How Much Can D2D Communication Reduce Content Delivery Latency in Fog Networks With Edge Caching?

Roy Karasik, *Student Member, IEEE*, Osvaldo Simeone, *Fellow, IEEE*, and Shlomo Shamai (Shitz), *Life Fellow, IEEE*

Abstract—A Fog-Radio Access Network (F-RAN) is studied in which cache-enabled Edge Nodes (ENs) with dedicated fronthaul connections to the cloud aim at delivering contents to mobile users. Using an information-theoretic approach, this work tackles the problem of quantifying the potential latency reduction that can be obtained by enabling Device-to-Device (D2D) communication over out-of-band broadcast links. Following prior work, the Normalized Delivery Time (NDT) — a metric that captures the high signal-to-noise ratio worst-case latency — is adopted as the performance criterion of interest. Joint edge caching, downlink transmission, and D2D communication policies based on compress-and-forward are proposed that are shown to be information-theoretically optimal to within a constant multiplicative factor of two for all values of the problem parameters, and to achieve the minimum NDT for a number of special cases. The analysis provides insights on the role of D2D cooperation in improving the delivery latency.

Index Terms—Caching, D2D communication, F-RAN, C-RAN, latency, zero-forcing, interference alignment.

I. INTRODUCTION

PROACTIVE caching of popular content at the Edge Nodes (ENs) is an effective way of reducing delivery time [1], [2]. 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. When requested contents are not cached at the edge, the ENs can satisfy the users' demands by leveraging fronthaul links to a Cloud Processor (CP) with full access to the content library. Fronthaul links can also enable cooperative transmission, as in a Cloud-Radio Access Network (C-RAN) architecture [3]. However, fronthaul transmissions generally entail additional latency. The Fog-RAN (F-RAN) architecture, illustrated in Fig. 1, makes use of both cloud and edge caching resources in order to carry out content delivery, hence potentially reaping the benefits of both edge caching and C-RAN [4]–[7].

Prior work, to be reviewed below, has studied the performance of F-RANs by assuming non-cooperative end users. In contrast, in this paper, motivated by the emergence of Device-to-Device (D2D) communication solutions [8]–[10], we study the impact of D2D communication on the delivery latency of F-RAN architectures. To this end, we consider a D2D-aided F-RAN, illustrated in Fig. 1, in which edge caching, fronthaul connectivity to a CP, and users' cooperation are leveraged to reduce content delivery time. We specifically 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: In prior work, the information-theoretic analysis of content delivery in F-RANs has been carried out in the high Signal-to-Noise-Ratio (SNR) regime in order to concentrate on the impact of interference. This line of work adopts as performance metric the Normalized Delivery Time (NDT), which measures the high-SNR worst-case latency relative to an ideal system with unlimited edge caching capability [4], [11]. The first related work is [12], which presents an upper bound on the NDT, or equivalently, on the reciprocal Degrees-of-Freedom (DoF), for a cache-aided interference channel with three users. Bounds on the NDT for arbitrary numbers of transmitters and receivers, where both transmitters and receivers have caching capabilities, were presented in [13] and in [14] under the constraint of linear precoders at the transmitters. A lower bound on the NDT was derived in [15] for any number of ENs and users, and it was shown to be tight for the setting of two ENs and two users. Upper and lower bounds on the NDT of a general interference channel with caches at all transmitters and receivers were presented in [16], and the achievable NDT was shown to be optimal in certain cache size regimes.

Including also fronthaul connections to the cloud, the NDT of a general F-RAN system was investigated in [11], 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
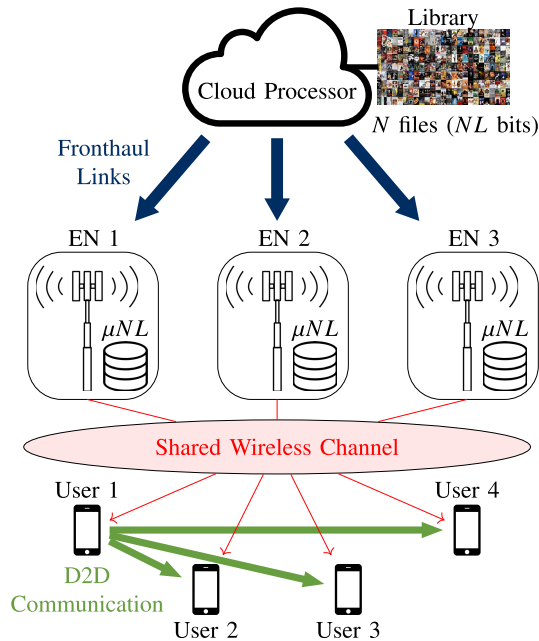
Fig. 1. Illustration of the D2D-aided F-RAN model under study with $M = 3$ ENs and $K = 4$ users.

as for other special cases. The F-RAN system with a shared multicast fronthaul link was studied in [17] and [18], where the advantages of coded multicast delivery were investigated. An F-RAN with heterogeneous contents was studied in [19], 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 [20] and in [21]. Under the constraints of linear precoding and uncoded fronthaul transmission, upper and lower bounds on the minimum NDT in an F-RAN were presented in [22], and the ratio between bounds was shown to be less than $3/2$ for all system parameters and equals to one for some special cases. This work was extended in [23] to include caches also at the users. An F-RAN with imperfect Channel State Information (CSI) at the CP was studied in [24], and a non-orthogonal transmission scheme was shown to improve the latency performance.

