Pipelined Fronthaul-Edge Content Delivery in Fog Radio Access Networks

Avik Sengupta
Hume Center, Department of ECE
Virginia Tech,
Blacksburg, VA 24060, USA
Email: aviksg@vt.edu

Ravi Tandon
Department of ECE
University of Arizona,
Tucson, AZ 85721 USA
Email: tandonr@email.arizona.edu

Osvaldo Simeone
CWiP, Department of ECE
New Jersey Institute of Technology,
Newark, NJ 07102 USA
Email: osvaldo.simeone@njit.edu

***Abstract*—In a Fog Radio Access Network (F-RAN), content delivery is carried out using both edge caching and cloud processing. A key design question for F-RANs hence concerns the optimal use of edge and cloud resources. In this work, this problem is addressed from an information theoretic viewpoint by investigating the fundamental limits of the normalized delivery time (NDT) metric, which captures the high signal-to-noise ratio (SNR) worst-case latency for delivering any requested content to the users. Specifically, unlike prior work, the NDT performance of an F-RAN is studied under *pipelined* fronthaul-edge transmission, whereby edge nodes are capable of simultaneously receiving fronthaul messages from the cloud on fronthaul links while transmitting to the mobile users over the wireless edge channel. Lower and upper bounds on the NDT are derived that yield insights into the trade-off between cache storage capacity, fronthaul capacity and delivery latency and on the impact of fronthaul-edge pipelining.**

I. Introduction

The Fog Radio Access Network (F-RAN) architecture harnesses the dual benefits of centralized cloud processing and of localized content caching at the network edge for content delivery [1]–[4] (see Fig. 1). Cloud processing enables the centralization of baseband functionalities from the base stations, or edge nodes (ENs), of a wireless system to a cloud processor. The latter has direct access to the content server but can communicate only with the ENs at the cost of the latency required for transmission on fronthaul links between the ENs and the cloud [5]. In a dual manner, edge caching allows low-latency content delivery without backhaul overhead, but only on content that was proactively stored at the ENs (see, e.g., [2], [3], [6]–[14]).

A key design question is how to operate the available resources at the cloud and at the edge in order to minimize the latency of content delivery. With cloud processing, cooperative transmission among ENs becomes possible, hence reducing the latency due to transmission on the edge, or the wireless channel but at the expense of adding the contribution of fronthaul latency. In contrast, the edge latency accrued by solutions based on edge caching is generally larger, due to the fact that cooperative transmission is limited to shared content in the caches of multiple ENs. However, no fronthaul latency is incurred. The goal of this paper is to address the interplay of cloud and edge processing from an information theoretic standpoint.

Interference-limited wireless systems for cache-aided content delivery were first investigated from an information-theoretic viewpoint in [8], where an upper bound on the

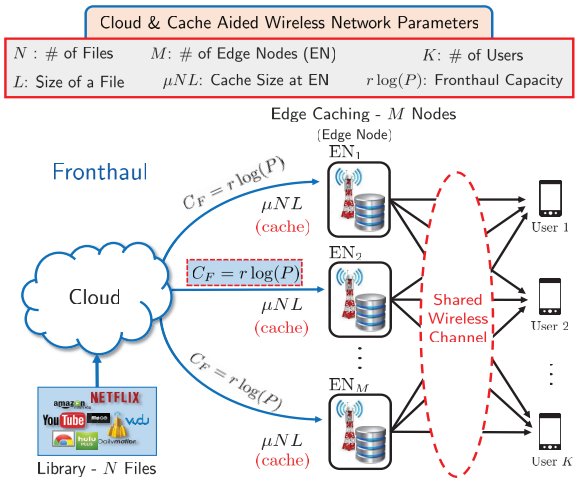


Fig. 1. Information-theoretic model for a cloud and cache-aided wireless system, referred to as Fog-Radio Access Network (F-RAN).

worst-case delivery latency, formulated in terms of degrees-of-freedom (DoF), is presented for $M = 3$ ENs and $K = 3$ users. Upper and lower bounds are derived in [10]–[12] by accounting for caching at both ENs and users. References [2], [3] instead consider both cloud processing and edge caching and present lower and upper bounds on the delivery latency in F-RANs, which is formalized in terms of a high Signal-to-Noise Ratio (SNR) metric defined as *Normalized Delivery Time* (NDT). In [2], [3], the operation of the F-RAN was assumed to follow a *serial* fronthaul-edge transmission schedule, where the ENs wait to receive the entire fronthaul transmission from the cloud before delivering content over the wireless edge.

The main premise of this work is that in practice, *pipelined* transmission across fronthaul and edge segments is possible and can further reduce the delivery latency in comparison to serial transmission (see e.g., [16]). However, no theoretical analysis of the advantages of pipelined fronthaul-edge transmission exists to the best of the authors' knowledge.

