Fundamental Limits of Wireless Caching Under Mixed Cacheable and Uncacheable Traffic

Hamdi Joudeh*, Eleftherios Lampiris*, Petros Elia† and Giuseppe Caire*
*Communications and Information Theory Group, Technische Universität Berlin, Germany
†Communication Systems Department, EURECOM, Sophia Antipolis, France
Email: *{h.joudeh, lampiris, caire}@tu-berlin.de, †elia@eurecom.fr

***Abstract*—We consider cache-aided wireless communication scenarios where each user requests both a file from an a-priori generated cacheable library (referred to as 'content'), and an uncacheable 'non-content' message generated at the start of the communication session. This scenario is easily found in real-world wireless networks, where the two types of traffic coexist and share limited radio resources. We focus our investigation on single-transmitter wireless networks with cache-aided receivers, where the wireless channel is modelled by a degraded Gaussian broadcast channel (GBC). For this setting, we study the (normalized) delay-rate trade-off, which characterizes the content delivery time and non-content communication rates that can be achieved simultaneously. We propose a scheme based on the separation principle, which isolates the coded caching problem from the physical layer transmission problem, and prove its information-theoretic order optimality up to a multiplicative factor of 2.01. A key insight emerging from our scheme is that substantial amounts of non-content traffic can be communicated while maintaining the minimum content delivery time, achieved in the absence of non-content messages; compliments of 'topological holes' arising from asymmetries in wireless channel gains.**

I. Introduction

Cache-aided architectures have emerged as an essential next step in the evolution of communication networks. The recent few years saw a heightened interest in studying the fundamental limits of cache-aided networks, initiated by Maddah-Ali and Niesen in [1]. For an idealized symmetric broadcast channel (BC), in which cache-equipped users (receivers) are connected to a server (transmitter) through a noiseless shared link, Maddah-Ali and Niesen showed that a novel *coded* caching and multicasting scheme can serve an arbitrarily large number of users with finite resources (e.g. time and bandwidth). The achievable performance in [1], characterized in terms of the shared link *normalized load*, was shown to be order optimal in the information-theoretic sense, within a multiplicative factor, which was tightened later on in [2].

Wireless caching: The bulk of data traffic nowadays is generated by wireless and mobile devices, a trend foreseen to continue and grow in forthcoming years. This has driven a surge of interest in extending the information-theoretic coded caching paradigm in [1] to wireless networks, including noisy and asymmetric broadcast channels (BCs) [3]–[7], multi-antenna BCs [8]–[12], device-to-device networks [13] and interference networks [14]–[16], amongst others.

All above works consider scenarios in which the network carries a single type of traffic that takes the form of content drawn from an a-priori generated library. This approach has been very successful in establishing the fundamental limits of cache-aided networks, and in gaining insights into the design of optimal and near optimal schemes. Nevertheless, wireless data traffic does not comprise of only *cacheable* content. Non-content traffic, generated from interactive applications, gaming, voice and video calls, to name a few examples, also constitutes a significant portion of traffic (estimated as 40 percent [17]). Moreover, content popularity profiles in reality are far from static and may change on a daily or even hourly basis [17]. Therefore, newly generated content can be both in very high demand as well as not yet available in caches.

This work: Motivated by the mixed nature of traffic, we initiate the study of cache-aided wireless networks with both *cacheable* and and *uncacheable* types of traffic, represented by pre-generated *content* files and instantaneously generated *non-content* messages, respectively. We focus on cache-aided networks where the wireless channel is modelled by a degraded Gaussian BC (GBC). The current treatment of content and non-content transmissions as two independent problems necessarily leads to orthogonal-access-based schemes, that schedule the different types of traffic on distinct (orthogonal) resource blocks. As we will see, this approach is rendered suboptimal due to the superposition and asymmetric nature of wireless channels, epitomized through the degraded GBC. In this work, we propose to treat these two problems jointly.

We study the trade-off between the content delivery time and non-content communication rates. For tractability, and to gain useful insights, we focus on generalized normalized delivery time (GNDT) and generalized degrees of freedom (GDoF) approximations [11], [18]. We derive an achievable GNDT-GDoF trade-off and prove its information-theoretic order optimality, up to a multiplicative factor of 2.01. The achievability scheme is based on the *separation principle* [16], where the coded caching side of the problem is separated from the physical-layer communication problem; while the converse is based on a non-trivial extension of the argument in [2].

Notation: For positive integers z_1 and z_2, with $z_1 \leq z_2$,

The work is supported by the European Research Council under the ERC grant agreement N. 789190 (project CARENET), and the ERC grant agreement N. 725929 (project DUALITY).

