b \in [B]$ is of the form

$$y_{g,b} = h_{g,b} \sum\nolimits_{b=1}^{B} \sqrt{\rho r_{g,b}^{-\eta}\alpha_b}\, s_b + z_{g,b}, \tag{3}$$

where $h_{g,b} \in \mathbb{C}$ represents the channel gain from the transmitter to user $\mathrm{U}_{g,b}$, $r_{g,b}$ denotes the distance between the transmitter and user $\mathrm{U}_{g,b}$, $\eta > 0$ is the pathloss exponent, and $z_{g,b} \sim \mathcal{CN}(0,1)$ represents the additive white Gaussian noise (AWGN).

Based on the design in Section II-A, *user* $\mathrm{U}_{g,b}$ *has cached the content requested by users from any other user-group* g' *with* $g' \in \mathcal{G}\setminus\{g\}$. We have that user $\mathrm{U}_{g,b}$ can decode its intended signal s_b at its single-link capacity and retrieve the desired message by leveraging its cached content. This is facilitated through a multi-rate transmission scheme that ensures the channel capacity region, as specified in Proposition 1, is achieved. Furthermore, because user $\mathrm{U}_{g,b}$ has a stronger link than user $\mathrm{U}_{g,b'}$ for any $b' < b$, it can also successfully decode signal $s_{b'}$ by combining it with its cached content, provided that $s_{b'}$ is transmitted to user $\mathrm{U}_{g,b'}$ at its single-link capacity. Note that user $\mathrm{U}_{g',b}$, for any $g' \in \mathcal{G}\setminus\{g\}$, can also decode $s_{b'}$ at its single-link capacity, thanks to the multi-rate transmission applied to $s_{b'}$. Under the SIC framework, user $\mathrm{U}_{g,b}$ first decodes the signals intended for weaker users, removing the corresponding interference. For signals intended for stronger users, user $\mathrm{U}_{g,b}$ treats them as noise. As a result, under basic Gaussian signalling assumptions, the achievable transmission rate for $\mathrm{U}_{g,b}$ is [13]

$$R_{g,b} \overset{\rho\to\infty}{=} \ln\left(1 + \frac{\alpha_b}{\sum_{i=b+1}^{B}\alpha_i}\right) + O\Big(\frac{1}{\rho}\Big) \tag{4}$$

where $H_{g,b} \triangleq r_{g,b}^{-\eta}|h_{g,b}|^2$ accounts for channel fading and pathloss, and $O(\cdot)$ is known as big-O notation. Specifically, $\lim_{\rho\to\infty} O(\rho^{-1}) = 0$. As $\mathrm{U}_{g,B}$ has the strongest link quality within the user-group g and can therefore decode the signals of all other users, the achievable rate for $\mathrm{U}_{g,B}$ is

$$R_{g,B} = \ln\left(1 + \rho\alpha_B H_{g,B}\right) = \ln\left(\rho\alpha_B H_{g,B}\right) + O\Big(\frac{1}{\rho}\Big). \tag{5}$$

The achievable (sum) rate $R_g = \sum_{b=1}^{B} R_{g,b}$ for the user-group $g \in \mathcal{G}$ takes the form

$$R_g = \ln(\rho\alpha_B H_{g,B}) + \sum_{b=1}^{B-1} \ln\left(1 + \frac{\alpha_b}{\sum_{i=b+1}^{B}\alpha_i}\right) + O\Big(\frac{1}{\rho}\Big).$$

After some simple mathematical manipulations, we can further simplify R_g as $R_g = \ln\left(\rho H_{g,B}\right) + O(\rho^{-1})$. This implies that the achievable rate for user group g approaches the rate achieved by always serving the strongest user in the high-SNR regime, thereby providing spatial diversity among the users within the group [13].

