Fundamental Limits on Latency in Small-Cell Caching Systems: An Information-Theoretic Analysis

Seyyed Mohammadreza Azimi, Osvaldo Simeone, and Ravi Tandon

Abstract—Caching of popular multimedia content at small-cell base stations (BSs) is a promising solution to reduce the traffic load of macro-BSs without relying on a high-speed backhaul architecture. While most prior work analyzed the effect of small-cell caching, or femto-caching, under the assumption of negligible interference between macro-BS and small-cell BS, this paper contributes to a more recent line of work in which the benefits of caching are reconsidered in the presence of interference on the downlink channel. In particular, a binary fading one-sided interference channel is considered in which the small-cell BS, whose transmission is interfered by the macro-BS, has a limited-capacity cache. An information-theoretic metric that captures the delivery latency is defined and fully characterized through information-theoretic achievability and converse arguments as a function of the cache capacity, as well as of the capacity of the backhaul link connecting cloud and small-cell BS.

Index Terms—Edge caching, interference channel, information theory, latency, cloud RAN.

I. INTRODUCTION

Caching of popular multimedia content at small-cell base stations (BSs) of a cellular system, also known as femto or edge-caching, has been widely studied in recent years as a low-latency means to deliver video files without relying on high-speed backhaul connections to the "cloud" [1], [2]. Most existing theoretical work on the performance advantages, in terms of latency, of edge caching has focused on wireless channel models in which small-cells BSs and macro-BSs cannot coordinate their transmissions and hence cannot manage their mutual interference (see [1], [2] and references therein). In contrast, recent work in [3], [4] endeavored to address the possibility of interference management among edge nodes, such as small-cell and macro-BSs, based on the respective cached contents.

The papers [3], [4] proposed caching and transmission schemes that enables coordination and cooperation at the BSs based on the cached contents for a system with three BSs and three users. The performance of these schemes was evaluated in terms of the information-theoretic high signal-to-noise ratio (SNR) metric of the degrees of freedom, or, more precisely, of its inverse, as a function of the cache capacity of the BSs. More recent research in [5] provided an operational meaning for the inverse of the degrees of freedom metric used in [3], [4] in terms of delivery latency, and derived a lower bound on this metric for a general system with any number of BSs and users. Furthermore, in [6], the system model studied in [3], [4], [5] was extended to encompass also a cloud server, which is connected to the BSs via finite-capacity backhaul links and can make up for partial caching of the library of files at the BSs. The mentioned high-SNR latency metric was fully characterized in [6] as a function of the cache and backhaul capacity by developing achievability and converse arguments for a number of special cases of interest. Related works that focus on signal processing aspects of the discussed cache and cloud-aided system include [7], [8], [9], [10].

In this work, we consider a set-up with a small-cell BS and a macro-BS, represented by Encoder 1 and Encoder 2, respectively, in Fig. 1. The small-cell BS (Encoder 1) is endowed with a cache of finite capacity and can serve a small-cell mobile user, represented by Decoder 1. The macro-BS (Encoder 2) can serve a macro-cell user, namely Decoder 2, as well as, possibly, also Decoder 1. When intended for Decoder 2, the transmission from the macro-BS (Encoder 2) hence causes interference to Decoder 1. It is noted that, unlike all mentioned prior work in which full wireless connectivity was assumed, in the practically relevant set-up of Fig. 1, the small-cell BS transmits with sufficiently small power so as not to create interference at Decoder 2, yielding a partially connected wireless channel.

The goal of this paper is to characterize the minimal delivery latency for the system in Fig. 1 as a function of the cache capacity at Encoder 1 and of the capacity of the backhaul link that connects the cloud to Encoder 1. To this end, we adopt a simplified channel model, namely the binary interference
channel first introduced in [11], which is further used in [12] in order to obtain insights into the performance of fading Gaussian one-sided interference channels.

The rest of the paper is organized as follows. In Sec. II we present the system model, including the definition of the key performance metric of Delivery Time per Bit (DTB). Sec. III and Sec. V characterize the DTB for the system in Fig. 1 in the absence and presence, respectively, of a backhaul link between cloud and Encoder 1, with Sec. IV detailing the proof of achievability for the set-up with no backhaul connectivity. Finally, Sec. VI concludes the paper.

