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].

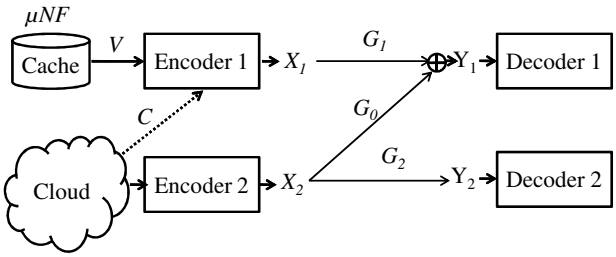

Figure 1: Cache and cloud-aided data delivery over binary fading interference channels.

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 fading

S. M. Azimi and O. Simeone are with the CWCSPR, Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA e-mail: (sa677, osvaldo.simeone@njit.edu). Their work was partially funded by U.S. NSF through grant CCF-1525629. R. Tandon is with University of Arizona, Tucson, AZ, USA.

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. VII concludes the paper.

Notation: Given $a > 0$, we define the set $[a] = \{1, 2, ..., \lceil a \rceil\}$. 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 μNF bits. The parameter μ 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:

$$\begin{aligned} Y_1(t) &= G_1(t)X_1(t) \oplus G_0(t)X_2(t) \\ Y_2(t) &= G_2(t)X_2(t), \end{aligned} \tag{1}$$

where $\mathbf{G}(t) = (G_0(t), G_1(t), G_2(t)) \in \{0,1\}^3$ 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 $G_1(t) \sim$ Bernoulli(ϵ_1) and $G_0(t), G_2(t) \sim$ Bernoulli(ϵ_2), are mutually independent and change independently over time. The parameters ϵ_1 and ϵ_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 $\mathbf{d} = (d_1, d_2) \in [N]^2$. 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:

$$V_i = \phi_i(W_i) \quad \forall i \in \{1, ..., N\}. \tag{2}$$

To satisfy cache storage constraint, it is required that

$$H(V_i) \leq \mu F. \tag{3}$$

The total cache content at encoder 1 is given by:

$$V = (V_1, ..., V_N). \tag{4}$$

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 $\mathbf{d}$ in each transmission interval given the current channel realization. It is defined by the following two functions.

- *Encoding*: Encoder 1 uses the encoding function

$$\psi_1 : [2^{\mu NF}] \times [N]^2 \times \{0,1\}^{3T} \to \{0,1\}^T, \tag{5}$$

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

$$\psi_2 : [2^{NF}] \times [N]^2 \times \{0,1\}^{3T} \to \{0,1\}^T, \tag{6}$$

  which maps the library $\mathcal{L}$ of all files, the demand vector $\mathbf{d}$, and the CSI vector $\mathbf{G}^T$ to the transmitted codeword $X_2^T = (X_2[1], ..., X_2[T]) = \psi_2(\mathcal{L}, \mathbf{d}, \mathbf{G}^T)$.
- *Decoding*: Each decoder $j \in \{1, 2\}$ is defined by the following mapping:

$$\eta_j : \{0,1\}^T \times [N]^2 \times \{0,1\}^{3T} \to [2^F], \tag{7}$$

  which outputs the detected message $\hat{W}_{d_j} = \eta_j(Y_j^T, \mathbf{d}, \mathbf{G}^T)$ 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

$$P_e^F = \max_{\mathbf{d} \in [N]^2} \max_{j \in \{1,2\}} \Pr(\hat{W}_{d_j} \neq W_{d_j}). \tag{8}$$

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

$$\lim_{F \to \infty} \frac{T}{F} = \delta \tag{9}$$

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

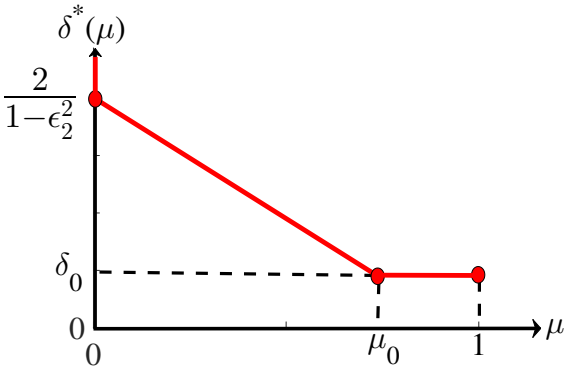

Figure 2: Minimum Delivery Time per Bit (DTB) $\delta^*(\mu)$ for the system in Fig. 1 with $C = 0$.