As the transmission stage for $\mathcal{G}$ ends when all users from the worst user-group in $\mathcal{G}$ decode their subfiles (cf. Section II-A), the achievable rate of the NACC scheme is

$$R_{\mathrm{NACC}} = |\mathcal{G}| \min\nolimits_{g\in\mathcal{G}} \left\{R_g\right\} \tag{6}$$

where the factor $|\mathcal{G}|$ is due to the fact that the transmitter serves $|\mathcal{G}|$ user groups at a time. Therefore, the average rate $\bar{R}_{\mathrm{ACC}}^{\mathrm{NOMA}}$ of NACC is of the form

$$\bar{R}_{\mathrm{NACC}} = |\mathcal{G}|\ln(\rho) + |\mathcal{G}|\mathbb{E}\Big\{ \ln\Big(\min_{g\in\mathcal{G}}\left\{H_{g,B}\right\}\Big)\Big\} + O\Big(\frac{1}{\rho}\Big). \tag{7}$$

Remark 2: In comparison to the original ACC scheme whose average rate takes the form (cf. [9])

$$\bar{R}_{\mathrm{ACC}} = |\mathcal{G}| \ln(\rho) + |\mathcal{G}| \mathbb{E}\left\{ \min_{g \in \mathcal{G}} \left\{ \frac{1}{B} \sum_{b=1}^{B} \ln(H_{g,b}) \right\} \right\} + O\left(\frac{1}{\rho}\right)$$

and considering $\ln(H_{g,B}) \geq \frac{1}{B}\sum_{b=1}^{B} \ln(H_{g,b})$, we always have that $\bar{R}_{\mathrm{NACC}} \geq \bar{R}_{\mathrm{ACC}}$. Thus, the NACC always outperforms the original ACC in high SNR.

III. Performance Analysis

We consider a single-cell setting in which a single-antenna 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. Considering the propagation models defined by the Third Generation Partnership Project (3GPP) [12], we have the instantaneous SNR at user $\mathrm{U}_{g,b}$ as

$$\mathrm{SNR}_{g,b} = \frac{P_t}{N_0 B_w N_f \beta_f L_c} |h_{g,b}|^2 r_{g,b}^{-\eta}, \tag{8}$$

where P_t, N_0, and B_w are the transmit power, the noise density, and the bandwidth, respectively. In the above, $h_{g,b}$ corresponds to the fast-fading channel coefficient, drawn from a zero-mean unit-variance complex Gaussian distribution. In (8), 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 $\mathrm{SNR}_{g,b}$ other than the distance and the effect of fast fading, which can be regarded as the transmit SNR in Section II-B. As the users are uniformly distributed within a ring, the probability density function (PDF) of $r_{g,b}$ is (cf. [21])

$$f_{r_{g,b}}(r) = \frac{2r}{D_2^2 - D_1^2}, \quad D_1 \leq r \leq D_2. \tag{9}$$

Similar to our previous works [7]–[10], we define the effective gain to reflect the real performance boost in *finite SNRs* brought about by the considered coded caching scheme.

Definition 1 (Effective coded caching gain): The effective coded caching gain of a particular coded caching scheme is the ratio of the average rate achieved by the said scheme over the average rate attained by simple uncoded TDM.[4]

We consider the affine approximation (with respect to $\ln x$) $\ln(1+x) \approx \ln x$ on the capacity, due to the fact that $\ln(1+x) = \ln x + O(x^{-1})$. We first derive the affine-approximated average rate of TDM. Recall that for any $g \in \mathcal{G}$ and $b \in [B]$, the average rate of uncoded TDM takes the form

$$\bar{R}_{\mathrm{TDM}} = \mathbb{E}\left\{ \ln\left(1 + \rho r_{g,b}^{-\eta} |h_{g,b}|^2\right) \right\}. \tag{10}$$

Lemma 1: The affine-approximated rate for TDM is

$$\tilde{R}_{\mathrm{TDM}} = \ln \rho - \xi + \frac{\eta}{2} - \eta \frac{D_2^2 \ln D_2 - D_1^2 \ln D_1}{D_2^2 - D_1^2} \tag{11}$$

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

Proof: Refer to the proof of [9, Lem. 1]. ■

We derive an analytical expression for $\tilde{R}_{\mathrm{NACC}}$ in Lemma 2, where Θ is defined as

$$\Theta \triangleq \int_0^\infty 1 - \left[1 - \left(1 - \frac{\Gamma\left(\frac{2}{\eta}, D_1^\eta e^{-x}\right) - \Gamma\left(\frac{2}{\eta}, D_2^\eta e^{-x}\right)}{\frac{1}{2}(D_2^2 - D_1^2)\eta \exp(-2x/\eta)}\right)^B\right]^{|\mathcal{G}|} \mathrm{d}x, \tag{12}$$

where $\Gamma(\cdot, \cdot)$ denotes the upper incomplete Gamma function [22, Eq. (8.350.2)].