Example 1. To exemplify the analysis put forth in this paper, consider an F-RAN set-up in which two ENs are deployed to serve two users over a shared wireless channel. There is a library of $N \geq 2$ popular files of equal size and each EN can cache at most a fraction $\mu \in [0, 1]$ of the library content, where μ is the *fractional cache size*. The ENs are connected to the cloud via fronthaul links whose capacity scales with the SNR P of the wireless edge links as $r\log(P)$,

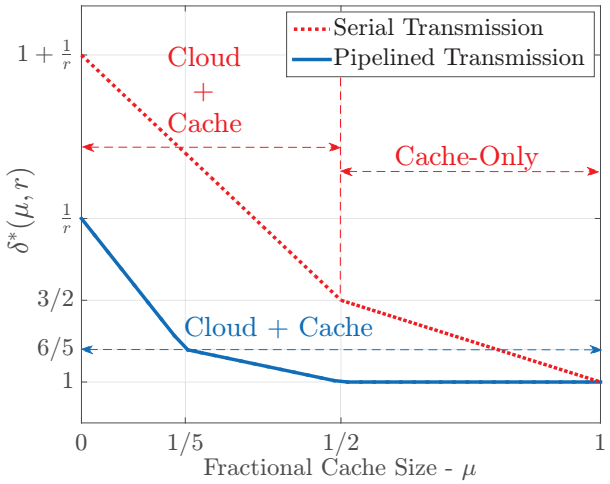


Fig. 2. Trade-off between normalized delivery time (NDT) and fractional cache size μ in the presence of full CSI at ENs, users and the cloud for $M = 2$ ENs and $K = 2$ users.

with $r \geq 0$ characterizing the fronthaul capacity. Full Channel State Information (CSI) is assumed as needed at all nodes. For this example, the *information-theoretically optimal* trade-off $\delta^*(\mu, r)$ between the latency metric NDT and the fractional cache size μ is shown in Fig. 2 for $r = 0.5$ for serial [2], [3] as well as for *pipelined* fronthaul-edge transmission, where the latter will be studied in this paper. The NDT measures the latency relative to an ideal system with unlimited caching and no interference, and hence we have $\delta^*(\mu, r) \geq 1$, with $\delta^* = 1$ corresponding to the performance of an ideal system.

With reference to this example, one of the results shown in this work reveals that, for pipelined fronthaul-edge transmission, cloud processing is instrumental in obtaining the minimum delivery latency for all values of μ, even when the fronthaul capacity is small. This is in contrast with serial fronthaul-edge transmission where, for low fronthaul gains ($r \leq 0.5$) and large enough cache capacity ($\mu \geq 1/2$), cloud-aided fronthaul transmission cannot improve the end-to-end latency [2], [3]. This is because, with pipelined transmission, the ENs need not wait for the fronthaul transmission to be completed before communicating to the users on the edge links, and hence the fronthaul latency contribution can be mitigated. For the same reason, pipelined fronthaul-edge transmission generally improves the NDT compared to serial transmission. In particular, even with partial caching, that is, with $\mu < 1$, the ideal NDT $\delta^* = 1$ is achievable with pipelined fronthaul-edge transmission, while this is not the case with serial transmission. □

Main Contributions: In this paper, general information-theoretic lower and upper bounds are derived on the NDT of F-RAN systems under pipelined fronthaul-edge operation. Leveraging these bounds, the optimal NDT is characterized to within a constant multiplicative factor of 2 for all values of problem parameters. Furthermore, using the developed bounds, as well as the results for serial fronthaul-edge tranmission presented in [2], [3], we highlight the improvement in NDT due to pipelining. Proofs of the main results are omitted for brevity and can be found in [3].

II. System Model

We consider an $M \times K$ F-RAN, shown in Fig. 1, where M ENs serve a total of K users through a shared wireless channel. The ENs can cache content from a library of N files, $F_1, \ldots, F_N$, where each file is of size L bits, for some $L \in \mathbb{N}^+$. The files F_n are assumed to be independent and identically distributed (i.i.d.) as:

$$F_n \sim \text{Unif}\left\{1, 2, \ldots, 2^L\right\}, \quad \forall n \in [1:N]. \tag{1}$$

Each EN is equipped with a cache in which it can store μNL bits, where the fraction μ, with $\mu \in [0, 1]$, is referred to as the *fractional cache size*. The cloud has full access to the library of N files, and each EN is connected to the cloud by a fronthaul link of capacity of C_F bits per symbol, where a symbol refers to a channel use of the downlink wireless, or edge, channel.

In a transmission interval, each user $k \in [1:K]$ requests one of the N files from the library. The demand vector is denoted by $\mathbf{D} \triangleq (d_1, \ldots, d_K) \in [1:N]^K$ and is known at the beginning of a transmission interval by both cloud and ENs. In this work, we assume *pipelined* or *parallel* operation of the fronthaul and wireless segments, whereby the ENs can simultaneously receive on fronthaul links and transmit on the wireless channel to the users. All the nodes have access to the global CSI about the wireless channels $\mathbf{H} = \left\{ \{h_{km}\} : \substack{k=[1:K] \\ m=[1:M]} \right\}$, where $h_{km} \in \mathbb{C}$, denotes the channel coefficient between user $k \in [1:K]$ and EN_m, $m \in [1:M]$. The coefficients are assumed to be drawn independent and identically distributed (i.i.d.) from a continuous distribution and to be time-invariant within each transmission interval. The design of the F-RAN entails the definition of caching and delivery policies, which are formalized next. Throughout, the time index t runs over the time intervals corresponding to channel uses of the edge channel.