Notation: Given $a > 0$, we define the set $[a] = \{1, 2, ..., [a]\}$. For any probability $p$, we define $\bar{p} = 1 - p$.

II. System Model

We study the cache and cloud-aided system depicted in Fig. 1. To elaborate, let $\mathcal{L} = \{W_1, ..., W_N\}$ be a library of $N$ files, which are independent and identically distributed according to uniform distribution, so that we have $W_i \sim \mathcal{U}([2^F])$, for $i \in [N]$, where $F$ is the file size in bits. Encoder 1, which models a small-cell BS, has a local cache and is able to store $\mu NF$ bits. The parameter $\mu$ is hence the fractional cache size and represents the portion of library that can be stored at the cache. Encoder 2, which models a macro-BS, has available the whole library $\mathcal{L}$ thanks to its direct connection to the cloud. Encoder 1 is also connected to the cloud, which stores the entire library $\mathcal{L}$, but only through a rate-limited link of capacity $C$ bits per channel use. The special case with $C = 0$, i.e., with encoder 1 only aided by its cache, is considered first, and the extension to the more general set-up with $C > 0$ will be discussed at Sec. V.

Two receivers are served by the encoders via a binary fading interference channel, previously studied in [11], [12]. As illustrated in Fig. 1, the signal received at Decoder 1 and Decoder 2 at time $t$ can be written as:

$$
\begin{align*}
Y_1(t) &= G_1(t)X_1(t) + G_0(t)X_2(t), \\
Y_2(t) &= G_2(t)X_2(t),
\end{align*}
$$

(1)

where $G(t) = (G_0(t), G_1(t), G_2(t)) \in \{0, 1\}^3$ represents the vector of binary channel coefficients at time $t$, and $X_1(t)$ and $X_2(t)$ are the binary transmitted signals from Encoder 1 and Encoder 2, respectively. In (1), all operations are in the binary field. The channel gains are distributed as $G_1(t) \sim \text{Bernoulli}(\epsilon_1)$ and $G_0(t), G_2(t) \sim \text{Bernoulli}(\epsilon_2)$, are mutually independent and change independently over time. The parameters $\epsilon_1$ and $\epsilon_2$ describes the average quality of the communication links originating at Encoder 1 and Encoder 2, respectively, and are hence in practice related to the transmission powers of Encoder 1 and Encoder 2. We remark that a more general model with different erasure probabilities for the links $G_0(t)$ and $G_2(t)$ could also be considered but at the expense of a more cumbersome notation and analysis, which is not further pursued here.

Each user or decoder $k$, requests a file $W_{d_k}$ from the library $\mathcal{L}$ at every transmission interval for $k = 1, 2$. The demand vector is defined as $d = (d_1, d_2) \in [N]^2$. The system operates according to the following two phases.

1) Placement phase: The placement phase is defined by functions $\phi_i(\cdot)$, at Encoder 1, which maps each file $W_i \in \mathcal{L}$ to its cached version $V_i$:

$$
V_i = \phi_i(W_i) \quad \forall i \in \{1, ..., N\}.
$$

(2)

To satisfy cache storage constraint, it is required that

$$
H(V_i) \leq \mu F.
$$

(3)

The total cache content at encoder 1 is given by:

$$
V = (V_1, ..., V_N).
$$

(4)

Note that, as in [5] and [13], we focus on caching that allows for arbitrary intra-file coding but not for inter-file coding as per (2). Furthermore, the caching policy is kept fixed over multiple transmission intervals and is thus independent of the receivers’ requests and of the channel realizations in the transmission intervals.

2) Delivery phase: The delivery phase is in charge of delivering the given request vector $d$ in each transmission interval given the current channel realization. It is defined by the following two functions.