Lemma 2: The affine-approximated average rate of NACC takes the form

$$\tilde{R}_{\mathrm{NACC}} = |\mathcal{G}| \ln(\rho) - |\mathcal{G}|\Theta. \tag{13}$$

The effective coded caching gain of NACC over uncoded TDM is approximated as

$$\tilde{G}_{\mathrm{NACC}} = \frac{\tilde{R}_{\mathrm{NACC}}}{\tilde{R}_{\mathrm{TDM}}} = |\mathcal{G}| \frac{\ln(\rho) - \Theta}{\tilde{R}_{\mathrm{TDM}}}. \tag{14}$$

Proof: The cumulative distribution function (CDF) of $H_{g,B} \triangleq \max_{b \in [B]}\{r_{g,b}^{-\eta} |h_{g,b}|^2\}$ can be derived as

$$\begin{aligned} F_{H_{g,B}}(x) &= \prod_{b=1}^{B} \mathbb{E}_{r_{g,b}} \left\{ F_{|h_{g,b}|^2}\left(x r_{g,b}^\eta\right) \right\} \\ &= \prod_{b=1}^{B} \left(1 - \frac{2}{D_2^2 - D_1^2} \int_{D_1}^{D_2} r \exp(-x r^\eta) \mathrm{d}r\right) \\ &= \left(1 - \frac{\Gamma\left(\frac{2}{\eta}, D_1^\eta x\right) - \Gamma\left(\frac{2}{\eta}, D_2^\eta x\right)}{\frac{1}{2}(D_2^2 - D_1^2)\eta x^{2/\eta}}\right)^B. \end{aligned} \tag{15}$$

Let $H_{\mathcal{G}} \triangleq -\ln\left(\min_{g \in \mathcal{G}}\{H_{g,B}\}\right)$. The CDF of $H_{\mathcal{G}}$ is derived in (16), shown at the top of the next page, where (a) follows from using the CDF of $H_{g,B}$. Following the same reason as [9, Eq. (18)], we can regard $H_{\mathcal{G}}$ as a non-negative random variable, and its expectation can be derived by $\Theta \triangleq \mathbb{E}\{H_{\mathcal{G}}\} = \int_0^\infty \left(1 - F_{H_{\mathcal{G}}}(x)\right)\mathrm{d}x$, which yields (12). We can then easily derive (13) by considering Θ in (7). The derivation of the effective gain in (14) is straightforward by considering (11) and (13) in Definition 1. ■

IV. Numerical Results

We validate our analytical results using Monte Carlo simulations for an urban Micro-cell environment based on 3GPP's 5G standards. In line with 5G specifications [12], we use the following parameters: carrier frequency $f_{\mathrm{GHz}} = 3.5$ GHz, bandwidth $B_w = 20$ MHz, noise spectral density

[4] Uncoded TDM refers to a scenario where users cache content using an uncoded caching strategy, and the transmitter sequentially sends the uncached portions of each requested file over time. In uncoded caching, each user can, for example, store the first $F\gamma$ bits of every file in the library.

$$F_{H_{\mathcal{G}}}(x) = \mathbb{P}\left(\min_{g\in\mathcal{G}}\{H_{g,B}\} \geq \exp(-x)\right) \overset{(a)}{=} \left[1 - \left(1 - \frac{\Gamma\left(\frac{2}{\eta}, D_1^{\eta}\exp(-x)\right) - \Gamma\left(\frac{2}{\eta}, D_2^{\eta}\exp(-x)\right)}{\frac{1}{2}(D_2^2 - D_1^2)\eta\exp(-2x/\eta)}\right)^B\right]^{|\mathcal{G}|} \tag{16}$$

$N_0 = -174$ dBm/Hz, and $N_f = L_c = 10$ dB. According to 3GPP [12], $\beta_f\,(\mathrm{dB}) = 32.4 + 20\log_{10} f_{\mathrm{GHz}}$ with a path loss exponent $\eta = 2.1$. For user placement, we follow the guidelines in [11], considering a ring-shaped area around the transmitter with distances between $D_1 = 10$ and $D_2 = 100$ meters. We note that the typical Micro-cell transmit power is around $P_t = 33$ dBm [11], [12]. To capture a broader operational range, we extend the power values in Figs. 1–2. In the ACC scheme, the number of users caching the same content (i.e., B) can be as large as 16 in an urban Micro-cell environment (cf. [9]).

