Cloud-Aided Edge Caching with Wireless Multicast Fronthauling in Fog Radio Access Networks

Jeongwan Koh*, Osvaldo Simeone†, Ravi Tandon‡ and Joonhyuk Kang*
* Department of EE, Korea Advanced Institute of Science and Technology
Email: jw_koh@kaist.ac.kr, jhkang@ee.kaist.ac.kr
†CWiP, ECE Department, New Jersey Institute of Technology
Email: osvaldo.simeone@njit.edu
‡ Department of ECE, University of Arizona
Email: tandon@email.arizona.edu

***Abstract*—In this paper, we investigate the total delivery latency across the fronthaul and wireless segments of a Fog Radio Access Network (F-RAN) under the assumption that cloud processor and edge nodes (ENs) are connected by a multicast fronthaul link. The total delivery latency is assessed via the *Normalized Delivery Time* (NDT) metric which provides a high signal-to-noise ratio (SNR) measure of the relative delivery worst-case latency with respect to an interference-free system. We derive upper and lower bounds on the achievable NDT for a F-RAN with two ENs and two users as a function of cache and fronthaul resources. The upper bound is obtained by studying the NDT achieved by delivery strategies that encompass both coded and uncoded multicast strategies on the fronthaul. The lower bound is instead derived by leveraging information theoretic converse arguments. Upper and lower bounds are shown to coincide, hence characterizing the minimum NDT, for a large range of problem parameters. Among the conclusions of this study, we demonstrate that coded multicasting is not useful for reducing the NDT for the mentioned range of parameters.**

***Index Terms*—Fog-RAN, Edge caching, Information theory.**

I. Introduction

Video delivery currently accounts for the majority of wireless traffic and its relevance is predicted to increase over the next years [1], [2]. In order to reduce backhaul overhead, delivery latency and network congestion, it has been recently proposed to equip edge nodes (ENs), such as base stations, with local caches so as to store popular contents at base stations during off-peak traffic times [3]. Assuming that all popular files can collectively be cached at the ENs, references [4], [5] studied the high signal-to-noise ratio (SNR) performance of cache-aided cellular system in terms of degrees of freedom. When the set of popular files is large enough that it cannot be fully cached by the ENs, content must be download from cloud to ENs by leveraging transport links, known as fronthaul, between cloud and ENs. Cellular systems with both edge cache and cloud connection to the ENs were studied in [6], [7] by accounting for fronthaul capacity limitation. The resulting architecture is referred to as Fog-Radio Access Networks (F-RAN). Specifically, the total delivery latency was investigated assuming *dedicated fronthaul wired links* between the cloud processor and ENs. Caching and delivery strategies were identified that achieve the optimal high-SNR performance within a factor of two.

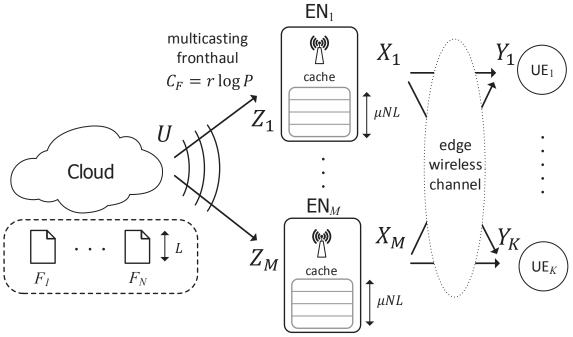

Fig. 1. Information-theoretic model of cloud-aided edge cache system, or F-RAN, with a multicast fronthaul.

Wired fronthauling, e.g., via fiber optic cables, becomes expensive as number of ENs increase. Therefore, an alternative option to use wireless fronthauling to reduce deployment costs, which is also being considered by industry [8], [9]. A key aspect of wireless fronthauling is the possibility to multicast messages from cloud to each ENs. Multicasting enables novel transmission strategies such as coded multicasting [10].

In this paper, we consider a F-RAN architecture with wireless multicast fronthauling as illustrated in Fig. 1. We investigate the total delivery latency over the fronthaul and wireless segments by adopting the high-SNR performance metric of *Normalized Delivery Time* (NDT) introduced in [6], [7]. We specifically derive upper and lower bounds on the achievable NDT for two ENs and two users as a function of cache and fronthaul resources. The upper bound is obtained by studying the NDT achieved by delivery strategies that encompass both coded and uncoded multicast strategies on the fronthaul, while the lower bound is derived through novel information theoretic arguments.

