Cloud-Aided Wireless Networks with Edge Caching: Fundamental Latency Trade-Offs in Fog Radio Access Networks

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

Osvaldo Simeone
CWCSPR, ECE Department
New Jersey Institute of Technology
osvaldo.simeone@njit.edu

***Abstract*—Fog Radio Access Network (F-RAN) is an emerging wireless network architecture that leverages caching capabilities at the wireless *edge* nodes, as well as edge connectivity to the *cloud* via fronthaul links. This paper aims at providing a *latency-centric* analysis of the degrees of freedom of an F-RAN by accounting for the total content delivery delay across the fronthaul and wireless segments of the network. The main goal of the analysis is the identification of optimal caching, fronthaul and edge transmission policies. The study is based on the introduction of a novel performance metric, referred to as the *Normalized Delivery Time* (NDT), which measures the total delivery latency as compared to an ideal interference-free system. An information-theoretically optimal characterization of the trade-off between NDT, on the one hand, and fronthaul and caching resources, on the other, is derived for a class of F-RANs with two edge nodes and two users. Using these results, the interplay between caching and cloud connectivity is highlighted, as well as the impact of both caching and fronthaul resources on the delivery latency.**

***Index Terms*—Cloud Radio Access Network (C-RAN), caching, 5G, degrees of freedom, latency.**

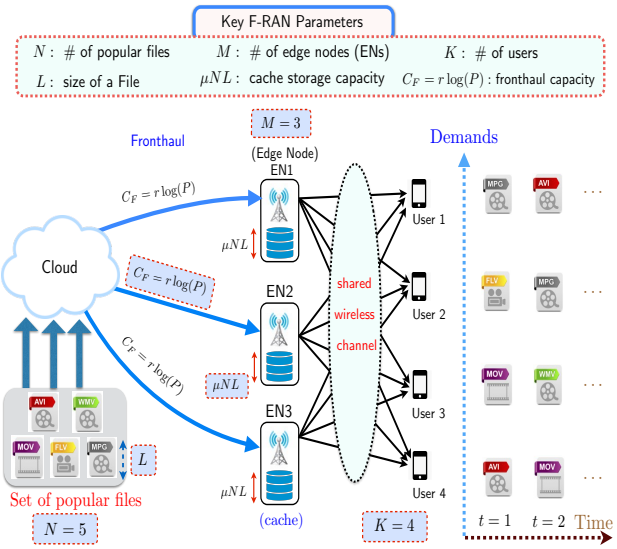

Fig. 1. Information-theoretic model for F-RAN.

I. Introduction

Edge processing and virtualization are two of the key emerging trends in the evolution of 5G networks [1], [2]. *Edge processing* refers to the localization of storage and computing resources at the network edge. Notably, edge nodes (ENs), such as base stations, can be equipped with *local* caches to store popular content, with the aim of reducing the delivery latency by limiting the need to communicate with remote content servers [1]. In a dual manner, *virtualization* allows the implementation of network functionalities at a *centralized* cloud processor. An important example is given by the *Cloud Radio Access Network* (C-RAN) architecture, in which the ENs are connected to a cloud processor by so called *fronthaul* links, so as to enable, among other performance gains, enhanced interference management capabilities thanks to the joint baseband processing in the cloud [2], [3].

Bridging the gap between these two complementary trends, the *Fog Radio Access Network* (F-RAN) architecture has been recently advocated that combines the benefits of both edge and cloud processing (see, e.g., [4]). In this architecture, as illustrated in Fig. 1, ENs may be endowed with caching capabilities, so as to serve local data requests of popular content with low latency, while at the same time being controllable from a central cloud processor, in order to serve arbitrary data requests with stronger interference management properties and less stringent delay constraints.

The design of F-RANs involves the following intertwined design questions: (a) *What to cache?*; (b) *How to utilize the limited capacity available on the fronthaul links?*; and (c) *How to deliver the requested content on the downlink wireless channel?* In this work, we set out to obtain fundamental insights into these questions by means of an information-theoretic approach. The proposed framework aims at providing a *latency-centric* analysis of the degrees of freedom of an F-RAN by accounting for the total content delivery delay across the fronthaul and wireless segments of the network.

Related Work: Questions (a) and (c) were recently tackled from an information-theoretic viewpoint by Maddah-Ali and Niesen in [5], [6], for a cache-aided scenario that allows for edge caching but not for cloud processing. Specifically, for a scenario with $M = 3$ ENs and $K = 3$ users, these works present an upper bound on the inverse of the number of achievable degrees of freedom (DoF). In [7], instead, a lower

bound is presented that proves the optimality of the caching and transmission schemes proposed in [5], [6] for a given range of values of the cache storage capacity and under the constraint that no inter-file coding is allowed.

Main Contributions: In contrast to the mentioned prior research, in this paper, we focus on the F-RAN architecture in Fig. 1, which allows for both edge caching and cloud processing by means of fronthaul connections. To this end, we first present an information-theoretic model of F-RANs that succinctly captures its new design aspects and constraints. We also develop a new performance measure, which we refer to as the *Normalized Delivery Time* (NDT). The NDT captures the worst-case latency incurred over the fronthaul and wireless access segments of the network for content delivery as compared to an ideal interference-free system in the high signal-to-noise ratio (SNR) regime. For the case of a caching-only system, with no fronthaul links, the NDT is related to the inverse of the DoF metric studied in [5], [6], as discussed in [7]. Under the constraint of uncoded inter-file caching, we characterize the optimal trade-off between NDT and caching and fronthaul resources for an F-RAN with $M = 2$ ENs and $K = 2$ users (see Fig. 3).