A. Multi-antenna Transmitter Serving a Benchmark

We consider an L-antenna transmitter to *simultaneously* serve K' single-antenna *cache-aided* users, uniformly distributed within a Micro-cell, under two different scenarios: uncoded unicasting and coded multicasting. In the uncoded unicasting scenario, users cache content based on an uncoded caching strategy (cf. Footnote 4). As a result, the transmitter must send the requested files separately to each user. Leveraging multiple antennas, the transmitter can send K' distinct signal symbols to the users in parallel, with each symbol corresponding to the content of a different file and being mapped to a specific user. To manage inter-user interference, an appropriate precoding scheme is employed. For coded multicasting, we adopt the standard XOR-based coded caching approach (cf. [2]) to deliver content efficiently.[5] These two scenarios serve as performance *benchmarks*.

Let $\mathbf{H} \in \mathbb{C}^{L\times K'}$ represent the instantaneous channel matrix from the L-antenna transmitter to the K' single-antenna users, where the k-th column of $\mathbf{H}$, denoted by $\mathbf{h}_k$, is the channel vector corresponding to the k-th user. Specifically, $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}_L, r_k^{-\eta}\mathbf{I}_L)$, where r_k, following the same distribution as $r_{g,b}$, is the distance from the transmitter to the k-th user.

In uncoded unicasting with equal power allocation for the users, the average (sum) rate of linear precoding is

$$\bar{R}_{\mathrm{UC}} = \sum_{k=1}^{K'} \mathbb{E}\left\{\ln\left(1 + \frac{\frac{\rho}{K'}|\mathbf{h}_k^T\mathbf{w}_k|^2}{1 + \frac{\rho}{K'}\sum_{k'\neq k}^{K'}|\mathbf{h}_k^T\mathbf{w}_{k'}|^2}\right)\right\}, \tag{17}$$

where $\mathbf{w}_k \in \mathbb{C}^{L\times 1}$ with unit-norm denotes the k-th column of the precoder matrix $\mathbf{W} \in \mathbb{C}^{L\times K'}$. Here, we consider the MMSE precoder where the precoding matrix (without power normalization) is $\mathbf{W}' = \mathbf{H}^H\left(\mathbf{H}^T\mathbf{H}^H + \frac{L}{\rho}\mathbf{I}_{K'}\right)^{-1}$, which will form the actual precoding matrix $\mathbf{W}$ by normalizing each column in $\mathbf{W}'$ into unit-norm [24].

In coded multicasting, the multicasting rate is always limited by the user with the worst channel link quality to guarantee successful decoding at the K' users served in a time. The averaged (sum) rate over the served K' users is

$$\bar{R}_{\mathrm{MC}} = \mathbb{E}\left\{K' \min_{k\in[K']}\left\{\ln\left(1 + \rho|\mathbf{h}_k^T\mathbf{f}_{\mathrm{MC}}|^2\right)\right\}\right\}. \tag{18}$$

As ACC only changes the content delivery way, K' should be equal to $|\mathcal{G}|$ under the same file subpacktization level. In (18), $\mathbf{f}_{\mathrm{MC}} \in \mathbb{C}^{L\times 1}$ denotes the applied beamformer. In this paper, we consider three different beamforming designs, as follows,

$$\mathbf{f}_{\mathrm{MC}}^{\mathrm{FD}} = \mathop{\arg\max}_{\mathbf{f}\in\mathbb{C}^{L\times 1}} \min_{k\in[K']}\left\{|\mathbf{h}_k^T\mathbf{f}|^2\right\}, \text{ s. t. } ||\mathbf{f}||^2 = 1, \tag{19}$$

