Latency Limits for Content Delivery in a Fog-RAN with D2D Communication

Roy Karasik*, Osvaldo Simeone†, and Shlomo Shamai (Shitz)*
*Department of Electrical Engineering, Technion, Haifa 32000, Israel
†Centre for Telecommunications Research, Department of Informatics, King's College London, London WC2R 2LS, U.K.
{royk@campus.technion.ac.il, osvaldo.simeone@kcl.ac.uk, sshlomo@ee.technion.ac.il}

***Abstract*—A Fog-Radio Access Network (F-RAN) with arbitrary number of edge nodes and users is studied in which the users are able to cooperate by communicating over out-of-band broadcast Device-to-Device (D2D) links. Placement and delivery strategies are proposed with the aim of minimizing the Normalized Delivery Time (NDT) — a metric that captures the high signal-to-noise ratio worst-case latency for delivering any subset of requested contents to the users. The proposed strategies, based on compress-and-forward, are shown to be optimal to within a constant multiplicative factor of two for all values of the problem parameters. The analysis provides insights on the role of D2D cooperation in improving the delivery latency.**

I. Introduction

The proactive caching of popular content at the Edge Nodes (ENs) is an effective way of reducing delivery time [1]. Apart from alleviating the need to access centralized network resources to fetch requested contents, edge caching also offers opportunities for cooperative transmission and interference management if there are common contents across the caches of multiple ENs. Even without common cached contents, cooperative transmission is possible if the ENs are connected over fronthaul links to a Cloud Processor (CP) with full access to the content library, as in Cloud-Radio Access Network (C-RAN). However, cooperative transmission in C-RAN comes at the cost of additional latency due to fronthaul transfer [2], [3]. When fronthaul capacity is also limited or not available, an alternative approach to mitigate the inter-user interference on the shared wireless channel is by allowing the receivers to cooperate over out-of-band Device-to-Device (D2D) links [4]. In such a scenario, the latency overhead caused by D2D communication must be taken into account in order to assess the benefits of D2D communication.

In this paper, we consider a D2D-aided Fog-RAN (F-RAN), illustrated in Fig. 1, in which edge caching, fronthaul connectivity to a CP, and receiver cooperation are leveraged for reducing content delivery time. Following [2], the term F-RAN is used to indicate the use of both cloud and edge caching resources. We aim at characterizing the potential latency reduction that may be achieved by utilizing out-of-band D2D links, while properly accounting for the latency overhead associated with D2D communications.

Related Work: Bounds on the high-Signal-to-Noise-Ratio (SNR) delay metric, the Normalized Delivery Time (NDT), were presented in [5] for a general interference channel with caches equipped at all transmitters and receivers, and the achievable NDT was shown to be optimal in certain cache size regimes. Content delivery in a multi-hop D2D caching network was studied in [6], where the per-node capacity scaling law was derived. The trade-off between cache storage and transmission rate was characterized in [7] for a cache-aided network where the users can demand multiple files. Optimal worst-case delay was derived in [8] for a multi-sender coded caching network with shared caches. The NDT of a general F-RAN system without D2D links was investigated in [9], 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 users, as well as for other special cases. An F-RAN with heterogeneous contents was studied in [10], and the NDT region was characterized for the case with two ENs and two users. A caching and delivery scheme was presented for a partially-connected F-RAN in [11] and in [12]. An F-RAN with imperfect Channel State Information (CSI) at the CP was studied in [13], and a non-orthogonal transmission scheme was shown to improve the latency performance. The only prior work on D2D-aided F-RAN are [14], [15], which derive the minimum NDT for the special case of an F-RAN with two ENs and two users.

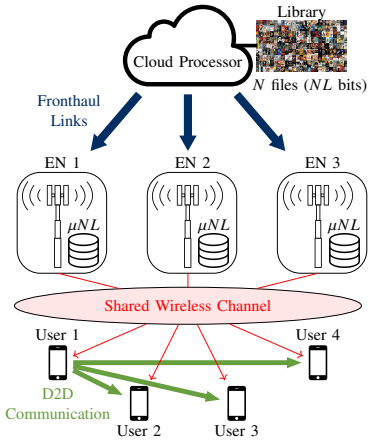

Fig. 1. Illustration of the D2D-aided F-RAN model under study with $M = 3$ ENs and $K = 4$ users.

