Fundamental Latency Limits for D2D-Aided Content Delivery in Fog Wireless Networks

Roy Karasik*, Osvaldo Simeone†, and Shlomo Shamai (Shitz)*
*EEE Dept., Technion - IIT, Technion City, Haifa, Israel
†Centre for Telecommunications Research, Dept. of Informatics, King's College London, UK
{royk@campus.technion.ac.il, osvaldo.simeone@kcl.ac.uk, sshlomo@ee.technion.ac.il}

***Abstract*—Device-to-Device (D2D) communication can support the operation of cellular systems by reducing the traffic in the network infrastructure. In this paper, the benefits of D2D communication are investigated in the context of a Fog-Radio Access Network (F-RAN) that leverages edge caching and fronthaul connectivity for the purpose of content delivery. Assuming offline caching, out-of-band D2D communication, and an F-RAN with two edge nodes and two user equipments, an information-theoretically optimal caching and delivery strategy is presented that minimizes the delivery time in the high signal-to-noise ratio regime. The delivery time accounts for the latency caused by fronthaul, downlink, and D2D transmissions. The proposed optimal strategy is based on a novel scheme for an X-channel with receiver cooperation that leverages tools from real interference alignment. Insights are provided on the regimes in which D2D communication is beneficial.**

I. Introduction

Device-to-Device (D2D) communication is a main enabler of novel applications such as mission critical communication, video sharing, and proximity-aware gaming and social networking. Furthermore, it can enhance conventional cellular services, including content delivery, by reducing the traffic at the cellular network infrastructure. D2D communication in cellular networks can be either out-of-band, whereby direct communication between the users takes place over frequency resources that are orthogonal with respect to the spectrum used for cellular transmission; or in-band, in which case the same frequency band is used for both D2D and cellular transmissions [1].

In this paper, we study the benefits of out-of-band D2D communications for the modern cellular architecture of a Fog-Radio Access Network (F-RAN) by focusing on content delivery [2], [3]. As illustrated in Fig 1, in an F-RAN, content delivery leverages both edge caching and fronthaul connectivity to a Cloud Processor (CP). In this work, we characterize the potential latency reduction that can be achieved by utilizing D2D links in an F-RAN, while properly accounting for the latency overhead associated with D2D communications.

Related Work: The cache-aided interference channel was first studied in [4], where an upper bound on the minimum delivery latency in the high signal-to-noise ratio (SNR) regime was derived for a system with three users. A lower bound on the Normalized Delivery Time (NDT), which measures the high-SNR worst-case latency relative to an ideal system with unlimited caching capability, was presented in [5] for any number of Edge Nodes (ENs) and User Equipments (UEs), and it was shown to be tight for the setting of two ENs and two UEs. Lower and upper bounds for arbitrary numbers of ENs and UEs, where both ENs and UEs have caching capabilities, were presented in [6] under the constraint of linear precoders at the ENs. The NDT of a general F-RAN system with fronthaul links was studied in [7], where the proposed schemes were shown to achieve the minimum NDT to within a factor of 2, and the minimum NDT was completely characterized for two ENs and two UEs, as well as for other special cases. In [8], it was shown that, for the interference channel with in-band cooperation, transmitter or receiver cooperation cannot increase the high-SNR performance in terms of sum Degrees of Freedom (DoF). The interference channel with out-of-band receiver cooperation was studied in [9], where receiver cooperation was shown to increase the Generalized DoF metric. Importantly, reference [9] only imposes a rate constraint on the D2D links, hence not accounting for the latency overhead caused by D2D communications, which is of central interest in this work.