II. Information Theoretic Modeling of F-RAN

As illustrated in Fig. 1, we consider an F-RAN architecture with M edge nodes (ENs), which can serve a set of K users over a shared wireless channel. Note that, in Sec. IV, we will specialize the set-up for our main results to the case $M = K = 2$, but we present here the model in its full generality in order to highlight more general research questions and open problems. We assume the presence of a library of N files, which represent the content that may be requested by the users, where each file is of size L bits. Time is organized into transmission intervals, as shown in Fig. 1 and Fig. 2 and further discussed below. The system model, notation and main assumptions are summarized as follows.

- The library of N contents, or files, that may be requested is denoted by $\mathcal{F} = \{F_1, F_2, \dots, F_N\}$, where each file is of size L bits. This library is assumed to be constant across many transmission intervals. At each transmission interval t, users issue a vector of requests $D(t) = (D_1(t), \dots, D_K(t))$, where $D_k(t) \in \{1, \dots, N\}$ indicates that file $F_{D_k(t)}$ is requested by user k at time t. We make no assumption on the nature of the time-variability of the demands made by the users.
- Each EN has a cache which can store μNL bits, where $\mu \in [0, 1]$ represents the *fractional cache size*.
- The cloud has access to all N files, and each EN is connected to the cloud via a fronthaul link. The fronthaul capacity is given by C_F bits per symbol for each EN, where a symbol refers to a channel use of the downlink wireless channel. The capacity C_F is assumed to be fixed, reflecting conventional scenarios in which fronthaul links correspond to dedicated wired connections (see, e.g., [2], [3]).
- The collective time-varying wireless channel state information (CSI) at transmission interval t is defined as $H(t) = \{\{H_{m,k}(t)\}_{k=1}^{K}\}_{m=1}^{M}$, where $H_{m,k}(t)$ denotes the channel coefficients that characterize the propagation between the mth EN and the kth user. The channel coefficients are assumed to be drawn in an independent identically distributed (i.i.d.) manner from a continuous distribution (as in [5], [6]).

Next, we define the *caching-fronthaul-transmission policy*. We focus on the case in which, in each transmission interval t, the ENs and cloud are aware of the channel realization $H(t)$, as well as of the users' demand vector $D(t)$, but not of any future channel realizations and demands $H(t')$ and $D(t')$ with $t' > t$. As detailed below, the caching-fronthaul-transmission policy decides each ENs' cache composition, which is kept fixed for many transmission intervals, as well as the duration and content of the transmissions across fronthaul and wireless segments for each transmission interval, as shown in Fig. 2.

Definition 1. ***(Caching-Fronthaul-Transmission Policy)*** *A caching-fronthaul-transmission policy $\pi = (\pi_C, \pi_F, \pi_E)$ is characterized by the following three encoding functions.*

a) Caching policy π_C: The caching policy is defined by a function $\mathcal{F} \to \{S_1, S_2, \dots, S_M\}$, which maps the set $\mathcal{F}$ of files into the cache content S_m of the mth EN for $m = 1, ..., M$, which in turn cannot exceed μNL bits. For the scope of this paper, we focus on the practically relevant class of caching policies that do not allow for inter-file coding but do include arbitrary intra-file coding. Within this class, the cache content S_m of the mth EN can be partitioned into N independent sub-caches, i.e., $S_m = (S_{mF_1}, S_{mF_2}, \dots, S_{mF_n})$, where S_{mF_n} can be any arbitrary function of file F_n, for $n = 1, \dots, N$. We emphasize that the caching policy is kept fixed for many transmission intervals and is only a function of the library $\mathcal{F}$ of files.

b) Fronthaul policy π_F: The fronthaul policy is defined by a function $\{\mathcal{F}, D(t), H(t)\} \to \{T_F(t), U_1^{T_F(t)}(t), U_2^{T_F(t)}(t), \dots, U_M^{T_F(t)}(t)\}$, which maps the set of files $\mathcal{F}$, instantaneous demands and channels to the duration $T_F(t)$ of the fronthaul transmission (see Fig. 2) and to the message $U_m^{T_F(t)}(t)$, of duration $T_F(t)$, sent on the mth link. In keeping with the definition of the fronthaul capacity C_F, all time intervals, including $T_F(t)$, are normalized to the symbol transmission time on the downlink wireless channel. Accordingly, the fronthaul message cannot exceed $T_F(t)C_F$ bits. We emphasize that the fronthaul policy, as well as the transmission policy discussed below, adapts to the instantaneous demands and to the CSI in each transmission interval t, unlike the caching policy.

c) Edge transmission policy π_E: The edge transmission policy, or transmission policy for short, is defined by the collection of functions $\pi_E = (\pi_{E_0}, \pi_{E_1}, \pi_{E_2}, \dots, \pi_{E_M})$, which characterize the transmission on the wireless downlink channel as a function of the current demands and CSI and of the fronthaul messages. Specifically, the function $\pi_{E_0} : \{\mathcal{F}, D(t), H(t), U_1^{T_F(t)}(t), \dots, U_M^{T_F(t)}(t)\} \to T_E(t)$ selects the transmission duration, in number of symbols, on the downlink wireless channel for all the ENs (see Fig. 2). Instead, the M local transmission functions $\pi_{E_m} : \{S_m, D(t), H(t), T_E(t), U_m^{T_F(t)}(t)\} \to X_m^{T_E(t)}(t)$, one for

each EN, determine the codeword $X_m^{T_E(t)}(t)$*, of duration* $T_E(t)$ *symbols, sent on the wireless channel by the* m*th EN, under an average power constraint given by the parameter* P*.*

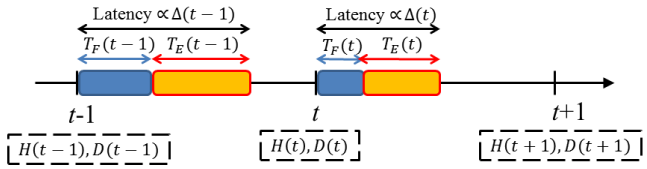

Fig. 2. Illustration of transmission intervals and of the definition of latency.