Main Contributions: In this paper, we study the general D2D-aided F-RAN system with M ENs and K users illustrated in Fig. 1. We first present a lower bound on the minimum

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).

NDT. Then, we propose an achievable strategy that uses a D2D cooperation scheme based on Compress-and-Forward (CF). Although this strategy is known to be generally suboptimal [14], we show that it achieves the minimum NDT to within a multiplicative factor of 2. This implies that the optimality gap does not scale with the size of the system.

II. System Model

We consider the F-RAN system with Device-to-Device (D2D) links depicted in Fig. 1, where $K \geq 2$ single-antenna users are served by $M \geq 2$ single-antenna Edge Nodes (ENs) over a downlink wireless channel. Each user is connected to all other users by an orthogonal out-of-band broadcast D2D link of capacity C_D bits per symbol. The model generalizes the set-up studied in [9] by including D2D communications. Each EN is connected to a Cloud Processor (CP) by a fronthaul link of capacity C_F bits per symbol. A symbol refers to a channel use of the downlink wireless channel.

Let $\mathcal{F} = \{f_1, \ldots, f_N\}$ denote a library of $N \geq K$ files, each of size L bits. The library is fixed for the considered time period. 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 cache capacity constraints.

After the placement phase, the system enters the delivery phase, which is organized in Transmission Intervals (TIs). In every TI, each user arbitrarily requests one of the N files from the library. The users' requests in a given TI are denoted by the demand vector $\mathbf{d} \triangleq (d_1, d_2, \ldots, d_K) \in [N]^K$, 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 users 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 user $k \in [K]$ receives a channel output given by

$$y_k[t] = \sum_{m=1}^{M} h_{km} x_m[t] + z_k[t], \tag{1}$$

where $x_m[t] \in \mathbb{C}$ is the baseband symbol transmitted from EN $m \in [M]$ 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 user k, which is assumed to be drawn independently and identically distributed (i.i.d.) from a continuous distribution and 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 i.i.d. across time and users. The Channel State Information (CSI) $\mathbf{H} \triangleq \{h_{km} : k \in [K], m \in [M]\}$ is assumed to be known to all nodes.

A. Caching, Delivery, and D2D Transmission

The operation of the system is defined by 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, $m \in [M]$, the caching policy is defined by functions $\pi^m_{c,n}(\cdot)$ that map each file f_n to its cached content $s_{m,n}$ as

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

Note that, as per (2), we consider policies where only coding within each file is allowed, i.e., no inter-file coding (e.g., [16]) is permitted. In order to satisfy the cache capacity constraints, we restrict the mappings to satisfy $H(s_{m,n}) \leq \mu L$, where $H(s_{m,n})$ denotes the entropy of $s_{m,n}$. 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, $m \in [M]$, the CP maps the library, $\mathcal{F}$, the demand vector $\mathbf{d}$, and CSI $\mathbf{H}$ to the fronthaul message

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

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 $\pi^m_e(\cdot)$ 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

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

4) D2D Interactive Communication Policies: After receiving the signals (1) over T_E symbols, in any TI, the users apply a D2D conferencing policy. For each user $k \in [K]$, this is defined by the interactive functions $\pi^k_{\text{D2D},t}(\cdot)$ 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 users $[K]\backslash\{k\}$ to the D2D message

$$v_k[t] = \pi^k_{\text{D2D},t}\left(\mathbf{y}_k, \mathbf{H}, \mathbf{v}^{t-1}_{[K]}\right), \tag{5}$$

where $t \in [T_D]$, with T_D being the duration of the D2D communication, and

$$\begin{aligned}\mathbf{v}^{t-1}_{[K]} \triangleq &(v_1[1], \ldots, v_1[t-1], v_2[1], \ldots, v_2[t-1], \ldots, \\ &v_K[1], \ldots, v_K[t-1]).\end{aligned} \tag{6}$$