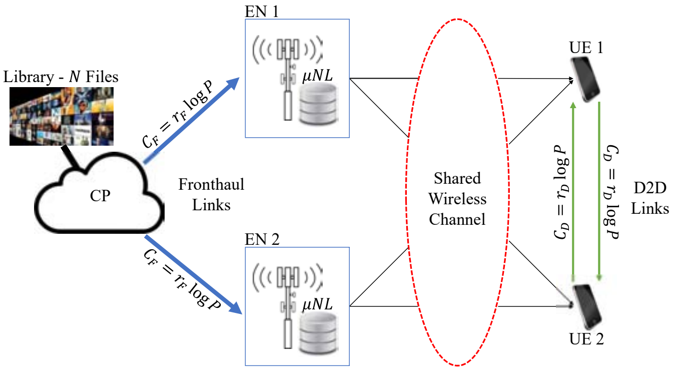

Fig. 1: Illustration of the D2D-aided F-RAN model under study.

Main Contributions: In this paper, we study the D2D-aided F-RAN system with two ENs and two UEs in Fig. 1, and put forth the following main contributions. First, in Sec. III, we

This work has been supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement Nos. 694630 and 725731).

978-1-5386-4780-6/18/$31.00©2018 IEEE

present a novel scheme that improves the NDT achievable on an X-channel with out-of-band D2D receiver cooperation. The proposed scheme enables interference cancellation at the receiver's side with minimal overhead on the D2D links. Second, in Sec. IV, we characterize the minimum NDT of the D2D-aided F-RAN illustrated in Fig. 1. The minimum NDT is used to identify the conditions under which D2D communication is beneficial, and to provide insights on the interplay between fronthaul and D2D resources.

II. System Model

We consider the F-RAN system with Device-to-Device (D2D) links depicted in Fig. 1, where two single-antenna User Equipments (UEs) are served by two single-antenna Edge Nodes (ENs) over a downlink wireless channel. The UEs are connected by two orthogonal out-of-band D2D links of capacity C_D bits per symbol. The model generalizes the set-up studied in [10] by including D2D communications. Each EN is connected to a Cloud Processor (CP) by a fronthaul link of capacity C_F bits per symbol. Throughout this paper, a symbol refers to a channel use of the downlink wireless channel.

Let $\mathcal{F}$ denote a library of $N \geq 2$ files, $\mathcal{F} = \{f_1, \ldots, f_N\}$, each of size L bits. The library is fixed for the considered time interval. The entire library is available at the CP, while the ENs can only store up to μNL bits each, where $0 \leq \mu \leq 1$ is the fractional cache size. During the placement phase, contents are proactively cached at the ENs, subject to the mentioned cache capacity constraints.

After the placement phase, the system enters the delivery phase, which is organized in Transmission Intervals (TIs). In every TI, each UE arbitrarily requests one of the N files from the library. The UEs' requests in a given TI are denoted by the demand vector $\mathbf{d} \triangleq (d_1, d_2) \in [N]^2$, where for any positive integer a, we define the set $[a] \triangleq \{1, 2, \ldots, a\}$. This vector is known at the beginning of a TI at the CP and ENs. The goal is to deliver the requested files to the UEs within the lowest possible delivery latency by leveraging fronthaul links, downlink channel and D2D links.

For a given TI, let T_E denote the duration of the transmission on the wireless downlink channel. At time $t \in [T_E]$, each UE $k \in [2]$ receives a channel output given by

$$y_k[t] = h_{k1}x_1[t] + h_{k2}x_2[t] + z_k[t], \tag{1}$$

where $x_m[t] \in \mathbb{C}$ is the baseband symbol transmitted from EN $m \in [2]$ at time t, which is subject to the average power constraint $\mathbb{E}|x_m[t]|^2 \leq P$ for some $P > 0$; coefficient $h_{km} \in \mathbb{C}$ denotes the quasi-static flat-fading channel between EN m to UE k, which is assumed to remain constant during each TI; and $z_k[t]$ is an additive white Gaussian noise, such that $z_k[t] \sim \mathcal{CN}(0, 1)$ is independent and identically distributed (i.i.d.) across time and UEs. The Channel State Information (CSI) $\mathbf{H} \triangleq \{h_{km} : k \in [2], m \in [2]\}$ is assumed to be drawn i.i.d. from a continuous distribution, and known to all nodes.

