Delivery Latency Trade-Offs of Heterogeneous Contents in Fog Radio Access Networks

Jasper Goseling
Stochastic Operations Research,
University of Twente, The Netherlands
j.goseling@utwente.nl

Osvaldo Simeone
Department of Informatics
King's College London, London, UK
osvaldo.simeone@kcl.ac.uk

Petar Popovski
Department of Electronic Systems,
Aalborg University, Denmark
petarp@es.aau.dk

***Abstract*—A Fog Radio Access Network (F-RAN) is a cellular wireless system that enables content delivery via the caching of popular content at edge nodes (ENs) and cloud processing. The existing information-theoretic analyses of F-RAN systems, and special cases thereof, make the assumption that all requests should be guaranteed the same delivery latency, which results in identical latency for all files in the content library. In practice, however, contents may have heterogeneous timeliness requirements depending on the applications that operate on them. Given per-EN cache capacity constraint, there exists a fundamental trade-off among the delivery latencies of different users' requests, since contents that are allocated more cache space generally enjoy lower delivery latencies. For the case with two ENs and two users, the optimal latency trade-off is characterized in the high-SNR regime in terms of the Normalized Delivery Time (NDT) metric. The main results are illustrated by numerical examples.**

***Index Terms*—Edge caching, Cloud Radio Access Network, Fog Radio Access Network, Normalized Delivery Time.**

I. Introduction

Fog networking is a novel paradigm in which computing, storage and communication functions are implemented at both cloud and edge nodes (ENs), such as base stations, of a wireless cellular system. As Fig. 1 shows, content delivery can benefit from fog networking via edge caching (storing popular content at the ENs), as well as via cloud processing, which enables the delivery of content fetched from a central content library.

The information-theoretic analysis of edge caching in [1]–[4] and of more general fog-assisted wireless networks, or Fog Radio Access Networks (F-RANs), in [5] makes the assumption that all files in the content library have the same timeliness constraint. Under this assumption, caching schemes in which all contents are allocated the same fraction of the ENs' caches were proven to be optimal or near-optimal in [5]. In practice, however, contents may have heterogeneous latency requirements; e.g. video chunks may be buffered to reduce the delay constraints, while information feeding an Augmented Reality (AR) application has stricter latency requirements. Reducing the delivery latency of a content type generally requires allocating a larger fraction of the ENs' cache capacity to it, which in turn increases the delivery latency of other contents.

As in [1]–[5], in this paper, users are assumed to make simultaneous requests from a library of contents, which may be partially cached at the ENs during an offline caching phase. Unlike prior work, in which all request sets experience the same delivery coding latency, here we study the trade-offs among the latencies that are achievable across different request sets, when one allows arbitrary allocation of cache capacity at the ENs across files. Leveraging a fine-grained understanding of these trade-offs makes it possible to analyze individual content latency constraints, including the average latency for a content type under a probabilistic popularity model.

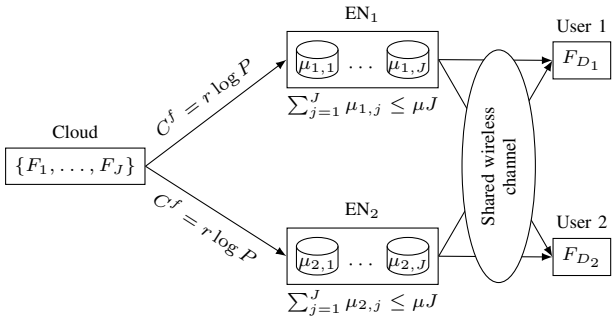


Fig. 1. Illustration of the F-RAN system under study for $M = 2$ and $K = 2$.

As in [5], as well as in [3], [6], delivery latencies are measured here in the high-Signal-to-Noise Radio (SNR) regime with respect to a reference interference-free system, yielding the performance metric of Normalized Delivery Time (NDT). In [5], the minimum NDT was characterized under the assumption that all users' requests should be guaranteed the same latency; upper and lower bounds that are within a multiplicative factor of 2 for any number of ENs and users.

Focusing on the special case with two ENs and two users the main contributions of this work are as follows: (*i*) The performance metric of the *NDT region* is introduced with the aim of analyzing the trade-off among the latencies achievable for individual users' requests under non-uniform cache partitions across files (Sec. II); (*ii*) Novel achievable schemes are presented that yield an inner bound to the NDT region (Sec. IV); (*iii*) Outer bounds on the NDT region are derived that conclusively characterize the NDT region (Sec. V); (*iv*) Numerical results corroborate the analysis (Sec. VI).

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

II. System Model

A. Model

We consider an F-RAN architecture with M edge nodes (ENs), which serve K users over a shared wireless channel, see Fig. 1. As in prior works [1]–[5], the system operates in two separate phases, namely an (offline) caching phase and an (online) delivery phase. In both phases, the content library $\mathcal{F} = \{F_1, \dots, F_J\}$ of $J \geq K$ files, where each file is of length L bits, is fixed and static. The assumption of equally-sized files simplifies the treatment, as in prior work, and should be alleviated in future studies. In the caching phase, each EN m can cache at most μJL bits from the library, where $0 \leq \mu \leq 1$ is referred to as the *fractional cache capacity*.