- **Encoding**: Encoder 1 uses the encoding function

$$
\psi_1: [2^\mu NF] \times [N]^2 \times \{0, 1\}^{3T} \rightarrow \{0, 1\}^T,
$$

(5)

which maps the cached content $V$, the demand vector $d$ and the CSI sequence $G^T = (G(1), ..., G(T))$ to the transmitted codeword $X_1^T = (X_1[1], ..., X_1[T]) = \psi_1(V, d, G^T)$. Note that $T$ represents the duration of transmission in channel uses. Encoder 2 uses the following encoding function:

$$
\psi_2: [2^\mu NF] \times [N]^2 \times \{0, 1\}^{3T} \rightarrow \{0, 1\}^T,
$$

(6)

which maps the library $\mathcal{L}$ of all files, the demand vector $d$, and the CSI vector $G^T$ to the transmitted codeword $X_2^T = (X_2[1], ..., X_2[T]) = \psi_2(\mathcal{L}, d, G^T)$.

- **Decoding**: Each decoder $j \in \{1, 2\}$ is defined by the following mapping:

$$
\eta_j: \{0, 1\}^T \times [N]^2 \times \{0, 1\}^{3T} \rightarrow [2^F],
$$

(7)

which outputs the detected message $\hat{W}_{d_j} = \eta_j(Y_j^T, d, G^T)$ where $Y_j^T = (Y_j[1], ..., Y_j[T])$ is the received signal (1) at receiver $j$.

We refer to a selection for the caching function and the encoding and decoding functions in (5)-(7) as a policy. The probability of error is evaluated with respect to the worst-case demand vector and decoder as

$$
P_e^F = \max_{d \in [N]^2} \max_{j \in \{1, 2\}} \Pr(\hat{W}_{d_j} \neq W_{d_j}).
$$

(8)

The delivery time per bit (DTB) of a code is defined as $T/F$ and is measured in channel symbols per bit. A DTB $\Delta$ is said to be achievable if there exists a sequence of policies indexed by the file size $F$ for which the limits

$$
\lim_{F \rightarrow \infty} \frac{T}{F} = \Delta
$$

(9)

and $P_e^F \rightarrow 0$ as $F \rightarrow \infty$ hold. The minimum DTB $\delta^*(\mu)$ is the infimum of all achievable DTB when the fractional cache capacity at encoder 1 is equal to $\mu$. 
III. MINIMUM DTB

In this section, we derive the minimum DTB $\delta^*(\mu)$ for the system in Fig. 1 by assuming $C = 0$.

**Proposition 1.** The minimum DTB for the cache and cloud-aided system in Fig. 1 with $C = 0$ is

$$\delta^*(\mu) = \begin{cases} \frac{2-\mu}{\delta_0} & \text{if } \mu \leq \mu_0 \\ \frac{2}{\delta_0} & \text{if } \mu \geq \mu_0, \end{cases}$$

with $\mu_0$ and $\delta_0$ given by

$$\mu_0 = \begin{cases} 1 - \epsilon_2 & \text{if } \epsilon_1 \epsilon_2 > \epsilon_2^2 \epsilon_1 \\ \frac{2(1-\epsilon_1)(\epsilon_2^2-\epsilon_1)}{2-\epsilon_1-\epsilon_2+\epsilon_1 \epsilon_2-\epsilon_2^2} & \text{if } \epsilon_1 \epsilon_2 \leq \epsilon_2^2 \epsilon_1 \end{cases}$$

and

$$\delta_0 = \max\left(1, \frac{2}{1-\epsilon_2}, \frac{2}{2-\epsilon_1-\epsilon_2+\epsilon_1 \epsilon_2-\epsilon_2^2}\right).$$