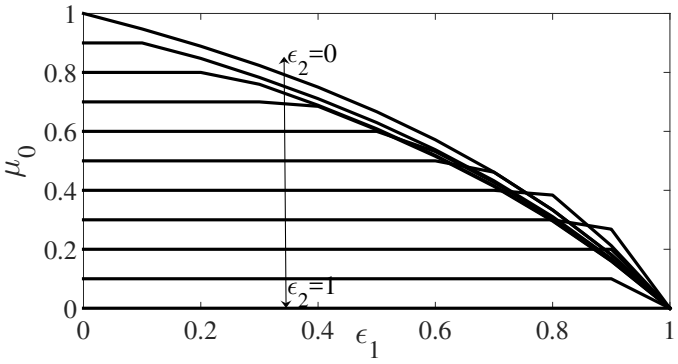

Figure 3: Optimum fractional cache size μ_0 as a function of ϵ_1 for different values of ϵ_2, which ranges from 0 to 1 with step size 0.1.

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*

$$\delta^*(\mu) = \begin{cases} \frac{2-\mu}{1-\epsilon_2^2} & \text{if } \mu \le \mu_0 \\ \delta_0 & \text{if } \mu \ge \mu_0, \end{cases} \tag{10}$$

with μ_0 and δ_0 are given by

$$\mu_0 = \begin{cases} 1-\epsilon_2 & \text{if } \bar{\epsilon}_1\epsilon_2 > \bar{\epsilon}_2^2\epsilon_1 \\ \frac{2(1-\epsilon_1)(\epsilon_2^2-\epsilon_2+1)}{2-\epsilon_1-\epsilon_2+\epsilon_1\epsilon_2-\epsilon_1\epsilon_2^2} & \text{if } \bar{\epsilon}_1\epsilon_2 \le \bar{\epsilon}_2^2\epsilon_1 \end{cases} \tag{11}$$

and

$$\delta_0 = \max\Big(\frac{1}{1-\epsilon_2}, \frac{2}{2-\epsilon_1-\epsilon_2+\epsilon_1\epsilon_2-\epsilon_1\epsilon_2^2}\Big). \tag{12}$$

Proof. See Sec. IV for achievability and Appendix A 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 ϵ_1 and ϵ_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 $\bar{\epsilon}_1 \ge \bar{\epsilon}_2$, 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 $\bar{\epsilon}_1 \le \bar{\epsilon}_2$, the resulting minimum DTB depends on both ϵ_1 and ϵ_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 μ_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 $\mu \ge \mu_0$ is sufficient to achieve the same DTB δ_0 as with full caching. Fig. 3 shows the value μ_0 as a function of ϵ_1 for different values of ϵ_2. It is observed that, for fixed ϵ_2, the fraction μ_0 decreases with ϵ_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 ϵ_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 μ. 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. Consider two policies that require fractional cache sizes μ_1 and μ_2 and achieve DTBs δ_1 and δ_2, respectively. Given a fractional cache size $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ for any $\alpha \in [0,1]$, the system can operate by splitting each file into two parts, one of size αF and the other of size $(1-\alpha)F$, while satisfying the cache constraints. The first fraction of the files is delivered following the first policy, while the second fraction is delivered using the second policy. Since the delivery time is additive over the two file fractions, the DTB $\delta = \alpha\delta_1 + (1-\alpha)\delta_2$ is achieved. □

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 $Y_2 = X_2$.