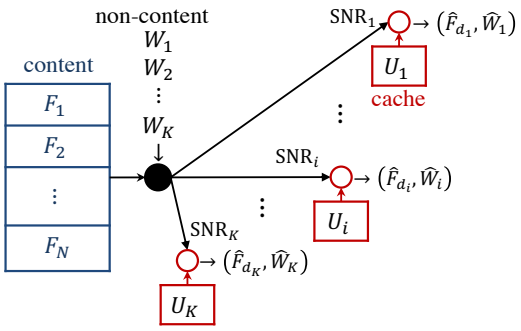

Fig. 1. Cache-aided degraded GBC with content and non-content.

the sets $\{1, 2, \ldots, z_1\}$ and $\{z_1, z_1+1, \ldots, z_2\}$ are denoted by $[z_1]$ and $[z_1 : z_2]$, respectively. $\binom{z_2}{z_1}$ denotes the binomial coefficient. For a real number a, $(a)^+ = \max\{0, a\}$. The tuple $(a_1, \ldots, a_y, b_1, \ldots, b_z)$ is denoted by $(a_i : i \in [y], b_i : i \in [z])$. For a set $\mathcal{A} \subseteq \mathbb{R}^K$, $\mathrm{cl}\{\mathcal{A}\}$ denotes its closure.

II. System Model

In the considered K-user wireless network, the transmitter has access to a content library of N files ($N \geq K$), denoted by $F_1, \ldots, F_N$, each of size B bits. Each user is equipped with a cache memory of size MB bits, where $M \in [0, N]$. The normalized cache size is defined as $\mu \triangleq \frac{M}{N}$. The network operates in two phases: *placement* and *delivery*.

1) Placement phase: During this phase, users have access to the entire library of files to fill the content of their caches. This occurs without knowledge of future file requests.

2) Delivery phase: Each user k requests a *content* file F_{d_k}, where $\mathbf{d} \triangleq (d_1, \ldots, d_K) \in [N]^K$ is the demand tuple. Moreover, the transmitter generates K *non-content* messages $W_1, \ldots, W_K$, intended to users $1, \ldots, K$, respectively. These messages are mutually independent and independent of the content library. During the delivery phase, the transmitter sends a codeword over the physical channel, while each user k receives a corresponding noisy signal and tries to recover (F_{d_k}, W_k) from this signal and the local cache content.

A. Physical channel

The physical channel is a degraded GBC. In the t-th use of the channel, the input-output relationship is described as:

$$Y_k(t) = h_k X(t) + Z_k(t) \tag{1}$$

where $X(t)$ is the input signal; while $Y_k(t), Z_k(t)$ and h_k are the output signal, zero-mean, unit-variance additive white Gaussian noise (AWGN) signal, and the channel coefficient of user k, respectively (all complex). Communication occurs over T channel uses, in which the transmitter is subject to a unit average power constraint given by $\frac{1}{T}\sum_{t=1}^{T} |X(t)|^2 \leq 1$.

For each user k, the signal-to-noise ratio (SNR) is given by $\mathsf{SNR}_k = |h_k|^2$. For GDoF-GNDT purposes, we have $\mathsf{SNR}_k = P^{\alpha_k}$, where the exponent α_k is known as the *channel strength level*, while $P > 1$ is a *nominal power parameter* which approaches infinity to define the GDoF limit—see [11], [18]. We assume, without loss of generality, that strengths are ordered as $0 < \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_K = 1$. The channel strength tuple is given by $\boldsymbol{\alpha} \triangleq (\alpha_1, \ldots, \alpha_K)$.

B. Codes and performance measures

Files $F_1, \ldots, F_N$ are i.i.d. random variables, each uniformly distributed over $[2^{\lfloor B \rfloor}]$. To define asymptotic limits, we scale B as $B = TR_F$. Messages $W_1, \ldots, W_K$ are independent, and each W_k is uniformly distributed over the set $[2^{\lfloor TR_k \rfloor}]$, where $\mathbf{R} \triangleq (R_1, \ldots, R_K)$ are the non-content message rates.

A code $(T, R_F, \mathbf{R}, M)$ consists of the above file and message sets, in addition to appropriately defined caching functions, as well as channel encoding and decoding functions. For any code, the probability of decoding error is defined as

$$P_{e,T} \triangleq \max_{\mathbf{d} \in [N]^K} \max_{k \in [K]} \Pr\left\{(\hat{F}_{d_k}, \hat{W}_k) \neq (F_{d_k}, W_k)\right\} \tag{2}$$

where $(\hat{F}_{d_k}, \hat{W}_k)$ are the estimates of (F_{d_k}, W_k) at receiver k. The error in (2) accounts for the worst-case demand scenario.