**Proof.** See Sec. IV for achievability and [14] for the converse.

To elaborate on the result in Proposition 1, consider first the setting in which Encoder 1 has no caching capabilities, i.e. $\mu = 0$. In this case, Encoder 2 needs to deliver the requested files to both decoders on a binary erasure broadcast channel. Considering the worst-case in which two different files are requested by two decoders, the minimum average time to serve both users is $T = 2F/(1-\epsilon_2^2)$, since with probability $1-\epsilon_2^2$ a bit can be delivered to either Decoder 1 or Decoder 2 by Encoder 2, yielding a minimum DTB of $\delta^*(0) = 2/(1-\epsilon_2^2)$. In contrast, when the entire library is available at Encoder 1, i.e., $\mu = 1$, depending on the relative values of $\epsilon_1$ and $\epsilon_2$, two different cases should be distinguished. Roughly speaking, if the channel between Encoder 2 and the Decoders is weaker on average than the channel between Encoder 1 and Decoder 1, or more precisely if $\epsilon_1 \geq \epsilon_2$, then the minimum DTB is limited by transmission delay to Decoder 2 and the minimum DTB is $\delta^*(1) = 1/(1-\epsilon_2)$. Instead, when the channel between Encoder 1 and Decoder 1 is weaker on average than the channel between Encoder 2 and both decoders, or $\epsilon_1 \leq \epsilon_2$, the resulting minimum DTB depends on both $\epsilon_1$ and $\epsilon_2$. In both cases, Encoder 2 serves a fraction $1-\mu_0$ of the requested file to Decoder 1, so that Encoder 1 only needs to deliver a fraction $\mu_0$ of the requested file by Decoder 1.

As it will be detailed in the next section, a key element of the transmission policies is that, in the channel state in which all three links are active, the presence of the cache at Encoder 1 allows the latter to coordinate its transmission with Encoder 2 and cancel the interference caused by Encoder 2 to Decoder 1. Furthermore, from the discussion above, a fractional cache size $\mu \geq \mu_0$ is sufficient to achieve the same DTB $\delta_0$ as with full caching. Fig. 3 shows the value $\mu_0$ as a function of $\epsilon_1$ for different values of $\epsilon_2$. It is observed that, for fixed $\epsilon_2$, the fraction $\mu_0$ decreases with $\epsilon_1$, showing that an Encoder 1 with a low channel quality cannot benefit from a large cache size. Furthermore, as the channel from Encoder 2 becomes more reliable, i.e., for small $\epsilon_2$, a larger cache at Encoder 1 enables the latter to coordinate more effectively with Encoder 2, hence improving the DTB.

IV. PROOF OF ACHIEVABILITY

In this section, we provide details on the policies that achieve the minimum DTB identified in Proposition 1. We start by proving that the minimum DTB $\delta^*(\mu)$ is a convex function of $\mu$. The proof leverages the splitting of files into subfiles delivered using different strategies via time sharing.

**Lemma 1.** The minimum DTB $\delta^*(\mu)$ is a convex function of $\mu \in [0, 1]$.

**Proof.** See [14].

By the convexity of $\delta^*(\mu)$ proved in Lemma 1, it suffices to prove that the corner points ($\mu = 0, \delta^*(0) = 2/(1-\epsilon_2^2)$) and ($\mu = \mu_0, \delta_0$) are achievable. In fact, the minimum DTB $\delta^*(\mu)$ can then be achieved, following the proof of Lemma 1, by file splitting and time sharing between the optimal policies for $\mu = 0$ and $\mu = \mu_0$ in the interval $0 \leq \mu \leq \mu_0$ and by using the optimal policy for $\mu = \mu_0$ in the interval $\mu_0 \leq \mu \leq 1$ (see Fig. 2).

In the following, we use the notation $(g_0,g_1,g_2) \in \{0,1\}^3$ to identify the channel realization $(G_0 = g_0, G_1 = g_1, G_2 = g_2)$. For instance, $(0,1,1)$ represents the channel realization in which $Y_1 = X_1$ and $Y_2 = X_2$, and $(1,0,1)$ that in which $Y_1 = X_2$ and $Y_2 = X_2$. 

![Figure 2](image-url) Figure 2: Minimum Delivery Time per Bit (DTB) $\delta^*(\mu)$ for the system in Fig. 1 with $C = 0$. 