The delivery phase consists of an arbitrary number of slots. In any slot, each user k requests a file F_{D_k} in $\mathcal{F}$ with index $D_k \in [1:J]$. We let $D = [D_1, \dots, D_K]$ denote the vector of requested files in a slot. We make the assumption that the requested files are distinct, as e.g. in [7]. In future work, we plan to alleviate this limitation. The main goal of this work is understanding the trade-offs achievable among the delivery latencies that are achievable for different request vectors D. As we discuss in Sec. VI, this fine-grained understanding of the trade-offs among the delivery latencies for different requests can be used to study individual latency requirements for different files under a given popularity distribution.

The channel from the ENs to the users is defined by:

$$Y_k = \sum_{m=1}^{M} H_{m,k} X_m + Z_k, \tag{1}$$

which is a standard quasi-static model, where $X_m \in \mathbb{C}^{n^e}$ is a codeword of length n^e symbols transmitted by the EN m; $H_{m,k} \in \mathbb{C}$ is the channel coefficient from EN m to user k; Z_k is complex Gaussian additive noise with unitary power, i.i.d. over time and users and also independent of the channel coefficients; and $Y_k \in \mathbb{C}^{n^e}$ is the received signal of length n^e symbols by user k. The channel coefficients are realizations of continuous random variables, and are i.i.d. over ENs and users. Using the notation $[1:K] = \{1, \dots, K\}$, let $\mathcal{H} = (H_{m,k})_{m\in[1:M],k\in[1:K]}$ denote the channel state information (CSI), which is assumed to be known throughout the network, i.e., at the cloud, the ENs and the users.

The cloud has orthogonal fronthaul links to each of the ENs. Using the parametrization in [5], the capacity is measured in bits per symbol, where a *symbol* is a channel use of the wireless channel. Furthermore, as in [5], the fronthaul capacity is written as $C^f = r \log(P)$, where r is defined as the *fronthaul rate* and P is the average high SNR of the wireless edge links. The fronthaul rate describes the ratio between the fronthaul capacity and the high-SNR capacity of each EN-to-user when used with no interference from other links.

In the *caching phase*, EN m stores an arbitrary function

$$S_{m,j} = \pi^c_{m,j}(F_j) \tag{2}$$

of each file $F_j, j \in [1:J]$,. We allow for an arbitrary partition of each EN's cache capacity. Denoting the entropy of the cached content for file F_j at EN m as $H(S_{m,j}) = \mu_{m,j} L$, with $0 \leq \mu_{m,j} \leq 1$, we impose that the cache partition $\{\mu_{m,j}\}_{j\in[1:J]}$ satisfies the per-EN cache capacity constraint

$$\sum_{j=1}^{J} \mu_{m,j} \leq \mu J, \tag{3}$$

for each m. We will refer to $\pi^c = (\pi^c_{m,j})_{m\in[1:M],j\in[1:J]}$ as the caching policy and to the matrix

$$\boldsymbol{\mu} = (\mu_{m,j})_{m\in[1:M],j\in[1:J]} \tag{4}$$

as the *cache partition matrix* for the given caching policy.

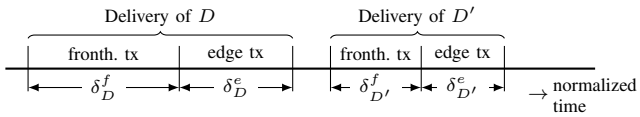


Fig. 2. Delivery consists of a fronthaul transmission and an edge transmission phase. The length of these phases depends on the files that are requested.

Each slot of the *delivery phase* consists of two subsequent subslots (Fig. 2). In the *first subslot*, the cloud sends information on the requested files to the ENs on the fronthaul links, while in the *second subslot* the ENs use the shared wireless channel to transmit to the users. To elaborate, in the first subslot, the cloud sends a message U_m to the EN m on the fronthaul as a function of the demand vector, the files and the CSI

$$U_m = \pi^f_m(D, \mathcal{F}, \mathcal{H}). \tag{5}$$

The first subslot has n^f_D symbols, where we make explicit the dependence on the vector D, and the entropy of message U_m must be bounded as $H(U_m) \leq C^f n^f_D$ in order to satisfy the fronthaul capacity constraints. We call $\pi^f = (\pi^f_1, \dots, \pi^f_M)$ the fronthaul policy. In the second subslot, the ENs transmit a codeword X_m, of n^e_D symbols, on the wireless channel as a function of the users' demand D, the cache content $S_m = \{S_{m,j}\}_{j\in[1:J]}$ of EN m, the fronthaul message U_m to EN m and the global CSI $\mathcal{H}$:

$$X_m = \pi^e_m(D, S_m, U_m, \mathcal{H}). \tag{6}$$

We call $\pi^e = (\pi^e_1, \dots, \pi^e_M)$ the edge transmission policy. After receiving Y_k in (1), user k decodes the requested file as

$$\hat{F}_{D_k} = \pi^d_k(Y_k, D, \mathcal{H}), \tag{7}$$

and we let $\pi^d = (\pi^d_1, \dots, \pi^d_K)$ denote the decoding policy. The error probability of a policy $\pi = (\pi^c, \pi^f, \pi^e, \pi^d)$ is defined as the worst-case error probability across requests and users

$$P_e = \max_{D} \max_{k\in[1:K]} P(\hat{F}_k \neq F_{D_k}). \tag{8}$$

A sequence of policies, parametrized by L and P, is defined as *feasible* if it satisfies the limit $\lim_{P\to\infty} \lim_{L\to\infty} P_e = 0$.

B. Problem Statement

For any sequence of feasible policies π parametrized by L and P, we now define the high-SNR delivery time metric for each demand vector D. To this end, we introduce the normalized durations of the first and second subslots in the given transmission interval as

$$\delta_D^x = \lim_{P\to\infty} \lim_{L\to\infty} \frac{n_D^x}{L/\log P}, \tag{9}$$

where $x = f$ for the first subslot (fronthaul transmission) and $x = e$ for the second subslot (edge transmission). In (9), the subslot durations are normalized by the high-SNR delivery time of a reference system in which each user is served on an interference-free dedicated channel by an EN, namely $L/\log P$. Note, in fact, that an interference-free channel has a high-SNR capacity of $\log P$ (see also [5] for additional discussion). We refer to δ_D^f and δ_D^e as the *fronthaul and edge NDTs*, respectively, for request D. The overall NDT for request D is hence given by $\delta_D = \delta_D^e + \delta_D^f$.

We are interested in characterizing the region $\Delta^*(\mu, r)$ of all achievable NDT tuples $\delta = (\delta_D)_{D\in\mathcal{D}}$ under the per-EN capacity constraint (3), which we refer to as *NDT region*. We impose that the same NDT δ_D be achieved for all permutations of the vector D. This allows us to obtain a characterization that depends only on the subset of files that are requested. Henceforth, with a slight abuse of notation, D represents a subset of $[1 : J]$. As a result, the NDT region is contained in the positive orthant of $\mathbb{R}^{\binom{J}{K}}$.

To study the NDT region $\Delta^*(\mu, r)$ it is convenient to analyze also the region $\Delta^*(\boldsymbol{\mu}, r)$ of all NDT tuples that are achievable with a given cache partition matrix $\boldsymbol{\mu}$ in (4). By definition, we have

$$\Delta^*(\mu, r) = \bigcup_{\boldsymbol{\mu}:\, \sum_{j=1}^{J} \mu_{m,j} \le \mu J, \forall m} \Delta^*(\boldsymbol{\mu}, r). \tag{10}$$

A first observation is summarized in the following lemma.

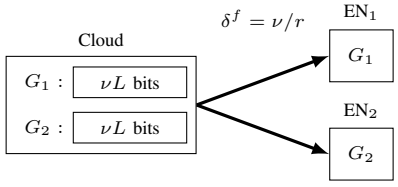
Lemma 1. *The NDT regions $\Delta^*(\mu, r)$ and $\Delta^*(\boldsymbol{\mu}, r)$ are convex.*

We finally note that the minimum NDT introduced in [5], corresponds to the minimum value δ in the NDT region $\Delta^*(\boldsymbol{\mu}, r)$, with equal cache partition $\mu_{m,j} = \mu$, such that the equality $\delta = \delta_D$ holds for all request subsets D. In the rest of this paper, we focus on the special case $K = M = 2$ and we write $\delta_D = \delta_{i,j}$ for any request subset $D = \{i, j\}$.

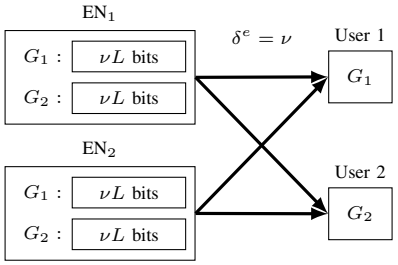
Due to space constraints proofs are omitted from this paper. All proofs appear in [8].