All users broadcast the D2D messages (5) to all other users over orthogonal broadcast channels of capacity C_D. Hence, 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 user $k \in [K]$ implements a decoding policy $\pi^k_d(\cdot)$ that maps the channel outputs, the D2D messages from users $[K]\backslash\{k\}$, the user demand, and the global CSI to an estimate of the requested file f_{d_k} given as

$$\hat{f}_{d_k} = \pi^k_d(\mathbf{y}_k, \mathcal{V}_k, d_k, \mathbf{H}), \tag{7}$$

where $\mathcal{V}_k \triangleq \{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_K\}$ is the set of D2D messages sent by users $k' \in [K]\backslash\{k\}$ and received by user k.

The probability of error is defined as

$$P_e \triangleq \max_{\mathbf{d} \in [N]^K} \max_{k \in [K]} \Pr(\hat{f}_{d_k} \neq f_{d_k}), \tag{8}$$

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

We adopt the Normalized Delivery Time (NDT), introduced in [9], as the performance metric of interest. The NDT is the high-SNR ratio between the worst-case delivery time per bit required to satisfy any possible demand vector $\mathbf{d}$ and the delivery time per bit for an ideal reference system in which each user can receive the desired file at the maximum high-SNR rate of $\log(P)$ [bits/symbol]. To formalize the NDT, we parametrize 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-user wireless link in the absence of interference; a similar interpretation holds for the D2D rate $r_D \geq 0$.

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 users use the out-of-band D2D links for a total time of T_D symbols. The corresponding NDT contributions are obtained by normalizing the above terms by the delivery time needed on the mentioned reference system:

$$\delta_F \triangleq \lim_{P\to\infty} \lim_{L\to\infty} \frac{\mathbb{E}[T_F]}{L/\log(P)}, \quad \delta_E \triangleq \lim_{P\to\infty} \lim_{L\to\infty} \frac{\mathbb{E}[T_E]}{L/\log(P)}, \tag{9}$$

and

$$\delta_D \triangleq \lim_{P\to\infty} \lim_{L\to\infty} \frac{\mathbb{E}[T_D]}{L/\log(P)}. \tag{10}$$

The factor $L/\log(P)$, used for normalizing the delivery times in (9)–(10), represents the minimal time to deliver a file in the reference system. The total NDT is hence defined as

$$\delta(\mu, r_F, r_D) \triangleq \delta_F + \delta_E + \delta_D, \tag{11}$$

where the notation emphasizes the dependence of the NDT on the fractional cache size μ, and the fronthaul and D2D rates r_F and r_D, respectively.

The minimum NDT is finally defined as the minimum over all NDTs achievable by some feasible policy:

$$\delta^*(\mu, r_F, r_D) \triangleq \inf\{\delta(\mu, r_F, r_D) : \delta(\mu, r_F, r_D) \text{ achievable}\}. \tag{12}$$

By construction, we have the lower bound $\delta^*(\mu, r_F, r_D) \geq 1$. Furthermore, the minimum NDT can be proved by means of file-splitting and cache-sharing arguments to be convex in μ for any fixed values of r_F and r_D [9, Lemma 1].

III. Bounds on the Minimum NDT

In this section, we provide lower and upper bounds on the minimum NDT for the $M \times K$ D2D-aided F-RAN described in the previous section.

A. Lower Bound

A general lower bound on the minimum NDT is given in Prop. 1. Following [9], the bound is derived by identifying subsets of information resources from which, for high-SNR, all requested files must be reliably decoded when a feasible policy is implemented. Specifically, for $l = 0, 1, \ldots, \min\{M, K\}$, we consider a subset that consists of the signals $\{\mathbf{y}_1, \ldots, \mathbf{y}_l, \mathcal{V}_1, \ldots, \mathcal{V}_l\}$ received by l users on the downlink and D2D channels, along with the cache contents and fronthaul messages $\{s_1, \ldots, s_{(M-l)}, \mathbf{u}_1, \ldots, \mathbf{u}_{(M-l)}\}$ of $(M-l)$ ENs.