The rest of the paper is organized as follows: We describe the system model in Sec. II. We study the performance of various baseline caching-fronthaul transmission policies in Sec. III and of conventional and per-block time-sharing in Sec. IV. Then, we partially characterize the minimum NDT in Sec. V, and we conclude the paper in Sec. VI.

II. System model

A. F-RAN with Multicast Fronthaul

Fig. 1 describes the considered information-theoretic model of an F-RAN system, that is, a cloud-aided edge caching system with an out-of-band wireless multicast fronthaul link between cloud and ENs. The system includes a cloud processor, M ENs, each with average transmit power constraint P, as well as K user equipments (UEs). We assume that the cloud can access a content server storing a library $\mathcal{F} = \{F_1, \ldots, F_N\}$ of N popular files, each of L bits, i.e., $H(F_i) = L$. The library is assumed to be static during many transmission intervals. All files are assumed to be equally popular, as it conventionally done in related analyses [6]. Each EN can cache μL bits from each file in the library, where μ, with $0 \leq \mu \leq 1$, denotes the fractional cache size.

Time is organized into transmission intervals. At each transmission interval, every user k request a file $F_{D_k} \in \mathcal{F}$, where $D_k \in \{1, \ldots, N\}$. Furthermore, at any transmission interval, the received signal by UE$_k$ in a channel use $t = 1, \cdots, T$ of the downlink edge channel is given as

$$Y_k(t) = \sum_{m=1}^{M} H_{m,k} X_m(t) + N_k(t), \qquad (1)$$

where $H_{m,k}$ is the complex wireless channel gain between EN$_m$ and UE$_k$, which is constant in each transmission interval, $X_m(t)$ denotes the transmitted symbol by EN$_m$, and $N_k(t) \sim \mathcal{CN}(0,1)$ is the additive white noise process at UE$_k$. We have the power constraint

$$T^{-1} \sum_{t=1}^{T} |X_m(t)|^2 \leq P. \qquad (2)$$

Based on the received signal (1) each kth users produces on estimate $\hat{F}_{D_k}$ of the requested file F_{D_k}.

The ENs transmit to the UEs based on the content of their caches and on the signal received on the fronthaul link. The received signal at EN$_m$ on the fronthaul link from the cloud at each channel use t can be written as

$$Z_m(t) = G_m U(t) + W_m(t), \qquad (3)$$

where G_m denotes the wireless channel between cloud and EN$_m$, $U(t)$ is the signal transmitted by the cloud in channel use t and the white additive noise $W_m(t)$ has i.i.d $\mathcal{CN}(0,1)$ entries. The cloud has a power constraint

$$T^{-1} \sum_{t=1}^{T} |U(t)|^2 \leq P^r, \qquad (4)$$

where $r \geq 0$ describes the power scaling of the fronthaul transmission as compared to wireless edge transmission in a manner akin to the parameterization used in the analysis of the generalized degrees of freedom [11]. We emphasize that prior works [6], [7] assumed orthogonal dedicated (wired) fronthaul link between cloud and ENs.

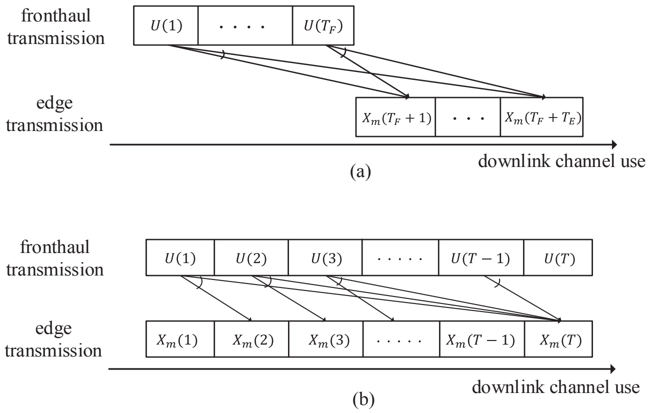

Fig. 2. (a) Serial fronthaul-edge transmission policies. (b) Parallel fronthaul-edge transmission policies. Edge transmission at channel use t depends on fronthaul transmission from channel use 1 to $(t-1)$.

B. Delivery Policy

We now define the general form of a caching-fronthaul-transmission delivery policy $f = (f_C, f_F, f_E)$ used to operate the described F-RAN system. As introduced above, we denote time within each transmission interval by using the index $t = 1, \ldots, T$ that runs over the channel uses of the downlink edge channel, where T is the overall delivery latency.