A. No Caching ($\mu = 0$)

We first consider the corner point $(\mu = 0, \delta^*(0) = 2/(1-\epsilon_2^2))$. 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 $X_1 = 0$. 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 T_1 required for Decoder 1 to decode successfully F bits. We can write this random variable as

$$T_1 = \sum_{k=1}^{F} T_{1,k}, \tag{13}$$

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 $k \in [F]$ and have a Geometric distribution with mean $(\Pr[\mathbf{G} = (1,0,0)] + \Pr[\mathbf{G} = (1,1,0)] + 1/2\Pr[\mathbf{G} = (1,0,1)] + 1/2\ \Pr[\mathbf{G} = (1,1,1)])^{-1} = 2/(1-\epsilon_2^2)$. By the strong law of large numbers we now have the limit

$$\lim_{F\to\infty} \frac{T_1}{F} = E[T_1] = \frac{2}{1-\epsilon_2^2} \tag{14}$$

with probability 1. In a similar manner, the resulting delivery time for Decoder 2 for any given bit has a Geometric distribution with mean $(\Pr[\mathbf{G} = (0,0,1)] + \Pr[\mathbf{G} = (0,1,1)] + \frac{1}{2}\Pr[\mathbf{G} = (1,0,1)] + \frac{1}{2}\ \Pr[\mathbf{G} = (1,1,1)])^{-1} = 2/(1-\epsilon_2^2)$; and, by the strong law of large numbers, we obtain that the time T_2 needed to transmit F bits to Decoder 2 satisfies the limit $\lim_{F\to\infty} \frac{T_2}{F} = E[T_2] = \frac{2}{1-\epsilon_2^2}$ almost surely. Using this limit along with (14) allows to conclude that there exists a sequence of policies with $T/F \to 2/(1-\epsilon_2^2)$ for any arbitrarily small probability of error.

B. Partial Caching ($\mu = \mu_0$) with $\bar{\epsilon}_1\epsilon_2 \geq \epsilon_1\bar{\epsilon}_2^2$

Next, we consider the corner point (μ_0, δ_0) under the condition $\bar{\epsilon}_1\epsilon_2 \geq \epsilon_1\bar{\epsilon}_2^2$. In this case, in which Encoder 1 has a better channel than Decoder 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 $(1-\mu_0)F$ bits to Decoder 1. Encoder 1 cache a fraction μ_0 of each file in the library and delivers $\mu_0 F$ 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 $\mathbf{G}$.

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

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

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

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

5) $\mathbf{G} = (1,0,1)$: 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) $\mathbf{G} = (1,1,0)$: Both channels from Encoder 1 and Encoder 2 to Decoder 1 are on. Encoder 1 transmits $X_1 = 0$ and Encoder 2 transmits 1 bit of information to Decoder 1.

7) $\mathbf{G} = (1,1,1)$: Encoder 2 transmits 1 bit X_2 of information to Decoder 2. Encoder 1 transmits $X_1 = \tilde{X}_1 \oplus X_2$, where $\tilde{X}_1$ is an information bit for Decoder 1. This form of coordination is enabled by the fact that Encoder 1 knows the bit X_2, since it is part of the $\mu_0 F$ 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

$$\begin{aligned} \delta_{22} &= \Big(\Pr[\mathbf{G} = (0,0,1) + \Pr[\mathbf{G} = (0,1,1)] \\ &\qquad + \Pr[\mathbf{G} = (1,0,1)] + \Pr[\mathbf{G} = (1,1,1)]\Big)^{-1} \\ &= \tfrac{1}{\bar{\epsilon}_2}. \end{aligned} \tag{15}$$

Furthermore, Encoder 2 transmits $(1-\mu_0)F$ bits to decoder 1 in the states at 4) and 6). The required normalized time for large F is hence

$$\delta_{21} = \frac{1-\mu_0}{\epsilon_2\bar{\epsilon}_2} \tag{16}$$

Finally, Encoder 1 transmits $\mu_0 F$ bits to Decoder 1 in the states at 2), 3) and 7). The required time is thus

$$\delta_{11} = \frac{\mu_0}{\bar{\epsilon}_1\bar{\epsilon}_2 + \bar{\epsilon}_1\epsilon_2^2} \tag{17}$$

It can be shown that $\delta_{11} \leq \delta_{21} = \delta_{22} = \delta_0$ under the given condition $\bar{\epsilon}_1\epsilon_2 \geq \epsilon_1\bar{\epsilon}_2^2$, and hence the DTB is given by $\max(\delta_{11}, \delta_{21}, \delta_{22}) = \delta_0$.