III. Preliminaries

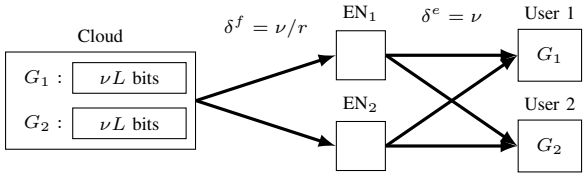
Here we review delivery strategies, see Fig. 3, for the fronthaul and edge channels from [5], which will be used as ingredients in the next section to propose a more general caching and delivery policy. (1) *Hard-transfer fronthauling* (HT, Fig. 3a): As shows, via the HT fronthaul delivery strategy, the cloud delivers a fraction ν of one of the requested files, say G_1, to EN 1 and a fraction of the other file G_2 to EN 2 on the respective fronthaul links. (2) *Zero-forcing beamforming* (ZF, Fig. 3b): If both ENs have both requested messages G_1 and G_2, or a fraction ν thereof, available in the respective caches, the edge delivery strategy of cooperative ZF beamforming can be carried out on this fraction to deliver G_i to user i, yielding parallel interference-free channels to both users. (3) *Soft-transfer fronthauling with zero-forcing beamforming* (ST+ZF, Fig. 3c): With the fronthaul-edge delivery strategy, the cloud implements ZF beamforming and transmits the resulting baseband signals to the ENs in quantized form. The ENs simply forward the quantized signals over the shared wireless channel [9]. (4) *X-channel interference alignment* (X-IA, Fig. 3d): If the ENs cache different fractions ν of each requested file, the resulting channel model for the delivery for this fraction is an X-channel, for which interference alignment (IA) edge delivery strategies were presented in [10].



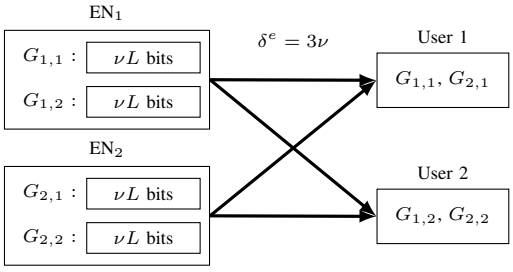
(a) Hard-transfer fronthauling (HT).



(b) Zero-forcing beamforming (ZF).

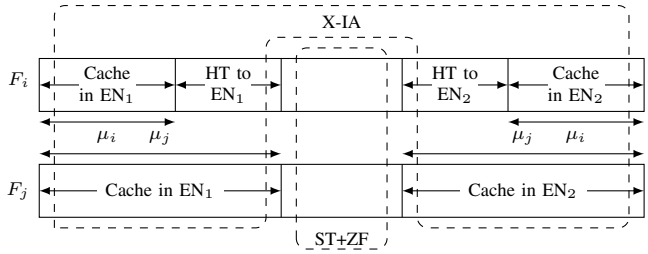


(c) Soft-transfer fronthauling with zero-forcing beamforming (ST+ZF).

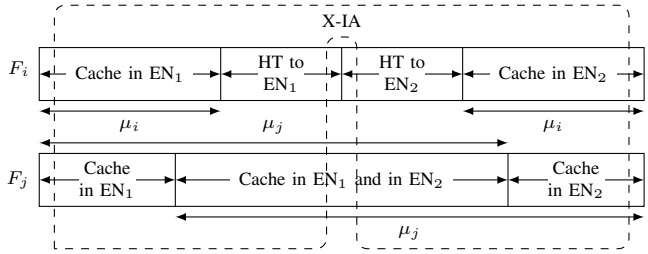


(d) X-channel interference alignment (X-IA).

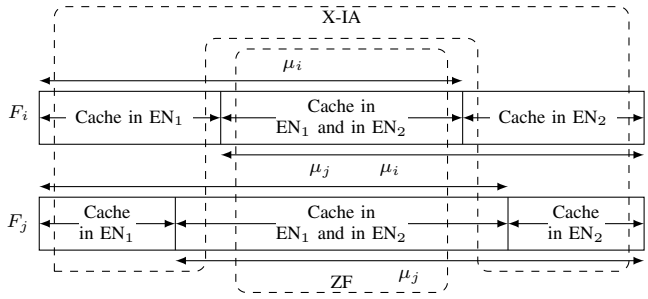
Fig. 3. Illustration of constituent delivery strategies.



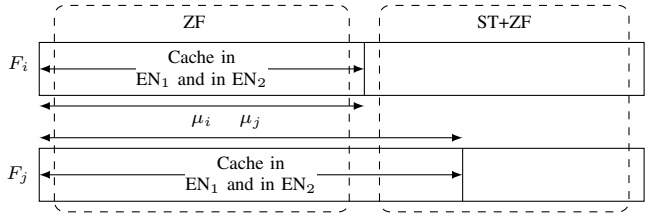
(a) Achievable strategy for $r \leq 1$, $\mu_i < 1/2$ and $\mu_j < 1/2$.



(b) Achievable strategy for $r \leq 1$ and $\mu_i \leq 1/2 \leq \mu_j$.



(c) Achievable strategy for $r \leq 1$ and $\mu_i > 1/2$.



(d) Achievable strategy for $r > 1$.