A. Caching, Delivery and D2D Transmission

The operation of the system is defined by the following policies that perform caching, as well as delivery via fronthaul, edge and D2D communication resources.

1) Caching Policy: During the placement phase, for EN m, $m \in [2]$, the caching policy is defined by functions $\pi_{c,n}^m(\cdot)$ that map each file f_n to its cached content $s_{m,n}$ as

$$s_{m,n} = \pi_{c,n}^m(f_n), \quad \forall n \in [N]. \tag{2}$$

Note that, as per (2), we consider policies where only coding within each file is allowed, i.e., no inter-file coding is permitted. We have the cache capacity constraint $H(s_{m,n}) \leq \mu L$. The overall cache content at EN m is given by $s_m \triangleq (s_{m,1}, s_{m,2} \ldots, s_{m,N})$.

2) Fronthaul Policy: In each TI of the delivery phase, for EN m, $m \in [2]$, the CP maps the library, $\mathcal{F}$, the demand vector $\mathbf{d}$ and CSI $\mathbf{H}$ to the fronthaul message

$$\mathbf{u}_m = (u_m[1], u_m[2], \ldots, u_m[T_F]) = \pi_f^m(\mathcal{F}, s_m, \mathbf{d}, \mathbf{H}), \tag{3}$$

where T_F is the duration of the fronthaul message. Note that the fronthaul message cannot exceed $T_F C_F$ bits, i.e., $H(\mathbf{u}_m) \leq T_F C_F$.

3) Edge Transmission Policies: After fronthaul transmission, in each TI, the ENs transmit using a function $\pi_e^m(\cdot)$ that maps the local cache content, s_m, the received fronthaul message $\mathbf{u}_m$, the demand vector $\mathbf{d}$ and the global CSI $\mathbf{H}$, to the output codeword

$$\mathbf{x}_m = (x_m[1], x_m[2], \ldots, x_m[T_E]) = \pi_e^m(s_m, \mathbf{u}_m, \mathbf{d}, \mathbf{H}) \tag{4}$$

4) D2D Interactive Communication Policies: After receiving the signals (1) over T_E symbols, in any TI, the UEs use a D2D conferencing policy. For each UE $k \in [2]$, this is defined by the interactive functions $\pi_{\text{D2D},i}^k(\cdot)$ that map the received signal $\mathbf{y}_k \triangleq (y_k[1], \ldots, y_k[T_E])$, the global CSI and the previously received D2D message from UE $k' \neq k \in [2]$ to the D2D message

$$v_k[i] = \\ \pi_{\text{D2D},i}^k(\mathbf{y}_k, \mathbf{H}, v_{k'}[1], \ldots, v_{k'}[i-1], v_k[1], \ldots, v_k[i-1]), \tag{5}$$

where $i \in [T_D]$, with T_D being the duration of the D2D communication. The total size of each D2D message cannot exceed $T_D C_D$ bits. i.e., $H(\mathbf{v}_k) \leq T_D C_D$, where $\mathbf{v}_k \triangleq (v_k[1], \ldots, v_k[T_D])$.

5) Decoding Policy: After D2D communication, each UE $k \in [2]$ implements a decoding policy $\pi_d^k(\cdot)$ that maps the channel outputs, the D2D messages from UE $k' \neq k \in [2]$, the UE demand and the global CSI to an estimate of the requested file f_{d_k} given as

$$\hat{f}_{d_k} = \pi_d^k(\mathbf{y}_k, \mathbf{v}_{k'}, d_k, \mathbf{H}). \tag{6}$$

The probability of error is defined as

$$P_e \triangleq \max_{\mathbf{d}} \max_{k \in [2]} \Pr(\hat{f}_{d_k} \neq f_{d_k}), \tag{7}$$

which is the worst-case probability of decoding error measured over all possible demand vectors $\mathbf{d}$ and over all users $k \in [2]$.

A sequence of policies, indexed by the file size L, is said to be feasible if, for almost all channel realization $\mathbf{H}$, we have $P_e \to 0$ when $L \to \infty$.