$$\mathbf{f}_{\mathrm{MC}}^{\mathrm{AD}} = \mathop{\arg\max}_{\mathbf{f}\in\mathbb{C}^{L\times 1}} \min_{k\in[K']}\left\{|\mathbf{h}_k^T\mathbf{f}|^2\right\}, \text{ s. t. } |\mathbf{f}(\ell)|^2 = \frac{1}{L}, \forall\ell\in[L], \tag{20}$$

$$\mathbf{f}_{\mathrm{MC}}^{\mathrm{MRT}} = \frac{1}{||\sum_{k\in[K']}\mathbf{h}_k^*||}\sum\nolimits_{k\in[K']}\mathbf{h}_k^*, \tag{21}$$

where $\mathbf{f}_{\mathrm{MC}}^{\mathrm{FD}}$ and $\mathbf{f}_{\mathrm{MC}}^{\mathrm{AD}}$ are the FD and analog (AD) beamformers optimized for the max-min fairness (MMF) respectively. In (20), $\mathbf{f}(\ell)$ denotes the ℓ-th element of $\mathbf{f}$. We note that both the optimized designs for $\mathbf{f}_{\mathrm{MC}}^{\mathrm{FD}}$ and $\mathbf{f}_{\mathrm{MC}}^{\mathrm{AD}}$ are non-convex NP-hard optimization problems (cf. [25], [26]), while the maximum ratio transmission (MRT) beamformer $\mathbf{f}_{\mathrm{MC}}^{\mathrm{MRT}}$ is widely considered as a simple but well-performed beamforming scheme (cf. [27], [28]).

Similar to reflecting the boost performance of coded caching in finite SNRs, we define the *effective gain* as the ratio of the average rate achieved by the said multi-antenna unicasting (or multicasting) scheme over the average rate attained by uncoded TDM (single-antenna). The effective gains in uncoded unicasting and coded multicasting are respectively

$$G_{\mathrm{UC}} \triangleq \frac{\bar{R}_{\mathrm{UC}}}{\bar{R}_{\mathrm{TDM}}}, \quad G_{\mathrm{MC}} \triangleq \frac{\bar{R}_{\mathrm{MC}}}{\bar{R}_{\mathrm{TDM}}}. \tag{22}$$

where $\bar{R}_{\mathrm{TDM}}$ is given by (10).

B. Numerical Comparisons

In Fig. 1, we plot the average rate of NACC, as well as the average rates of the original (TDM-based) ACC, uncoded TDM (single-antenna) and MMSE precoding with $L = 8$ transmit antennas in uncoded unicasting. We also plot the results derived by using Lemma 2 in dashed lines. Specifically, the number of users served in a time under MMSE precoding

[5] Although ACC can be extended to work with multi-antenna transmitters, this paper focuses on the single-antenna case, with multi-antenna ACC design left for future work. We refer to [23] for additional details.

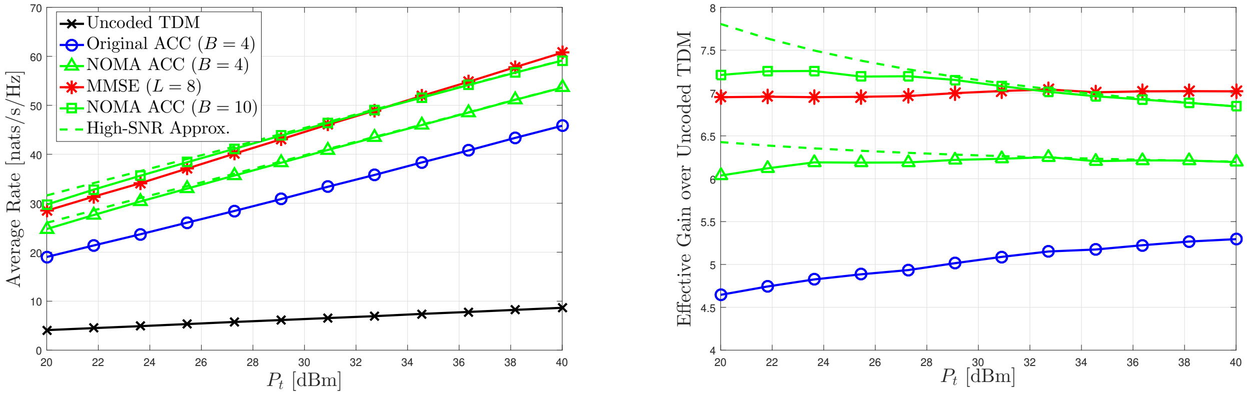

Fig. 1: Performance comparison between single-antenna ACC and multi-antenna uncoded unicasting for $|\mathcal{G}| = 6$ and $L = 8$, where the multiplexing gain (i.e., number of users served in a time) of MMSE precoding is optimized.