Fig. 4. Achievable strategies

Lemma 2. *[5] The following fronthaul and edge NDTs are achievable using the delivery strategies summarized above.*
HT: *Let G_1 and G_2 be messages of νL bits that are available in the cloud. HT requires the fronthaul NDT $\delta^f = \frac{\nu}{r}$ to transmit G_1 to EN_1 and G_2 to EN_2.*
ZF: *Let both ENs have messages G_1 and G_2 of νL bits available. ZF requires the edge NDT $\delta^e = \nu$ to transmit G_1 to user 1 and G_2 to user 2.*
ST+ZF: *Let G_1 and G_2 be messages of νL bits that are available in the cloud. ST+ZF requires the fronthaul and edge NDTs $\delta^f = \frac{\nu}{r}$ and $\delta^e = \nu$ to transmit G_1 to user 1 and G_2 to user 2.*
X-IA: *Let $G_{i,1}$ and $G_{i,2}$ be messages of νL bits that are available at EN_i, $i = 1, 2$. X-IA requires the edge NDT $\delta^e = 3\nu$ to transmit $G_{1,1}$ and $G_{2,1}$ to user 1 and $G_{1,2}$ and $G_{2,2}$ to user 2.*

IV. Achievable NDT Region

In this section, we present achievable strategies that yield an inner bound on the NDT region $\Delta^*(\mu, r)$. To this end, we consider policies with cache partitions $\boldsymbol{\mu}$ such that the two ENs cache the same number of bits for each file, i.e., $\mu_{1,j} = \mu_{2,j} = \mu_j$. As a result, each file F_j is generally allocated a different cache fraction μ_j at the ENs. We will show in the next section that this restriction comes with no loss of optimality.

Theorem 1. *An inner bound on the NDT region is given by the inclusion*

$$\Delta^*(\mu, r) \supseteq \Delta^{(\mathsf{in})}(\mu, r) = \bigcup_{\substack{\boldsymbol{\mu}:\mu_{1,i}=\mu_{2,i}=\mu_i, \\ \sum_{j=1}^{J} \mu_j \leq \mu J}} \Delta^{(\mathsf{in})}(\boldsymbol{\mu}, r), \quad (11)$$

where the region

$$\Delta^{(\mathsf{in})}(\boldsymbol{\mu}, r) = \left\{ \delta_D \middle| \delta_{i,j} \geq \delta_{i,j}^{(\mathsf{in})}(\boldsymbol{\mu}, r), \forall \{i,j\} \in \mathcal{D} \right\} \quad (12)$$

is included in $\Delta^*(\boldsymbol{\mu}, r)$, and we have

$$\delta_{i,j}^{(\mathsf{in})}(\boldsymbol{\mu}, r) = \begin{cases} \delta_{i,j}^{(\mathsf{in},1)}(\boldsymbol{\mu}, r), & \text{if } r \leq 1, \mu_i < \frac{1}{2} \text{ and } \mu_j < \frac{1}{2}, \\ \delta_{i,j}^{(\mathsf{in},2)}(\boldsymbol{\mu}, r), & \text{if } r \leq 1 \text{ and} \\ & (\mu_i \leq \frac{1}{2} \leq \mu_j, \text{ or } \mu_j \leq \frac{1}{2} \leq \mu_i), \\ \delta_{i,j}^{(\mathsf{in},3)}(\boldsymbol{\mu}, r), & \text{if } r \leq 1, \mu_i > \frac{1}{2} \text{ and } \mu_j > \frac{1}{2}, \\ \delta_{i,j}^{(\mathsf{in},4)}(\boldsymbol{\mu}, r), & \text{if } r > 1, \end{cases} \quad (13)$$

with the definitions

$$\delta_{i,j}^{(\mathsf{in},1)}(\boldsymbol{\mu}, r) = 1 + \frac{1}{r} - \left(\frac{1}{r} - 1\right) \max\{\mu_i, \mu_j\} - \frac{1}{r} \min\{\mu_i, \mu_j\}, \quad (14)$$

$$\delta_{i,j}^{(\mathsf{in},2)}(\boldsymbol{\mu}, r) = \frac{3}{2} + \frac{1}{r}\left(\frac{1}{2} - \min\{\mu_i, \mu_j\}\right), \quad (15)$$

$$\delta_{i,j}^{(\mathsf{in},3)}(\boldsymbol{\mu}, r) = 2 - \min\{\mu_i, \mu_j\}, \quad (16)$$

$$\delta_{i,j}^{(\mathsf{in},4)}(\boldsymbol{\mu}, r) = 1 + \frac{1}{r} - \frac{1}{r} \min\{\mu_i, \mu_j\}. \quad (17)$$

In the remainder of this section, we present the achievable strategies that yield the inner bound in the previous theorem at an intuitive level. To this end, we will present two different caching policies for the cases $r \leq 1$ and $r > 1$. Note that the caching policy cannot depend on the demand $D = \{i, j\}$, unlike the delivery policy. In the following, we set $\mu_i \leq \mu_j$ without loss of generality.
Caching policy for $r \leq 1$: As seen in Figures 4a–4c, each file F_i is cached so that the bits indexed by $1, \ldots, \mu_i L$ are stored in EN_1 and bits $(1 - \mu_i)L, \ldots, L$ are stored in EN_2, i.e., we minimize the overlap in the cached content in EN_1 and EN_2 by storing the first part of the file in EN_1 and the last part

of the file in EN_2. If $\mu_i \geq 1/2$ some overlap will occur and some bits will be stored in both ENs.