It is instructive to work with the reciprocal of the content rate R_F, which enjoys desirable analytical properties, see [14], [16]. To this end, we define the *delivery time* (or *delay*) as

$$\mathcal{T} \triangleq \frac{1}{R_F} = \frac{T}{B} \tag{3}$$

which is the number of physical channel uses required to communicate one bit of content to each user. Given a memory size M, a delay-rate trade-off tuple is denoted by $(\mathcal{T}, \mathbf{R}; M)$, which is achievable if there exists a sequence of $(T, 1/\mathcal{T}, \mathbf{R}, M)$ codes such that $P_{e,T} \to 0$ as $T \to \infty$. To define the GDoF-GNDT limit, the dependency of the rates and delivery time on P is highlighted, i.e. $(\mathcal{T}(P), \mathbf{R}(P); M)$.

We denote a GDoF tuple by $\mathbf{r} \triangleq (r_1, \ldots, r_K)$, where r_k is the GDoF of user k, while the GNDT is denoted by τ. A GNDT-GDoF trade-off $(\tau, \mathbf{r}; M)$ is achievable if there exists an achievable sequence $(\mathcal{T}(P), \mathbf{R}(P); M)$, $\forall P$, such that

$$r_k = \lim_{P \to \infty} \frac{R_k(P)}{\log(P)}, \ \forall k \in [K] \tag{4}$$

$$\tau = \lim_{P \to \infty} \mathcal{T}(P) \log(P). \tag{5}$$

For any $(\mathbf{r}; M)$, the optimal GNDT is defined as

$$\tau^\star(\mathbf{r}; M) \triangleq \inf\left\{\tau : (\tau, \mathbf{r}; M) \text{ is achievable}\right\}. \tag{6}$$

Similarly, for any $(\tau; M)$, the GDoF region is defined as:

$$\mathcal{D}(\tau; M) = \mathrm{cl}\left\{\mathbf{r} : (\tau, \mathbf{r}; M) \text{ is achievable}\right\}. \tag{7}$$

Remark 1. *The characterizations obtained in this work all depend on the normalized memory size μ instead of M. This is reflected in the arguments of the performance measures in the following sections, where we also highlight the dependency on channel strength levels, e.g. $\tau^\star(\mathbf{r}; \mu, \boldsymbol{\alpha})$ and $\mathcal{D}(\tau; \mu, \boldsymbol{\alpha})$.*

III. Main Result and Insights

We start by defining an upper bound for the GNDT.

Definition 1. *For any μ, $\boldsymbol{\alpha}$ and $\mathbf{r}$, where the components of the GDoF tuple $\mathbf{r}$ satisfy $\sum_{i\in[k]} r_i \le \alpha_k$ for all $k \in [K]$, we define $\tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha})$ as*[1]

$$\max_{k\in[K]}\left\{\frac{1}{\left(\alpha_k-\sum_{i\in[k]} r_i\right)}\cdot \mathrm{conv}\left(\frac{\binom{K}{K\mu+1}-\binom{K-k}{K\mu+1}}{\binom{K}{K\mu}}\right)\right\} \quad (8)$$

where $\mathrm{conv}\big(f(K\mu)\big)$ denotes the lower convex envelope of the points $\big\{\big(K\mu, f(K\mu)\big) : K\mu \in [0:K]\big\}$.

We are now ready to state the main theorem of this work.

Theorem 1. *The GNDT-GDoF trade-off described by $\tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha})$ in (8) is achievable and is within a multiplicative factor of 2.01 from the optimal trade-off, that is*

$$\frac{1}{2.01}\cdot\tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha}) \le \tau^{\star}(\mathbf{r};\mu,\boldsymbol{\alpha}) \le \tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha}). \quad (9)$$

The achievability of Theorem 1 is described in Sections IV and V, while the converse is presented in Section VI. The proofs and discussion in this paper focus on the case of integer $K\mu$, due to space limitation. The extension to non-integer $K\mu$ is treated in a longer version of this paper [19].

Next, we draw some insights from the main result.

1) Separation principle: The achievability of $\tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha})$ employs a separation-based strategy, which isolates the coded caching problem from the physical layer transmission problem [16]. Caching and generating coded multicast messages (XORs) are carried out at the bit level in the original Maddah-Ali and Niesen manner [1]. The physical channel sees $\binom{K}{K\mu+1}$ multicast messages (coded content) and K unicast messages (non-content), and communicates them jointly using a scheme based on superposition coding. Different GNDT-GDoF trade-offs are achieved by tuning the underlying physical layer power allocation and GDoF assignment problems.