C. Partial Caching ($\mu = \mu_0$) with $\bar{\epsilon}_1\epsilon_2 \leq \epsilon_1\bar{\epsilon}_2^2$

Finally, we consider the corner point (μ_0, δ_0) under the complementary condition $\bar{\epsilon}_1\epsilon_2 \leq \epsilon_1\bar{\epsilon}_2^2$, in which Encoder 2 has better channels to the decoders. In this case, as above, Encoder 1 caches a fraction μ_0 of all files. Transmission take place as described in Sec. IV-B except for state 5) which is modified as follows:

5) $\mathbf{G} = (1, 0, 1)$: Encoder 2 transmits 1 bit of information to either Decoder 1 or Decoder 2 with probabilities $\alpha = (1 - \bar{\epsilon}_1\epsilon_2/\epsilon_1\bar{\epsilon}_2^2)/2$ 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

$$\begin{aligned}\delta_{22} &= \Big(\Pr[\mathbf{G} = (0,0,1) + \Pr[\mathbf{G} = (0,1,1)] \\ &\quad + \Pr[\mathbf{G} = (1,1,1)] + (1-\alpha)\Pr[\mathbf{G} = (1,0,1)]\Big)^{-1} \\ &= \tfrac{2}{2-\epsilon_1-\epsilon_2+\epsilon_1\epsilon_2-\epsilon_1\epsilon_2^2}.\end{aligned} \tag{18}$$

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

$$\delta_{21} = \frac{1 - \mu_0}{\epsilon_2\bar{\epsilon}_2 + \frac{1}{2}(1 - \bar{\epsilon}_1\epsilon_2)}. \tag{19}$$

It can be shown that $\delta_{11} = \delta_{21} = \delta_{22} = \delta_0$, where δ_{11} is given in (17) under the given condition $\bar{\epsilon}_1\epsilon_2 \leq \epsilon_1\bar{\epsilon}_2^2$, 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 μ 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

$$\psi_C : [2^{NF}] \times [N]^2 \times \{0,1\}^{3T} \to [2^{T_C C}], \tag{20}$$

which maps the library $\mathcal{L}$ of all files, the demand vector $\mathbf{d}$ and the CSI vector $\mathbf{G}^T$ to the signal $U^{T_C} = (U_1, ..., U_{T_C}) = \psi_C(\mathcal{L}, \mathbf{d}, \mathbf{G}^T)$ 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

$$\psi_1 : [2^{\mu NF}] \times [2^{T_C C}] \times [N]^2 \times \{0,1\}^{3T} \to \{0,1\}^T, \tag{21}$$

which maps the cached content V, the received signal U^{T_C}, the demand vector $\mathbf{d}$ and the CSI sequence $\mathbf{G}^T = (\mathbf{G}(1), ..., \mathbf{G}(T))$ to the transmitted codeword $X_1^T = (X_1[1], ..., X_1[T]) = \psi_1(V, U^{T_C}, \mathbf{d}, \mathbf{G}^T)$. 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 Δ 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

$$\lim_{F\to\infty} \frac{T + T_C}{F} = \delta \tag{22}$$

and $P_e^F \to 0$ as $F \to \infty$ 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 μ 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*

$$C \leq 1 - \epsilon_2^2; \tag{23}$$

otherwise, it is given by

$$\delta^*(\mu, C) = \begin{cases} \frac{2-\mu}{C} + \left(1 - \frac{1-\epsilon_2^2}{C}\right)\delta_0 & \text{if } \mu \leq \mu_0 \\ \delta_0 & \text{if } \mu \geq \mu_0, \end{cases} \tag{24}$$

with μ_0 and δ_0 is defined in (11) *and* (12)*, respectively.*

Proof. See Sec. V-C and Appendix B. □

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 $C \leq 1 - \epsilon_2^2$, 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 $1/(1 - \epsilon_2^2)$ (see Sec. III), then it is optimal to neglect Encoder 1 and operate as discussed in Sec. IV-A. Instead, if $C \geq 1 - \epsilon_2^2$, 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 Propositon 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 δ_0 corresponding to full caching and the Cloud-to-Encoder 1 link can be neglected with no loss of optimality.