Caching policy for $r > 1$: As seen in Figure 4d, each file F_i is cached so that bits $1, \ldots, \mu_i L$ in both EN_1 and EN_2, i.e., we cache only the first part of the file and we maximize the overlap between the content that is cached in the ENs.

Delivery strategy for $\mu_j < 1/2$ and $r \leq 1$: This case is illustrated in Figure 4a and achieves $\delta_{i,j}^{(\mathsf{in},1)}(\boldsymbol{\mu}, r)$. The delivery proceeds in three phases: a) we use HT to deliver bits $\mu_i L, \ldots \mu_j L$ and $(1-\mu_j)L, \ldots, (1-\mu_i)L$ of file F_i to EN_1 and EN_2, respectively; b) we use X-IA to transmit bits $1, \ldots, \mu_j L$ and $(1-\mu_j)L, \ldots, L$ of the files from the ENs to the users; c) we use ST+ZF to transmit bits $\mu_j L, \ldots, (1-\mu_j)L$ directly from the cloud. Note that the strategy in [5] does not require step a). In fact, interestingly, the optimal policy in [5] did not make any use of HT fronthauling. The optimality results presented in the next section demonstrate that, instead, when $\mu_i \neq \mu_j$, the joint use of both HT and ST are instrumental in achieving the optimal NDT performance.

Delivery strategy for $\mu_i \leq 1/2 \leq \mu_j$ and $r \leq 1$: This case is illustrated in Figure 4b and achieves $\delta_{i,j}^{(\mathsf{in},2)}(\boldsymbol{\mu}, r)$. The delivery proceeds in two phases: a) we use HT to deliver bits $\mu_i L, \ldots L/2$ and $L/2, \ldots, (1-\mu_i)L$ of file F_i to EN_1 and EN_2, respectively; and b) we use X-IA to deliver both complete files from the ENs to the users. We remark that this scenario is not relevant for the special case from [5]. We emphasize the important role of HT for deriving an achievable strategy, which is used here but not in [5],

Delivery strategy for $\mu_i > 1/2$ and $r \leq 1$: This case is illustrated in Figure 4c and achieves $\delta_{i,j}^{(\mathsf{in},3)}(\boldsymbol{\mu}, r)$. The delivery proceeds in two phases: a) we use X-IA to deliver bits $1, \ldots, (1-\mu_i)L$ and $\mu_i L, \ldots, L$ of the files from the ENs to the users; and b) we use ZF to transmit bits $(1-\mu_i)L, \ldots, \mu_i L$ from the ENs to the users. Note that this strategy does not make use of the fronthaul during the delivery.

Delivery strategy for $r > 1$: The first $\min\{\mu_i, \mu_j\}L$ bits of both files, which are stored at both ENs, are delivered using ZF. The remaining $(1-\min\{\mu_i, \mu_j\})L$ bits are delivered using ST+ZF. The strategy is illustrated in Fig. 4d and achieves $\delta_{i,j}^{(\mathsf{in},4)}(\boldsymbol{\mu}, r)$.

V. Characterization of the NDT Region

In this section, we present an outer bound on the NDT region $\Delta^*(\mu, r)$ and we prove that the inner bound from the previous section is in fact tight, hence characterizing the NDT region. The first result of this section provides an outer bound on the achievable NDT tuple region for a fixed, and generic, cache partition $\boldsymbol{\mu}$.

Theorem 2. *For any cache partition $\boldsymbol{\mu}$, we have the outer bound $\Delta^*(\boldsymbol{\mu}, r) \subseteq \Delta^{(\mathsf{out})}(\boldsymbol{\mu}, r)$, where*

$$\Delta^{(\mathsf{out})}(\boldsymbol{\mu}, r) = \left\{ \delta \;\middle|\; \delta_{i,j} \geq \delta_{i,j}^{(\mathsf{out})}(\boldsymbol{\mu}, r), \forall \{i,j\} \in \mathcal{D} \right\}, \quad (18)$$

with

$$\delta_{i,j}^{(\mathsf{out})}(\boldsymbol{\mu}, r) = \begin{cases} \max_{\ell=1,\ldots,3} \left\{ \delta_{i,j}^{(\mathsf{out},\ell)}(\boldsymbol{\mu}, r) \right\}, & \text{if } r \leq 1, \\ \delta_{i,j}^{(\mathsf{out},4)}(\boldsymbol{\mu}, r), & \text{if } r > 1, \end{cases} \quad (19)$$

