Cloud RAN and Edge Caching: Fundamental Performance Trade-Offs

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

***Abstract*—A wireless network architecture is studied in which edge nodes (ENs), such as base stations, are connected to a cloud processor by dedicated fronthaul links, while also being endowed with caches, in which popular content, such as multimedia files, can be proactively stored. Cloud processing enables the centralized implementation of cooperative transmission by the ENs, albeit at the cost of an increased latency due to fronthaul transfer. In contrast, edge caching allows for the low-latency delivery of the cached files, but with generally limited cooperation among the ENs. The interplay between cloud processing and edge caching is studied from an information-theoretic viewpoint by investigating the fundamental limits of a metric, termed normalized delivery time (NDT), which captures the worst-case latency for delivering any requested content to the users. 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.**

I. Introduction

The cloud radio access network (C-RAN) architecture enables the virtualization of baseband functionalities from the base stations, or edge nodes (ENs), of a wireless system to a centralized processor. C-RAN is known to enhance the spectral efficiency, but at the cost of a potentially large latency, due to the need to communicate on fronthaul links between ENs and cloud [1]. In a dual manner, edge caching allows the low-latency delivery of multimedia content with no backhaul overhead, by proactively storing popular files at the ENs (see, e.g., [2]–[9]). In this work, a hybrid architecture is considered, referred to here as Fog-RAN (F-RAN), in which ENs are connected to a cloud processor, as in a C-RAN, while also being equipped with local caches (see Fig. 1).

The design of F-RAN networks involves two key design questions: a) *what to cache* at the ENs, under the constraint that the caches cannot be updated for long periods encompassing multiple transmission intervals; and b) *how to deliver* the requested content to the users over the wireless channel in each transmission interval by leveraging both cloud processing and edge caching. The two questions are strongly intertwined and determine the content delivery latency. In fact, with cloud processing, content delivery on the wireless channel can benefit from cooperative transmission among the ENs. For instance, in a C-RAN, cooperative transmission is realized by means of encoding and precoding at the cloud followed by compression of the resulting baseband signals, which are forwarded to the ENs for transmission on the wireless channel [1]. Thus, cloud processing incurs the latency entailed by fronthaul communication from the cloud to the

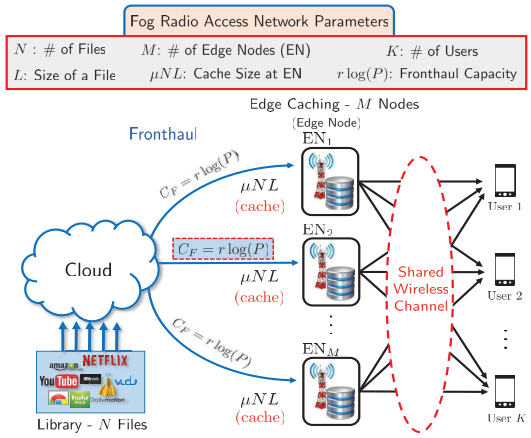

Fig. 1: Information-theoretic model for a cloud and cache-aided wireless system, referred to as F-RAN.

ENs. In contrast, cooperative transmission on cached content is limited to files that are shared among the caches of multiple ENs, but no fronthaul latency is incurred.

Interference-limited cache-aided wireless systems were first investigated from an information-theoretic viewpoint in [5], where an upper bound on the 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 [6] by accounting for caching at both ENs and users, and linear delivery strategies on the wireless channel. Reference [7] instead presents a lower bound on the delivery latency, which is formalized in terms of a high Signal-to-Noise Ratio (SNR) metric defined as *Normalized Delivery Time* (NDT), proving the optimality of the scheme proposed in [5] for a given regime of cache capacity values. A cloud- and cache-aided wireless system, or F-RAN, was first studied in [8], for the specific case of $M = 2$ ENs and $K = 2$ users, and the minimum NDT was characterized along with the optimal caching and delivery policies. Caching and precoding optimization of F-RANs were also investigated in [10], [11].