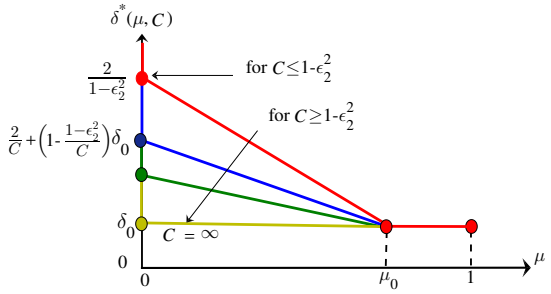

Figure 4: Minimum Delivery Time per Bit (DTB) $\delta^*(\mu, C)$ for the system in Fig. 1.

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 $C \leq 1-\epsilon_2^2$, achievability follows from Proposition 1, and hence we can focus on the case $C \geq 1-\epsilon_2^2$. We first observe that the minimum DTB $\delta^*(\mu, C)$ is a convex function of μ for any value of C. The proof follows as in Lemma 1 by file splitting and time sharing and is hence omitted.

Lemma 2. *The minimum DTB $\delta^*(\mu, C)$ is a convex function of $\mu \in [0,1]$ for any given value of $C \geq 0$.*

By the convexity of $\delta^*(\mu, C)$ in Lemma 2, and by the achievability of the DTB in Proposition 1 with $C = 0$, and hence also for $C \geq 0$, it suffices to prove that the corner point $\delta^*(0, C) = 2/C + (1 - (1-\epsilon_2^2)/C)\delta_0$ is achievable for $C \geq 1-\epsilon_2^2$. To this end, we consider the worst case in which each decoder requests a different file, and we adopt the following policy.

The Cloud-to-Encoder 1 link is used for a normalized time $\delta_C = T_C/F = (2 - \delta_0(1-\epsilon_2^2))/C$ to transmit ρF bits from the file requested by Encoder 1, with

$$\rho = 2 - \delta_0(1-\epsilon_2^2). \tag{25}$$

Of these bits, $\rho F \bar{\epsilon}_1\epsilon_2/(\bar{\epsilon}_1\epsilon_2 + \bar{\epsilon}_1\bar{\epsilon}_2^2)$ bits are sent to Encoder 1 by the Cloud in an uncoded form. Instead, the remaining $\rho F \bar{\epsilon}_1\bar{\epsilon}_2^2/(\bar{\epsilon}_1\epsilon_2 + \bar{\epsilon}_1\bar{\epsilon}_2^2)$ bits are transmitted by XORing each bit of the file with the corresponding bit of the file requested by Decoder 2. The mentioned ρF bits are sent to Decoder 1 by Encoder 1, while the remaining $(1-\rho)F$ bits are sent by Encoder 2 to Decoder 1, as discussed next.

The transmission strategy follows the approach described in Sec. IV-B for all states except for case 5) in which the encoders operate according to Sec. IV-B if $\bar{\epsilon}_1\epsilon_2 > \epsilon_1\bar{\epsilon}_2^2$ or to Sec. IV-C if $\bar{\epsilon}_1\epsilon_2 \geq \epsilon_1\bar{\epsilon}_2^2$. As for (17) the transmission of uncoded bits from Encoder 1 to Decoder 1 requires a normalized time on the channel

$$\delta_{11}^u = \frac{\rho}{\bar{\epsilon}_1\epsilon_2 + \bar{\epsilon}_1\bar{\epsilon}_2^2}. \tag{26}$$

while the transmission of coded bits requires time

$$\delta_{11}^c = \frac{\rho}{\bar{\epsilon}_1\epsilon_2 + \bar{\epsilon}_1\bar{\epsilon}_2^2}. \tag{27}$$

Similar to (16) and (19), the time required for Encoder 2 to transmit to Decoder 1 is

$$\delta_{21} = \begin{cases} \frac{1-\rho}{\epsilon_2\bar{\epsilon}_2} & \text{if } \bar{\epsilon}_1\epsilon_2 > \bar{\epsilon}_2^2\epsilon_1 \\ \frac{1-\rho}{\epsilon_2\bar{\epsilon}_2 + \frac{1}{2}(1-\bar{\epsilon}_1\epsilon_2)} & \text{if } \bar{\epsilon}_1\epsilon_2 \leq \bar{\epsilon}_2^2\epsilon_1 \end{cases} \tag{28}$$