![Figure 3](image-url) Figure 3: Optimum fractional cache size $\mu_0$ as a function of $\epsilon_1$ for different values of $\epsilon_2$, which ranges from 0 to 1 with step size 0.1.
A. No Caching (µ = 0)

We first consider the corner point (µ = 0, δ*(0) = 2/(1 - e₂)). In this setting, in which Encoder 1 has no caching capabilities, the model reduces to a broadcast erasure channel from Encoder 2 to both decoders. The worst-case demand vector is any one in which the decoders request different files. In fact, if the same file is requested, it can always be treated as two distinct files achieving the same latency as for a scenario with distinct files. Focusing on this worst-case scenario, we adopt the following delivery policy.

Encoder 1 always transmits X₁ = 0. Encoder 2 transmits 1 bit of information to Decoder 1 in the states (1, 0, 0) and (1, 1, 0), in which the channel from Encoder 2 to Decoder 1 is on while the channel to Decoder 2 is off. It transmits 1 bit of information to Decoder 2 in the states (0, 0, 1) and (0, 1, 1), in which the channel to Decoder 2 is on while the channel to decoder 1 is off. Instead, in states (1, 0, 1) and (1, 1, 1), in which both channels to Decoder 1 and Decoder 2 are on, Encoder 2 transmits 1 bit of information to Decoder 1 or to Decoder 2 with equal probability.

Consider now the time T₁ required for Decoder 1 to decode successfully F bits. We can write this random variable as

$$T₁ = \sum_{k=1}^{F} T_{1,k}, \quad (13)$$

where T_{1,k} denotes the number of channel uses required to transmit the kth bit. Given the discussion above, the variables T_{1,k} are independent for k ∈ [F] and have a Geometric distribution with mean \( \Pr(G = (1, 0, 0)) + \Pr(G = (1, 1, 0)) \) = 1/2 and

$$\Pr(G = (1, 0, 1)) + 1/2 \Pr(G = (1, 1, 1)) = 2/(1 - e₂).$$

By the strong law of large numbers we now have the limit

$$\lim_{F \to \infty} \frac{T₁}{F} = \mathbb{E}[T₁] = \frac{2}{1 - e₂} \quad (14)$$

with probability 1. In a similar manner, the resulting delivery time for Decoder 2 for any given bit has a Geometric distribution with mean \( \Pr(G = (0, 0, 1)) + \Pr(G = (0, 1, 1)) \) = 1/2 and

$$\Pr(G = (1, 0, 1)) + 1/2 \Pr(G = (1, 1, 1)) = 2/(1 - e₂);$$

and, by the strong law of large numbers, we obtain that the time T₂ needed to transmit F bits to Decoder 2 satisfies the limit

$$\lim_{F \to \infty} \frac{T₂}{F} = \mathbb{E}[T₂] = \frac{2}{1 - e₂} \quad (15)$$

almost surely. Using this limit along with (14) allows to conclude that there exists a sequence of policies with \( T/F \to 2/(1 - e₂) \) for any arbitrarily small probability of error.

B. Partial Caching (µ = µ₀) with \( \bar{ε}_1 \varepsilon_2 \geq \varepsilon_1 \varepsilon_2^2 \)

Next, we consider the corner point (µ₀, δ₀) under the condition \( \bar{ε}_1 \varepsilon_2 \geq \varepsilon_1 \varepsilon_2^2 \). In this case, in which Encoder 1 has a better channel than Encoder 2 in the average sense discussed above, our findings show that Encoder 2 should communicate to Decoder 1 only in the channel states in which the channel to Decoder 2 is off. Using these states, Encoder 2 sends \((1 - \mu₀)F\) bits to Decoder 1. Encoder 1 cache a fraction \( \mu₀ \) of each file in the library and delivers \( \mu₀F \) bits of the requested file to Decoder 1. For this purpose, coordination between Encoder 1 and Encoder 2 is needed to manage interference in the state (1, 1, 1) in which all links are on.

A detailed description of the transmission strategy is provided below as a function of the channel state \( G \).

1) \( G = (0, 0, 1) \): Only the channel between Encoder 2 and Decoder 2 is active, and Encoder 2 transmits 1 bit of information to Decoder 2.

2) \( G = (1, 0, 0) \): The only active channel is between Encoder 1 and Decoder 1, and Encoder 1 transmits 1 information bit to Decoder 1.