Definition 1 (*Policy*)**.** A pipelined caching, fronthaul, edge transmission, and decoding policy $\pi_{\text{P}} = (\pi_c, \pi_f, \pi_e, \pi_d)$ is characterized by the following functions.

a) Caching Policy π_c: The caching policy at each edge node EN_m, $m \in [1:M]$, is defined by a function $\pi_c^m(\cdot)$ that maps each file F_n to its cached content $S_{m,n}$ as

$$S_{m,n} \triangleq \pi_c^m\left(F_n\right), \quad \forall n \in [1:N]. \tag{2}$$

The mapping is such that the entropy of each fractional component $H(S_{m,n}) \leq \mu L$ in order to satisfy the cache capacity constraints. The overall cache content at EN_m is given by $S_m = (S_{m,1}, S_{m,2}, \ldots, S_{m,N})$. Note that the caching policy π_c allows for arbitrary coding within each file, but it does not allow for inter-file coding. Furthermore, the caching policy is kept fixed over multiple transmission intervals and is thus agnostic to the demand vector $\mathbf{D}$ and the global CSI $\mathbf{H}$.

b) Fronthaul Policy π_f: A fronthaul policy is defined by a function $\pi_f(\cdot)$, which maps the set of files $F_{[1:N]}$, the demand vector $\mathbf{D}$ and CSI $\mathbf{H}$ to the fronthaul message

$$\mathbf{U}_m^T = (U_m[t])_{t=1}^T = \pi_f^m\left(\{F_{[1:N]}\}, \mathbf{D}, \mathbf{H}\right), \tag{3}$$

which is transmitted to EN_m via the fronthaul link of capacity C_F bits per symbol. In 3, T represents the total end-to-end delivery latency in the number of channel uses of the edge channel. The fronthaul message cannot exceed TC_F bits.

c) Edge Transmission Policy π_e: Each edge node EN_m starts transmitting at the beginning of the transmission interval using an edge transmission policy $\pi_e^m(\cdot)$, such that, at any time instant t, the EN maps the demand vector $\mathbf{D}$, the global CSI $\mathbf{H}$, the local cache content S_m and the fronthaul messages received up to time $t-1$, to the transmitted signal at time t as $\mathbf{X}_m^T = (X_m[t])_{t=1}^T$, where

$$X_m[t] = \pi_e^m\Big(S_m, U_m[1], U_m[2], \ldots, U_m[t-1], \mathbf{D}, \mathbf{H}\Big), \quad (4)$$

which is transmitted to the users on the shared wireless link. An average power constraint of P is imposed for each codeword $\mathbf{X}_m^T$. Note that, unlike the caching policy, π_c, the fronthaul policy, π_f and the edge transmission policy, π_e, can adapt to the instantaneous demands and CSI at each transmission interval.

d) Decoding Policy π_d: Each user $k \in [1:K]$, receives a channel output given by:

$$\mathbf{Y}_k^T = (Y_k[t])_{t=1}^T = \sum_{m=1}^{M} h_{km}\mathbf{X}_m^T + \mathbf{n}_k^T, \quad (5)$$

where the noise $\mathbf{n}_k^T = (n_k[t])_{t=1}^T$ is such that $n_k[t] \sim \mathcal{CN}(0,1)$ is i.i.d. across time and users. Each user $k \in [1:K]$, implements a decoding policy $\pi_d(\cdot)$, which maps the channel outputs, the receiver demands and the channel realization to the estimate

$$\widehat{F}_{d_k} \triangleq \pi_d^k\Big(\mathbf{Y}_k^T, d_k, \mathbf{H}\Big) \quad (6)$$

of the requested file F_{d_k}. The caching, fronthaul, edge transmission and decoding policies together form the policy $\pi_{\mathsf{P}} = (\pi_c^m, \pi_f^m, \pi_e^m, \pi_d^k)$ that defines the operation of the P-FRAN system. The probability of error of a policy π_{P} is defined as

$$P_e = \max_{\mathbf{D}} \max_{k\in[1:K]} \mathbb{P}\left(\widehat{F}_{d_k} \neq F_{d_k}\right), \quad (7)$$

which is the worst-case probability of decoding error measured over all possible demand vectors $\mathbf{D}$ and over all users $k \in [1:K]$. A sequence of policies, indexed by the file size L, is said to be *feasible* if, for almost all channel realizations $\mathbf{H}$, i.e., with probability 1, we have $P_e \to 0$ when $L \to \infty$.

We next define normalized delivery time (NDT) by first introducing the notion of delivery time per bit.