and the definitions

$$\delta_{i,j}^{(\mathsf{out},1)}(\boldsymbol{\mu}, r) = 1 + \frac{1}{r} - \min\{\mu_{1,i}, \mu_{2,i}, \mu_{1,j}, \mu_{2,j}\} - \frac{1}{2}\left(\frac{1}{r} - 1\right)(\mu_{1,i} + \mu_{2,i} + \mu_{1,j} + \mu_{2,j}), \quad (20)$$

$$\delta_{i,j}^{(\mathsf{out},2)}(\boldsymbol{\mu}, r) = \frac{3}{2} + \frac{1}{2r} - \min\{\mu_{1,i}, \mu_{2,i}, \mu_{1,j}, \mu_{2,j}\} - \frac{1}{2}\left(\frac{1}{r} - 1\right)\min\{\mu_{1,i} + \mu_{2,i}, \mu_{1,j} + \mu_{2,j}\}, \quad (21)$$

$$\delta_{i,j}^{(\mathsf{out},3)}(\boldsymbol{\mu}, r) = 2 - \min\{\mu_{1,i}, \mu_{2,i}, \mu_{1,j}, \mu_{2,j}\}, \quad (22)$$

$$\delta_{i,j}^{(\mathsf{out},4)}(\boldsymbol{\mu}, r) = 1 + \frac{1}{r} - \frac{1}{r}\min\{\mu_{1,i}, \mu_{2,i}, \mu_{1,j}, \mu_{2,j}\}. \quad (23)$$

Using the outer bound in the previous theorem, we show that the inner bound from the previous section is tight. This result implies that, in order to exhaust the NDT region, it is sufficient to consider cache partitions in which $\mu_{1,j} = \mu_{2,j}$ for all files $j \in [1:J]$.

Theorem 3. *The NDT region is given as $\Delta^*(\mu, r) = \Delta^{(\mathsf{in})}(\mu, r)$.*

VI. Numerical Example

Consider a set-up in which the set of popular files is partitioned into two disjoint classes as $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$, where class $\mathcal{F}_i$ has J_i files. We first illustrate the NDT region derived above and we then discuss how this can be used to obtain optimal trade-offs among the individual latencies of different files under a popularity distribution.

We first illustrate a slice of the NDT region in which we impose that the same NDT $\delta_{(i),(j)}$ be achieved for all subsets D for which one file is in the class $\mathcal{F}_i$ and the other in $\mathcal{F}_j$. Recall that we assume that the requested files are distinct (even if requested from the same class). The considered slice of the NDT region is three-dimensional with axes given by $\delta_{(1),(1)}$, $\delta_{(2),(2)}$ and $\delta_{(1),(2)}$. To further reduce the dimensionality, we let $\delta_{(2),(2)}$ be arbitrary, so as to focus only on the plane $(\delta_{(1),(2)}, \delta_{(1),(1)})$. In order to evaluate the boundary of this slice of the NDT region, it can be argued that it is sufficient to consider cache partitions such as all files within the same class, which we denote as $\mu_{(1)}$ and $\mu_{(2)}$ for the files of class 1 and 2, respectively. With this choice, the cache capacity constraint (3) reduces to $J_1\mu_{(1)} + J_2\mu_{(2)} \leq \mu(J_1 + J_2)$.

The slice of the NDT region at hand is illustrated in Figure 5 for $r = 1/5$, $J_1 = J_2$ and various values of μ. The figure also indicates the values of the cache allocations $(\mu_{(1)}, \mu_{(2)})$ that are required to obtain various points on the boundary of the region as well as the delivery strategy that should be used at various segments of the boundary. As it can be seen, the slice

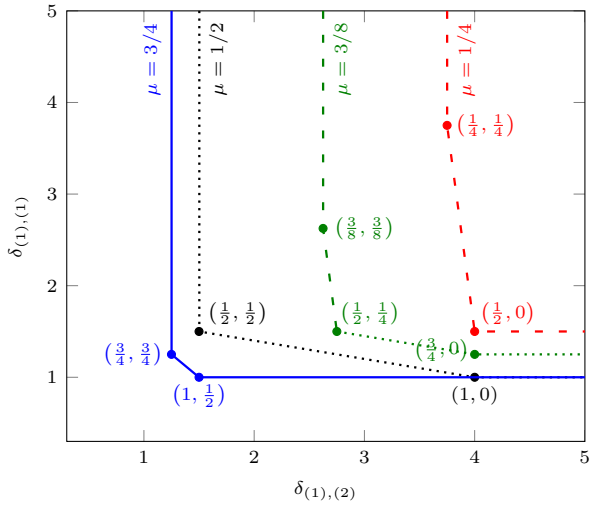


Fig. 5. Slice of the NDT region as a function of the fractional cache capacity μ ($r = 1/5$, $J_1 = J_2$). The labels indicate the cache allocations $(\mu_{(1)}, \mu_{(2)})$ that are required at specific points. The line styles indicate the strategy to be used, with the dashed line corresponding to $\delta_{i,j}^{(\text{in},1)}$, the dotted lines to $\delta_{i,j}^{(\text{in},2)}$, and the solid line to $\delta_{i,j}^{(\text{in},3)}$.

of the NDT region is a polyhedron and each linear portion of the boundary corresponds to a different delivery strategy as indicated in the figure (see Sec. III for a correspondence between strategies and NDT tuples $\delta^{(\text{in},\ell)}$). For instance, it is seen that, for $\mu = 3/8$, one has to use a different strategy depending on the operating point: as $\delta_{(1),(2)}$ increases, one needs to switch between the strategies that achieve the NDTs $\delta_{i,j}^{(\text{in},1)}$ and $\delta_{i,j}^{(\text{in},2)}$.