Upon the transmission by the ENs at a transmission interval t, the kth user receives the signal

$$Y_k^{T_E(t)}(t) = \sum_{m=1}^{N} H_{m,k}(t) X_m^{T_E(t)}(t) + N_k^{T_E(t)}(t), \quad (1)$$

on each channel use of the wireless downlink channel, where $H_{m,k}(t)$ is the channel coefficient between the kth user and the mth EN; $N_k^{T_E(t)}(t)$ denotes the additive noise at the kth user, which is assumed to be complex Gaussian with unitary power, i.i.d. over time and users and also independent of the channel coefficients. Each user k maps its received signal $Y_k^{T_E(t)}(t)$ to an estimate $\hat{F}_{D_k(t)}$ of the demanded file (i.e., $F_{D_k(t)}$), incurring a probability of error $\mathbb{P}(\hat{F}_{D_k(t)} \neq F_{D_k(t)})$. The probability of error of the policy π is then defined for the worst-case request vector as

$$P_e = \max_{D(t)} \max_{k\in\{1,\dots,K\}} \mathbb{P}(\hat{F}_{D_k(t)} \neq F_{D_k(t)}). \quad (2)$$

A policy π is said to be *feasible* if, for almost all realizations $H(t)$ of the channel, i.e., with probability 1, we have $P_e \to 0$ when $L \to \infty$.

III. Latency-Centric DoF Analysis: Normalized Delivery Time (NDT)

We now define the proposed delivery metric by first introducing the notion of delivery time per bit.

Definition 2. *(Delivery time per bit) For a given fractional cache size* μ*, fronthaul capacity* C_F*, and an EN power constraint* P*, consider a sequence of feasible policies* π *indexed by the file size* L*. Denote as* $T_F^{(D,H,L)}$ *and* $T_E^{(D,H,L)}$ *the durations of the fronthaul transmission and edge transmission as in Definition 1 when* $D(t) = D$ *and* $H(t) = H$ *for a file size* L*. The average achievable delivery time per bit for a given sequence of feasible policies is then defined as*

$$\Delta(\mu, C_F, P) = \max_{D} \lim_{L\to\infty} \frac{1}{L} \mathbb{E}_H \left(T_F^{(D,H,L)} + T_E^{(D,H,L)} \right) \quad (3)$$

where the expectation is over the channel realizations H*.*

The delivery time per bit (3) captures the latency within each transmission interval, which is depicted in Fig. 2, as evaluated for the worst-case users' requests and on average over the channel distribution. It is noted that, in order to obtain vanishing probabilities of error, as required by Definition 2, the latencies $T_F^{(D,H,L)}$ and $T_E^{(D,H,L)}$ need to scale with L, and it is this scaling that is measured by (3). We also observe that the definition of delivery time per bit in (3) is akin to the completion time studied in [8], [9] for standard channel models, such as broadcast and multiple access.

The optimal latency performance is in principle obtained by minimizing the delivery time per bit (3) over all possible policies π. This optimization is generally prohibitive and it is also dependent on all parameters (μ, C_F, P). With the aim of obtaining analytical insights, we next propose a novel tractable metric that retains the key dependence of latency on cache size and fronthaul capacity while adopting a high-SNR approximation in the vein of the by now standard DoF analysis of interference networks [10]. To this end, we let the fronthaul capacity scale with the SNR parameter P as $C_F = r\log(P)$, where r is a parameter that 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)*

$$\delta(\mu, r) = \lim_{P\to\infty} \frac{\Delta(\mu, r\log(P), P)}{1/\log(P)} \quad (4)$$

is said to be achievable. For a given pair (μ, r) *the minimum NDT is defined as*

$$\delta^*(\mu, r) = \inf\{\delta(\mu, r) : \delta(\mu, r) \text{ is achievable}\}. \quad (5)$$

In (5), the delivery time per bit (3) is normalized by the term $1/\log(P)$, which measures for the delivery time per bit, at high SNR, of a *baseline system with no interference and unlimited caching, in which each user can be served by a dedicated EN that has all files*. As such, *an NDT of* δ^* *indicates that the worst-case time required to serve any possible request is* δ^* *times larger than the time that would be needed by the baseline system.*

Based on the definitions above, our goal is to characterize the novel metric NDT, $\delta^*(\mu, r)$ that captures the interplay between the normalized cache storage μ and the fronthaul multiplexing gain r. We note that the minimum NDT (5) generally depends on the number N of files, although we do not make this dependence explicit to simplify the notation. We close this section with a key property of NDT that lends further evidence to its suitability as a performance metric for the analysis of F-RANs.

Remark 1 (Time Sharing between Policies and Convexity of Minimum NDT). Consider two (sequences of) policies π_1 and π_2, requiring caching and fronthaul resources (μ_1, r_1) and (μ_2, r_2) and achieving NDTs $\delta^*(\mu_1, r_1)$ and $\delta^*(\mu_2, r_2)$, respectively. An F-RAN is now given with caching and fronthaul resources characterized as $(\mu, r) = (\alpha\mu_1 + (1-\alpha)\mu_2, \alpha r_1 + (1-\alpha)r_2)$ for some parameter $\alpha \in [0, 1]$. On this F-RAN, one could then operate with policy π_1 on a fraction α of the cache storage, fronthaul capacity and spectral resources (i.e., time or frequency), and with policy π_2 on the remaining parts. It can be readily shown, based on additivity arguments, that this

"time- and memory-sharing" strategy achieves an NDT equal to the convex combination $\alpha\delta^*(\mu_1, r_1) + (1-\alpha)\delta^*(\mu_2, r_2)$, which is lower bounded by $\delta^*(\mu, r)$, i.e.,

$$\delta^*(\alpha\mu_1 + (1-\alpha)\mu_2, \alpha r_1 + (1-\alpha)r_2) \leq \alpha\delta^*(\mu_1, r_1) + (1-\alpha)\delta^*(\mu_2, r_2). \quad (6)$$

The above argument demonstrates that the NDT performance measure $\delta^*(\mu, r)$ is jointly convex in (μ, r). We note that a similar observation was made in [5], [6] for a system with caching only (i.e., $r = 0$). This observation motivated the authors of [5], [6] to study the inverse of the DoF metric, instead of the DoF itself, as the performance criterion of interest, since the DoF is shown not to have the same convexity properties (see [7, Remark 2]).