Main Contributions: The contribution of the paper is twofold. First, a general information-theoretic lower bound on the delivery latency, or NDT, of F-RAN systems is developed for any number of ENs and users. Then, an upper bound is derived for the same system by considering the time-sharing between two cooperative schemes that use only cloud or only cache resources, respectively. The two bounds demonstrate the optimality of this scheme in the absence of caching at the ENs in terms of NDT. Furthermore, using the developed bounds,

as well as the results in [5], we partially characterize the NDT trade-off for an F-RAN with $M = 3$ ENs and $K = 3$ users.

Notation: For any integers a and b with $a \leq b$, we define $[a : b] = (a, a+1, \ldots, b)$. We use $b \in [a, c]$ to imply $a \leq b \leq c$ and $b \in (a, c]$ to imply $a < b \leq c$. We define $(x)^+ = \max\{0, x\}$.

II. System Model

We consider an $M \times K$ F-RAN system, shown in Fig. 1 and first introduced in [8], 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 $L \in \mathbb{N}^+$. Formally, the files F_n are independent and identically distributed (i.i.d.) as:

$$F_n \sim \text{Unif}\{1, 2, \ldots, 2^L\}, \quad \forall n = 1, \ldots, N. \tag{1}$$

Each EN is equipped with a cache in which it can store μNL bits, where the fraction $\mu \in [0, 1]$, is referred to as the *fractional cache size* i.e., the fraction of each file which can be cached at an EN. 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 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 $\mathbf{D} \triangleq (d_1, \ldots, d_K) \in [1 : N]^K$. These requests are then served by the ENs' transmissions, which are based on the local cached content as well as on the signals received from the cloud via the fronthaul. All ENs, as well as the cloud, have access to the global channel state information (CSI) $\mathbf{H} = \{\{h_{km}\} : {}^{k=1:K}_{m=1:M}\}$, where $h_{km} \in \mathbb{C}$, denotes the wireless channel between user $k \in [1 : K]$ and EN_m, $m \in [1 : M]$. The coefficients are assumed to be drawn i.i.d. from a continuous distribution and to be time-invariant within each transmission interval.

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

a) Caching Policy π_c: The caching policy at each edge node EN_m, $m = [1 : M]$, is defined by a function, $\pi_c^m(\cdot)$, which maps each file to its cache storage

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

The mapping is such that $H(S_{m,n}) \leq \mu L$ in order to satisfy the cache capacity constraints. The total 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 $F_1, \ldots, F_N$, the demand vector $\mathbf{D}$ and CSI $\mathbf{H}$ to the fronthaul message

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

which is transmitted to EN_m via the fronthaul link of capacity C_F bits per symbol. Here, T_F is the duration of the fronthaul message. In keeping with the definition of fronthaul capacity C_F, all time intervals, including T_F, are normalized to the symbol transmission time on the downlink wireless channel. Thus, the fronthaul message cannot exceed $T_F C_F$ bits.

c) Edge Transmission Policy π_e: During the final delivery phase of a transmission interval, each edge-node EN_m uses an edge transmission policy, $\pi_e^m(\cdot)$, which maps the demand vector $\mathbf{D}$ and global CSI $\mathbf{H}$, along with its local cache content and the received fronthaul message to output a codeword

$$\mathbf{X}_m^{T_E} = (X_m[t])_{t=1}^{T_E} = \pi_e^m\Big(S_m, \mathbf{U}_m^{T_F}, \mathbf{D}, \mathbf{H}\Big), \tag{4}$$

which is transmitted to the users. Here, T_E is the duration of the transmission on the wireless channel, on which an average power constraint of P is imposed for each codeword $\mathbf{X}_m^{T_E}$. Note that the fronthaul policy, π_f and the edge transmission policy, π_e, can adapt to the instantaneous demands and CSI at each transmission interval, unlike the caching policy, π_c, which remains unchanged over multiple transmission intervals.