1) Caching policy: The caching policy is characterized by an encoding function $f_C : \mathcal{F} \rightarrow \{S_1, \ldots, S_M\}$, where S_m denotes cache contents stored at EN$_m$. We assume that the cache content S_m of each EN$_m$ is divided into N independent sub-cache contents as $S_m = \{S_{m,1}, \ldots, S_{m,N}\}$, where $S_{m,n}$ denotes any arbitrary function of file F_n. In order not to exceed the cache storage capacity of each EN$_m$, we have the constraint $H(S_{m,n}) \leq \mu L$, i.e., no more than μL bits can be stored for each file. The cache contents are kept fixed for many transmission intervals, that is, for as long as the library of popular files does not change.

2) Fronthaul policy: In each transmission interval, the cloud transmits a signal $\{U(1), \ldots, U(T)\}$ to the ENs based on the specific demand vector $D = \{D_1, \ldots, D_M\}$ of the users and of the current CSI $\{G_m\}$ and $\{H_{m,k}\}$. We express the encoding function of fronthaul policy as $f_F : \{\mathcal{F}, D, \{G_m\}, \{H_{m,k}\}\} \rightarrow U^T$, where U^T denotes the fronthaul message sent on the multicast fronthaul with power constraint (4).

3) Edge transmission policy: As illustrated in Fig. 2, each transmission interval, any EN$_m$ transmits using a policy $f^t_{E_m} : \{S_m, D, \{G_m\}, \{H_{m,k}\}, U^{t-1}\} \rightarrow X_m(t)$, which maps the fronthaul messages $U^{t-1} = \big(U(1), \ldots, U(t-1)\big)$ received prior to time t on the fronthaul link, as well as the cache content S_m and CSI, to the symbol $X_m(t)$ transmitted by EN$_m$ at symbol t under the power constraint (2). Note that the assumed set-up allows for parallel fronthaul-edge transmission (See [7]). As we depicted in Fig. 2, unlike the serial fronthaul-edge transmission, the parallel fronthaul-edge transmission allows to edge transmission at channel use t depend on fronthaul transmission from channel use 1 to $(t-1)$.

C. Normalized Delivery Time (NDT)

We now introduce the NDT performance metric [7]. We say that a sequence of policies $f=(f_C,f_F,f_E)$ is feasible if the probability of error $P_e=\mathbb{P}(\{\hat{F}_{D_k}\neq F_{D_k}\})$ goes to 0 with probability 1 with respect to the channel realization when $L\rightarrow\infty$ for all possible requests vectors $D\in\mathcal{F}^k$.

Definition 1. (NDT) For a given sequence of feasible policies $f=(f_C,f_F,f_E)$ indexed by the file length L, the average delivery time per bit is defined as

$$\Delta(\mu,C_F,P)=\limsup_{L\rightarrow\infty}\frac{T}{L}. \tag{5}$$

Furthermore, the NDT

$$\delta(\mu,r)=\lim_{P\rightarrow\infty}\frac{\Delta(\mu,C_F,P)}{1/\log(P)} \tag{6}$$

is said to be achievable for any achievable sequence $\Delta(\mu,C_F,P)$ as a function of P. The minimum NDT $\delta^*(\mu,r)$ is the infimum of (6) over all delivery policies.

According to Definition 1, the NDT measures the normalized delivery time with respect to that of a baseline system with unlimited caching and no mutual interference, whose delivery time per bits at high SNR is given by $1/\log(P)$. Therefore, NDT $\delta=1$ corresponding to the delivery time of an ideal interference-free system.

III. Delivery Strategies

In this section, we study the performance of various baseline caching-fronthaul transmission policies. We proceed by first considering cache-aided policies that disregard edge caching (Sec. III-A), then covering strategies based only on edge processing (Sec. III-B), and then studying techniques that leverage both edge and cloud processing by means of coded multicasting (Sec. III-C), pipelining (Sec. III-D) or time-sharing (Sec. IV-A). Throughout this paper, we consider the case $M=K=2$ and $N\geq K$ files.

A. Cache-Aided Polices

Here, following [7, Sec. IV], we review policies that deliver content based only on edge caching, without using the connection to the cloud.

1) Cache-Aided Zero-Forcing (ZF) Beamforming: When $\mu=1$, each EN has the entire file library in the cache. Therefore, any request vector D of the UEs can be supported with no mutual interference by means of cooperative ZF beamforming at the ENs. As a result the NDT $\delta=1$ is achievable. For reference, we summarize the performance of cache-aided ZF beamforming as

$$\mu=1;\quad\delta_F=0;\quad\delta_E=1, \tag{7}$$

for a total NDT $\delta=1$, where δ_F and δ_E measure the normalized delivery times on fronthaul and edge segments, respectively, to be formally defined below in (9). Note that we have $\delta_F=0$ since the fronthaul segment is not used.