Proposition 1: For a D2D-aided F-RAN with M ENs, each with a fractional cache size $\mu \in [0, 1]$, K users, a library of $N \geq K$ files, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$, the minimum NDT is lower bounded as $\delta^*(\mu, r_F, r_D) \geq \delta_{\text{lb}}(\mu, r_F, r_D)$, with $\delta_{\text{lb}}(\mu, r_F, r_D)$ being the minimum value of the following linear program

$$\begin{aligned}
\text{minimize} \quad & \delta_F + \delta_E + \delta_D && \text{(13a)}\\
\text{subject to} \quad & l\delta_E + (M-l) r_F \delta_F + g(l) r_D \delta_D \\
& \qquad \geq K - (M-l)(K-l)\mu, && \text{(13b)}\\
& \delta_E \geq \frac{K}{\min\{M, K\}}, && \text{(13c)}\\
& \delta_F \geq 0, \ \delta_D \geq 0, && \text{(13d)}
\end{aligned}$$

where (13b) is a set of constraints with $l = 0, 1, \ldots, \min\{M, K\}$, and

$$g(l) \triangleq \begin{cases} 0 & \text{for } l = 0, \\ K-1 & \text{for } l = 1, \\ K & \text{for } l = 2, \ldots, \min\{M, K\}. \end{cases} \tag{14}$$

Proof: Omitted for brevity. See [17, Appendix B]. ∎

Note that, without D2D communication, i.e., $r_D = 0$, the linear program (13) is identical to that of [9, Proposition 1]. For $r_D > 0$, the additional term $g(l) r_D \delta_D$ in (13b) reflects the novel trade-off between the D2D NDT δ_D and the edge and fronthaul NDTs δ_E and δ_F, respectively. This will be further discussed below.

B. Upper Bounds and Achievable Strategy

We first consider the special case in which the fractional cache size is $\mu = 1/M$, and present a D2D-based delivery scheme that make use of Compress-and-Forward (CF), and does not require the use of the cloud infrastructure. Note that this is possible since a cache capacity of $\mu = 1/M$ guarantees that the entire library $\mathcal{F}$ is available across the caches of all ENs.

Lemma 1: For a D2D-aided F-RAN with M ENs, each with a fractional cache size $\mu = 1/M$, K users, a library of $N \geq K$ files, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$, the minimum NDT is upper bounded as $\delta^*(\mu = 1/M, r_F, r_D) \leq \delta_{\text{D2D-CF}}$, where the NDT

$$\delta_{\text{D2D-CF}} \triangleq \frac{K}{\min\{M, K\}} \left(1 + \frac{1}{r_D}\right) \tag{15}$$

is achieved by means of CF-based D2D communication and Zero-Forcing (ZF) equalization at the devices.

Proof: See [17, Sec. III-B]. ∎

The NDT (15) is achieved by the following scheme. Consider first the case $M \geq K$. Under this assumption, K out

of the M ENs are active at any time to transmit, so that each active EN transmits part of the requested file to a given user. K users are hence served simultaneously, each by a different EN. Each user compresses and forwards its received signal to all other users over the D2D links. Then, each user collects the K received signals and carries out ZF equalization in order to recover the desired signal with no interference from other signals. By quantizing with a rate equal to $\log(P)$ bits per downlink symbol, ZF equalization achieves an ideal edge NDT of $\delta_E = 1$ given that the SNR after compression scales linearly with P (see [9, Prop. 3]). Due to the use of the D2D links, a latency overhead of $\delta_D = \delta_E / r_D$ is added to the delivery time, and hence the total NDT is (15). For the complementary case in which $M < K$, only M users can be served simultaneously, and hence the edge NDT is multiplied by K/M.