Definition 2 (*Delivery time per bit*)**.** A *delivery time per bit* $\Delta(\mu, C_F, P)$ is achievable if there exists a sequence of feasible policies such that

$$\Delta_{\mathsf{P}}(\mu, C_F, P) = \limsup_{L\to\infty} \frac{T}{L}. \quad (8)$$

The delivery time per bit accounts for the latency within each transmission interval. As in [2], [3], we next define a more tractable metric that reflects the latency performance in the high SNR regime. To this end, we let the fronthaul capacity scale with the SNR parameter P as $C_F = r\log(P)$, where r measures the multiplexing gain of the fronthaul links.

Definition 3 (*NDT*)**.** For any sequence of achievable $\Delta_{\mathsf{P}}(\mu, C_F, P)$ as function of P, with $C_F = r\log(P)$, the *normalized delivery time* (NDT), is defined as

$$\delta_{\mathsf{P}}(\mu, r) = \lim_{P\to\infty} \frac{\Delta_{\mathsf{P}}(\mu, r\log(P), P)}{1/\log P} \quad (9)$$

$$= \lim_{P\to\infty} \limsup_{L\to\infty} \frac{T}{L/\log P}. \quad (10)$$

Moreover, for any given pair (μ, r), the minimum NDT is defined as

$$\delta_{\mathsf{P}}^*(\mu, r) = \inf\left\{\delta_{\mathsf{P}}(\mu, r) : \delta_{\mathsf{P}}(\mu, r) \text{ is achievable}\right\}. \quad (11)$$

Remark 1 (*Operational significance of NDT*)**.** As introduced in [2], [3], to define the NDT, the delivery time per bit (8) is normalized by the term $1/\log P$. The latter is the delivery time per bit in the high SNR regime for an ideal baseline system with no interference and unlimited caching, in which each user can be served by a dedicated EN which has locally stored all the files. An NDT of δ^* hence indicates that the worst-case time required to serve any possible request vector $\mathbf{D}$ is δ^* times larger than the time needed by this ideal baseline system. ⋄

Lemma 1 (*Convexity of Minimum NDT*)**.** *The minimum NDT, $\delta_{\mathsf{P}}^*(\mu, r)$, is a convex function of μ for every value of $r \geq 0$.*

Proof. The proof follows from a *file-splitting and cache-sharing* argument, whereby files are split into two fractions, with the two fractions being served by different policies that share the cache resources and whose delivery times add up to yield the overall NDT. The proof of the Lemma is omitted for brevity and is provided in [3]. □

Remark 2 (*Pipelined vs. Serial Transmission*)**.** With the serial fronthaul-edge transmission policies discussed in [2], [3], in each transmission interval, the ENs can only begin transmission after the fronthaul message has been received in its entirety. This class of strategies is easily seen to be included as special a case in the class of pipelined fronthaul-edge transmission schemes introduced in Definition 1. As a result, the minimum NDT $\delta_{\mathsf{P}}^*(\mu, r)$ under pipelined operation can be no larger than that under serial operation. The following lemma bounds the improvement in NDT that can be achieved by the use of pipelining as opposed to serial transmission. ⋄

Lemma 2 (*Pipelined vs. Serial Fronthaul-Edge Transmission*)**.** *For an $M \times K$ cloud and cache-aided F-RAN, pipelined fronthaul-edge transmission can improve the minimum NDT as compared to serial transmission by a factor of at most 2, i.e.,*

$$\delta_{\mathsf{P}}^*(\mu, r) \geq \frac{\delta_{\mathsf{S}}^*(\mu, r)}{2}, \quad (12)$$

where $\delta_{\mathsf{S}}^*(\mu, r)$ denotes the serial NDT.

Proof. For the case of pipelined fronthaul-edge transmission, consider an optimal policy π_{P}^* that achieves the minimum NDT $\delta_{\mathsf{P}}^*(\mu, r)$. We use this policy π_{P}^* to construct a policy π under serial fronthaul-edge transmission as follows: the caching and fronthaul policies for π are the same as for π_{P}^*; and the edge-transmission policy for π is the same as for π_{P}^* with the caveat that the ENs start transmitting only after the fronthaul transmission is complete. The NDT $\delta_{\mathsf{S}}(\mu, r)$ achieved by the serial policy π is no larger than $2\delta_{\mathsf{P}}^*(\mu, r)$ since the durations of fronthaul and edge transmission for π_{P}^* are by definition of the NDT (11), both limited by $\delta_{\mathsf{P}}^*(\mu, r)$ when normalized by $L/\log(P)$ in the limit of large L and P. This concludes the proof. □

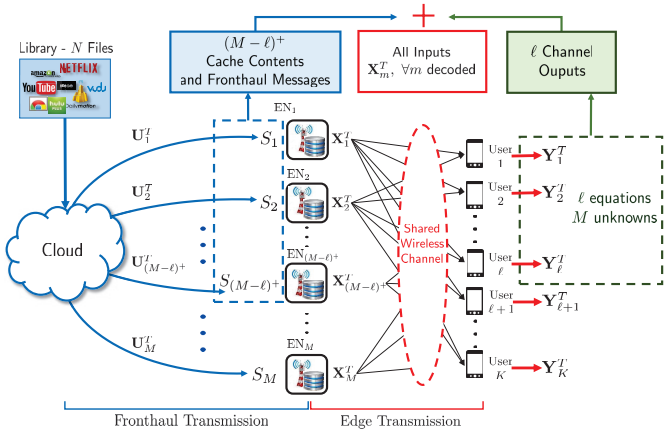


Fig. 3. Illustration of the F-RAN set-up for the proof of Proposition 1.