2) Absence of non-content messages: As a special case of Theorem 1, we recover the result in [7], where it was shown that in the absence of non-content messages, one can achieve

$$\tau^{\mathrm{ub}}(\mathbf{0};\mu,\boldsymbol{\alpha}) = \max_{k\in[K]}\left\{\frac{1}{\alpha_k}\cdot\frac{\binom{K}{K\mu+1}-\binom{K-k}{K\mu+1}}{\binom{K}{K\mu}}\right\}. \quad (10)$$

The order optimality of $\tau^{\mathrm{ub}}(\mathbf{0};\mu,\boldsymbol{\alpha})$ up to a multiplicative factor of 4.02 is also proved in [7], which we tighten in Theorem 1. In addition to strengthening the result of [7], our new achievability proof gives an operational interpretation of $\tau^{\mathrm{ub}}(\mathbf{0};\mu,\boldsymbol{\alpha})$ in terms of separation and the multiple multicast GDoF region of the underlying degraded GBC.

3) Achievable GDoF under minimum GNDT: Theorem 1 suggests that in scenarios with asymmetric channel strengths, the order-optimal GNDT in (10), achieved by eliminating non-content messages, can be maintained while simultaneously achieving non-zero GDoF for (some) non-content messages. To see this, let us define user $k^\star$ (bottleneck user in [7]) as

$$k^\star \triangleq \arg\max_{k\in[K]}\left\{\frac{\binom{K}{K\mu+1}-\binom{K-k}{K\mu+1}}{\alpha_k}\right\}. \quad (11)$$

For example, in the illustrations shown in Fig. 2 (where $K=3$ and $K\mu+1=2$), we have $k^\star=1$ when $\frac{\alpha_1}{2}\le\frac{\alpha_3}{3}$ (subfig. (a)), and $k^\star=2$ when $\frac{\alpha_1}{2}>\frac{\alpha_3}{3}$ (subfig. (b)).

From (8), it follows that achieving $\tau^{\mathrm{ub}}(\mathbf{0};\mu,\boldsymbol{\alpha})$ through the proposed strategy requires setting $r_k=0$ for all $k\in[k^\star]$. However, users in $[k^\star+1:K]$ can achieve non-zero non-content GDoF without affecting the GNDT in (10), by communicating through the '*topological holes*' arising from the asymmetry in channel strength levels. These achievable non-content GDoF tuples are described below.

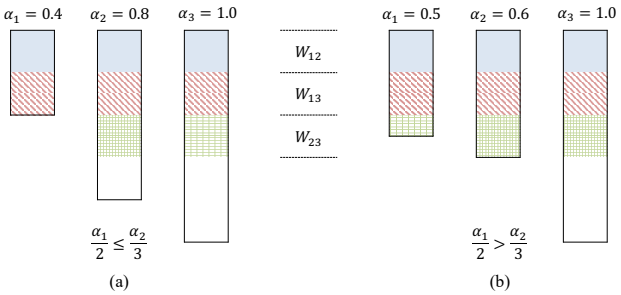

Fig. 2. Received signal power levels in a 3-user degraded GBC with a multiple multicast message set, where each message W_{ij} is intended to users i and j (group size 2). Top levels represent signals transmitted with higher powers, received by all users above their respective noise levels (bottom end of each bar). Bottom levels represent signals transmitted with lower powers, heard by sufficiently strong users and corrupted by noise (hence clipped) at weaker users. Multicast messages (coloured levels) can carry coded content (e.g. $\mu = 1/3$). Uncoloured signal levels are unoccupied, representing *topological holes* for communicating non-content messages (Corollary 1).

Corollary 1. *A trade-off $\tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha}) = \tau^{\mathrm{ub}}(\mathbf{0};\mu,\boldsymbol{\alpha})$ is achievable for all $\mathbf{r}$ that satisfy $r_k=0, \forall k\in[k^\star]$, and*

$$r_{k^\star+1}+\cdots+r_k \le \alpha_k-\alpha_{k^\star}\cdot\left(\frac{\binom{K}{K\mu+1}-\binom{K-k}{K\mu+1}}{\binom{K}{K\mu+1}-\binom{K-k^\star}{K\mu+1}}\right), \quad \forall k\in[k^\star+1:K]. \quad (12)$$

Examples that illustrate Corollary 1 in terms of signal power levels (i.e. exponents of P) are shown in Fig. 2.

4) GDoF region: Theorem 1 yields the following corollary.