Remark 1: For a D2D-aided F-RAN with $M = 2$ ENs and $K = 2$ users, a different D2D-based scheme was presented in [14], which achieves an NDT equal to $1 + 1/(2r_D) < \delta_{\text{D2D-CF}}$. This scheme is based on real interference alignment [18] and is hence strongly dependent on the assumption of perfect CSI at the transmitters. In contrast, the CF-based scheme discussed above requires only CSI at the receivers in order to perform the ZF equalization.

The CF-strategy can be combined with previously proposed delivery techniques studied in [9] by means of file-splitting and cache-sharing [9, Lemma 1]. That is, all files are split in the same way into a number of fragments, and each fragment is delivered through a different policy. In order to obtain a policy that applies for any value of fractional cache size μ, we combine the D2D-based CF scheme (Lemma 1) with the best-known general strategies for an F-RAN model with no D2D cooperation. These are: (*i*) cache-aided ZF [9, Lemma 2], whereby fragments cached by all ENs are delivered via cooperative ZF precoding; (*ii*) cache-aided EN coordination [9, Lemma 3], in which fragments cached by only one EN are delivered via interference alignment [18]; and (*iii*) cloud-aided soft-transfer [9, Proposition 3], whereby ZF precoding is carried out at the cloud, and the fronthaul links are used to convey quantized ZF-precoded signals to the ENs, such that no cache resources are required. To formulate the main result, we define the threshold values

$$r_F^{\text{th}} \triangleq \frac{K(M-1)}{M(\min\{M,K\}-1)}, \tag{16}$$

and

$$r_D^{\text{th}} \triangleq \max\left\{\frac{\max\{M,K\}}{\min\{M,K\}-1}, \frac{M^2 r_F}{(M-1)\min\{M,K\}}\right\}. \tag{17}$$

Proposition 2: For a D2D-aided F-RAN with M ENs, each with a fractional cache size $\mu \in [0,1]$, K users, a library of $N \geq K$ files, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$, the minimum NDT is upper bounded as $\delta^*(\mu, r_F, r_D) \leq \delta_{\text{ach}}(\mu, r_F, r_D)$, where the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ is obtained by combining the mentioned schemes as follows:

- Low cache, low fronthaul, and low D2D regime ($\mu \leq 1/M$, $r_F \leq r_F^{\text{th}}$, and $r_D \leq r_D^{\text{th}}$): Combining EN coordination and soft-transfer policies yields the NDT

$$\delta_{\text{ach}}(\mu, r_F, r_D) = (M+K-1)\mu + (1-\mu M)\left[\frac{K}{\min\{M,K\}} + \frac{K}{M r_F}\right]. \tag{18}$$

- High cache, low fronthaul, and low D2D regime ($\mu > 1/M$, $r_F \leq r_F^{\text{th}}$, and $r_D \leq r_D^{\text{th}}$): Combining EN coordination and ZF precoding policies yields the NDT

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{K}{\min\{M,K\}}\left(\frac{\mu M - 1}{M-1}\right) + (1-\mu)\frac{M+K-1}{M-1}. \tag{19}$$

- High fronthaul and low D2D regime ($\mu \in [0,1]$, $r_F > r_F^{\text{th}}$, and $r_D \leq r_D^{\text{th}}$): Combining ZF precoding and soft-transfer policies yields the NDT

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{K}{\min\{M,K\}} + \frac{(1-\mu)K}{M r_F}. \tag{20}$$

- Low cache and high D2D regime ($\mu \leq 1/M$, $r_F \geq 0$, and $r_D > r_D^{\text{th}}$): Combining soft-transfer and CF policies yields the NDT

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{K(1+\mu M/r_D)}{\min\{M,K\}} + \frac{(1-\mu M)K}{M r_F}. \tag{21}$$

- High cache and high D2D regime ($\mu > 1/M$, $r_F \geq 0$, and $r_D > r_D^{\text{th}}$): Combining CF and ZF precoding policies yields the NDT

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{K}{\min\{M,K\}}\left(1 + \frac{(1-\mu)M}{(M-1)r_D}\right). \tag{22}$$

Proof: See [17, Appendix A]. ∎

IV. Characterization of the Minimum NDT

In this section, based on the lower and upper bounds of Section III, we discuss the optimality properties of the CF-based strategy. We start with the main result in the following proposition, which shows that the achievable strategy of Prop. 2 is optimal to within a multiplicative factor of 2.