2) Cache-Aided Interference Alignment (IA): When $\mu=1/2$, all ENs can cache half of each file. Any file F_n can hence be divided into two non-overlapping sub-files as $F_n=(F_{n,1},F_{n,2})$, each of size $L/2$ bits, where subfile $F_{n,m}$ is cached at EN$_m$. The stored contents in the cache of EN$_m$ are then $\{F_{n,m}\}_{n=1}^N$. For any pair of requested files, the ENs can adopt the IA scheme proposed in [12] for the X-channel which provides a high-SNR rate of around $2/3\log(P)$ for each user, so that an NDT $\delta=3/2$ is achievable by the definition (5). We summarize the performance of cache-aided IA as:

$$\mu=\frac{1}{2};\quad\delta_F=0;\quad\delta_E=\frac{3}{2}, \tag{8}$$

for a total NDT of $\delta=3/2$.

B. Serial Cloud-Aided Policies

Here, we consider cloud-aided policies, that is, policies that neglect the cached contents and perform delivery based on the fronthaul connection. Specifically, we study delivery policies whereby fronthaul transmission is followed by edge transmission in a serial manner as illustrated in Fig. 2 (a). To elaborate, it is convenient to define T_F and T_E as the duration of fronthaul and edge transmission as illustrated in Fig. 2 (b) and to introduce the fronthaul NDT and edge NDT, respectively, as

$$\delta_F=\lim_{P\rightarrow\infty}\frac{T_F\log(P)}{L}\quad\text{and}\quad\delta_E=\lim_{P\rightarrow\infty}\frac{T_E\log(P)}{L} \tag{9}$$

for given feasible policies. We discuss separately hard and the soft-transfer fronthauling schemes described in [7, Sec. IV] for the case of dedicated fronthaul link.

1) Hard-Transfer Fronthauling: With hard-transfer fronthauling, the two requested files are transmitted in raw form on the fronthaul link. In the worst case in which the requested files in D are distinct, this requires to send $2L$ bits on the multicast fronthaul link which results in $T_F=2L/\big(r\log(P)\big)$. Note that this is because, due to the power constraint (2), the fronthaul channel carries $r\log(P)$ bits per channel use to both ENs in the high-SNR regime.

As a result of the fronthaul transmission, the ENs can communicate without mutual interference to the UEs by means of ZF beamforming. Hence, from (9), the performance of hard-transfer fronthauling can be summarized as

$$\mu=0;\quad\delta_F=\frac{2}{r};\quad\delta_E=1. \tag{10}$$

2) Soft-Transfer Fronthauling: With soft-transfer fronthauling, ZF beamforming is carried out at the cloud, which quantizes the encoded baseband samples with a properly chosen quantization rate following the showed C-RAN approach. Specifically, as discussed in [7, Sec. IV], a resolution of $\log(P)$ bits per sample is needed in order for the quantization noise not to limit the performance, and the performance of the soft-transfer fronthauling is given as

$$\mu=0;\quad\delta_F=\frac{2}{r};\quad\delta_E=1, \tag{11}$$

for a total NDT of $\delta = 2/r + 1$. Comparing (10) and (11), we note that, unlike in the case with dedicated fronthaul link in [7], in which soft-transfer fronthauling outperform hard-transfer fronthauling, with a multicast fronthaul, the two schemes have the same performance.

C. Coded Multicasting Fronthauling

The schemes studied in the previous subsection use the multicast fronthaul by partitioning its capacity between the two ENs. Here, instead, we explore the potential benefits of coded multicasting whereby the cloud transmits coded files on the fronthaul link to the ENs in order to make use of the side information at the ENs due to the cached content. We emphasize that coded multicasting is not applicable in the F-RAN model in [7], in which there are dedicated fronthaul links from the cloud to the ENs. We focus on the case $\mu = 1/2$, since other value of μ can be dealt with by time-sharing as discussed Sec. IV-A.

When $\mu = 1/2$, each file F_n is partitioned into two parts as $\{F_{n,1}, F_{n,2}\}$, where subfile $F_{n,m}$ is cached the EN$_m$, and the length of the each subfiles is $L/2$. Each EN$_m$ hence caches subfiles $\{F_{n,m}\}_{n=1}^{N}$ as for the cache-aided IA scheme discussed in Sec. III-A2. Considering the worst-case in which two different file F_i and F_j are requested from the users, the cloud transmits the coded files $(F_{i,2} \oplus F_{j,1})$ and $(F_{j,2} \oplus F_{i,1})$ on the fronthaul in order to leverage the side information on the multicast fronthaul link. The total number of bits transmitted on the fronthaul link is $2 \times L/2 = L$, which yields a fronthaul latency $T_F = L/(r \log(P))$.