Finally, we consider individual latency constraints for different files under a given popularity profile. To this end, we let a and $1-a$ denote the probabilities that a file is requested from class $\mathcal{F}_1$ and from class $\mathcal{F}_2$, respectively. We then have that the probability p_{11} that two files from class $\mathcal{F}_1$ are selected is $p_{11} = a^2$; the probability p_{12} that one file from each class is requested is $p_{12} = 2a(1-a)$; and the probability p_{22} that two files from class $\mathcal{F}_2$ are requested is $p_{22} = (1-a)^2$. The average latency for a file from a given class is:

$$\bar{\delta}_{(1)} = \mathbb{E}[\delta_{(1)}] = \frac{p_{11}}{p_{11}+p_{12}}\delta_{(1),(1)} + \frac{p_{12}}{p_{11}+p_{12}}\delta_{(1),(2)}, \quad (24)$$

$$\bar{\delta}_{(2)} = \mathbb{E}[\delta_{(2)}] = \frac{p_{22}}{p_{22}+p_{12}}\delta_{(1),(1)} + \frac{p_{12}}{p_{22}+p_{12}}\delta_{(1),(2)}. \quad (25)$$

Note that $\bar{\delta}_{(i)}$ is the average latency for files of class $\mathcal{F}_{(i)}$ when averaged over the second requested file.

In Figure 6, we illustrate the optimal trade-off between the average NDTs $\bar{\delta}_{(1)}$ and $\bar{\delta}_{(2)}$ that arises from adjusting the cache allocations among the two classes. In the figure we have set $\mu = 3/8$, $r = 1/5$, $J_1 = J_2$ and considered various values of a. The figure confirms that obtaining lower average delivery latencies for some files entails a larger average delivery latencies for other files due to the limited cache capacities.

VII. Conclusions

This work characterized the set of delivery latencies supported by an F-RAN with two ENs and two users in the high SNR regime, when allowing for any cache partition across the files in a set of popular contents. Various aspects call for further investigation, including the explicit minimization of the average delivery latency as a function of the content popularity profile, the extension of the main results to any number of ENs and users (see [5] for the case of uniform file popularity) and the derivation of an extended NDT region in which the same contents may be requested by multiple users.

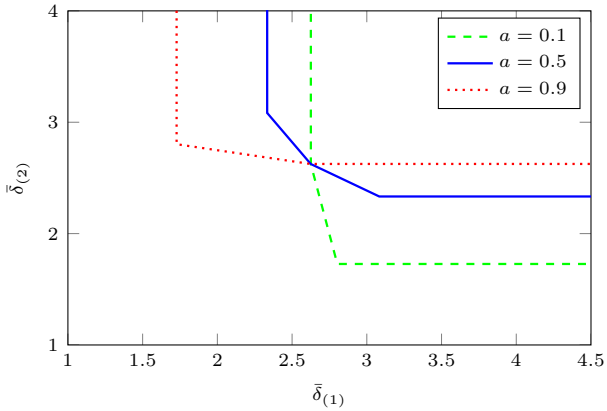


Fig. 6. Optimal trade-off between the average delivery latencies for the files of two classes for different popularity profiles defined by a ($\mu = 3/8$,$r = 1/5$, $J_1 = J_2$).

Acknowledgement

The work of P. Popovski has been in part supported by the European Research Council (ERC Consolidator Grant Nr. 648382 WILLOW) within the Horizon 2020 Program. O. Simeone has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 725731).

References

[1] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE ISIT*, June 2015, pp. 809–813.

[2] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, 2017.

[3] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug 2017.

[4] J. Hachem, U. Niesen, and S. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *arXiv:1606.03175*, 2016.

[5] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency trade-offs," *IEEE Transactions on Information Theory*, vol. PP, 2017.

[6] X. Yi and G. Caire, "Topological coded caching," in *Proc. IEEE ISIT*, 2016, pp. 2039–2043.

[7] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, 2016.

[8] J. Goseling, O. Simeone, and P. Popovski, "Delivery latency trade-offs of heterogeneous contents in fog radio access networks," *arXiv:1701.06303*, 2017.

[9] 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, no. 1, pp. 1–10, 2009.

[10] 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, 2008.