III. General Bounds on the Minimum NDT

In this section, we provide an information theoretic lower bound as well as an upper bound on the minimum NDT $\delta_{\mathsf{P}}^*(\mu, r)$. The proposed upper bounds are shown to be tight for some regimes of the fractional cache size μ. Since the model under study reduces to the cache-only system discussed in [3] when $r = 0$, we focus here only on the case of $r > 0$.

A. Lower Bound on the Minimum NDT

The following proposition provides a general lower bound on the minimum NDT for the $M \times K$ F-RAN with pipelined fronthaul-edge transmission.

Proposition 1. *For an F-RAN with M ENs, each with a fractional cache size $\mu \in [0, 1]$, K users, a library of $N \geq K$ files and a fronthaul capacity of $C_F = r \log(P)$ bits per symbol, the minimum NDT for pipelined fronthaul-edge transmission is lower bounded as*

$$\delta_{\mathsf{P}}^*(\mu, r) \geq \max\left\{ \max_{\ell \in [0:\min\{M,K\}]} \frac{K - (M-\ell)^+(K-\ell)^+\mu}{\ell + (M-\ell)^+ r}, \ 1 \right\}. \tag{13}$$

The proof of the first term inside the max function is based on a cut-set-like argument, which is illustrated in Fig. 3. Specifically, it can be argued that, for all sequence of feasible policies guaranteeing a vanishing probability of error, in the high-SNR regime, any K requested files must be decodable with low error probability from the received signal of ℓ users along with the cache contents and fronthaul messages of the remaining $(M-\ell)^+$ ENs. This is because, any $\ell \leq \min\{M, K\}$ received signals $\mathbf{Y}_{[1:\ell]}^T$ are functions of M channel inputs $\mathbf{X}_{[1:M]}^T$, which in turn are functions of the M user caches and their corresponding fronthaul messages $\mathbf{U}_{[1:M]}^T$. Thus, using these ℓ signals and the contents of $(M-\ell)^+$ caches, $S_{[1:(M-\ell)^+]}$ and associated fronthaul messages $\mathbf{U}_{[1:(M-\ell)^+]}^T$, all the inputs can be almost surely decoded using the invertible linear system of the form of (5), neglecting the noise in the high-SNR regime. The proposition is proved by carefully bounding the joint entropy of these random variables, which upper bounds the amount of information that can be reliably conveyed in the given time interval T or NDT $\delta_{\mathsf{P}}^*(\mu, r)$.

We also observe that the lower bound (13) is strictly smaller than the lower bound in [13, Theorem 1] derived under serial operation in accordance with the discussion in Remark 2. Next, we consider achievable schemes that yield upper bounds on the minimum NDT for the pipelined fronthaul-edge transmission model.

B. Upper Bounds on the Minimum NDT

We start the analysis of achievable schemes by considering strategies that operate under the serial fronthaul-edge transmission as introduced in Remark 2. We define as T_F and T_E as the number of channel uses in which a strategy uses the fronthaul, and edge channels respectively. Furthermore, we define $\delta_F = \lim_{P,L\to\infty} T_F \log(P)/L$ and $\delta_E = \lim_{P,L\to\infty} T_E \log(P)/L$ as the fronthaul and edge NDTs, respectively. Note that the achievable serial NDT is the sum of the fronthaul and edge NDTs, i.e., $\delta_{\mathsf{S,Ach}} = \delta_E + \delta_F$ [2], [3].

1) Cache-Aided EN Coordination via Interference Alignment: When each EN has fractional cache capacity $\mu = 1/M$, each file can be split into M non-overlapping fragments $F_n = (F_{n,1}, F_{n,2}, \ldots, F_{n,M})$, each of size L/M bits. The fragment $F_{n,m}$ is stored in the cache of EN_m for $n \in [1 : N]$. For any file d_k requested by a user k, each of the ENs has a fragment $F_{d_k,m}$ to transmit to the user, and, as a result, the $M \times K$ wireless edge becomes an X-channel, over which interference alignment (IA) yields [3, Sec. IV.A]

$$\delta_F = 0; \quad \delta_E = \delta_{\mathsf{Ca-IA}} = \frac{M + K - 1}{M}. \tag{14}$$