Proposition 3: For a D2D-aided F-RAN with M ENs, each with a fractional cache size $\mu \in [0,1]$, K users, a library of $N \geq K$ files, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$, the strategy of Prop. 2 achieves the minimum NDT to within a factor of 2, i.e.,

$$\frac{\delta_{\text{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \leq 2. \tag{23}$$

Proof: See [17, Appendix D]. ∎

The key result in Prop. 3 is that the multiplicative suboptimality factor of the CF-based D2D approach defined in the previous section does not scale with the size of the system. This is illustrated in Fig. 2, where we plot the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ and the lower bound $\delta_{\text{lb}}(\mu, r_F, r_D)$ as a function of the number of ENs and users, with $M = K$, fractional cache size $\mu = 1/M$, fronthaul rate $r_F = 1$, and D2D rate $r_D = 1.25$. As seen, the suboptimality gap can be, in practice, significantly smaller than two.

While the CF-based scheme is approximately optimal as proved by Prop. 3, the gap identified in (23) is generally not

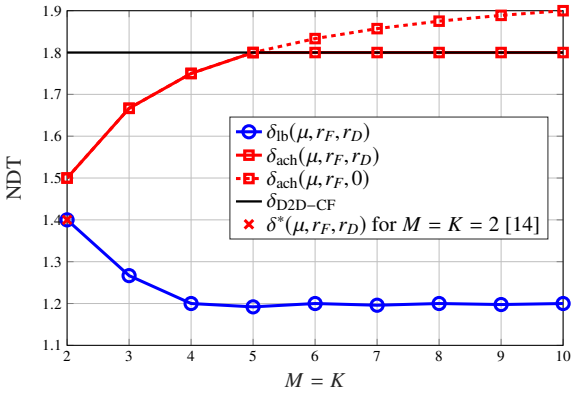

Fig. 2. Lower and upper bounds on the min. NDT as a function of the number of ENs and users $M = K$ for $r_F = 1$, $\mu = 1/M$, and $r_D = 1.25$ or $r_D = 0$.

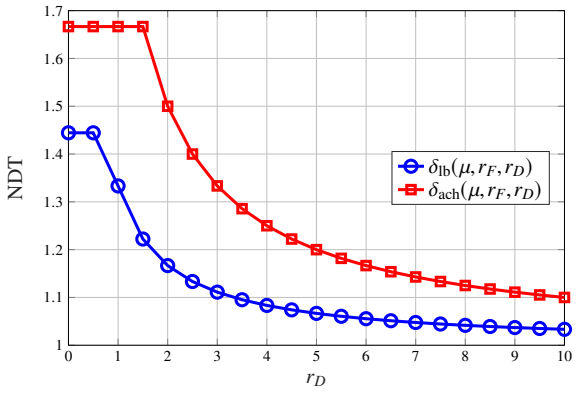

Fig. 3. Lower and upper bounds on the min. NDT as a function of r_D for $r_F = 1$, $M = K = 3$, and $\mu = 1/3$.

zero. As a notable example, for a D2D-aided F-RAN with $M = 2$ ENs and $K = 2$ users, the lower bound (13), illustrated in Fig. 2, is tight, and the CF-based strategy is suboptimal [14, Sec. III] (see Remark 1). This said, the next corollary states that CF is close to optimal for sufficiently high D2D rate r_D.