3) \( G = (0, 1, 1) \): The cross channel is off, and each encoder transmits 1 bit of information to its decoder.

4) \( G = (1, 0, 0) \): Only the channel to Decoder 2 is on, and Encoder 2 transmits 1 bit of information to Decoder 2.

5) \( G = (1, 1, 0) \): The direct channel between Encoder 1 and Decoder 1 is off, while two other channels are on. Encoder 2 transmits 1 bit of information to Decoder 2.

6) \( G = (1, 1, 0) \): Both channels from Encoder 1 and Encoder 2 to Decoder 1 are on. Encoder 1 transmits \( X₁ = 0 \) and Encoder 2 transmits 1 bit of information to Decoder 1.

7) \( G = (1, 1, 1) \): Encoder 2 transmits 1 bit \( X₂ \) of information to Decoder 2. Encoder 1 transmits \( \bar{X}_1 \oplus X₂ \), where \( \bar{X}_1 \) is an information bit for Decoder 1. This form of coordination is enabled by the fact that Encoder 1 knows the bit \( X₂ \), since it is part of the \( \mu₀F \) cached bits from the file requested by Decoder 2. In this way, interference from Encoder 2 is cancelled at Decoder 1.

From the discussion above, Encoder 2 transmits 1 bit of information to Decoder 2 in the states 1), 3), 5) and 7). For large \( F \), the normalized transmission delay for transmitting the requested file to Decoder 2 is then equal to

$$δ₂₂ = \left( \Pr(G = (0, 0, 1)) + \Pr(G = (0, 1, 1)) \right) + \Pr(G = (1, 0, 1)) + \Pr(G = (1, 1, 1)) \right)^{-1} \quad (16)$$

Furthermore, Encoder 2 transmits \((1 - \mu₀)F\) bits to decoder 1 in the states at 4) and 6). The required normalized time for large \( F \) is hence

$$δ₂₁ = \frac{1 - \mu₀}{\varepsilon_1 \varepsilon_2} \quad (17)$$

Finally, Encoder 1 transmits \( \mu₀F \) bits to Decoder 1 in the states at 2), 3) and 7). The required time is thus

$$δ₁₁ = \frac{\mu₀}{\bar{ε}_1 \varepsilon_2 + \varepsilon_1 \varepsilon_2^2} \quad (18)$$

It can be shown that \( \delta₁₁ \leq δ₂₁ = δ₂₂ = δ₀ \) under the given condition \( \bar{ε}_1 \varepsilon_2 \geq \varepsilon_1 \varepsilon_2^2 \), and hence the DTB is given by \( max(δ₁₁, δ₂₁, δ₂₂) = δ₀ \).

C. Partial Caching (µ = µ₀) with \( \bar{ε}_1 \varepsilon_2 \leq \varepsilon_1 \varepsilon_2^2 \)

Finally, we consider the corner point (µ₀, δ₀) under the complementary condition \( \bar{ε}_1 \varepsilon_2 \leq \varepsilon_1 \varepsilon_2^2 \), in which Encoder 2 has better channels to the decoders. In this case, as above, Encoder 1 caches a fraction \( \mu₀ \) of all files. Transmission take place as described in Sec. IV-B except for state 5) which is modified as follows:
5) $G = (1, 0, 1)$: Encoder 2 transmits 1 bit of information to either Decoder 1 or Decoder 2 with probabilities $\alpha = (1 - \bar{e}_1 \bar{e}_2 / \bar{e}_1 \bar{e}_2^2)/2$ and $1 - \alpha$, respectively.