To the best of our knowledge, F-RANs with D2D communication have not yet been considered, apart from the conference versions of this work [25]–[27]. Content delivery in a multi-hop D2D caching network was instead studied in [28], where the per-node capacity scaling law was derived. D2D-aided task offloading in fog computing networks was investigated in [29], where a novel resource allocation scheme was shown to improve system utility compared to other baseline schemes. In [30], it was shown that in-band transmitter or receiver cooperation cannot increase the sum DoF of an interference channel. In contrast, out-of-band D2D receiver cooperation was proven in [31] to increase the Generalized DoF metric for an interference channel. Extending this work, the trade-off between the sum DoF and the rate of D2D communication was recently investigated in [32] for the K-user interference channel with multiple antennas per node and with out-of-band D2D transmitter or receiver cooperation. Importantly, references [31], [32] only impose 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.

Main Contributions: In this work, we study the general D2D-aided F-RAN system with M ENs and K users illustrated in Fig. 1. First, we propose two caching and delivery strategies based on a novel form of interference alignment and on compress-and-forward. The first strategy is developed for the special case $M = K = 2$ and is shown to be optimal. The approach is however difficult to scale to a larger system and suffers from the typical lack of robustness to imperfect CSI of interference alignment [33]. For the general case of arbitrary number of M and K, we prove that a more practical D2D strategy based on compress-and-forward achieves the minimum NDT to within a multiplicative factor of 2. This implies that the optimality gap of this strategy does not scale with the size of the system. Based on these results, we identify regimes in terms of fronthaul and cache capacities under which D2D communication is beneficial in reducing delivery latency. This work was presented in part in the conference papers [25]–[27], where we covered, without providing technical proofs, the special case of two ENs and users, as well as a general D2D-aided F-RAN with serial delivery policies. In this paper, we provide a novel characterization of the minimum NDT under pipelined delivery policies for any number of ENs and users (see Section VI) apart from unifying and formally proving these results.

Organization: The rest of the paper is organized as follows. In Section II, we present an information-theoretic model for a general D2D-aided F-RAN under serial or pipelined delivery policies. In addition, the metric of interest, namely the NDT, is defined. In Section III, we describe the proposed D2D-based caching and delivery strategies. In Section IV, upper and lower bounds on the minimum NDT under serial delivery are derived. In Section V, we present an exact characterization of the minimum NDT for the special case with $M = K = 2$ and a finite-gap characterization for arbitrary M and K. In Section VI, we discuss pipelined delivery policies. Lower and upper bounds on the minimum NDT along with a finite-gap characterization are presented. Finally, in Section VII we conclude the paper and highlight some open problems.

Notation: For any positive integer A, we define the set $[A] \triangleq \{1, 2, \ldots, A\}$.

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. The users can cooperate by utilizing out-of-band D2D links whereby direct communication between the users takes place over radio resources that are orthogonal with respect to the spectrum used for ENs' transmission. D2D links are usually implemented with low-range transmission technologies, such as Bluetooth and WiFi, that are based on near-orthogonal transmission strategies, such as frequency hopping and Carrier-Sense Multiple-Access with Collision Avoidance (CSMA/CA). Therefore, we assume that each user is connected to all other users by an orthogonal broadcast D2D link of capacity C_D bits per symbol. A symbol

refers to a channel use of the downlink wireless channel. The model generalizes the set-up studied in [11] by including D2D communications. Each EN is connected to a Cloud Processor (CP) by a fronthaul link of capacity C_F bits per symbol.

Let $\mathcal{F}$ denote a library of $N \geq K$ files, $\mathcal{F} = \{f_1, \ldots, f_N\}$, each of size L bits. The library is fixed for the considered time period. The entire library is available at the CP, whereas the ENs can only store up to $\mu N L$ 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 user arbitrarily requests one of the N files from the library. The users' requests in a given TI are denoted by the demand vector $\mathbf{d} \triangleq (d_1, d_2, \ldots, d_K) \in [N]^K$. 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

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

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. For the delivery phase, there are generally two types of transmission policies, serial and pipelined. As detailed below, we first consider the serial transmission mode illustrated in Fig. 2a, and then, in Section II-C, we describe the necessary adjustments to the delivery policies for allowing the pipelined simultaneous transmission mode illustrated in Fig. 2b.