Furthermore, as a result of the fronthaul transmission, the ENs can support delivery to the UEs with no mutual interference by means of ZF beamforming. As a result, the performance of the coded multicasting fronthauling is summarized as

$$\mu = \frac{1}{2}; \quad \delta_F = \frac{1}{r}; \quad \delta_E = 1, \tag{12}$$

for a total NDT of $\delta = 1/r + 1$.

Comparing the NDT $\delta = 1/r + 1$ with the NDT achievable by cache-aided IA, namely $\delta = 3/2$ (see (8)), which requires the same cache capacity μ, we see that the coded multicasting fronthauling outperforms IA for $2/3 \le r < 2$. This confirms the potential benefits of coded multicasting fronthauling.

D. Pipelined Fronthaul-Edge Transmission via Block-Markov Coding

The strategies discussed above either do not use the fronthaul, as the cache-aided policies studied in Sec. III-A, or they use fronthaul and edge channels in a serial manner as discussed in Sec. III-B and Sec. III-C (See Fig. 2). Here, we show how to convert serial policies into pipelined policies that leverage the possibility to communicate simultaneously on fronthaul and edge segments via block-Markov coding. To elaborate, following [7, Sec. VII], we partition each file in the library into B blocks, each of size L/B bits. We also divide the overall delivery time T into $B+1$ slots, each of duration $T/(B+1)$. Fixing a given serial policy, for each slot b, the cloud operates the fronthaul as in the selected policy to transmit the bth block, while the ENs use the edge channel according to the given policy to transmit the $(b-1)$th block based on the signal received on the fronthaul on the $(b-1)$th block. As shown in [7, Sec. VII], the resulting NDT is given as

$$\delta = \max\big(\delta_F, \delta_E\big), \tag{13}$$

where δ_F and δ_E are the NDTs (9) of the selected serial policy. Based on (13), the NDT that can be achieved by the schemes described above when implemented in a pipelined fashion can be readily computed by using (7), (8), (10), (11) and (12). We finally note that, as shown in [7, Sec. VII], pipelining can reduce the NDT of serial policies by a factor of at most 2.

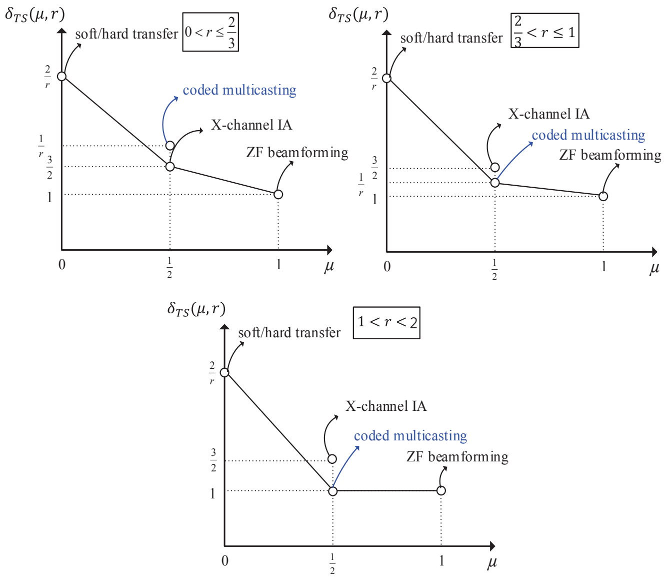

Fig. 3. NDT achievable by conventional time-sharing for M = 2-ENs, K = 2-users as a function of μ. We only consider the regime of $0 < r < 2$, since $\delta^*(\mu, r) = 1$ for $r > 2$.

IV. Time Sharing

In this section, we discuss how different policies can be combined by means of time-sharing in order to obtain achievable NDTs for all values of μ. We first analyze the NDT of conventional time-sharing and then we discuss the NDT of a more sophisticated form of time-sharing, namely per-block time sharing,

A. Conventional Time-Sharing

Fix any two feasible policies, f_1 and f_2, with NDTs δ_1 and δ_2 and required fractional cache size μ_1 and μ_2. If the fractional cache size is $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ with $\alpha \in [0, 1]$, using policy f_1 is for an α-fraction of the library while f_2 is applied for the remaining part, according to [7, Sec. II], we achieve the NDT

$$\delta = \alpha\delta_1 + (1-\alpha)\delta_2. \tag{14}$$

The next proposition describes the minimum NDT that can be achieved by means of conventional time-sharing among all