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

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

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

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

of the requested file F_{d_k}. The caching, fronthaul, edge transmission and decoding policies together form a policy $\pi = (\pi_c^m, \pi_f^m, \pi_e^m, \pi_d^k)$. The probability of error of a policy π is defined as

$$P_e = \max_{\mathbf{D}} \max_{k \in \{1, \ldots, K\}} \mathbb{P}\left(\widehat{F}_{d_k} \neq F_{d_k}\right). \tag{7}$$

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, and for any demand vector $\mathbf{D}$, we have $P_e \to 0$ when $L \to \infty$. Note that the fronthaul and edge transmission durations, T_F and T_E, respectively, generally depend on L.

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

$$\Delta(\mu, C_F, P) = \limsup_{L \to \infty} \frac{T_F + T_E}{L}. \tag{8}$$

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 achievable $\Delta(\mu, C_F, P)$, with $C_F = r \log(P)$, the *normalized delivery time* (NDT), is defined as

$$\delta(\mu, r) = \lim_{P \to \infty} \frac{\Delta(\mu, r \log(P), P)}{1/\log P}. \tag{9}$$

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

$$\delta^*(\mu, r) = \inf\{\delta(\mu, r) : \delta(\mu, r) \text{ is achievable}\}. \tag{10}$$

Remark 1. The delivery time per bit (8) is normalized by the term $1/\log P$. This 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^* indicates that the worst-case time required to serve any possible request $\mathbf{D}$, is δ^* times larger than the time needed by this ideal baseline system (see also [7], [8]).

III. General Bounds on the Minimum NDT

In this section, we provide lower and upper bounds for the NDT of a general $M \times K$ F-RAN system as described above under the assumption that perfect CSI is available at all ENs and the cloud. The bounds are shown to be tight for a system with no caching, i.e., $\mu = 0$, hence identifying the optimal operation of cloud processing in terms of delivery latency in this regime. The bounds are further used in the next section to tackle the special case with $M = 3$ and $K = 3$.

A. A Lower Bound on the Minimum NDT

The following theorem provides an information-theoretic lower bound on the NDT.

Theorem 1. *For an $M \times F$ F-RAN, with each EN having a fractional cache size $\mu \in [0, 1]$, a library of $N \geq K$ files and fronthaul gain $r \geq 0$, the NDT is lower bounded as $\delta^*(\mu, r) \geq \delta_{LB}(\mu, r)$, where $\delta_{LB}(\mu, r)$ is the minimum value of the following linear program (LP)*

$$\delta_{LB}(\mu, r) = \min \delta_F + \delta_E \tag{11}$$

subject to:

$$\ell\delta_E + (M - \ell)^+ r\delta_F \geq K - (M - \ell)^+(K - \ell)^+\mu, \tag{12}$$

$$\delta_F \geq 0,\ \delta_E \geq 1 \tag{13}$$

where $\ell \in [0 : \min\{M, K\}]$ in (12).

In Theorem 1, the variables δ_F and δ_E capture the normalized latencies associated with fronthaul and edge transmissions, respectively. The family of constraints in (12) are obtained by generalizing the approach that was first used in [7] for a scenario with no fronthauling ($r = 0$). The proof, which is not provided here due to space constraints (see [12]), is based on a cut-set-like argument. Specifically, it can be argued that, for all sequence of feasible policies, 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. The proposition can be proved by carefully bounding the joint entropy of these random variables, which upper bounds the amount of information that can be reliably conveyed in given time intervals T_E and T_F.

B. An Upper Bound on the Minimum NDT

The following Lemma gives an upper bound on the minimum NDT, which is attained by a specific policy that is described below.