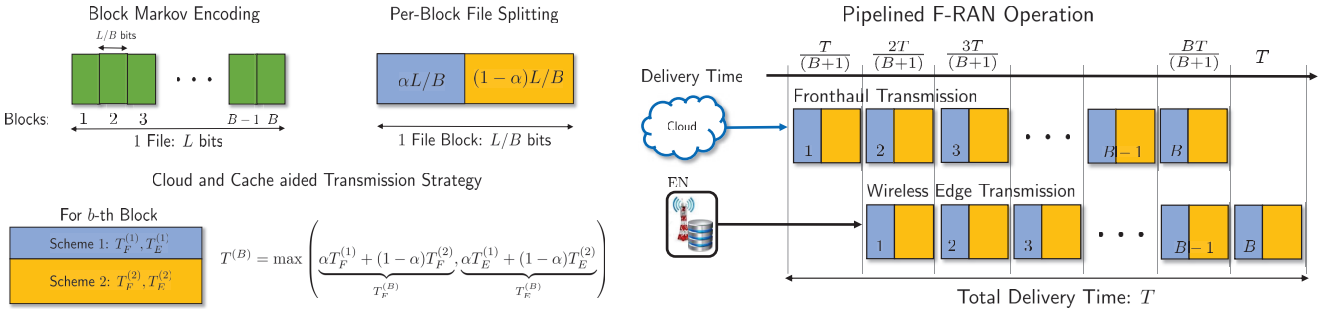
2) Cache-Aided EN Cooperation via Zero-Forcing Beamforming: When $\mu = 1$, each EN can store the entire library of N files and the resulting system can be treated as a multi-antenna broadcast channel with M co-located transmit antennas. Transmitter cooperation in the form of zero-forcing (ZF) beamforming can be carried out with high probability with respect to the channel realizations, yielding interference-free transmission to the K users and the achievable NDTs [3, Sec. IV.A]

$$\delta_F = 0; \quad \delta_E = \delta_{\mathsf{Ca-ZF}} = \frac{K}{\min\{M, K\}}. \tag{15}$$

3) Cloud-Aided Soft-Transfer Fronthauling: With soft-transfer fronthauling, as first proposed in [15], the cloud implements ZF-beamforming and quantizes the resulting encoded signals. Using a resolution of $\log(P)$ bits per downlink baseband sample, it can be shown that the effective SNR in the downlink scales proportionally to the power P. This scheme yields [3, Sec. IV.B]

$$\delta_F = \frac{(1-\mu)K}{Mr}; \quad \delta_E = \delta_{\mathsf{Ca-ZF}} = \frac{K}{\min\{M, K\}}. \tag{16}$$

As explained next, the proposed achievable scheme leverages pipelined fronthaul-edge transmission by means of *block-Markov encoding* to convert serial transmission policies to pipelined policies. We further integrate block-Markov encoding with *per-block file splitting* to time-share between two transmission policies within each block.



(a) (b)

Fig. 4. Pipelined F-RAN operation: (a) File-splitting and block Markov encoding using B blocks; file-splitting enables the use of two constituent schemes to deliver content; (b) pipelined transmission where a serial transmission strategy is used within each block.

• *Block-Markov Encoding*: To convert a serial policy into a pipelined policy, we split each file in the library into B blocks, so that each block is of size L/B bits. Correspondingly, we also divide the total delivery time T into $B+1$ slots, each of duration $T/(B+1)$. In each slot $b \in [1:B]$, the cloud operates the fronthaul according to the serial policy to deliver the bth blocks of the requested files, while the ENs apply the corresponding edge delivery policy to deliver the $(b-1)$th blocks of the requested files, as illustrated in Fig. 4(b).

Let $T_F^{(B)}$ denote the per-block fronthaul time and $T_E^{(B)}$ denote the per-block edge time required by the selected policies in each block. These times are related to the total fronthaul and edge delivery times T_F and T_E of the serial policy as $T_F^{(B)} = T_F/B$ and $T_E^{(B)} = T_E/B$, since in each block, only a fraction L/B of a file is transmitted. The total delivery time per bit is hence given by

$$\begin{aligned}\Delta_{\mathsf{P}}(\mu, C_F, P) &= \limsup_{L\to\infty} \frac{(B+1)\max\left(T_F^{(B)}, T_E^{(B)}\right)}{L} \\ &= \limsup_{L\to\infty} \frac{(B+1)}{B}\frac{\max\left(T_F, T_E\right)}{L}. \end{aligned} \tag{17}$$

The corresponding NDT (11) is computed as

$$\begin{aligned}\delta_{\mathsf{P,Ach}}(\mu, r) &= \lim_{B\to\infty}\lim_{P\to\infty}\limsup_{L\to\infty} \frac{(B+1)}{B}\frac{\max\left(T_F, T_E\right)}{L/\log(P)} \\ &= \max\left(\delta_F, \delta_E\right), \end{aligned} \tag{18}$$

where δ_F and δ_E are the fronthaul and edge NDTs of the serial transmission scheme. Thus, under the limit of an arbitrarily large number of blocks B, the achievable NDT under pipelined fronthaul-edge transmission is the *maximum* of the edge and fronthaul NDTs of the serial policy.