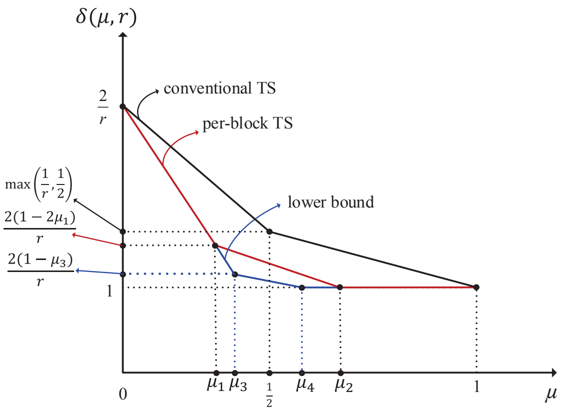

Fig. 4. NDT achievable via per-block time-sharing for $M = 2$-EN and $K = 2$-user as a function of μ along with a lower bound on the minimum NDT. ($\mu_1 = (2-r)/(4+r), \mu_2 = 1 - r/2, \mu_3 = 2/(3r+4)$, and $\mu_4 = 1 - r$.)

strategies that are introduced above. Note that, since pipelining is beneficial for all schemes, except for cache-aided strategies, the pipelined version of all strategies is implicitly adopted in the following.

Lemma 1 (Achievable NDT via conventional time-sharing)**.** *With $M = 2$-ENs, $K = 2$-UEs, $N \geq K$ files and $r > 0$, the following NDT is achievable by conventional time-sharing:*

- *Low fronthaul* - $0 < r \leq \dfrac{2}{3}$:

$$\delta_{TS}(\mu, r) = \begin{cases} \mu(3 - \frac{4}{r}) + \frac{2}{r}, & \text{for } 0 \leq \mu < \frac{1}{2}, \\ 2 - \mu, & \text{for } \frac{1}{2} \leq \mu \leq 1. \end{cases} \quad (15)$$

- *Intermediate fronthaul 1* - $\dfrac{2}{3} < r \leq 1$:

$$\delta_{TS}(\mu, r) = \begin{cases} -\frac{2}{r}\mu + \frac{2}{r}, & \text{for } 0 \leq \mu < \frac{1}{2}, \\ 2\mu(1 - \frac{1}{r}) + \frac{2}{r} - 1, & \text{for } \frac{1}{2} \leq \mu \leq 1. \end{cases} \quad (16)$$

- *Intermediate fronthaul 2* - $1 < r < 2$:

$$\delta_{TS}(\mu, r) = \begin{cases} 2(1 - \frac{2}{r})\mu + \frac{2}{r}, & \text{for } 0 \leq \mu < \frac{1}{2}, \\ 1, & \text{for } \frac{1}{2} \leq \mu \leq 1. \end{cases} \quad (17)$$

- *High fronthaul* - $r \geq 2$:

$$\delta_{TS}(\mu, r) = 1, \qquad \text{for } 0 \leq \mu \leq 1. \quad (18)$$

In (15), the NDT is achieved by time-sharing between hard/soft-transfer and X-channel IA for $0 \leq \mu < 1/2$; and X-channel IA and cache-aided ZF for $1/2 \leq \mu \leq 1$. In (16) and (17), the NDT is achieved by time-sharing between hard/soft transfer and coded multicasting for $0 \leq \mu < 1/2$; and coded multicasting and cache-aided ZF for $1/2 \leq \mu \leq 1$. For high fronthaul regime, the NDT is obtained by hard/soft transfer.

Proof. The proof of Lemma 1 is based on combining all pairs of schemes whose NDTs are given by (14) with (7), (8), (10), (11) and (12) and choosing in each fronthaul regime the pair that yields the minimum NDT. □

We illustrate Lemma 1 in Fig. 3, where the solid line denotes the NDT described in the proposition and circles are used to mark the NDT of the constituent schemes. It is observed that the coded multicasting fronthauling provides the minimum NDT for the regime $2/3 \leq r < 2$ of intermediate fronthaul capacity.

B. Per-Block Time Sharing

We now consider per-block time sharing first proposed in [7], whereby between two policies are carried out on a per-block basis. This is in the sense that each block is divided into two parts, which are allocated to the two strategies. As shown in [7, Appendix IX], if we denote as NDT are $\delta_{F,1}$ and $\delta_{E,1}$ and $\delta_{F,2}$ and $\delta_{E,2}$ the NDTs achievable by the two policies to be used for per-block time-sharing, the resulting NDT is given as

$$\delta = \max\left(\alpha\delta_{F,1} + (1-\alpha)\delta_{F,2},\ \ \alpha\delta_{E,1} + (1-\alpha)\delta_{E,2}\right). \quad (19)$$

The following proposition expresses the minimum NDT obtained by means of per-block time-sharing based on the strategies discussed above.