Lemma 1. *For an $M \times K$ F-RAN with $\mu \in [0, 1]$, a library of $N \geq K$ files and fronthaul gain $r \geq 0$, the NDT is upper bounded as $\delta^*(\mu, r) \leq \delta_{UB}(\mu, r)$, where*

$$\delta_{UB}(\mu, r) = \frac{K}{\min\{M, K\}} + (1 - \mu)\frac{K}{Mr}. \tag{14}$$

The lemma is proved by considering the NDT of a policy that performs time-sharing between two schemes that leverage only cache or only cloud resources, respectively. The first, *cache-based*, policy operates by caching the same fraction μ of all the files at all the ENs. Note that this choice satisfies the cache capacity constraint. The ENs can then transmit interference-free these fractional files to any subset of $\min\{M, K\}$ users by leveraging zero-forcing (ZF) beamforming. Since interference-free transmission implies an NDT of one by definition, this scheme can be seen to achieve an NDT of $\mu K/\min\{M, K\}$. Note that no fronthaul latency is incurred by this cache-based scheme. The remaining fraction $(1-\mu)$ of the requested files is instead sent by using a *cloud-based* strategy that follows the C-RAN principle. In particular, the cloud encodes and precodes the fractional files at hand for subsets of $\min\{M, K\}$ users at a time by means of ZF-beamforming. Then, the resulting baseband signals are compressed and sent to the ENs on the respective fronthaul links, so as to allow the ENs to simultaneously transmit to the users on the wireless channel. This scheme obtains a normalized edge transmission latency of $(1 - \mu)K/\min\{M, K\}$, similar to the cache-based scheme, but it also requires the normalized fronthaul latency $(1 - \mu)K/Mr$. The latter follows from the fact that the baseband signals need to be compressed with at least $\log(P)$ bits per symbol in order for the quantization noise not to be the limiting factor on the achievable performance. Details can be found in [12].

C. Discussion

In comparing the upper and lower bounds derived above, a first observation is that, as the number M of ENs increases, the achievable NDT in (14) approaches the ideal NDT lower bound of $\delta(\mu, r) = 1$ for any value of μ. We next use the lower bound in Theorem 1 to show the optimality of the proposed achievable NDT in Lemma 1 for the case $\mu = 0$, i.e., with no caching. To this end, we observe that any lower bound on the optimal value of the LP in Theorem 1 is also a valid lower bound on the NDT. Summing the constraints in (12) with $\ell = M$ and $\ell = 0$ yields a lower bound on the optimal value of the LP, which can be seen to equal (14) for $M \leq K$; while summing the constraint in (12) with $\ell = 0$ and the constraint $\delta_E \geq 1$ yields (14) for $M \geq K$, hence concluding the proof.

IV. The Case $M = K = 3$

In this section, we investigate in detail the case of an F-RAN with $M = 3$ ENs and $K = 3$ users by using the results put forth in the previous section as well as in [5]. Note that reference [8] presents a related study for the simpler case with $M = 2$ and $K = 2$.

Corollary 1. *For an F-RAN with $M = 3$ ENs, $K = 3$ users and $N \geq 3$ files, the minimum NDT is characterized as:*

- *Low Fronthaul ($r \in [0, 1/2]$):*

$$\delta^*(\mu, r) = \begin{cases} 1 + 2\mu + \dfrac{1 - 3\mu}{r} & \text{for } \mu \in [0, 1/3], \\ 3/2 - \mu/2 & \text{for } \mu \in [2/3, 1], \end{cases}$$

$$\delta^*(\mu,r) \quad \begin{cases} \geq \max\left(3-4\mu, \dfrac{3-\mu}{2}\right) & \textit{for} \quad \mu \in [1/3, 2/3]. \\ \leq 13/6 - 3\mu/2 \end{cases} \tag{15}$$