Encoder 2 hence transmits 1 bit of information to Decoder 2 in the states at cases 1), 3) and 7) and also with probability $1 - \alpha$ in case 5). For large $F$, the normalized transmission delay for transmitting the requested file to Decoder 2 tends to

$$\delta_{22} = \left( \Pr(G = (0, 0, 1)) + \Pr(G = (0, 1, 1)) \right)^{-1}$$

$$= \frac{2 - \bar{e}_1 - \bar{e}_2 + \bar{e}_1 \bar{e}_2}{2 - \bar{e}_1 - \bar{e}_2 - \bar{e}_1 \bar{e}_2}.$$  \hspace{1cm} (18)

In addition, Encoder 2 transmits 1 bit to Decoder 1 in cases 4) and 6) as well as in case 5) with probability $\alpha$. The required time to transmit $(1 - \mu_0)F$ bits from Encoder 2 to Decoder 1 is hence

$$\delta_{21} = \frac{1 - \mu_0}{\bar{e}_2 \bar{e}_2 + \frac{1}{2}(1 - \bar{e}_1 \bar{e}_2)}.$$  \hspace{1cm} (19)

It can be shown that $\delta_{21} \leq \delta_{11} = \delta_{22} = \delta_0$, where $\delta_{11}$ is given in (17) under the given condition $\bar{e}_1 \bar{e}_2 \leq \bar{e}_1 \bar{e}_2^2$, yielding the DTB $\max(\delta_{11}, \delta_{21}, \delta_{22}) = \delta_0$. This concludes the proof of achievability.

V. CLOUD-AIDED SMALL-CELL BS

In this section, we reconsider the cloud and cache-aided system in Fig. 1 by allowing a rate-limited connection with capacity $C > 0$ between Cloud and Encoder 1, which represents a small-cell BS as discussed in Sec. I. The system model is first revised to include the Cloud-to-Encoder 1 link, and then the minimum DTB is derived as a function of both cache size $\mu$ and fronthaul capacity $C$.

A. System Model

The same system model as described in Sec. II is adopted with the following caveats. In the delivery phase, the Cloud implements an encoding function

$$\psi_C: [2^{NF}] \times [N]^2 \times \{0, 1\}^{3T} \rightarrow [2^{TC}],$$  \hspace{1cm} (20)

which maps the library $L$ of all files, the demand vector $d$ and the CSI vector $G^T$ to the signal $U_{TC} = (U_1, ..., U_{TC}) = \psi_C(L, d, G^T)$ to be delivered to encoder 1. Here, parameter $T_C$ represents the duration of the transmission from Cloud to Encoder 1 in terms of number of channel uses. We have the inequality $H(U_i) \leq C$ for $i \in [T_C]$ by the capacity limitations on the Cloud-to-Encoder 1 link. Furthermore, Encoder 1 uses the encoding function

$$\psi_1: [2^{NF}] \times [2^{TC}] \times [N]^2 \times \{0, 1\}^{3T} \rightarrow [0, 1]^T,$$  \hspace{1cm} (21)

which maps the cached content $V$, the received signal $U_{TC}$, the demand vector $d$ and the CSI sequence $G^T = (G(1), ..., G(T))$ to the transmitted codeword $X_T^{\gamma} = (X_1[1], ..., X_1[T]) = \psi_1(V, U_{TC}, d, G^T)$. Note that $T$ represents, as above, the duration of transmission on the binary fading channel in channel uses.

Decoding and probability of error are defined as in Sec. II. Instead, a DTB $\Delta$ is said to be achievable if there exists a sequence of policies, defined by (2), (6), (7), (20) and (21) and indexed by $F$, such that the limits

$$\lim_{F \rightarrow \infty} \frac{T + T_C}{F} = \Delta$$  \hspace{1cm} (22)

and $P^F \rightarrow 0$ as $F \rightarrow \infty$ hold. The minimum DTB $\delta^*(\mu, C)$ is the infimum of all achievable DTBs when the fractional cache size at Encoder 1 is equal to $\mu$ and the Cloud-to-Encoder 1 capacity is equal to $C$.

B. Minimum DTB

In this section, we derive the minimum DTB $\delta^*(\mu, C)$ for the system in Fig. 1.