while $\delta_{22} = \delta_0$ is sufficient to communicate to Decoder 2. Under the channel conditions $\bar{\epsilon}_1\epsilon_2 > \bar{\epsilon}_2^2\epsilon_1$, from (25), (26) and (28), it can be shown that $\delta_{11}^u = \delta_{11}^c \leq \delta_{21} = \delta_{22} = \delta_0$. Therefore, the normalized time required on the edge channel is δ_E=max$(\delta_{11}^u, \delta_{11}^c, \delta_{21}, \delta_{22}) = \delta_0$. Instead, under the condition $\bar{\epsilon}_1\epsilon_2 \leq \bar{\epsilon}_2^2\epsilon_1$, using the same equations, it can be seen that $\delta_{11}^c = \delta_{11}^u = \delta_{21} = \delta_{22} = \delta_0$. It follows that δ_E=max$(\delta_{21}, \delta_{11}^c, \delta_{11}^u, \delta_{22}) = \delta_0$. We can conclude that DTB is $\delta_C + \delta_E = \delta_0 + (2 - \delta_0(1-\epsilon_2^2))/C$, 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 fading interference 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

[1] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[2] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[3] M. A. Maddah-Ali and U. Niesen, "Cache aided interference channels," Oct. 2015. [Online]. Available: http://arxiv.org/abs/1510.06121

[4] ——, "Cache aided interference channels," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 809-813, Hong Kong, Jul. 2015.

[5] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: tradeoffs between storage and latency," in *Proc. IEEE CISS 2016*, Princeton, NJ, Mar. 2016 (arXiv:1512.07856).

[6] ——, "Cloud and cache-aided wireless networks: Fundamental latency trade-offs," Submitted. (arXiv1605.01690v3).

[7] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE PIMRC*, pp. 1370-1374, Washington, DC, USA, Sep. 2014.

[8] S.-H. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," Submitted. (arXiv1601.02460v1).

[9] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud ran," Submitted. (arXiv:1512.06938v1).

[10] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, "Hypergraph-based analysis of clustered cooperative beamforming with application to edge caching," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 84–87, Feb. 2016.

[11] Y. Zhu and D. Guo, "Ergodic fading Z-interference channels without state information at transmitters," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 2627–2647, May 2011.
[12] A. Vahid, M. A. Maddah-Ali, and A. S. Avestimehr, "Capacity results for binary fading interference channels with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6093–6130, Oct. 2014.
[13] R. Tandon and O. Simeone, "Cloud aided wireless networks with edge caching : fundamental latency trade offs in fog radio access networks," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, Jul. 2016.

APPENDIX A
PROOF OF CONVERSE FOR PROPOSITION 1

Consider any request vector $\mathbf{d}$ containing two arbitrary, different files W_1 and W_2, and any coding scheme satisfying $P_e^F \to 0$ as $F \to \infty$. The following set of inequalities is based on the fact that, under any such coding scheme, a hypothetical decoder provided with the CSI vector $\mathbf{G}^T$, with the cached contents V_1 and V_2 in (2) relative to files W_1 and W_2, and with the signal $\tilde{G}^T X_2^T$, to be described below, must be able to decode both messages W_1 and W_2. The signal $\tilde{G}^T X_2^T = (\tilde{G}(1)X_2(1), ..., \tilde{G}(T)X_2(T))$ is such that $\tilde{G}(t) = 0$ if $G_0(t) = G_2(t) = 0$ and $\tilde{G}(t) = 1$ otherwise. Note, therefore, that $\tilde{G}(t)X_2(t) = X_2(t)$ as long as either or both $G_0(t)$ and $G_2(t)$ are equal to one. The intuition here is that from $\tilde{G}^T X_2^T$ and G^T, the hypothetical decoder can recover Y_2^T and hence W_2; while from $\tilde{G}^T X_2^T$, G^T and V_1, the decoder can reconstruct Y_1^T and hence decode W_1. Details are as follows:

$$\begin{aligned}
2F &= H(W_1, W_2) \\
&= I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T) \\
&\quad + H(W_1, W_2 | \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T) \\
&= I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T) \\
&\quad + H(W_1 | \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T) \\
&\quad + H(W_2 | \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T, W_1) \\
&\overset{(a)}{=} I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T) \\
&\quad + H(W_1 | \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T, Y_1^T) \\
&\quad + H(W_2 | \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T, W_1, Y_2^T) \\
&\overset{(b)}{\leq} I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, \mathbf{G}^T) + F\gamma_F \\
&= I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2 | \mathbf{G}^T) + F\gamma_F \\
&\overset{(c)}{\leq} H(V_1) + H(\tilde{G}^T X_2^T | \mathbf{G}^T) + F\gamma_F \\
&\overset{(d)}{\leq} \mu F + T(1 - \epsilon_2^2) + F\gamma_F,
\end{aligned} \tag{29}$$

where γ_F indicates any function that satisfies $\gamma_F \to 0$ as $F \to \infty$. In above derivation, (a) follows from the facts that: (*i*) Y_1^T is a function of V_1, V_2, $\mathbf{G}^T$, and $\tilde{G}^T X_2^T$, since X_1^T can be assumed to depend on without loss of generality only on V_1 and V_2, and the vector $G_0^T X_2^T$ can be obtained from $\tilde{G}^T X_2^T$ and $\mathbf{G}^T$; (*ii*) Y_2^T is a function of $(\mathbf{G}^T, \tilde{G}^T X_2^T)$; (b) follows from Fano's inequality; inequality (c) follows from the fact that the messages are independent of channel realization and from Fano inequality $H(V_2 | \tilde{G}^T X_2^T, \mathbf{G}^T) \leq F\gamma_F$; (d) hinges on the cache constraint (3) and by the following bounds:

$$\begin{aligned}
H(\tilde{G}^T X_2^T | \mathbf{G}^T) &\leq \sum_{t=1}^{T} H(\tilde{G}(t)X_2(t) | \mathbf{G}(t)) \\
&\leq T \sum_{\mathbf{g} \in \mathcal{G}} p(\mathbf{g}) \max_{p(X_2)} H(\tilde{G}X_2 | \mathbf{G} = \mathbf{g}) \\
&\leq T(1 - \epsilon_2^2),
\end{aligned} \tag{30}$$

where $\mathcal{G}$ is the set of all channel states and the last inequality follows from the fact that the entropy in all states $\mathbf{G} = \mathbf{g}$ is maximized for $X_2 \sim \text{Bernoulli}(1/2)$. For $F \to \infty$, (29) yields the bound on the minimum DTB:

$$\delta^*(\mu) \geq \frac{2 - \mu}{1 - \epsilon_2^2}. \tag{31}$$

Based on the fact that requested files should be retrieved from the received signals, another bound can be derived as follows:

$$\begin{aligned}
2F &= H(W_1, W_2) \\
&= I(W_1, W_2; Y_1^T, Y_2^T, \mathbf{G}^T) + H(W_1, W_2 | Y_1^T, Y_2^T, \mathbf{G}^T) \\
&\overset{(a)}{\leq} I(W_1, W_2; Y_1^T, Y_2^T, \mathbf{G}^T) + F\gamma_F \\
&\overset{(b)}{\leq} I(W_1, W_2; Y_1^T, Y_2^T | \mathbf{G}^T) + F\gamma_F \\
&= H(Y_1^T, Y_2^T | \mathbf{G}^T) + F\gamma_F \\
&\overset{(c)}{\leq} T \sum_{\mathbf{g} \in \mathcal{G}} p(\mathbf{g}) \max_{p(X_1, X_2)} H(Y_1, Y_2 | \mathbf{G} = \mathbf{g}) + F\gamma_F \\
&\overset{(d)}{=} T(2 - \epsilon_1 - \epsilon_2 + \epsilon_1\epsilon_2 - \epsilon_1\epsilon_2^2) + F\gamma_F,
\end{aligned} \tag{32}$$

where (a) follows from Fano's inequality; (b) follows from the fact that channel gains are independent from files; (c) follows in a manner similar to (30); and (d) is due to the fact that the entropy terms in the previous step are maximized by choosing X_1 and X_2 to be independent and identically distributed as Bernoulli$(1/2)$. With $F \to \infty$, we obtain the bound:

$$\delta^*(\mu) \geq \frac{2}{2 - \epsilon_1 - \epsilon_2 + \epsilon_1\epsilon_2 - \epsilon_1\epsilon_2^2}. \tag{33}$$