Corollary 1: For a D2D-aided F-RAN with M ENs, each with fractional cache size $\mu \in [0,1]$, K users, a library of $N \geq K$ files, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq \max\{r_D^{\text{th}}, 1/\epsilon\}$ with r_D^{th} in (17) and $\epsilon > 0$, the achievable strategy of Prop. 2 is close to optimal in the sense that we have

$$\frac{\delta_{\text{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \leq 1 + \epsilon. \tag{24}$$

Proof: Follows from the proof of Prop. 3. See [17]. ■

Corollary 1 is illustrated in Fig. 3. where we plot the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ and the lower bound $\delta_{\text{lb}}(\mu, r_F, r_D)$ as a function of the D2D rate r_D, for $M = 3$ ENs, $K = 3$ users, fractional cache size $\mu = 1/3$, and fronthaul rate $r_F = 1$. As r_D increases, $\delta_{\text{ach}}(\mu, r_F, r_D)$ is seen to approach the lower bound $\delta_{\text{lb}}(\mu, r_F, r_D)$. E.g., for $r_D \geq 1/\epsilon = 10$, the gap to optimality is smaller than $\epsilon = 0.1$. This is because, for arbitrarily large D2D rate, the latency overhead caused by D2D communications is negligible, and an ideal NDT of one can be achieved by means of ZF-equalization at the users. The figure also highlights the gains that can be achieved with sufficiently high D2D rate.

V. Conclusions

In this work, we have studied the benefits of out-of-band broadcast Device-to-Device (D2D) communication for content delivery in a general Fog-Radio Access Network (F-RAN) with arbitrary number of Edge Nodes (ENs) and users. Focusing on the normalized delivery time (NDT) metric, a strategy based on compress-and-forward D2D communication was shown to be approximately optimal to within a constant factor of 2 for all values of the problem parameters. For sufficiently high D2D capacity, the proposed strategy was proved to achieve a significantly lower delivery latency than the minimum NDT for F-RAN without D2D communication. Similar results for a D2D-aided F-RAN under pipelined delivery policies, whereby simultaneous transmissions on fronthaul, edge and D2D channels are enabled, can be found in [17].

References

[1] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, 2018.

[2] 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.

[3] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Int. Things J.*, vol. 3, no. 6, pp. 854–864, Dec 2016.

[4] I. H. Wang and D. N. C. Tse, "Interference mitigation through limited receiver cooperation," *IEEE Int. Things J.*, vol. 57, no. 5, pp. 2913–2940, May 2011.

[5] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov 2017.

[6] S. W. Jeon, S. N. Hong, M. Ji, G. Caire, and A. F. Molisch, "Wireless multihop device-to-device caching networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1662–1676, March 2017.

[7] A. Sengupta and R. Tandon, "Improved approximation of storage-rate tradeoff for caching with multiple demands," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 1940–1955, May 2017.

[8] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of caching in heterogeneous networks with uncoded prefetching," *arXiv preprint arXiv:1811.06247*, 2018.

[9] 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.

[10] J. Goseling, O. Simeone, and P. Popovski, "Delivery latency trade-offs of heterogeneous contents in fog radio access networks," in *Proc. IEEE Global Conf. Communications (GLOBECOM)*, Dec. 2017, pp. 1–6.

[11] A. Roushdy, A. S. Motahari, M. Nafie, and D. Gündüz, "Cache-aided fog radio access networks with partial connectivity," in *Proc. IEEE Wireless Communications and Networking (WCNC)*, April 2018, pp. 1–6.

[12] K. Wan, D. Tuninetti, M. Ji, and G. Caire, "Novel inter-file coded placement and D2D delivery for a cache-aided Fog-RAN architecture," *arXiv preprint arXiv:1811.05498*, 2018.

[13] J. Zhang and O. Simeone, "Cloud-edge non-orthogonal transmission for fog networks with delayed CSI at the cloud," in *Proc. IEEE Inform. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[14] R. Karasik, O. Simeone, and S. Shamai, "Fundamental latency limits for D2D-aided content delivery in fog wireless networks," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Vail, USA, Jun. 2018.

[15] ——, "Information-Theoretic analysis of D2D-aided pipelined content delivery in Fog-RAN," in *Proc. IEEE Int. Symp. Wireless Commun. Sys. (ISWCS)*, Lisbon, Portugal, Aug. 2018.

[16] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[17] R. Karasik, O. Simeone, and S. Shamai, "How much can D2D communication reduce content delivery latency in fog networks with edge caching?" *arXiv preprint arXiv:1904.01256*, 2019.

[18] 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.