B. Performance Metric

As discussed, in each TI, the CP first sends the fronthaul messages to the ENs for a total time of T_F symbols; then, the ENs transmit on the wireless shared channel for a total time of T_E symbols; and, finally, the UEs use the out-of-band D2D links for a total time of T_D symbols. For any sequence of feasible policies, the delivery time per bit $\Delta(\mu, C_F, C_D, P)$ is hence defined as the limit

$$\Delta(\mu, C_F, C_D, P) \triangleq \limsup_{L\to\infty} \frac{\mathbb{E}(T_F + T_E + T_D)}{L}. \quad (8)$$

The notation emphasizes the dependence on the fractional cache size μ, the fronthaul and D2D capacities C_F and C_D, respectively, and the average power constraint P.

We adopt the Normalized Delivery Time (NDT), introduced in [7], as the performance metric of interest. To this end, we evaluate the performance in the high-SNR regime by parameterizing fronthaul and D2D capacities as $C_F = r_F \log(P)$ and $C_D = r_D \log(P)$. With this parametrization, the fronthaul rate $r_F \geq 0$ represents the ratio between the fronthaul capacity and the high-SNR capacity of each EN-to-UE wireless link in the absence of interference; and a similar interpretation holds for the D2D rate $r_D \geq 0$.

For any given tuple (μ, r_F, r_D), the NDT of a sequence of achievable policies is defined as

$$\delta(\mu, r_F, r_D) \triangleq \lim_{P\to\infty} \frac{\Delta(\mu, r_F \log(P), r_D \log(P), P)}{1/\log(P)}. \quad (9)$$

The factor $1/\log(P)$, used for normalizing the delivery time in (9), represents the minimal time to deliver one bit over an EN-to-UE wireless link in the high-SNR regime and in the absence of interference. The minimum NDT is finally defined as the minimum over all achievable policies

$$\begin{aligned}&\delta^*(\mu, r_F, r_D) \triangleq \\ &\qquad \inf\{\delta(\mu, r_F, r_D) : \delta(\mu, r_F, r_D) \text{ is achievable}\}. \quad (10)\end{aligned}$$

By construction, we have the lower bound $\delta^*(\mu, r_F, r_D) \geq 1$.

III. The Two-User X-Channel with Receiver Cooperation

In this section, we present a result of independent interest that will be used in Sec. IV to derive the minimum NDT (10). Specifically, we develop a new delivery scheme for the special case in which no fronthaul communication is enabled, i.e., $r_F = 0$, and the fractional cache size is $\mu = 1/2$. In this regime, each EN can only store half of each file in the library. Under the mentioned caching strategy, in the worst-case scenario in which the UEs request different files, the set-up is equivalent to a two-user Gaussian X-channel with receiver cooperation. In this channel, as illustrated in Fig. 2, each UE needs to download half of the requested file from one EN and the other half from the second EN. The proposed scheme achieves the NDT detailed in the following Proposition.

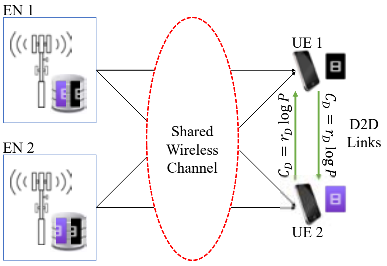

Fig. 2: X-channel with receiver cooperation studied in Sec. III, which represents an F-RAN system with no fronthaul, i.e., with $r_F = 0$, and fractional cache size $\mu = 1/2$.

Proposition 1: For $\mu = 1/2$, $r_F = 0$ and $r_D \geq 0$, the minimum NDT is upper bounded as

$$\delta^* \left(\mu = \frac{1}{2}, r_F = 0, r_D\right) \leq \delta_X \triangleq 1 + \frac{1}{2r_D}. \quad (11)$$

Proposition 1 is proved in the next two subsections by first proposing a novel scheme for the deterministic X-channel, and then adapting it for the Gaussian counterpart. The scheme is based on layered transmission and successive interference cancellation at the receivers.