IV. Optimum NDT Trade-Off for $M = K = 2$

In this section, we present the optimum NDT as introduced in Definition 3 for the special case of $M = 2$ and $K = 2$.

Theorem 1. *The optimal NDT trade-off for the $M = 2$-EN, $K = 2$-user F-RAN with number of files $N \geq 2$ is given as*

$$\delta^*(\mu, r) = \begin{cases} \max\left(1 + \mu + \frac{1-2\mu}{r}, 2 - \mu\right) & \text{for } 0 \leq r \leq 1 \\ 1 + \frac{1-\mu}{r} & \text{for } r > 1. \end{cases} \quad (7)$$

As illustrated in Fig. 3, the NDT trade-off analysis in Theorem 1 identifies two distinct regimes in terms of the fronthaul capacity, namely a *low-fronthaul capacity regime* with $r \leq 1$ and a *high-fronthaul capacity regime* with $r > 1$. In the latter case, the use of both fronthaul and caching resources is necessary in order to obtain the optimal NDT performance, while, in the former, if the cache capacity is sufficiently large, namely if $\mu \geq 1/2$, it is sufficient to leverage the caching resources to achieve the optimal performance. We next discuss the caching-fronthaul-transmission policies that achieve the NDT trade-off in Theorem 1. For the converse, we refer to Appendix A.

A. Achievability of Minimum NDT

Here we describe the specific policies that achieve the NDT trade-off curve in Fig. 3. Without loss of generality, we focus on the corner points in both low-fronthaul and high-fronthaul capacity regimes. This is because all the remaining points on the trade-off curve can be achieved by time- and memory-sharing between the policies corresponding to the corner points as per Remark 1.

• Achieving $\delta^*(\mu, r) = 1$ for $\mu = 1$: With $\mu = 1$, each EN can cache all files. This enables full cooperation between the ENs, since each EN can cache the entire library. Thus, the set of ENs forms a virtual multiple-input single-output (MISO) broadcast channel, and zero-forcing beamforming yields parallel interference-free channels to the two users. Therefore, the latency equals the time needed by the mentioned reference interference-free channel with full caching, resulting in the achievable NDT $\delta = 1$.

• Achieving $\delta^*(\mu, r) = 3/2$ for $\mu = 1/2$: With $\mu = 1/2$, each EN can only cache at most half of each file. The caching policy is to split each file F_i into two equal sized sub-files, i.e., $F_i = (F_{i1}, F_{i2})$ for $i = 1, \ldots, N$. EN1 stores the first sub-file, $F_{i1}, i = 1, \ldots, N$, and EN2 stores the second sub-file F_{i2}, for all files $i = 1, \ldots, N$. For any request of distinct files, say F_i by user 1, and F_j by user 2, the transmission problem on the wireless channel is then equivalent to an X-network with four *virtual* messages, namely F_{i1}, F_{j1} available at EN1, and F_{i2}, F_{j2} available at EN2. Thus, we can use the interference alignment scheme proposed in [11], which achieves an equal rate of $2/3\log(P)$ towards each of the two users, ignoring $o(\log(P))$ terms. The delivery time is, neglecting $o(\log(P))$ terms, given by the edge transmission time $T_E = 3L/2\log(P)$, and the delivery time per bit in Definition 2 is then approximately $\Delta = 3/(2\log(P))$, yielding an achievable NDT of $\delta = 3/2$. Interestingly, as shown in Theorem 1, in the low-fronthaul regime when $r \leq 1$, if $\mu \geq 1/2$, the usage of the fronthaul resource cannot further reduce the NDT and this scheme achieves the minimum NDT.

• Achieving $\delta^*(\mu, r) = 1 + 1/r$ for $\mu = 0$: The case $\mu = 0$ corresponds to the setting in which the ENs have no caching capability. A finite NDT can hence only be achieved by using the fronthaul resources. The fronthaul links can be utilized in two distinct ways, referred to here as *hard-* and *soft-transfer* modes. With the *hard-transfer mode*, the cloud can directly transmit both requested files to each EN, and then the ENs can use the same fully cooperative zero-forcing approach adopted above for $\mu = 1$. Since the fronthaul links have capacities $C_F = r\log(P)$ each, the fronthaul delivery time is $T_F = 2L/(r\log(P))$, while the edge transmission time, following the same arguments as for the first corner point, is approximately $T_E = L/\log(P)$. This yields an approximate

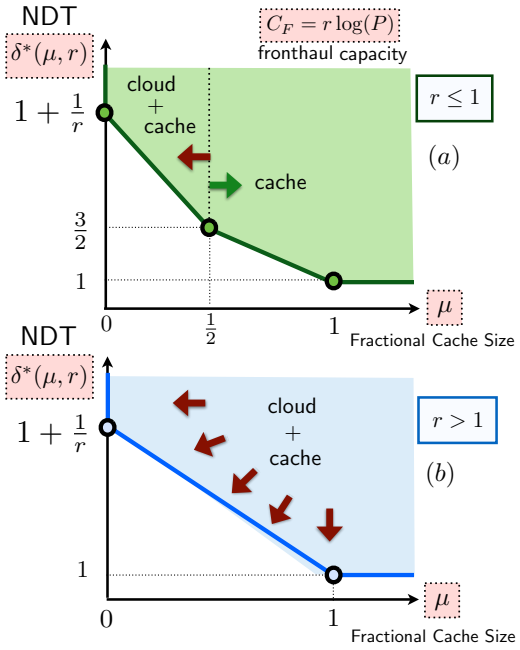

Fig. 3. Optimal NDT trade-off for the $M = 2$-EN, $K = 2$-user F-RAN as a function of μ (fractional cache size per EN) and for fronthaul capacity $C_F = r\log(P)$. The trade-off has distinct regimes of operations: (a) low-fronthaul capacity regime, with $r \leq 1$; (b) high-fronthaul capacity regime, with $r > 1$.