Corollary 2. *The GDoF region $\mathcal{D}(\tau;\mu,\boldsymbol{\alpha})$ satisfies:*

$$\mathcal{D}_{\mathrm{in}}(\tau;\mu,\boldsymbol{\alpha}) \subseteq \mathcal{D}(\tau;\mu,\boldsymbol{\alpha}) \subseteq \mathcal{D}_{\mathrm{in}}(2.01\cdot\tau;\mu,\boldsymbol{\alpha}) \quad (13)$$

where $\mathcal{D}_{\mathrm{in}}(\tau;\mu,\boldsymbol{\alpha})$ is the set of all tuples $\mathbf{r}\in\mathbb{R}_+^K$ satisfying

$$\sum_{i\in[k]} r_i + \frac{1}{\tau}\cdot\mathrm{conv}\left(\frac{\binom{K}{K\mu+1}-\binom{K-k}{K\mu+1}}{\binom{K}{K\mu}}\right) \le \alpha_k,\ \forall k\in[K]. \quad (14)$$

As one would hope, the above GNDT-GDoF-based results serve as an initial step towards approximate characterizations of the optimal delay-rate trade-off at finite P (see [19]).

[1] Throughout this work, we use the convention $\binom{n}{k}=0$, for all $n<k$.

IV. Physical Channel

Here we focus on the degraded GBC with no caches and with a unicast message set and a multiple multicast message set. The latter message set is referred to as the σ-multicast message set, where $\sigma \in [2:K]$ is the size of the corresponding multicast groups. This channel model is at the heart of the separation architecture—unicast messages carry non-content traffic, while multicast messages carry coded content traffic. The results in this section will help us establish the GNDT of the scheme we present the in the following section.

A. Unicast and σ-multicast message sets

The unicast message set is given by $\{W_k : k \in [K]\}$, where each message W_k has a GDoF of r_k; while the σ-multicast message set is given by $\{W_{\mathcal{S}} : \mathcal{S} \subseteq [K], |\mathcal{S}| = \sigma\}$, where each message $W_{\mathcal{S}}$ has a GDoF of $r_{\mathcal{S}}$. For any σ and $\boldsymbol{\alpha}$, the GDoF region of this channel is denoted by $\mathcal{D}^{\mathrm{PHY}}(\sigma, \boldsymbol{\alpha})$. We defined $\Sigma \triangleq \{\mathcal{S} \subseteq [K] : |\mathcal{S}| = \sigma\}$ as the set of all σ-multicast groups, where $|\Sigma| = \binom{K}{\sigma}$. Moreover, we introduce a family of subsets of Σ given by $\{\Sigma_i : i \in [K-\sigma+1]\}$, where Σ_i is defined as:

$$\Sigma_i \triangleq \{\mathcal{S} \in \Sigma : \min\{\mathcal{S}\} = i\}. \tag{15}$$

It can be verified that $\{\Sigma_i : i \in [K-\sigma+1]\}$ is a partition of Σ, that is $\cup_{i\in[K-\sigma+1]}\Sigma_i = \Sigma$ and $\Sigma_i \cap \Sigma_j = \emptyset,\ \forall i \neq j$. We are now ready to present a characterization of $\mathcal{D}^{\mathrm{PHY}}(\sigma, \boldsymbol{\alpha})$.

Theorem 2. *The GDoF region $\mathcal{D}^{\mathrm{PHY}}(\sigma, \boldsymbol{\alpha})$ is given by all tuples $(r_k : k \in [K],\ r_{\mathcal{S}} : \mathcal{S} \in \Sigma) \in \mathbb{R}_+^{K+\binom{K}{\sigma}}$ that satisfy*

$$\begin{aligned} \sum_{i\in[k]} r_i + \sum_{\mathcal{S}\in\cup_{i\in[k]}\Sigma_i} r_{\mathcal{S}} \leq \alpha_k,\ \forall k \in [K-\sigma+1] \\ \sum_{i\in[k]} r_i + \sum_{\mathcal{S}\in\Sigma} r_{\mathcal{S}} \leq \alpha_k,\ \forall k \in [K-\sigma+2:K]. \end{aligned} \tag{16}$$

The above GDoF region is achieved using a scheme based on power control with superposition coding and successive decoding. The proof is presented in the longer version [19].

Theorem 2 is intuitively interpreted as follows. User 1 recovers all messages in $\{W_1, W_{\mathcal{S}} : \mathcal{S} \in \Sigma_1\}$, bounding the sum-GDoF of such messages by α_1. Due to the degradedness, user 2 can recover whatever user 1 recovers, and must also decode for messages in $\{W_2, W_{\mathcal{S}} : \mathcal{S} \in \Sigma_2\}$. This bounds the sum-GDoF of $\{W_1, W_2, W_{\mathcal{S}} : \mathcal{S} \in \Sigma_1 \cup \Sigma_2\}$ by α_2. The same argument applies to all users up to user $K-\sigma+1$. Beyond user $K-\sigma+1$, each user k in $[K-\sigma+2:K]$ can recover $\{W_1, \ldots, W_{k-1}, W_{\mathcal{S}} : \mathcal{S} \in \Sigma\}$, and must additionally decode for W_k, yielding the bounds in the second line of (16).