Considering decoder 2, the file W_2 should be decodable from Y_2^T, leading to the following bounds:

$$\begin{aligned}
F = H(W_2) &= I(W_2; Y_2^T, \mathbf{G}^T) + H(W_2 | Y_2^T, \mathbf{G}^T) \\
&\overset{(a)}{\leq} I(W_2; Y_2^T | \mathbf{G}^T) + F\gamma_F \\
&\leq H(Y_2^T | \mathbf{G}^T) + F\gamma_F \\
&\overset{(b)}{\leq} T(1 - \epsilon_2) + F\gamma_F,
\end{aligned} \tag{34}$$

where (a) follows from Fano's inequality and (b) follows in a manner similar to (30) and the independence of channel gains from files. Therefore, based on (34) as $F \to \infty$, we obtain the bound

$$\delta^*(\mu) \geq \frac{1}{1 - \epsilon_2}. \tag{35}$$

Combining (31), (33) and (35) yields the desired lower bound.

APPENDIX B
PROOF OF CONVERSE FOR PROPOSITION 2

Let us denote $\delta_C = T_C/F$ the normalized latency on the cloud-to-encoder 1 link and $\delta_E = T/F$ the normalized latency on the channel between encoders and decoders. We first observe that, following the same argument as in (32)-(35), we have the bound

$$\delta_E \geq \delta_0 \tag{36}$$

for any sequence of feasible policies. We now obtain a lower bound on both normalized delays δ_E and δ_C by observing

that a hypothetical decoder provided with the CSI vector $\mathbf{G}^T$, with the cached content V_1 and V_2 in (2), with the cloud-aided message U^{T_C}, and with the signal $\tilde{G}^T X_2^T$ described in Appendix A can decode both messages W_1 and W_2. Details are as follows:

$$\begin{aligned} 2F &= H(W_1, W_2) \\ &= I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, U^{T_C}, \mathbf{G}^T) \\ &\quad + H(W_1, W_2 | \tilde{G}^T X_2^T, V_1, V_2, U^{T_C}, \mathbf{G}^T) \\ &\overset{(a)}{\leq} I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, U^{T_C}, \mathbf{G}^T) + F\gamma_F \\ &= I(W_1, W_2; \tilde{G}^T X_2^T, V_1, V_2, U^{T_C} | \mathbf{G}^T) + F\gamma_F \\ &\overset{(b)}{\leq} H(V_1) + H(U^{T_C}) + H(\tilde{G}^T X_2^T | \mathbf{G}^T) + F\gamma_F \\ &\overset{(c)}{\leq} \mu F + T_C C + T(1-\epsilon_2^2) + F\gamma_F, \end{aligned} \tag{37}$$

where, as in Appendix A, γ_F indicates any function that satisfies $\gamma_F \to 0$ as $F \to \infty$. In above derivation, steps (a)-(b) follow as steps (a)-(b) in (29), where we note that the inequality $H(V_2|\tilde{G}^T X_2^T, \mathbf{G}^T) \leq F\gamma_F$ by Fano inequality, while (c) hinges on the cache constraint (3) and the bound $H(U^{T_C}) \leq \sum_{i=1}^{T_C} H(U_i) \leq T_C C$ due to the capacity constraint on the cloud-to-encoder 1 link. As $F \to \infty$, the inequality (37) yields the bound on the latency components δ_c and δ_E:

$$\frac{1-\epsilon_2^2}{C}\delta_E + \delta_C \geq \frac{2-\mu}{C}. \tag{38}$$

To complete the proof, we combine bounds (36) and (38) as follows.

- For $C \leq 1-\epsilon_2^2$, the bound (38), directly yields

$$\delta^*(\mu, C) = \delta_E + \delta_C \geq \delta_E + \frac{C}{1-\epsilon_2^2}\delta_C \geq \frac{2-\mu}{1-\epsilon_2^2}. \tag{39}$$

  DTB for $C = 0$.
- For $C \geq 1-\epsilon_2^2$, two scenarios are possible. If $\mu \leq \mu_0$, multiplying (36) by the positive coefficient $1-(1-\epsilon_2^2)/C$ and summing the result with (38), provides the corresponding result in (24). Instead, if $\mu \geq \mu_0$, from (24), we directly obtain $\delta^*(\mu, C) \geq \delta_E \geq \delta_0$.