total delivery time per bit of $\Delta = (1 + 2/r)/(\log(P))$, and hence the achievable NDT $\delta = 1+2/r$. However, hard transfer turns out to be suboptimal in this scenario. The optimal NDT is in fact achieved through a *soft-transfer mode* approach typical of C-RAN (see, e.g., [3]): the cloud implements zero-forcing beamforming and quantizes the resulting baseband signals [12]. 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 (see [12, Eq. (5)]). As a result, this scheme entails a fronthaul latency T_F that equals the edge latency T_E of the zero-forcing beamforming scheme, namely $T_E = L/(\log(P))$, multiplied by the time needed to carry each baseband sample on the fronthaul link, namely $\log(P)/(r\log(P))$, yielding the NDT $\delta = 1 + 1/r$ (see also Appendix B for details).

V. Conclusions

In this paper, we presented an information-theoretic analysis of Fog Radio Access Networks (F-RANs), an emerging wireless architecture that encompasses both edge caching and cloud processing. The study aims at providing a latency-centric understanding of the degrees of freedom, in the high-SNR regime, of F-RAN networks by accounting for the available limited resources in terms of fronthaul capacity, cache storage sizes, as well as power and bandwidth on the wireless channel. We detailed a general model and a novel performance metric, referred to as Normalized Delivery Time (NDT), which captures the worst-case delivery latency with respect to an ideal interference-free system. For the special case of $M = 2$ edge nodes and $K = 2$ users, we fully characterized the trade-off between the NDT and the fronthaul and caching resources of the system. This result reveals optimal caching-fronthaul-transmission policies as a function of the system resources. Ongoing work focuses on extending the NDT trade-off to the more general setting introduced here.

References

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

[2] 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 Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, First Quarter 2014.

[3] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud Radio Access Network: Virtualizing Wireless Access for Dense Heterogeneous Systems," *ArXiv e-prints*, Dec. 2015. [Online]. Available: http://arxiv.org/abs/1512.07743

[4] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog Computing based Radio Access Networks: Issues and Challenges," *ArXiv e-prints*, Jun. 2015. [Online]. Available: http://arxiv.org/abs/1506.04233

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

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

[7] A. Sengupta, R. Tandon, and O. Simeone, "Cache Aided Wireless Networks: Tradeoffs between Storage and Latency," *ArXiv e-prints*, Dec. 2015. [Online]. Available: http://arxiv.org/abs/1512.07856

[8] Y. Liu and E. Erkip, "Completion time in multi-access channel: An information theoretic perspective," in *Proc. IEEE Information Theory Workshop*, Paraty, Brazil, 2011, pp. 708–712.

[9] ——, "Completion time in broadcast channel and interference channel," in *Proc. 49th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, 2011, pp. 1694–1701.

[10] S. A. Jafar, "Interference alignment– a new look at signal dimensions in a communication network," *Foundations and Trends in Communications and Information Theory*, vol. 7, no. 1, pp. 1–134, 2010. [Online]. Available: http://dx.doi.org/10.1561/0100000047

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

[12] 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, p. 3, Feb. 2009.

VI. Appendix A: Lower Bounds on NDT (Converse)

In this Appendix, we settle the optimality of the NDT trade-off in Theorem 1 by proving a matching lower bound on the achievable NDT. We recall that S_1 and S_2 denote the contents of the caches of EN1 and EN2, respectively. As in Sec. II, we also denote $U_1^{T_F}$ and $U_2^{T_F}$ the outputs sent on the fronthaul links and $Y_1^{T_E}$ and $Y_2^{T_E}$ denote the vector of channel outputs (1) at the users. Since we focus on a given transmission interval t, we thus drop index t throughout this section. The goal here is to obtain an information theoretic lower bound on $T_E + T_F$ for any given sequence of feasibile policies, from which we can then bound $\delta^*(\mu, r)$. We consider first the case with $N = 2$ files, and a demand vector in which user 1 requests file F_1, and user 2 requests file F_2. We detail later how to extend the proof for $N \geq 2$ files.

The converse is based on the idea of considering *subsets of information resources* across the caching, fronthaul and wireless segments of the F-RAN, with the property that the information in each subset is sufficient to decode one or more of the requested files for any sequence of feasible policies in the high-SNR regime. In particular, the first subset encompasses caching, fronthaul and wireless resources and yields a lower bound on a linear combination of (T_E, T_F) as a function of μ and r. The second subset concerns the caching and fronthaul resources of F-RAN, yielding a lower bound on T_F as a function of μ and r. The third subset pertains only to the wireless segment of F-RAN and yields a lower bound on T_E. The resulting lower bounds are then subsequently used to obtain the lower bound on NDT presented in Theorem 1. Throughout this appendix, we denote as ϵ_L any function that vanishes for $L \to \infty$, and as ϵ_P any function such that $\epsilon_P/\log(P) \to 0$ as $P \to \infty$ (i.e., ϵ_P is any $o(\log(P))$ function).