• *Per-Block File Splitting*: To further improve the performance of the block-Markov coding, we propose a per-block *file-splitting* strategy in order to time-share between any two serial fronthaul-edge policies. To elaborate, for some $\alpha \in [0,1]$ fraction of each file block (of size L/B bits), a (serial) policy requiring total fronthaul and edge NDTs $\delta_F^{(1)}$ and $\delta_E^{(1)}$ is used, and for the remaining $(1-\alpha)$ fraction of each file block, a (serial) policy requiring NDTs $\delta_F^{(2)}$ and $\delta_E^{(2)}$ is used (see Fig. 4(a)). Based on the discussion above, this yields an achievable NDT of

$$\delta_{\mathsf{P,Ach}} = \max\left(\alpha\delta_F^{(1)} + (1-\alpha)\delta_F^{(2)}, \alpha\delta_E^{(1)} + (1-\alpha)\delta_E^{(2)}\right). \tag{19}$$

The following proposition gives an achievable NDT obtained by leveraging block-Markov coding and per-block file-splitting.

Proposition 2. *For an $M \times K$ F-RAN with a fronthaul gain of $r > 0$, the minimum NDT for pipelined fronthaul-edge transmission is upper bounded as $\delta_{\mathsf{P}}^*(\mu, r) \le \delta_{\mathsf{P,Ach}}(\mu, r)$, where*

$$\delta_{\mathsf{P,Ach}}(\mu, r) = \begin{cases} \delta_{\mathsf{P-IA}} & \text{for } \mu \in [0, \mu_1], \\ \delta_{\mathsf{P-FS}} & \text{for } \mu \in [\mu_1, \mu_2], \\ \delta_{\mathsf{P-ZF}} & \text{for } \mu \in [\mu_2, 1], \end{cases} \tag{20}$$

with $\mu_1 \le \mu_2 \le 1$. For $\mu \in [0, \mu_1]$, the NDT

$$\delta_{\mathsf{P-IA}} = \frac{(1-\mu M)K}{Mr} \tag{21}$$

is achieved by per-block file-splitting between cloud-aided soft-transfer fronthauling and cache-aided EN coordination via X-channel based interference alignment. For $\mu \in [\mu_1, \mu_2]$, the NDT

$$\delta_{\mathsf{P-ZF}} = \frac{K}{\min\{M, K\}} \tag{22}$$

is achieved by per-block file-splitting between cloud-aided soft-transfer fronthauling and cache-aided EN cooperation via ZF-beamforming. For $\mu \in [\mu_2, 1]$, the NDT

$$\delta_{\mathsf{P-FS}} = \frac{K}{Mr}\left[1 - \mu_2 - [\mu_1 M - \mu_2]\left(\frac{\mu_2 - \mu}{\mu_2 - \mu_1}\right)^+\right] \tag{23}$$

is achieved by per-block file-splitting between the schemes achieving $\delta_{\mathsf{P-IA}}$ at $\mu = \mu_1$ and $\delta_{\mathsf{P-ZF}}$ at $\mu = \mu_2$ respectively, where

$$\mu_1 = \left(\frac{K - \max\{M,K\}r}{KM + Mr\left[\min\{M,K\} - 1\right]}\right)^+, \tag{24}$$

$$\mu_2 = \left(1 - \frac{Mr}{\min\{M,K\}}\right)^+. \tag{25}$$

Proof. The proof is omitted for brevity and is given in [3]. □

C. Minimum NDT for a Cloud and Cache-Aided F-RAN

We next provide a partial characterization of the minimum NDT for a general cloud and cache-aided F-RAN with pipelined fronthaul-edge transmission. Specifically, the following proposition gives the minimum NDT for the low cache regime with $\mu \in [0, \mu_1]$; for the high cache regime with $\mu \in [\mu_2, 1]$; and for the high fronthaul regime with $r \ge ((1-\mu)\min\{M,K\})/M$.

Proposition 3. *For a general $M \times K$ F-RAN, with pipelined fronthaul-edge transmission and with fronthaul gain $r > 0$, we have*

$$\delta^*_{\mathsf{P}}(\mu, r) = \begin{cases} \delta_{\mathsf{P-IA}}, & \text{for} \quad \mu \in [0, \mu_1], \\ \delta_{\mathsf{P-ZF}}, & \text{for} \quad \mu \in [\mu_2, 1], \end{cases} \tag{26}$$

where $\delta_{\mathsf{P-IA}}$ and $\delta_{\mathsf{P-ZF}}$ are defined in (20) *and the fractional cache sizes μ_1, μ_2 are defined in* (24)*. Furthermore, for any fractional cache size $\mu \in [0, 1]$, we have*

$$\delta^*_{\mathsf{P}}(\mu, r) = \delta_{\mathsf{P-ZF}}, \quad \text{for} \quad r \geq \frac{(1-\mu)\min\{M, K\}}{M}. \tag{27}$$

Proof. The proof is presented in [3]. □

Remark 3. Proposition 3, along with Proposition 2, demonstrate that, even with partial caching, i.e., with $\mu < 1$, it is possible to achieve the same performance as in a system with full caching or ideal fronthaul, namely $\delta = \delta_{\mathsf{P-ZF}} = K/\min\{M, K\}$. This is the case as long as either the fronthaul capacity is large enough (see (27)) or the fronthaul capacity is positive and the cache capacity μ is sufficiently large (see (26)). We observe that this is not true for serial fronthaul-edge transmission, in which case no policy can achieve the NDT $\delta = K/\min\{M, K\}$ for $\mu < 1$ and finite fronthaul capacity. The intuition behind this result is that, with pipelined transmission, cloud resources can be leveraged to make up for partial caching by transmitting on the fronthaul while edge transmission takes place. ◇

We finally provide an approximate characterization of the minimum NDT for a general $M \times K$ F-RAN with pipelined fronthaul-edge transmission by showing that the lower bound in Proposition 1 and the upper bound in Proposition 2 are within a constant multiplicative gap, independent of problem parameters for any fronthaul gain $r > 0$, in the intermediate cache regime with $\mu \in [\mu_1, \mu_2]$, where the minimum NDT is not characterized by Proposition 3.