• *Intermediate Fronthaul 1* ($r \in [1/2, 6/7]$)*:*

$$\delta^*(\mu,r) \quad \begin{cases} \geq 1 + \dfrac{2}{3}\mu + \dfrac{3-7\mu}{r} & \\ \leq 1 + 2\mu + \dfrac{1-3\mu}{r} & \textit{for} \quad \mu \in [0, 1/3], \end{cases}$$

$$\delta^*(\mu,r) \quad \begin{cases} \geq \max\left(1 + \dfrac{2}{3}\mu + \dfrac{3-7\mu}{3r}, \dfrac{3-\mu}{2}\right) & \\ \leq 13/6 - 3\mu/2 & \textit{for} \quad \mu \in [1/3, 2/3], \end{cases}$$

$$\delta^*(\mu,r) = 3/2 - \mu/2, \qquad \textit{for} \quad \mu \in [2/3, 1]. \tag{16}$$

• *Intermediate Fronthaul 2* ($r \in [6/7, 2]$)*:*

$$\delta^*(\mu,r) \quad \begin{cases} \geq \max\left(1 + \dfrac{2}{3}\mu + \dfrac{3-7\mu}{3r}, \dfrac{3-\mu}{2}\right) & \\ \leq 1 + \dfrac{\mu}{4} + \dfrac{2-3\mu}{2r} & \textit{for} \quad \mu \in [0, 2/3], \end{cases}$$

$$\delta^*(\mu,r) = 3/2 - \mu/2, \qquad \textit{for} \quad \mu \in [2/3, 1]. \tag{17}$$

• *High Fronthaul* ($r \geq 2$)*:*

$$\delta^*(\mu,r) = 1 + \frac{1-\mu}{r}, \qquad \textit{for} \quad \mu \in [0, 1]. \tag{18}$$

Corollary 1 provides a partial characterization of the minimum NDT of a 3×3 F-RAN by identifying upper and lower bounds for all values of μ and r as well as conclusive results for specific regimes of the parameters. As discussed next, in these regimes, we can characterize optimal policies and hence the optimal interplay between cloud and edge processing. The lower bounds are obtained from Theorem 1, while the upper bounds are derived by considering a more general time-sharing scheme than the one used in the proof of Lemma 1. In this policy, the constituent schemes are the cloud-aided ZF-beamforming strategy described above and the cached-based scheme that uses interference alignment presented in [5].

To aid the interpretation of the main results, Fig. 2 shows the NDT trade-off bounds presented in Corollary 1 for four values of r, namely $\{0.25, 0.75, 1.25, 2\}$, which lie in the different intervals as defined in Corollary 1. The figures partition the values of μ into two distinct intervals: for smaller values of μ, the policy used in the proof of Corollary 1 leverages both cloud and cache resources, whereas for larger values of μ only cache resources are employed for transmission on the wireless channel. In the *low fronthaul* regime of $r \leq 1/2$, time-sharing between cache- and cloud-based schemes is optimal for $\mu \leq 1/3$. Instead, for larger cache storage, using cloud resources in addition to cache resources may only provide a marginal decrease of the NDT. A similar behavior is observed also for *intermediate fronthaul*, here $r = \{0.75, 1.25\}$, which fall in the second and third intervals described in Corollary 1. In particular, for $\mu \geq 2/3$, optimal F-RAN operation does not require the use of the cloud. Finally, in the *high fronthaul* regime $r \geq 2$, achieving the optimal NDT performance requires the use of both cloud and caching resources. In the rest of this section, we provide some details on the proof of Corollary 1.

1) Upper Bounds: As discussed, the upper bounds on the minimum NDT reported in Corollary 1 are obtained by performing time-sharing between the cloud-only scheme described in the proof of Lemma 1, which yields an NDT of $1 + 1/r$ (obtained by setting $\mu = 0$ and $K = M$ in (14)), and the policy presented in [5], which achieves the NDT

$$\delta(\mu,r) \leq \begin{cases} 13/6 - 3\mu/2 & \text{for } \mu \in [1/3, 2/3], \\ 3/2 - \mu/2 & \text{for } \mu \in [2/3, 1]. \end{cases} \tag{19}$$

We next look at different regimes of r to characterize the achievable NDT in Corollary 1.