**Proposition 2.** The minimum DTB for the cache and cloud-aided system in Fig. 1 is equal to (10) if

$$C \leq 1 - \bar{e}_2^2;$$  \hspace{1cm} (23)

otherwise, it is given by

$$\delta^*(\mu, C) = \text{min} \left\{ \begin{array}{ll} \frac{2 - \bar{e}_1 - \bar{e}_2 + \bar{e}_1 \bar{e}_2^2}{\bar{e}_1 \bar{e}_2} + \left(1 - \frac{1 - \bar{e}_2^2}{\bar{e}_1 \bar{e}_2^2}\right) \delta_0 & \text{if } \mu \leq \mu_0 \\ \frac{2 - \bar{e}_1 - \bar{e}_2 + \bar{e}_1 \bar{e}_2^2}{\bar{e}_1 \bar{e}_2} & \text{if } \mu \geq \mu_0 \end{array} \right.$$  \hspace{1cm} (24)

with $\mu_0$ and $\delta_0$ is defined in (11) and (12), respectively.

**Proof.** See Sec. V-C and [14].

To shed light on the results in Proposition 2, consider first the setting in which Encoder 1 has no caching capability, i.e., $\mu = 0$. In this case, unlike the scenario studied in the previous section, Encoder 1 can deliver part of the file requested by Decoder 1 through the connection to the Cloud. Nevertheless, if $C \leq 1 - \bar{e}_2^2$, that is, if the average delay for transmission of 1 bit from cloud to Encoder 1, namely $1/C$, is larger than the corresponding delay between Encoder 2 and both decoders, namely $1/(1 - \bar{e}_2^2)$ (see Sec. III), then it is optimal to neglect Encoder 1 and operate as discussed in Sec. IV-A. Instead, if $C \geq 1 - \bar{e}_2^2$, it is optimal for Encoder 1 to transmit parts of the requested files, or functions thereof, which are received from the cloud. In fact, as discussed below, it is necessary for the cloud to transmit a coded signal obtained from both the files requested by the users in order to obtain the DTB in Proposition 2. Furthermore, if the fractional cache size satisfies the inequality $\mu \geq \mu_0$, then the cache size at Encoder 1 is sufficient to achieve the DTB $\delta_0$ corresponding to full caching and the Cloud-to-Encoder 1 link can be neglected with no loss of optimality.

C. Proof of Achievability

In this section, we provide details on the policies that achieve the minimum DTB identified in Proposition 2. We start by noting that for $C \leq 1 - \bar{e}_2^2$, achievability follows from Proposition 1, and hence we can focus on the case $C \geq 1 - \bar{e}_2^2$. We first observe that the minimum DTB $\delta^*(\mu, C)$ is a convex function of $\mu$ for any value of $C$. The proof follows as in Lemma 1 by file splitting and time sharing and is hence omitted.
Therefore, the normalized time required on the edge channel while the transmission of coded bits requires time from Encoder 1 to Decoder 1 requires a normalized time on

\[ \delta_{E} = \max(\delta_{11}, \delta_{11}, \delta_{21}, \delta_{22}) = \delta_{0}. \]

Instead, under the condition \( \bar{\epsilon}_{1} \bar{\epsilon}_{2} \leq \bar{\epsilon}_{3} \bar{\epsilon}_{4} \), using the same equations, it can be seen that \( \delta_{21} \leq \delta_{11} = \delta_{0} = \delta_{22} = \delta_{0}. \) It follows that \( \delta_{E} = \max(\delta_{21}, \delta_{11}, \delta_{11}, \delta_{22}) = \delta_{0}. \) We can conclude that DTB is \( \delta_{C} + \delta_{E} = \delta_{0} + (2 - \delta_{0}(1 - \bar{\epsilon}_{2})) / C_{2}, \)

which is equal to \( \delta^{*}(0, C) \) in (24).

VI. CONCLUSIONS

As cache and cloud-aided wireless network architectures begin to assume a prominent role in the deployment of next-generation wireless systems, it becomes imperative to understand the potential of interference management as a function of the caching and backhaul capacity limitations. This work contributed to this study by investigating an one-sided interference scenario modeling a macro-BS coexisting with a cache and cloud-aided small-cell BS. Using an original information-theoretic framework that centers on the evaluation of a minimum delivery latency metric, the trade-off between latency and system resources has been studied, and a full characterization has been provided under a simplified binary interference fading channel and in the presence of full CSI. Interesting extensions include the analysis of the impact of imperfect CSI as well as of a more general channel model.

REFERENCES