Proposition 1 (Achievable NDT via per-block time-sharing)**.** *With $M = 2$-ENs, $K = 2$-UEs, $N \geq K$ files and $0 < r < 2$, the minimum NDT via per-block time-sharing is*

$$\delta_{TS-B}(\mu, r) = \begin{cases} \frac{2(1-2\mu)}{r}, & \text{for } 0 \leq \mu \leq \mu_1 = \frac{2-r}{4+r}, \\ \frac{2(2-\mu)}{2+r}, & \text{for } \mu_1 < \mu \leq \mu_2 = 1 - \frac{r}{2}, \\ 1, & \text{for } \mu_2 < \mu \leq 1. \end{cases} \quad (20)$$

In (20), the NDT is obtained by time-sharing between hard/soft transfer and X-channel IA for $0 \leq \mu \leq \mu_1$; and X-channel IA and cache-aided ZF for $\mu_1 < \mu \leq \mu_2$; and hard/soft transfer and cache-aided ZF for $\mu_2 < \mu \leq 1$.

Proof. The NDT in Proposition 1 is obtained by considering all pairs of scheme, where NDTs are given by (7), (8), (10), (11) and (12), and choosing the pair that yields the minimum NDT for any given value of μ. In particular, we observe that the value μ_1 is obtained by equality $2(1-2\mu_1)/r = 1 + \mu_1$ achievable by per-block time-sharing between hard/soft transfer fronthauling and X-channel IA, and the value μ_2 is obtained by equality $2(1-\mu_2) = 1$ achievable by per-block time-sharing between X-channel IA and cache-aided ZF. □

We illustrate Proposition 1 in Fig. 4. Comparing with the NDT of conventional time-sharing, it is seen that per-block time-sharing decreases the NDT. Furthermore, we emphasize that the NDT in Proposition 1 is obtained by per-block time-sharing between pairs of techniques that do not include the coded multicasting strategy introduced in Sec. III-C.

V. Minimum NDT

In this section, we partially characterize the minimum NDT.

Proposition 2 (Minimum NDT)**.** *With $M = 2$-ENs, $K = 2$-UEs, $N \geq K$ files and $0 < r < 2$, the minimum NDT satisfies*

$$\delta^*(\mu,r)=\delta_{TS-B}(\mu,r),\ \ \textit{for}\ \ 0\le\mu\le\mu_1,\qquad\qquad \textit{and}\ \mu_2\le\mu\le 1,\ \ (21)$$

$$\begin{cases} \frac{2(1-2\mu)}{r}\le\delta^*(\mu,r)\le\delta_{TS-B}(\mu,r), & \textit{for}\ \ \mu_1<\mu\le\mu_3,\\ \frac{2-\mu}{1+r}\le\delta^*(\mu,r)\le\delta_{TS-B}(\mu,r), & \textit{for}\ \ \mu_3<\mu\le\mu_4,(22)\\ 1\le\delta^*(\mu,r)\le\delta_{TS-B}(\mu,r), & \textit{for}\ \ \mu_4<\mu\le\mu_2, \end{cases}$$

where $\delta_{TS-B}(\mu,r),\mu_1$ and μ_2 are defined in Proposition 1, and we have $\mu_3=2/(3r+4)$ and $\mu_4=1-r$.

Proof. The proof is based on combining (20) with the lower bound

$$\delta^*(\mu,r)\ge\max\left(\frac{2(1-2\mu)}{r},\frac{2-\mu}{1+r},1\right),\qquad (23)$$

which is proved in Appendix A. μ_3 and μ_4 is obtained from $2(1-2\mu_3)/r=(2-\mu_3)/(1+r)$ and $(2-\mu_4)/(1+r)=1$. □

Proposition 2 is illustrated in Fig. 4, which shows that the achievable NDT in Proposition 1 provides the optimal NDT except in the interval of fractional cache capacity value $\mu_1<\mu<\mu_2$, which decreases as $r\to 0$. From Proposition 2, recalling that the NDT in Proposition 1 does not require coded multicasting, we can conclude that except in the range $\mu_1<\mu<\mu_2$ coded multicasting is not required to obtain the minimum NDT.

VI. Conclusions

In this paper, we investigated the total delivery latency over fronthaul and wireless link in a F-RAN with a wireless multicast fronthaul. Specifically, we studied the minimum delivery latency as a function of cache and fronthaul resources by deriving upper and lower bounds on the minimum NDT, which is a high-SNR measure of content delivery. Among the main conclusions, we have observed that, unlike for the receiver-side caching problem [10], coded multicasting is not useful to reduce the NDT for a large range of system parameters in the presence of two users and two ENs.