• *Low and Intermediate Fronthaul 1* ($r \leq 6/7$): For this regime, the upper bound

$$\delta(\mu,r) \leq 1 + 2\mu + \frac{1-3\mu}{r}, \quad \text{for } \mu \in [0, 1/3], \tag{20}$$

is obtained by file-splitting between the cloud-aided scheme discussed in Section III and the policy in [4] for $\mu = 1/3$, which achieves an NDT of $5/3$ by (19). We recall the latter converts the system into an X-channel, i.e., a channel in which each transmitter has a distinct message intended to all receivers, by placing a different third of each file in the caches of the ENs; it then performs interference alignment on the resulting X-channel. To elaborate, for a fraction 3μ of the files, the mentioned scheme in [4] is used, hence satisfying the cache capacity constraint; while for the remaining $(1-3\mu)$ fraction of the files, the cloud-based scheme is used. For $\mu \geq 1/3$, instead, it can be seen that transmitting a part of the files by means of the cloud-only scheme does not improve the NDT, and the achievable NDT is given by (19). This validates the achievable NDT in (15)-(16).

• *Intermediate Fronthaul 2* ($r \in [6/7, 2]$): In this case, file-splitting between the cloud-only scheme and the scheme from [4] with $\mu = 2/3$, which yields and NDT of $7/6$ from (19), gives the improved achievable NDT:

$$\delta(\mu,r) \leq 1 + \frac{\mu}{4} + \frac{2-3\mu}{2r}, \quad \text{for } \mu \in [0, 2/3]. \tag{21}$$

For $\mu \geq 2/3$, the achievable NDT is given by (19). This validates the achievable NDT given in (17).

• *High Fronthaul* ($r \geq 2$): In this regime, the achievable NDT is given by file-splitting between the cloud-only policy and cache-aided ZF-beamforming, yielding (14) for $M = K$ and validating the achievability result in (18).

2) Lower Bounds: The converse is derived from Theorem 1 by considering linear combinations of the constraints (12)-(13) to obtain lower bounds on the optimal value of the LP, i.e., on the minimum NDT. The constraints can be written as:

$$\mathsf{Ineq\ 1}: \quad (\delta_E + 2r\delta_F) \geq (3 - 4\mu) \tag{22}$$

$$\mathsf{Ineq\ 2}: \quad (2\delta_E + r\delta_F) \geq (3 - \mu) \tag{23}$$

$$\mathsf{Ineq\ 3}: \quad \delta_F \geq (1 - 3\mu)/r \tag{24}$$

$$\mathsf{Ineq\ 4}: \quad \delta_E \geq 1. \tag{25}$$

Ineq 1, Ineq 2 and Ineq 3 are obtained from (12) by substituting $\ell = 1, 2$ and 0 respectively, while Ineq 4 follows directly from (13). We next utilize these inequalities to prove the converse.