B. Symmetric σ-multicast GDoF

We are interested in a lower dimensional projection of $\mathcal{D}^{\mathrm{PHY}}(\sigma, \boldsymbol{\alpha})$, denoted by $\mathcal{D}^{\mathrm{PHY}}_{\mathrm{sym}}(\sigma, \boldsymbol{\alpha})$, capturing the symmetric σ-multicast GDoF $r_{\mathrm{sym}} \triangleq \min_{\mathcal{S}\in\Sigma} r_{\mathcal{S}}$. From Theorem 2, $\mathcal{D}^{\mathrm{PHY}}_{\mathrm{sym}}(\sigma, \boldsymbol{\alpha})$ is given by $(r_k : i \in [K],\ r_{\mathrm{sym}})$ that satisfy:

$$\begin{aligned} \sum_{i\in[k]} r_i + \left|\bigcup_{i\in[k]} \Sigma_i\right| \cdot r_{\mathrm{sym}} \leq \alpha_k, \forall k \in [K-\sigma+1] \\ \sum_{i\in[k]} r_i + |\Sigma| \cdot r_{\mathrm{sym}} \leq \alpha_k, \forall k \in [K-\sigma+2:K]. \end{aligned} \tag{17}$$

It can be verified that the following identity holds

$$\left|\bigcup_{i\in[k]} \Sigma_i\right| = \sum_{i\in[k]} |\Sigma_i| = \binom{K}{\sigma} - \binom{K-k}{\sigma}, \forall k \in [K-\sigma+1]$$

which leads us to the following corollary.

Corollary 3. *The GDoF region $\mathcal{D}^{\mathrm{PHY}}_{\mathrm{sym}}(\sigma, \boldsymbol{\alpha})$ is described by:*

$$\sum_{i\in[k]} r_i + \left[\binom{K}{\sigma} - \binom{K-k}{\sigma}\right] \cdot r_{\mathrm{sym}} \leq \alpha_k, \forall k \in [K]. \tag{18}$$

From $\mathcal{D}^{\mathrm{PHY}}_{\mathrm{sym}}(\sigma, \boldsymbol{\alpha})$ in the above corollary, it follows that for any feasible unicast GDoF tuple $\mathbf{r} = (r_k : k \in [K])$, the maximum achievable symmetric multicast GDoF is given by

$$r^{\star}_{\mathrm{sym}} = \min_{k\in[K]} \left\{ \frac{\left(\alpha_k - \sum_{i\in[k]} r_i\right)}{\binom{K}{\sigma} - \binom{K-k}{\sigma}} \right\}. \tag{19}$$

V. Achievability

Equipped with the GDoF characterization for the unicast and σ-multicast physical channel, the achievability part of Theorem 1 will follow from a scheme that adheres to the separation principle. Recall that we focus on integer values of $K\mu$, drawn from $[0:K]$.

1) Cache placement: Each file F_n is divided into $\binom{K}{K\mu}$ equal sized sub-files, i.e. $\left\{F_n^{\mathcal{S}'} : \mathcal{S}' \subseteq [K], |\mathcal{S}'| = K\mu\right\}$. Each user k then fills its cache memory as in [1], that is:

$$U_k = \left\{F_n^{\mathcal{S}'} : n \in [N], \mathcal{S}' \subseteq [K], |\mathcal{S}'| = K, k \in \mathcal{S}'\right\}. \tag{20}$$

2) Coded multicast messages: Once the K demands are revealed, the transmitter prepares $\binom{K}{K\mu+1}$ coded multicast messages, each intended to a unique subset of $K\mu+1$ users. We use the physical channel notation and set the multicast group size as $\sigma = K\mu + 1$. For each subset of users $\mathcal{S} \in \Sigma$ of size σ, the coded multicast message generated as

$$W_{\mathcal{S}} = \bigoplus_{k\in\mathcal{S}} F_{d_k}^{\mathcal{S}\setminus\{k\}}. \tag{21}$$

It follows from [1] that, for each user k, the requested file F_{d_k} can be successfully recovered from the cache content U_k and the set of coded multicast messages $\{W_{\mathcal{S}} : \mathcal{S} \in \Sigma, k \in \mathcal{S}\}$.