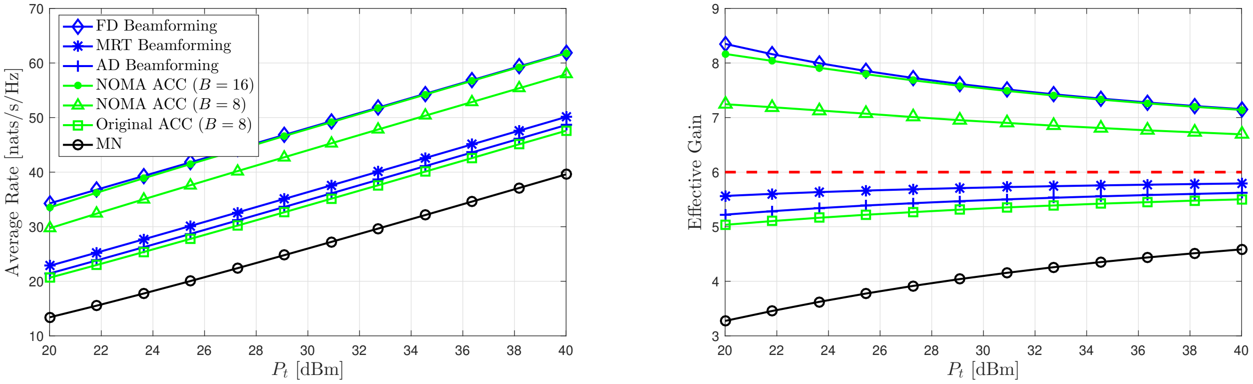

Fig. 2: Performance comparison between single-antenna ACC and multi-antenna coded multicasting for $|\mathcal{G}| = K' = 6$ and $L = 32$. where the red dashed line in the right figure represents the multiplexing gain in the high-SNR limit.

is optimized by maximizing the average rate in (17) w.r.t. K'. Since the Micro-cell operates in the high-SNR regime, we do not optimize the power-splitting factors in NOMA. Following the approach in [13], we set $\alpha_b = \frac{B-b+1}{\bar{\alpha}}$, where $\bar{\alpha}$ is chosen to satisfy $\sum_{b=1}^{B} \alpha_b = 1$. The results in Fig. 1 show that the NACC scheme achieves significantly higher spectral efficiency than the original ACC and uncoded TDM schemes. Moreover, the performance of the NACC, even with a single transmit antenna, approaches the spectral efficiency of the MMSE precoding with 8 transmit antennas in multi-user unicasting. This highlights the potential of the NACC to deliver competitive performance with far fewer antennas, showing its efficiency in modern wireless networks.

In the following, we perform some numerical results to compare the delivery performance between single-antenna ACC separately using TDM and NOMA schemes and multi-antenna coded multicasting in an urban Micro-cell in Fig. 2. We note that the MN scheme represents a typical XOR-based coded caching approach. As highlighted in [7]–[10], ACC provides a highly efficient transmission method for the content delivery process in coded caching over realistic wireless channels, clearly demonstrating that XOR-based transmission is *no longer suitable* for such scenarios. To further validate this fact, we also use the (single-antenna) MN scheme as a benchmark. In Fig. 2, the plotted lines for FD and AD beamformers are generated via numerically solving the corresponding MMF optimization problems (cf. (19) and (20)) in each channel and location realization. We note that there are several efficient methods tailored to numerically solving the aforementioned MMF problem. For simplification, we use the built-in function “fminimax” in MATLAB to numerically solve the MMF problems. For clarity, we omit the analytical results from Lemma 2, as they closely match the simulations.