• *Low Fronthaul* ($r \leq 1/2$): In this regime, Ineq 1 + (1 −

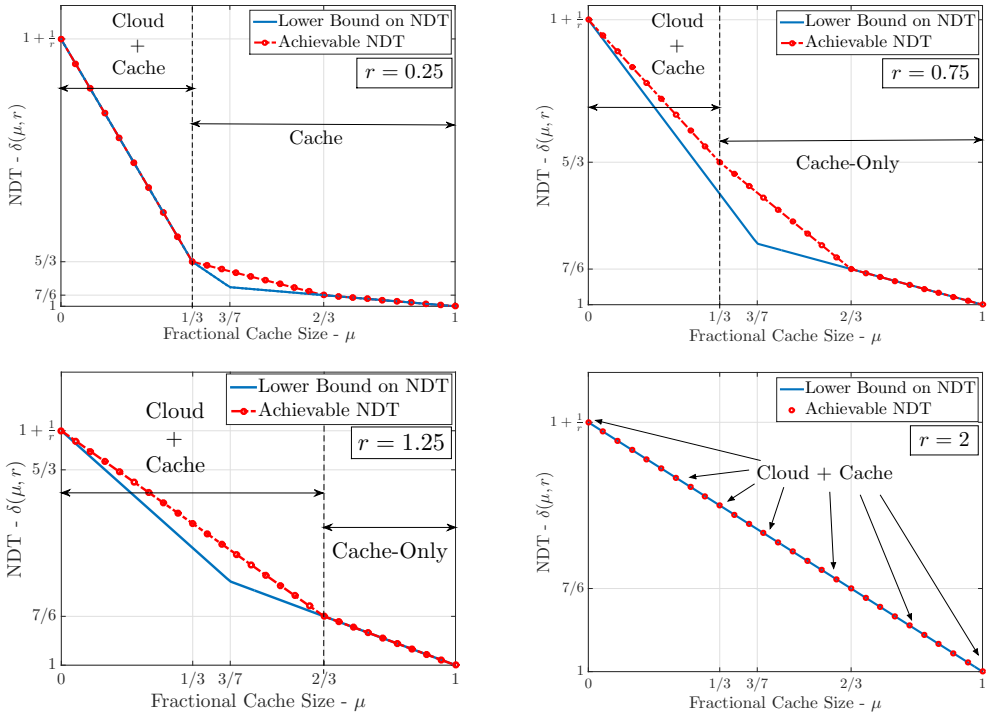

Fig. 2: NDT tradeoff for an F-RAN with $M = K = 3$ for $r = 0.25$, $r = 0.75$, $r = 1.25$ and $r = 2$.

$2r) \times$ Ineq 3 gives the lower bound:

$$\delta^*(\mu, r) \geq 1 + 2\mu + \frac{1 - 3\mu}{r}. \tag{26}$$

Also, considering Ineq 1 and using $r \leq 1/2$, we have

$$\delta^*(\mu, r) \geq 3 - 4\mu. \tag{27}$$

• *Low & Intermediate Fronthaul* ($r \leq 2$): Considering Ineq 2, and using $r \leq 2$, we have the desired lower bound:

$$\delta^*(\mu, r) \geq \frac{3 - \mu}{2}. \tag{28}$$

• *Intermediate Fronthaul* ($r \in [1/2, 2]$): In this regime, $\left(\frac{2-r}{3r}\right) \times$ Ineq 1 $+ \left(\frac{2r-1}{3r}\right) \times$ Ineq 2 yields the lower bound:

$$\delta^*(\mu, r) \geq 1 + \frac{2}{3}\mu + \frac{3 - 7\mu}{r}. \tag{29}$$

• *High Fronthaul* ($r \geq 2$): In this regime, Ineq $2 + (r - 2) \times$ Ineq 4 gives us the lower bound:

$$\delta^*(\mu, r) \geq 1 + \frac{1 - \mu}{r}. \tag{30}$$

V. Conclusions

In this paper, we have considered an emerging wireless network architecture that enables both virtualized RAN, by means of cloud processing and fronthauling, and edge caching. We have studied the fundamental trade-off between delivery latency and system resources, namely fronthaul and cache capacities, from an information-theoretic viewpoint. We have developed lower bounds on the normalized delivery time (NDT), which captures the high-SNR worst-case latency in delivering content to users. Based on this result, we have identified the optimal operation of cloud and caching resources in various regimes, pointing to cloud-based compressed precoding as well as edge-based interference management as key techniques in order to achieve minimum delivery latency.

References

[1] 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

[2] 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(12), pp. 8402–8413, Dec. 2013.

[3] 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

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

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

[6] 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

[7] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," *arXiv:1512.07856*, Dec 2015. [Online]. Available: http://arxiv.org/pdf/1512.07856v1.pdf

[8] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in Fog radio access Networks," *to appear IEEE International Symposium on Information Theory*, Jul 2016.

[9] 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

[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 Communications Letters*, 2015.

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

[12] A. Sengupta, R. Tandon, and O. Simeone, "Cache and cloud aided wireless networks: Fundamental latency trade-offs," *In preparation*, Feb 2016.