3) Transmission: The problem now reduces to delivering the set of $\binom{K}{\sigma}$ coded multicast messages, as well as the set of K unicast messages. This is exactly the joint unicast and multicast problem discussed in Section IV. For any achievable tuple $(\mathbf{r}, r_{\mathrm{sym}}) \in \mathcal{D}^{\mathrm{PHY}}_{\mathrm{sym}}(\sigma, \boldsymbol{\alpha})$, each of the non-content unicast messages achieves its corresponding GDoF in $\mathbf{r}$, while the achievable content GNDT is given by $\tau = \frac{1}{r_{\mathrm{sym}} \cdot \binom{K}{\sigma-1}}$. The

normalization factor in τ appears since each coded multicast message $W_{\mathcal{S}}$ has a (normalized) size of $1/\binom{K}{\sigma-1}$. Combining with (19), the GNDT $\tau^{\mathrm{ub}}(\mathbf{r};\mu,\boldsymbol{\alpha})$ is achieved.

VI. Converse

For any distinct demands $\mathbf{d}$, each user k in $[K]$ must recover both W_k and F_{d_k} from Y_k^T and U_k, with a decoding error that vanishes as T grows large. Therefore, Fano's inequality implies $H\big(W_k, F_{d_k}|Y_k^T, U_k\big) \leq 1 + P_{e,T}(TR_k + B) = T\epsilon_T$. We define a side information variable S_k which is independent of W_k. The side information S_k is provided to user k through a genie, and will be specified further on. It follows that

$$\begin{aligned} &TR_k + H\big(F_{d_k}|U_k, S_k\big) = H\big(W_k\big) + H\big(F_{d_k}|U_k, S_k\big) \\ &= H\big(W_k, F_{d_k}|U_k, S_k\big) \\ &\leq I\big(W_k, F_{d_k}; Y_k^T|U_k, S_k\big) + T\epsilon_T. \end{aligned} \tag{22}$$

We ignore $T\epsilon_T$ for brevity. From the above single-user bound, we obtain a bound for any subset of users $[s]$, $s \in [K]$, as

$$\sum_{k\in[s]} H\big(F_{d_k}|U_k, S_k\big) \leq \sum_{k\in[s]} \Big[I\big(W_k, F_{d_k}; Y_k^T|U_k, S_k\big) - TR_k\Big]. \tag{23}$$

Next, we apply a symmetrization procedure over file demands and user orders, required to bound the left-hand-side in (23).

Let $p : [s] \to [s]$ be a permutation over the subset of users $[s]$, and $\mathcal{P}_s$ be the corresponding set of all $s!$ user permutations. Similarly, $q : [N] \to [N]$ is a permutation over $[N]$, and $\mathcal{P}_N$ is the corresponding set of all $N!$ file permutations. For any pair of permutations $(p, q) \in \mathcal{P}_s \times \mathcal{P}_N$, suppose that each user $p(k)$ demands the file $F_{q(k)}$. From (23), and by taking an average over all possible permutations $(p, q) \in \mathcal{P}_s \times \mathcal{P}_N$, we obtain

$$\begin{aligned} &\frac{1}{s!N!}\sum_{(p,q)}\sum_{k\in[s]} H\big(F_{q(k)}|U_{p(k)}, S_{p(k)}\big) \leq \frac{1}{s!N!}\sum_{(p,q)} \\ &\sum_{k\in[s]} \Big[I\big(W_{p(k)}, F_{q(k)}; Y_{p(k)}^T|U_{p(k)}, S_{p(k)}\big) - TR_{p(k)}\Big]. \end{aligned} \tag{24}$$

We set the side information $S_{p(k)}$ for each user $p(k)$ as

$$S_{p(k)} = \big(W_{p(i)}, F_{q(i)}, U_{p(i)} : i \in [k-1]\big) \tag{25}$$

consisting of intended messages, demanded files and cache contents of all users that precede $p(k)$ in the permutation order. The independence between $S_{p(k)}$ and $W_{p(k)}$ is preserved. Next, we separately bound each side of (24).

A. Bounding the right-hand-side of (24)

To this end, we present the following lemma.

Lemma 1. *For any k and j in $[K]$, such that $k \leq j$, we have*

$$I\big(W_k, F_{d_k}; Y_k^T|U_k, S_k\big) \leq I\big(W_k, F_{d_k}; Y_j^T|U_k, S_k\big). \tag{26}$$

The inequality in (26) holds due to degradedness. The proof is omitted for brevity. Using Lemma 1 while focusing on an arbitrary permutation pair $(p, q) \in \mathcal{P}_s \times \mathcal{P}_N$, the corresponding term on the right-hand-side of (24) is bounded as:

$$\begin{aligned} &\sum_{k\in[s]} I\big(W_{p(k)}, F_{q(k)}; Y_{p(k)}^T|U_{p(k)}, S_{p(k)}\big) \\ &\leq \sum_{k\in[s]} I\big(W_{p(k)}, F_{q(k)}; Y_s^T|U_{p(k)}, S_{p(k)}\big) \\ &\leq \sum_{k\in[s]} h\big(Y_s^T|U_{p(k)}, S_{p(k)}\big) - h\big(Y_s^T|U_{p(k+1)}, S_{p(k+1)}\big) \\ &= h\big(Y_s^T|U_{p(1)}\big) - h\big(Y_s^T|U_{p(s)}, S_{p(s)}, W_{p(s)}, F_{q(s)}\big) \\ &= h\big(Y_s^T|U_{p(1)}\big) - h\big(Z_s^T\big) \\ &= I\big(X^T; Y_s^T|U_{p(1)}\big) \\ &\leq T\log(1 + P^{\alpha_s}). \end{aligned} \tag{27}$$