References

[1] Cisco, "Global mobile data traffic forecast update, 2013–2018," *White paper*, 2014.

[2] Ericsson, "5G radio access - research and vision," *White Paper, [Online] http://goo.gl/Hug0b6*, 2012.

[3] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *in Proc. IEEE INFOCOM*, 1107–1115, Mar. 2012.

[4] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," *in Proc. IEEE Intern. Symposium on Information Theory (ISIT)*, 809–813, Oct. 2015.

[5] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *arXiv preprint arXiv:1602.04207*, 2016.

[6] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Intern. Symposium on Information Theory (ISIT)*, 2016, [Online] Available: http://arxiv.org/abs/1605.01690.

[7] A. Sengupta, R. Tandon, and O. Simeone, "Cloud and cache-aided wireless networks: Fundamental latency trade-offs," *arXiv:1605.01690*, 2016.

[8] NGMN Alliance, "Further study on critical C-RAN technologies," *[Online] Available: http://www.ngmn.org*, Mar. 2015.

[9] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks—a technology overview," *IEEE Commu. Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, Sept. 2015.

[10] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. on Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, Mar. 2014.

[11] S. A. Jafar and S. Vishwanath, "Generalized degrees of freedom of the symmetric gaussian k user interference channel," *in Proc. IEEE Intern. Symposium on Information Theory (ISIT)*, Aug. 2012.

[12] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. on Inform. Theory*, vol. 54, no. 8, pp. 3457–3470, July 2008.

Appendix A: Lower Bound on NDT (Converse)

Here we provide a sketch of the proof of (23). As in [6], [7], the converse is based on considering subsets of information resources such that the information in each subset is sufficient to decode the requested files for any feasible policies in the high SNR regime. Throughout this appendix, we denote ϵ_L and ϵ_P as any function that satisfies $\epsilon_L\to 0$ for $L\to\infty$ and $\epsilon_P/\log(P)\to 0$ for $P\to\infty$. We first consider the resource subset $\{Y_1^T,S_2,Z_2^T\}$ and we write

$$\begin{aligned} 2L &= H(F_1,F_2) && \text{(24a)}\\ &= I(F_1,F_2;Y_1^T,S_2,Z_2^T)+H(F_1,F_2|Y_1^T,S_2,Z_2^T) && \text{(24b)}\\ &\le T\log(P)+T\epsilon_P+\mu L+L\epsilon_L+h(Z_2^T|F_1) && \text{(24c)}\\ &\le T\log(P)+T\epsilon_P+\mu L+L\epsilon_L+h(U_2^T) && \text{(24d)}\\ &= T\log(P)+T\epsilon_P+\mu L+L\epsilon_L+Tr\log(P), && \text{(24e)} \end{aligned}$$

where (24c) is derived as [6, Eq. (8a)-(8d), (9a)-(9c)] ; (24d) follows from $h(Z_2^T|F_1)=h(G_2U^T+W_2^T|F_1)\le h(U^T|F_1)+T\epsilon_P\le h(U^T)+T\epsilon_P$. Next, we consider the subset $\{S_1,S_2,U^T,Z_1^T,Z_2^T\}$ and obtain

$$\begin{aligned} 2L &= H(F_1,F_2) && \text{(25a)}\\ &= I(F_1,F_2;S_1,S_2,U^T,Z_1^T,Z_2^T)\\ &\qquad +H(F_1,F_2|S_1,S_2,U^T,Z_1^T,Z_2^T) && \text{(25b)}\\ &\le h(U^T)+H(S_1)+H(S_2)+L\epsilon_L && \text{(25c)}\\ &\le Tr\log(P)+4\mu L+L\epsilon_L, && \text{(25d)} \end{aligned}$$

where (25c) follows as in [6, Eq. (11a) and (11b)] . Lastly, by Fano's inequality, we obtain

$$L=H(F_1)\le I(F_1;Y_1^T)+L\epsilon_L\le T\log(P)+L\epsilon_L.\qquad (26)$$

Combining (24e), (25d), and (26) and using following Definition 1 completes the proof.

Acknowledgment

This work was supported by ICT R&D program of MSIP/IITP. [B0114-16-0009, A research on a novel communication system using storage as wireless communication resource]. The work of O. Simeone was partially supported by U.S. NSF via grant CCF-1525629. The work of R. Tandon was supported by the U.S. NSF via grant CCF-1559758.