As compared to existing schemes that are applicable for $\mu = 1/2$ and $r_F = 0$, real interference alignment [11] achieves an NDT of $3/2$ without using the D2D links [10]. Therefore, the proposed D2D-based scheme of Proposition 1 is useful only when the D2D capacity is sufficiently large, i.e., when $r_D > 1$. Furthermore, the scheme in [9] has an NDT lower bounded by 2, since the latency due to D2D communications equals the transmission time on the downlink channel. Hence, the scheme is not advantageous in terms of NDT. Finally, as an alternative policy, one could have each UE compress and forward the received signal to the other UE, allowing each UE to carry out Zero Forcing (ZF) linear equalization. By quantizing with a rate equal to $\log(P)$, one can ensure that the SNR scales linearly with P, and that the approach achieves an NDT equal to $1 + 1/r_D > \delta_X$ (see [7] [10] for similar arguments).

A. The Deterministic Approach

We start by considering a deterministic approximation of the X-channel in order to facilitate the explanation of the main ideas behind the proposed scheme. We recall that, according to [12], in high SNR, the channel (1) is approximated by the deterministic model

$$\begin{aligned} y_1[t] &= x_1[t] + S^{n_d - n_c} x_2[t] \\ y_2[t] &= S^{n_d - n_c} x_1[t] + x_2[t], \end{aligned} \quad (12)$$

where summations and multiplications are over the binary field $\mathbb{F}_2$; n_d and n_c represent the number of direct and cross signal levels, respectively, with $n_d > n_c$; $x_i[t]$ and $y_i[t] \in \mathbb{F}_2^{n_d}$ for $i \in [2]$ are the binary vectors representing the inputs and outputs of the deterministic channel, respectively; and S is the $n_d \times n_d$ shift matrix with all zeros except in the first lower diagonal, which contains all ones. The number of levels is selected as $n_d = \lceil \log(P) \rceil$, while n_c will be taken to satisfy the limit $n_c/n_d \to 1$ when $n_d \to \infty$ in order to approximate the high-SNR behavior of the assumed channel model (1), as explained in [12, Appendix B]. Following this model, we set the D2D link capacity $C_D = r_D \log(P)$ to equal $r_D n_d$ signal levels between the UEs.

Consider, without loss of generality, the case where $n_c = n_d - 1$ and n_d is odd. EN 1 and EN 2 at each time t transmit independent bits $x_1 = [a_1, \ldots, a_{n_d}]^T$ and $x_2 = [b_1, \ldots, b_{n_d}]^T$ on the n_d levels, where we have dropped the dependence on t. By (12), the received signals at the UEs are $y_1 = [a_1, a_2 \oplus b_1, \ldots, a_{n_d} \oplus b_{n_d-1}]^T$ and $y_2 = [b_1, b_2 \oplus a_1, \ldots, b_{n_d} \oplus a_{n_d-1}]^T$. UE 2 uses its D2D link to convey the bits received on the even-numbered levels

$$v_2 = \{b_2 \oplus a_1, b_4 \oplus a_3, \ldots, b_{n_d-1} \oplus a_{n_d-2}\} \tag{13}$$

to UE 1, which consists of $(n_d - 1)/2$ bits. UE 1 is thus able to decode the bits $\{a_1, b_2, a_3, b_4, a_5, \ldots, b_{n_d-1}, a_{n_d}\}$ from $\{y_1, v_2\}$ by means of successive interference cancellation. To this end, it starts by decoding a_1 from $y_{1,1} = a_1$; then, it uses a_1 together with $b_2 \oplus a_1$ in (13) to decode b_2; next, it uses b_2 and $y_{1,3} = a_3 \oplus b_2$ to decode a_3; and so on, until all the desired bits are decoded. Similarly, UE 2 is able to decode bits $\{b_1, a_2, b_3, a_4, b_5, \ldots, a_{n_d-1}, b_{n_d}\}$ from y_2 and $v_1 = \{a_2 \oplus b_1, a_4 \oplus b_3, \ldots, a_{n_d-1} \oplus b_{n_d-2}\}$.