We now consider an information subset containing the following subset of wireless, cache, and fronthaul resources $\{Y_1^{T_E}, S_2, U_2^{T_F}\}$. The following sequence of bounds holds for any sequence of feasible policies:

$$\begin{aligned} 2L &= H(F_1, F_2) \\ &= I(F_1, F_2; Y_1^{T_E}, S_2, U_2^{T_F}) + H(F_1, F_2 | Y_1^{T_E}, S_2, U_2^{T_F}) \\ &= I(F_1, F_2; Y_1^{T_E}, S_2, U_2^{T_F}) \\ &\quad + H(F_1 | Y_1^{T_E}, S_2, U_2^{T_F}) + H(F_2 | Y_1^{T_E}, F_1, S_2, U_2^{T_F}) \end{aligned}$$

$$
\begin{aligned}
&\le I(F_1,F_2;Y_1^{T_E},S_2,U_2^{T_F}) \\
&\quad + L\epsilon_L + H(F_2|Y_1^{T_E},F_1,S_2,U_2^{T_F}) && \text{(8a)}\\
&= I(F_1,F_2;Y_1^{T_E},S_2,U_2^{T_F}) \\
&\quad + L\epsilon_L + H(F_2|Y_1^{T_E},F_1,X_2^{T_E},S_2,U_2^{T_F}) && \text{(8b)}\\
&= I(F_1,F_2;Y_1^{T_E},S_2,U_2^{T_F}) \\
&\quad + L\epsilon_L + H(F_2|Y_2^{T_E},Y_1^{T_E},F_1,X_2^{T_E},S_2,U_2^{T_F}) + T_E\epsilon_P && \text{(8c)}\\
&\le I(F_1,F_2;Y_1^{T_E},S_2,U_2^{T_F}) + L\epsilon_L + T_E\epsilon_P, && \text{(8d)}
\end{aligned}
$$

where (8a) follows from Fano's inequality by noticing that file F_1 must be decoded based on $Y_1^{T_E}$; equality (8b) follows due to the fact that $X_2^{T_F}$ is a function of $(S_2,U_2^{T_F})$; equality (8c) follows from the equality $I(F_2;Y_2^{T_E}|Y_1^{T_E},F_1,X_2^{T_E},S_2,U_2^{T_F}) = T_E\epsilon_P$ since $X_1^{T_E}$, and then $Y_2^{T_E}$, can be reconstructed from $(Y_1^{T_E},X_2^{T_E})$ subject to additional noise whose variance does not scale with P (also see Lemma 2 in [7] for a similar derivation). Finally, (8d) follows from Fano's inequality using the decodability requirement of file F_2 from $Y_2^{T_E}$.

Now, we bound the first term in (8d) as

$$
\begin{aligned}
&I(F_1,F_2;Y_1^{T_E},S_2,U_2^{T_F}) \\
&\le I(F_1,F_2;Y_1^{T_E},F_1,S_2,U_2^{T_F}) \\
&= I(F_1,F_2;Y_1^{T_E},F_1) + I(F_1,F_2;S_2,U_2^{T_F}|Y_1^{T_E},F_1) \\
&= I(F_1,F_2;Y_1^{T_E}) + I(F_1,F_2;F_1|Y_1^{T_E}) \\
&\quad + I(F_1,F_2;S_2,U_2^{T_F}|Y_1^{T_E},F_1) \\
&\le T_E\log(P) + T_E\epsilon_P + L\epsilon_L + I(F_1,F_2;S_2,U_2^{T_F}|Y_1^{T_E},F_1) && \text{(9a)}\\
&\le T_E\log(P) + T_E\epsilon_P + H(S_2,U_2^{T_F}|Y_1^{T_E},F_1) + L\epsilon_L \\
&\le T_E\log(P) + T_E\epsilon_P + H(S_2,U_2^{T_F}|F_1) + L\epsilon_L \\
&= T_E\log(P) + T_E\epsilon_P + H(S_{2F_1},S_{2F_2},U_2^{T_F}|F_1) + L\epsilon_L && \text{(9b)}\\
&= T_E\log(P) + T_E\epsilon_P + H(S_{2F_2},U_2^{T_F}|F_1) + L\epsilon_L \\
&\le T_E\log(P) + T_E\epsilon_P + H(S_{2F_2},U_2^{T_F}) + L\epsilon_L \\
&\le T_E\log(P) + T_E\epsilon_P + H(S_{2F_2}) + H(U_2^{T_F}) + L\epsilon_L \\
&\le T_E\log(P) + (T_E+T_F)\epsilon_P + \mu L + rT_F\log(P) + L\epsilon_L, && \text{(9c)}
\end{aligned}
$$

where (9a) follows by bounding the first mutual information by $T_E\log(P) + T_E\epsilon_P$ (see Lemma 1 in [7] for a similar derivation) and the second term by Fano's inequality; (9b) follows from the uncoded caching assumption, i.e., cache of EN2 can be expressed as $S_2 = (S_{2F_1}, S_{2F_2})$, where S_{2F_1} is a function of file F_1, and S_{2F_2} is a function of file F_2; and, in (9c), we invoked the caching storage constraint and the fact that the fronthaul capacity is bounded by $r\log(P)$. Plugging (9c) into (8d) and rearranging the resulting inequality, we obtain a bound on a linear combination of (T_E,T_F):

$$
(T_E+rT_F)\log(P) + (T_E+T_F)\epsilon_P \ge (2-\mu)L - L\epsilon_L. \tag{10}
$$

We now consider a second subset of resources that include only caching and fronthaul, namely $\{S_1,S_2,U_1^{T_F},U_2^{T_F}\}$. Again, the following sequence of inequalities holds for any sequence of feasible policies:

$$
\begin{aligned}
2L &\le I(F_1,F_2;S_1,S_2,U_1^{T_F},U_2^{T_F}) + L\epsilon_L && \text{(11a)}\\
&\le H(S_1,S_2,U_1^{T_F},U_2^{T_F}) + L\epsilon_L \\
&\le H(S_1) + H(S_2) + H(U_1^{T_F}) + H(U_2^{T_F}) + L\epsilon_L \\
&\le 4\mu L + 2rT_F\log(P) + L\epsilon_L, && \text{(11b)}
\end{aligned}
$$

where (11a) follows from Fano's inequality, in a manner similar to (8b), since the channel inputs of both the ENs can be obtained from $(S_1,S_2,U_1^{T_F},U_2^{T_F})$ and the ENs must be able to collectively decode the files; and (11b) follows from the cache storage constraint and the constraint on fronthaul capacity. The above inequality gives a lower bound on T_F as

$$
T_F\log(P) \ge \frac{(1-2\mu)L}{r} - \frac{L\epsilon_L}{r}. \tag{12}
$$

Finally, we consider the subset that includes only the wireless resource consisting of the received signal $Y_1^{T_E}$, from which file F_1 must be decodable, yielding the inequalities

$$
L = H(F_1) \le I(F_1;Y_1^{T_E}) + L\epsilon_L \le T_E\log(P) + L\epsilon_L, \tag{13}
$$

which gives the bound $T_E\log(P) \ge L - L\epsilon_L$.