In Fig. 2, we can see that the fully optimized FD beamformer always has the best delivery performance, followed by the figures for NACC with $B = 16$ and $B = 8$, the MRT beamformer and the fully optimized AD beamformer, while the effective gain *over uncoded TDM* of the (single-antenna) MN scheme is the lowest. This advantage of the fully optimized FD beamformer is partly due to its ability to employ fully connected radio frequency (RF) chains, albeit at a cost (not recorded here) of a much larger power consumption and a much larger computational complexity from having to continuously solve an NP-hard non-convex optimization problem. It is worth noting that the NACC delivery performance almost converges to the delivery performance of the fully optimized multi-antenna FD beamformer over the entire SNR regime, especially when we increase the number B of users per cache state (from $B = 8$ to $B = 16$). Despite using TDM in the original ACC scheme, it is as powerful as the fully optimized AD beamformer which, similar to the FD case, requires continuously solving an NP-hard non-convex MMF problem.[6]

V. Conclusions and Discussions

We proposed NOMA-aided Aggregated Coded Caching (NACC), a new class of coded caching schemes that mitigates both uneven channel bottlenecks and shared-cache limitations. NACC efficiently serves users with unequal channel strengths and, for the first time, uses non-orthogonal multiple access (NOMA) to serve multiple users sharing the same cache state. NACC effectively integrates two fundamentally different technologies, coded caching and NOMA, providing an efficient transmission solution for coded caching in realistic wireless channels. Our analysis shows that NACC can achieve, with a single transmit antenna, spectral efficiency comparable to systems with many more antennas. This highlights its potential to reduce resource usage, such as antennas and power, in wireless networks.

We recognize the practical limitations regarding the number of SIC layers used in NOMA (typically 1–3 layers). To address this, we propose initially applying NOMA to pairs of users sharing the same cache state. Once these two users decode their intended subfiles, they can be seamlessly replaced by another pair from the same user group without disrupting the multi-rate transmission, following the TDM approach used in the original ACC scheme. This hybrid strategy combines the benefits of both NOMA and ACC, optimizing the use of resources. Furthermore, to compensate for the reduced spatial diversity when B is small, a few transmit antennas can be employed. Extending this design to multi-antenna ACC will be explored in future work.

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] S. P. Shariatpanahi *et al.*, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[3] 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.

[4] F. Engelmann and P. Elia, "A content-delivery protocol, exploiting the privacy benefits of coded caching," in *Proc. 15th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, 2017.

[5] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7635–7651, 2020.

[6] H. B. Mahmoodi, M. Salehi, and A. Tölli, "Low-complexity multi-antenna coded caching using location-aware placement delivery arrays," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 12 687–12 702, 2024.

[7] H. Zhao and A. Bazco-Nogueras and P. Elia, "Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5450–5466, Jul. 2022.

[8] 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.

[9] H. Zhao, A. Bazco-Nogueras, and P. Elia, "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.

[10] 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.

[11] "5G Implementation Guidelines," GSMA, Tech. Rep. version 2.0, Jul. 2019.

[12] "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.

[13] Z. Ding *et al.*, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[14] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[15] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[16] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4854–4876, 2018.

[17] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020.

[18] 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.

[19] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Vector coded caching multiplicatively increases the throughput of realistic downlink systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2683–2698, 2023.

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

[21] 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.

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

[23] H. Zhao, "High performance cache-aided downlink systems: Novel algorithms and analysis," Ph.D. dissertation, Sorbonne University, 2022.

[24] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[25] Z. Wang, Q. Liu, M. Li, and W. Kellerer, "Energy efficient analog beamformer design for mmwave multicast transmission," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 2, p. 552–564, Jun. 2019.

[26] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[27] M. Sadeghi *et al.*, "Reducing the computational complexity of multicasting in large-scale antenna systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2963–2975, May 2017.

[28] M. He, H. Zhao, X. Miao, S. Wang, and G. Pan, "Secure rate-splitting multiple access transmissions in LMS systems," *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 19–23, Jan. 2024.

[6] We also conducted numerical simulations in a mmWave channel environment (cf. [23]), where NACC demonstrated performance advantages over both the original ACC scheme and the multi-antenna transmitter.