The number of channel uses required on the downlink channel to satisfy the UEs' demands is $L/(n_d - 1)$. For each channel use, each UE has to convey $(n_d - 1)/2$ bits using a D2D link of capacity $r_D n_d$. Therefore, the resulting NDT (9), if we let the number n_d of levels be arbitrary, is

$$\lim_{n_d \to \infty} \frac{n_d}{n_d - 1} \left(1 + \frac{(n_d - 1)/2}{r_D n_d}\right) = \delta_X. \tag{14}$$

Next, we show how to achieve the same NDT for the original model (1).

B. Real Interference Alignment with Receiver Cooperation

In order to convert the proposed scheme from the deterministic model to the X-channel (1), we follow the *real interference alignment* approach of [11]. Accordingly, in a manner similar to the deterministic model, each transmitter uses n_d signal layers, where n_d is odd. The signal transmitted by the ENs at each symbol can be written as

$$x_1 = \sum_{i=1}^{n_d} g_{1,i} a_i \text{ and } x_2 = \sum_{i=1}^{n_d} g_{2,i} b_i, \tag{15}$$

where $\{g_{m,i}\}$, with $m \in [2]$ and $i \in [n_d]$, are precoder gains, and the values a_i and b_i are chosen from a discrete constellation, so that we have $a_i, b_i \in A\mathbb{Z}_Q \triangleq \{0, A, 2A, \ldots, A(Q-1)\}$. Each layer i is coded using random coding with rate R bits per symbol. It is shown in [13, Appendix A] that, by choosing parameters $\{g_{m,i}\}$, A, Q and R properly, UE 1 can decode the symbols $\{a_1, a_2 + b_1, \ldots, a_{n_d} + b_{n_d-1}, b_{n_d}\}$, while UE 2 decodes $\{b_1, b_2 + a_1, \ldots, b_{n_d} + a_{n_d-1}, a_{n_d}\}$. The UEs now exchange the even-numbered layers as in the deterministic model, so that UE 1 transmits the message $v_1 = \{a_2 + b_1, a_4 + b_3, \ldots, a_{n_d-1} + b_{n_d-2}\}$ to UE 2, while UE 2 transmits $v_2 = \{b_2 + a_1, b_4 + a_3, \ldots, b_{n_d-1} + a_{n_d-2}\}$ to UE 1. As a result, UE 1 can decode $\{a_1, b_2, a_3, b_4, \ldots, a_{n_d}, b_{n_d}\}$ while UE 2 decodes $\{b_1, a_2, b_3, a_4, \ldots, b_{n_d}, a_{n_d}\}$.

As detailed in [13, Appendix A], for high SNR, the rate can be selected as $R \approx \log Q \approx \log(P)/(n_d + 1)$. Furthermore, in a manner similar to the deterministic model, the resulting NDT (9) is

$$\delta_{n_d} \triangleq \frac{n_d + 1}{n_d - 1} \cdot \left(1 + \frac{(n_d - 1)/2}{r_D(n_d + 1)}\right), \tag{16}$$

and hence, by increasing the number of layers n_d, we have the limit $\lim_{n_d \to \infty} \delta_{n_d} = \delta_X$.

IV. Minimum NDT

In this section, we derive the minimum NDT by presenting a novel achievable scheme and an information-theoretic lower bound. The achievable scheme leverages the D2D cooperative strategy introduced above along with the scheme in [10], which is optimal in the absence of D2D links.