To summarize, we have the following three inequalities corresponding to the three mentioned information subsets:

- Inequality 1:
$$
(T_E+rT_F)\log(P) + (T_E+T_F)\epsilon_P \ge (2-\mu)L - L\epsilon_L; \tag{14}
$$
- Inequality 2:
$$
T_F\log(P) \ge \frac{(1-2\mu)L}{r} - \frac{L\epsilon_L}{r}; \tag{15}
$$
- Inequality 3:
$$
T_E\log(P) \ge L - L\epsilon_L. \tag{16}
$$

We next show how to use the above three inequalities to obtain a lower bound on NDT, which matches Theorem 1.

(a) For the low-fronthaul capacity regime, i.e., $r \le 1$, we combine the inequalities above as (Inequality 1) + $(1-r)\times$ (Inequality 2), which yields

$$
\begin{aligned}
&(T_E+T_F)\log(P) \\
&= (T_E+rT_F)\log(P) + (1-r)T_F\log(P) && \text{(17)}\\
&\ge L\left(1+\mu+\frac{1-2\mu}{r}\right) - (L\epsilon_L/r + (T_E+T_F)\epsilon_P). && \text{(18)}
\end{aligned}
$$

From the above, we thus have the following:

$$
\frac{(T_E+T_F)\log(P)}{L} \ge \frac{\left(1+\mu+\frac{1-2\mu}{r}\right) - \epsilon_L/r}{1+\epsilon_P/\log(P)} \tag{19}
$$

obtain a lower bound on NDT as follows

$$\begin{aligned}\delta^*(\mu,r) &= \lim_{P\to\infty}\lim_{L\to\infty}\frac{(T_E+T_F)\log(P)}{L}\\ &\geq \lim_{P\to\infty}\lim_{L\to\infty}\frac{\left(1+\mu+\frac{1-2\mu}{r}\right)-\epsilon_L/r}{1+\epsilon_P/\log(P)}\\ &= 1+\mu+\frac{1-2\mu}{r}. \end{aligned} \tag{20}$$

For $r \leq 1$, we also use Inequality 1 as follows:

$$(T_E+T_F)\log(P) \geq (T_E+rT_F)\log(P) \tag{21}$$

$$\geq (2-\mu)L - L\epsilon_L - (T_E+T_F)\epsilon_P \tag{22}$$

leading to

$$\frac{(T_E+T_F)\log(P)}{L} \geq \frac{(2-\mu)-\epsilon_L}{1+\epsilon_P/\log(P)}. \tag{23}$$

Using the above inequality, we obtain a lower bound on NDT as follows

$$\begin{aligned}\delta^*(\mu,r) &= \lim_{P\to\infty}\lim_{L\to\infty}\frac{(T_E+T_F)\log(P)}{L}\\ &\geq \lim_{P\to\infty}\lim_{L\to\infty}\frac{(2-\mu)-\epsilon_L}{1+\epsilon_P/\log(P)}\\ &= 2-\mu. \end{aligned} \tag{24}$$

Hence, for $r \leq 1$, the NDT is compactly lower bounded as the minimum of the above two lower bounds:

$$\delta^*(\mu,r) \geq \max\left(1+\mu+\frac{1-2\mu}{r}, 2-\mu\right). \tag{25}$$

It can be readily seen that the first bound is active when $\mu \leq 1/2$, whereas the second one is active for $\mu > 1/2$.

(b) For the low-fronthaul capacity regime, i.e., $r > 1$, we use (Inequality 1) + $(r-1)\times$ (Inequality 3) to obtain

$$\begin{aligned}&r(T_E+T_F)\log(P)\\ &= (T_E+rT_F)\log(P) + (r-1)T_E\log(P)\\ &\geq (2-\mu)L + (r-1)L - (rL\epsilon_L + (T_E+T_F)\epsilon_P)\\ &= (r+1-\mu)L - (rL\epsilon_L + (T_E+T_F)\epsilon_P),\end{aligned}$$

which implies

$$\begin{aligned}&(T_E+T_F)\log(P)\\ &\geq \frac{(r+1-\mu)L}{r} - (L\epsilon_L + (T_E+T_F)\epsilon_P/r),\end{aligned} \tag{26}$$

and hence

$$\delta^*(\mu,r) \geq 1+\frac{1-\mu}{r}. \tag{27}$$

In conclusion, we obtained the lower bound on the NDT

$$\delta^*(\mu,r) \geq \begin{cases}\max\left(1+\mu+\frac{1-2\mu}{r}, 2-\mu\right) & r\leq 1\\ 1+\frac{1-\mu}{r} & r>1.\end{cases} \tag{28}$$

While this bound has been proved above under the assumption of $N = 2$ files, we conclude the proof by noting that the same bound holds to the more general case of $N > 2$ files. To this end, we can use the fact that all files are independent of each other, and hence, in the very first steps used to obtain the Inequalities 1-3, we can introduce the remaining $(N-2)$ files in the conditioning, i.e., $H(F_1,F_2) = H(F_1,F_2|F_3,\ldots,F_N)$. All the remaining steps follow directly, by adding the remaining files $(F_3,\ldots,F_N)$ in the conditioning in all the entropy and mutual information expressions.