As (27) holds for all permutations $(p, q) \in \mathcal{P}_s \times \mathcal{P}_N$, and since in (24) we have $\sum_{k\in[s]} R_{p(k)} = \sum_{k\in[s]} R_k$ for any such permutation, it follows that each of the inner sums (over k) on the right-hand-side of (24) is bounded by the same term. Therefore, the right-hand-side of (24) is bounded above by

$$T\Big[\log(1 + P^{\alpha_s}) - \sum_{k=1}^{s} R_k\Big]. \tag{28}$$

B. Bounding the left-hand-side of (24)

Each term $H\big(F_{q(k)}|U_{p(k)}, S_{p(k)}\big)$ is equal to

$$H\big(F_{q(k)}|U_{p(1)}, \ldots, U_{p(k)}, F_{q(1)}, \ldots, F_{q(k-1)}\big), \tag{29}$$

which holds since messages are independent of files and cache contents (see (25)). From (29), it can be seen that the left-hand-side of (24) is in fact a lower bound on the number of bits that must be delivered (i.e load) in a conventional share-link setting with s users [2, eq. (30)]. We hence employ the results and techniques of [2] to obtain:

$$\begin{aligned} &\frac{1}{s!N!}\sum_{(p,q)}\sum_{k\in[s]} H\big(F_{q(k)}|U_{p(k)}, S_{p(k)}\big) \\ &\geq \frac{B}{2.01}\cdot\left(\frac{N-M}{M}\big(1-(1-M/N)^s\big)\right) \qquad (30) \\ &\geq \frac{B}{2.01}\cdot \mathrm{conv}\left(\frac{\binom{K}{K\mu+1} - \binom{K-s}{K\mu+1}}{\binom{K}{K\mu}}\right). \qquad (31) \end{aligned}$$

Going to within a multiplicative factor of 2.01 from the decentralized load in (30) holds due to [2, Lem. 3] and [2, Lem. 1], while (31) follows from the results in [20] (see also [2, Appendix G] where a similar step is used).

C. Combining bounds

From (24), (28) and (31), and by taking the appropriate limits $T \to \infty$ and $P \to \infty$, we obtain

$$\sum_{k=1}^{s} r_k + \frac{1}{2.01\cdot\tau}\cdot \mathrm{conv}\left(\frac{\binom{K}{K\mu+1} - \binom{K-s}{K\mu+1}}{\binom{K}{K\mu}}\right) \leq \alpha_s \tag{32}$$

which holds for any $s \in [K]$. By rearranging the terms in (32) and taking the tightest of such bounds over all s, we obtain the desired lower bound on the GNDT in Theorem 1.

VII. Conclusion

In this work, we introduced the problem of wireless coded caching under mixed cacheable content and uncacheable non-content types of traffic in the context of the degraded GBC. The extension of this result to other networks, including multi-transmitter and multi-antenna networks, is of high interest.

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] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 647–663, Jan. 2019.

[3] 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.

[4] 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.

[5] M. M. Amiri and D. Gündüz, "On the capacity region of a cache-aided Gaussian broadcast channel with multi-layer messages," in *Proc. IEEE ISIT*, Jun. 2018, pp. 1909–1913.

[6] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," in *Proc. IEEE ISIT*, Jun. 2017, pp. 401–405.

[7] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, "Fundamental limits of wireless caching under uneven-capacity channels," *arXiv:1908.04036*, 2019.

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

[9] E. Lampiris, J. Zhang, and P. Elia, "Cache-aided cooperation with no CSIT," in *Proc. IEEE ISIT*, Jun. 2017, pp. 2960–2964.

[10] 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.

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

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

[13] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[14] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE ISIT*, Jun. 2015, pp. 809–813.

[15] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[16] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5359–5380, Jul. 2018.

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

[18] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5534–5562, Dec. 2008.

[19] H. Joudeh, E. Lampiris, P. Elia, and G. Caire, "Fundamental limits of wireless caching under mixed cacheable and uncacheable traffic," *arXiv:2002.07691*, 2020.

[20] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1281–1296, Feb. 2018.