Theorem 1: The minimum NDT for the 2×2 F-RAN system with number of files $N \geq 2$, fractional cache size $\mu \geq 0$, fronthaul rate $r_F \geq 0$ and D2D rate $r_D \geq 0$ is given as

$$\delta^*(\mu, r_F, r_D) = \tag{17}$$
$$\begin{cases} \max\left\{1 + \mu + \frac{1-2\mu}{r_F}, 2 - \mu\right\} & \text{for } 0 \leq r_F, r_D \leq 1 \\ 1 + \frac{1-\mu}{r_F} & \text{for } r_F \geq \max\{1, r_D\} \\ \max\left\{1 + \frac{\mu}{r_D} + \frac{1-2\mu}{r_F}, 1 + \frac{1-\mu}{r_D}\right\} & \text{for } r_D > \max\{1, r_F\}. \end{cases}$$

Before sketching the proof, we use the result in Theorem 1 in order to draw conclusions on the role of D2D cooperation in improving the delivery latency. We start by observing that, for $r_D \leq \max\{1, r_F\}$, the minimum NDT (17) is identical to the minimum NDT without D2D links derived in [10, Theorem 1]. Therefore, D2D communication provides a latency reduction only when we have $r_D > \max\{1, r_F\}$. This is illustrated in Fig. 3, where we plot the minimum NDT (17) as a function of the fractional cache size μ for fixed fronthaul rate r_F and D2D rate r_D. For any $r_D \leq \max\{1, r_F\}$, the minimum NDT is not affected by the value of r_D, whereas a larger r_D yields a reduced minimum NDT.

The minimum useful value $\max\{1, r_F\}$ for the D2D rate r_D increases with fronthaul rate r_F. This demonstrates that there exists a trade-off between fronthaul and D2D resources for the purpose of interference management, although their role is not symmetric. The use of fronthaul links is in fact necessary to obtain a finite NDT when the library is not fully available at the ENs, i.e., when $\mu < 1/2$. D2D links can instead only reduce the NDT in regimes where fronthaul and edge resources would already be sufficient for content delivery with a finite NDT. In particular, as summarized in Fig. 3, when $r_D > \max\{1, r_F\}$, D2D communication reduces the minimum NDT for all values $0 < \mu < 1$. Furthermore, when

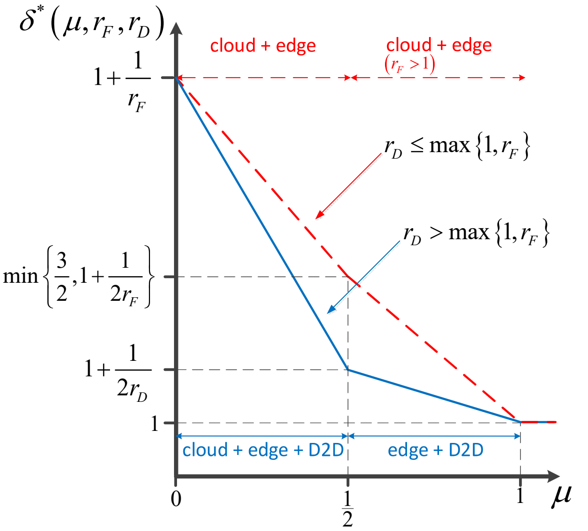

Fig. 3: Minimum NDT for the 2×2 F-RAN with D2D links as a function of μ: when $r_D \leq \max\{1, r_F\}$, D2D communication cannot reduce the delivery latency, while a reduction of the NDT is obtained when $r_D > \max\{1, r_F\}$.

$\mu > 1/2$, irrespective of the value of r_F, the minimum NDT is achieved by leveraging only edge caching and D2D links, without having to rely on fronthaul resources, thus reducing the traffic at the network infrastructure. This is in contrast to the case $r_D \leq \max\{1, r_F\}$, where, by [10], fronthaul transmission is needed to obtain the minimum NDT unless $r_F \leq 1$.