VII. Appendix B: Achievable NDT Under the Soft-Transfer Fronthaul Mode

In this Appendix, we present achievability results regarding the NDT that can be obtained by means of soft-transfer mode fronthauling for a general F-RAN with M ENs and K users. Specifically, we discuss the NDT performance of a scheme that uses fronthaul and wireless channels in the standard *serial* fashion that is adopted, for instance, in the CPRI fronthaul interface in C-RAN [2], [3]. In this scheme, the cloud quantizes the encoded baseband samples, and the all the ENs *simultaneously* transmit the quantized baseband signals.

Lemma 1. *The NDT*

$$\delta(\mu,r) = \frac{K}{\min\{M,K\}}\left(1+\frac{1}{r}\right) \tag{29}$$

is achievable by means of soft-transfer fronthauling for any fractional cache size $\mu \geq 0$ and any $r > 0$.

To interpret (29), we note that the NDT $\delta = K/\min\{M,K\}$ can be achieved by means of zero-forcing beamforming in an ideal system in which there is either full caching, i.e., $\mu = 1$, or no fronthaul capacity limitations, i.e., $r \to \infty$. In fact, in such systems, full cooperation is possible at the ENs for any users' demand vector, including the worst case in which users request distinct files, and hence transmission at the maximum per-user multiplexing gain $\min\{M,K\}/K$ can be attained. The achievable NDT (29) hence shows a multiplicative penalty term equal to $1+1/r$ due to fronthaul capacity limitations.

The proof of Lemma 1 relies on the use of the following fronthaul and transmission policies. Note that caching is not used, in accordance with the assumption that μ may be zero. The cloud encodes the signals using zero-forcing beamforming under a power constraints smaller than P that will be specified below. The resulting baseband signals are quantized and sent to the ENs on the fronthaul links. The ENs transmit simultaneously the respective received quantized samples on the wireless channel. Reception at the users is affected by the fronthaul quantization noise, as well as by the channel noise. If the quantization rate is properly chosen, it can be proved that the achievable NDT is (29), where the term $K/\min\{M,K\}$ captures the latency on the wireless channel, which is the same as for the ideal zero-forcing scheme, and $K/(r\min\{M,K\})$ accounts for the delay on the fronthaul. A more detailed discussion is provided next.

In the cloud-based scheme under study, the cloud performs zero-forcing precoding, producing signal $\bar{X}_i$ for each ENi with

power constraint $\bar{P} = E[|\bar{X}_i|^2]$. The signal $\bar{X}_i$ is quantized to obtain the signal X_i that is to be transmitted by ENi as

$$X_i = \bar{X}_i + Z_i, \tag{30}$$

where $Z_i \sim \mathcal{CN}(0, \sigma^2)$ represents the quantization noise with variance σ^2. In order to satisfy the power constraint P, we enforce the condition

$$P = \bar{P} + \sigma^2. \tag{31}$$

Furthermore, denoting as B the number of bits used on the fronthaul link for each baseband sample, from rate-distortion arguments and (30), we obtain the condition $I(X_i; \bar{X}_i) = \log_2(1 + \bar{P}/\sigma^2) = B$, and hence we have

$$\sigma^2 = \frac{\bar{P}}{2^B - 1}. \tag{32}$$

Therefore, from (31) and (32), we obtain the mentioned power constraint on the precoded signal as

$$\bar{P} = P(1 - 2^{-B}) \tag{33}$$

and the quantization noise power as

$$\sigma^2 = P2^{-B}. \tag{34}$$

The quantization noise terms Z_i for all ENs $i = 1, ..., M$ contribute to raising the noise level at each user. In particular, for any user k, the power of the effective noise on the received signals in (1) is given by $E[|N_k|^2] + \sigma^2 \sum_{m=1}^{N} |H_{m,k}|^2 = 1 + \sigma^2 G$, where $G = \sum_{m=1}^{N} |H_{m,k}|^2$. Normalizing the received signal (1) so that the variance of the effective noise is one, we hence obtain, using (33) and (34), an equivalent signal model in which the effective power constraint is

$$\frac{\bar{P}}{1 + \sigma^2 G} = \frac{P(1 - 2^{-B})}{1 + P2^{-B}G}. \tag{35}$$

Now, setting $B = \log(P)$, as indicated in the text, the effective power becomes $(P-1)/(1+G)$, which scales proportionally to P.

Dropping the dependence on D, H and L in order to simplify the notation, we denote as T_E the edge transmission latency for this scheme. It follows that the fronthaul latency is $T_F = BT_E/C_F$, since BT_E bits need to be sent on each fronthaul link at rate $C_F = r \log(P)$ to represent the quantized signals. It follows that the total latency of this scheme is

$$\begin{aligned} T_E + T_F &= T_E \left(1 + \frac{B}{C_F}\right) \\ &= T_E \left(1 + \frac{1}{r}\right), \end{aligned} \tag{36}$$

where we have used the choice $B = \log(P)$. Furthermore, we have the following limit

$$\lim_{P\to\infty} \lim_{L\to\infty} \frac{T_E \log((P-1)(1+G))}{L} = \frac{K}{\min\{M, K\}}, \tag{37}$$

due to the achievability of the NDT $K/\min\{M, K\}$ in the ideal zero-forcing system mentioned above and to the effective power $(P-1)/(1+G)$ for the scheme at hand.

We now conclude the proof by computing the NDT

$$\begin{aligned} &\lim_{P\to\infty} \lim_{L\to\infty} \frac{(T_E + T_F)\log(P)}{L} \\ &= \left(1 + \frac{1}{r}\right) \lim_{P\to\infty} \lim_{L\to\infty} \frac{T_E \log(P)}{L} \\ &= \left(1 + \frac{1}{r}\right) \frac{K}{\min\{M, K\}}, \end{aligned}$$

where the second equality follows due to (37).