Proposition 4. *For a general $M \times K$ F-RAN with pipelined fronthaul-edge transmission and with fronthaul gain $r > 0$, we have*

$$\frac{\delta_{\mathsf{P,Ach}}(\mu, r)}{\delta^*_{\mathsf{P}}(\mu, r)} \leq 2, \quad \text{for} \quad \mu \in [\mu_1, \mu_2]. \tag{28}$$

Proof. The proof is presented in [3]. □

Remark 4 (*2×2 F-RAN*)**.** The minimum NDT for a 2×2 F-RAN is derived in [3] by leveraging the bounds in Propositions 1 and 2, and is shown in Fig. 2 for fronthaul gain $r = 0.5$. The optimal strategy uses block-Markov encoding with cloud-aided soft transfer fronthaul in conjunction with cache-aided EN cooperation or coordination as for Proposition 2. We observe that, in contrast to serial fronthaul-edge transmission [2], [3], the optimal strategy leverages cloud resources for any given fronthaul gain $r > 0$. Furthermore, in line with the discussion in Remark 3, by using cloud resources, it is possible here to obtain the minimum NDT $\delta^*_{\mathsf{P}}(\mu, r) = 1$ for all $\mu \geq 1/2$. In general, $\delta^*_{\mathsf{P}}(\mu, r) = 1$ for all $\mu \geq \mu_2$ when $r < 1$ and for all $\mu \in [0, 1]$ when $r \geq 1$. ◇

IV. Conclusions

In this paper, we presented a latency-centric study of the fundamental information-theoretic limits of cloud and cache-aided wireless networks, namely Fog Radio Access Networks (F-RANs) with pipelined fronthaul-edge transmissions. We presented a general lower bound on the NDT and proposed achievable schemes which were shown to be approximately optimal in terms of NDT to within a constant multiplicative factor of 2. The analysis shows that piplined fronthaul-edge transmission always improves on the serial case and is able to benefit from cloud processing to reduce the delivery latency for all values of fractional cache size.

References

[1] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog computing based radio access networks: Issues and challenges," *arXiv:1506.04233*, 2015. [Online]. Available: http://arxiv.org/abs/1506.04233

[2] 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*, July 2016.

[3] A. Sengupta, R. Tandon, and O. Simeone, "Cloud and cache-aided wireless networks: Fundamental latency trade-offs," *arXiv: 1605.01690*, May 2016. [Online]. Available: http://arxiv.org/abs/1605.01690

[4] S. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," *arXiv:1601.02460*, Jan. 2016.

[5] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *arXiv:1512.07743*, Dec 2015. [Online]. Available: http://arxiv.org/abs/1512.07743

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

[7] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," *arXiv: 1512.02385*, 2015. [Online]. Available: http://arxiv.org/abs/1512.02385

[8] M. A. Maddah-Ali and U. Niesen, "Cache aided interference channels," in *Proc. IEEE International Symposium on Information Theory*, June 2015, pp. 809–813.

[9] ——, "Cache-aided interference channels," *arXiv: 1510.06121*, Oct 2015. [Online]. Available: http://arxiv.org/abs/1510.06121

[10] N. Naderializadeh, M. A. Maddah-Ali, and A. Salman Avestimehr, "Fundamental limits of cache-aided interference management," *arXiv:1602.04207*, Feb. 2016. [Online]. Available: http://arxiv.org/pdf/1602.04207v1

[11] F. Xu, M. Tao, and K. Liu, "Fundamental Tradeoff between Storage and Latency in Cache-Aided Wireless Interference Networks," *arXiv: 1605.00203*, May 2016. [Online]. Available: http://arxiv.org/pdf/1605.00203v1.pdf

[12] J. Hachem, U. Niesen, and S. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *arXiv: 1606.03175*, June 2016. [Online]. Available: https://arxiv.org/abs/1606.03175

[13] A. Sengupta, R. Tandon, and O. Simeone, "Cloud ran and edge caching: Fundamental performance trade-offs," in *Proc. IEEE International workshop on Signal Processing advances in Wireless Communications (SPAWC)*, July 2016.

[14] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *arXiv:1512.06938*, 2015. [Online]. Available: http://arxiv.org/abs/1512.06938

[15] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 3:1–3:10, Feb 2009.

[16] M. Leconte, G. S. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," *arXiv: 1601.03926*, 2016. [Online]. Available: http://arxiv.org/abs/1601.03926