Achievability: The strategy that achieves (17) is based on time- and memory-sharing [10, Remark 1] between the policies for the corner points $\mu = 0, 1/2$ and 1. For $\mu = 1$, we apply cache-aided cooperative ZF at the ENs by leveraging the fact that the ENs can both store the entire library of files. This achieves the NDT $\delta^*(\mu = 1, r_F, r_D) = 1$ [10, Sec. IV.A]. For $\mu = 0$, we apply the cloud-aided soft-transfer scheme of [7] and [10], which uses fronthaul links to convey quantized ZF-precoded signals, achieving the NDT $1 + 1/r_F$. Finally, for $\mu = 1/2$, we use one of the following three schemes: *(i)* EN coordination via interference alignment, which results in an NDT of 3/2 without using either fronthaul or D2D links [10, Sec. IV.A]; *(ii)* time- and memory-sharing between cloud-aided soft transfer and cache-aided cooperative ZF, which leverages fronthaul and cache resources, and results in an NDT of $1+1/(2r_F)$ [10, Theorem 1]; *(iii)* the proposed D2D-based delivery scheme, which results in an NDT of δ_X by leveraging edge and D2D resources.

Converse: The proof of the lower bound can be found in [13, Appendix B]. The proof leverages the approach of [10], which is based on a variation of cut-set arguments. Accordingly, subsets of information resources are identified, from which, in the high-SNR regime, the requested files can be reliably decoded when a feasible policy is implemented. In particular, the first subset, $\{s_1, s_2, \mathbf{u}_1, \mathbf{u}_2\}$, yields a lower bound on T_F as a function of μ and r_F; the seconds subset, $\{\mathbf{y}_1, \mathbf{y}_2\}$, yields a lower bound on T_E; and the third subset, $\{s_1, \mathbf{u}_1, \mathbf{y}_1, \mathbf{v}_2\}$, yields a lower bound on a linear combination of (T_F, T_E, T_D) as a function of μ, r_F and r_D. Note that, only the latter bound differs with respect to [10]. These bounds are then linearly combined according to the values of the fronthaul rate r_F and D2D rate r_D to show that (17) is a lower bound on the minimum NDT.

Remark: Although the definition of the D2D conferencing policy (5) allows for UEs' interactions, the optimal scheme described above uses two simultaneous one-shot D2D communications, whereby the D2D messages of the UEs are based only on the CSI and the respective own received signal.

V. Conclusions

In this work, fundamental insights were provided on the benefits of D2D communication for content delivery in an F-RAN. Considering the Normalized Delivery Time (NDT) metric, an optimal strategy for utilizing the fronthaul and D2D links, as well as the downlink wireless channel, was presented. This strategy is based on a novel scheme for the X-channel with receiver cooperation. It was demonstrated that, for sufficiently large D2D and cache capacities, D2D communication can reduce the traffic on the fronthaul links, and hence help reducing the load on the network infrastructure.

References

[1] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, 2014.

[2] S. C. Hung, H. Hsu, S. Y. Lien, and K. C. Chen, "Architecture harmonization between cloud radio access networks and fog networks," *IEEE Access*, vol. 3, pp. 3019–3034, 2015.

[3] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, August 2016.

[4] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, June 2015, pp. 809–813.

[5] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. Information Science and Systems (CISS)*, March 2016, pp. 320–325.

[6] 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, May 2017.

[7] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct 2017.

[8] C. Huang and S. A. Jafar, "Degrees of freedom of the MIMO interference channel with cooperation and cognition," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4211–4220, Sept 2009.

[9] I. H. Wang and D. N. C. Tse, "Interference mitigation through limited receiver cooperation," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 2913–2940, May 2011.

[10] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, July 2016, pp. 2029–2033, (full version at https://web.njit.edu/~simeone/ISIT_Tandon_16.pdf).

[11] A. S. Motahari, S. Oveis-Gharan, M. A. Maddah-Ali, and A. K. Khandani, "Real interference alignment: Exploiting the potential of single antenna systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4799–4810, Aug 2014.

[12] C. Huang, V. R. Cadambe, and S. A. Jafar, "Interference alignment and the generalized degrees of freedom of the X channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5130–5150, Aug 2012.

[13] R. Karasik, O. Simeone, and S. Shamai, "Fundamental latency limits for D2D-aided content delivery in fog wireless networks," Jan 2018. [Online]. Available: https://arxiv.org/abs/1801.00754