1) Caching Policy: During the placement phase, for EN m, $m \in [M]$, 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

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

Note that, as per (2), we consider policies where only coding within each file is allowed, i.e., no inter-file coding (e.g., [34]) is permitted. In order to satisfy the cache capacity constraints, we restrict the mappings to satisfy $H(s_{m,n}) \leq \mu L$.
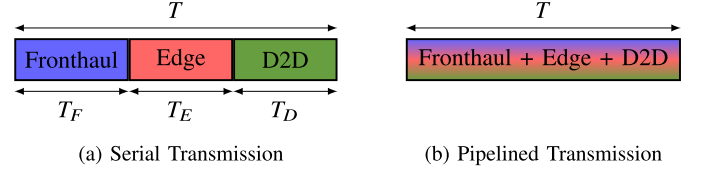

Fig. 2. Transmission interval structure for either serial or pipelined delivery policies.

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

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

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

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

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 $\pi_{\text{D2D},t}^k(\cdot)$ 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 users $[K] \backslash \{k\}$ to the D2D message

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

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

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

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 $\pi_d^k(\cdot)$ 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

$$\hat{f}_{d_k} = \pi_d^k(\mathbf{y}_k, \mathcal{V}_k, d_k, \mathbf{H}), \qquad (7)$$

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

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

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 [11], 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, under serial delivery, 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 these terms by the delivery time needed on the mentioned reference system:

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

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

$$\delta(\mu, r_F, r_D) \triangleq \delta_F + \delta_E + \delta_D, \tag{10}$$

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:

$$\delta^*(\mu, r_F, r_D) \triangleq \inf\{\delta(\mu, r_F, r_D) : \delta(\mu, r_F, r_D) \text{ achievable}\}. \tag{11}$$

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 [11, Lemma 1].

C. Pipelined Transmission

The system defined above is based on serial delivery as illustrated in Fig. 2a. Here we describe an alternative model,

whereby, as seen in Fig. 2b, simultaneous transmissions on fronthaul, edge, and D2D channels are enabled. Specifically, the ENs can simultaneously receive messages over the fronthaul links and transmit on the wireless channel; and the users can receive on the wireless channel while, at the same time, transmitting messages on the D2D links. Following [11], we refer to this model as enabling pipelined delivery.

To elaborate, at time instant $t \in [T]$, where T denotes the delivery latency in a TI, each EN and user transmits using the information received at times $1, \ldots, t-1$, in a causal way. Mathematically, each EN $m \in [M]$ at time $t \in [T]$ uses a function $\pi^m_{\text{P},e,t}(\cdot)$ to map the local cache content, the fronthaul messages received up to time $t-1$, the demand vector, and the global CSI to the output symbol

$$x_m[t] = \pi^m_{\text{P},e,t}(s_m, u_m[1], u_m[2], \ldots, u_m[t-1], \mathbf{d}, \mathbf{H}). \tag{12}$$

Furthermore, user $k \in [K]$ transmits using the function $\pi^k_{\text{P,D2D},t}(\cdot)$ that maps the received edge signal up to time $t-1$, global CSI, and the previously received D2D messages from users $[K] \backslash k$ to the D2D message

$$v_k[t] = \pi^k_{\text{P,D2D},t}\left(y_k[1], \ldots, y_k[t-1], \mathbf{H}, \mathbf{v}^{t-1}_{[K]}\right). \tag{13}$$

Similar to the serial transmission case, the NDT and minimum NDT under pipelined delivery are defined as $\delta_{\text{P}}(\mu, r_F, r_D) \triangleq \lim_{P \to \infty} \lim_{L \to \infty} \mathbb{E}[T] \log(P)/L$, and $\delta^*_{\text{P}}(\mu, r_F, r_D) \triangleq \inf\{\delta_{\text{P}}(\mu, r_F, r_D) : \delta_{\text{P}}(\mu, r_F, r_D)$ is achievable$\}$, respectively. Furthermore, we have the lower bound $\delta^*_{\text{P}}(\mu, r_F, r_D) \geq 1$, and the minimum NDT is a convex function of μ for any fixed values of r_F and r_D. Finally, since serial delivery is a special case of pipelined delivery, by the definition of the minimum NDT, we have the inequality $\delta^*_{\text{P}}(\mu, r_F, r_D) \leq \delta^*(\mu, r_F, r_D)$. This implies that pipelined delivery policies always outperform serial policies in terms of delivery latency. However, pipelined policies are generally more complex due to transmission, reception, and processing taking place at the same time. As a result, for some systems, implementing pipelined policies may not be feasible. The pipelined delivery model is studied in Section VI.

III. DELIVERY STRATEGIES FOR EDGE CACHING WITH D2D COOPERATION

In this section, we start by developing delivery schemes for the special case in which the fractional cache size is $\mu = 1/M$ and the fronthaul capacity is not used. This scenario corresponds to the important special case in which the edge cache capacity is the minimum necessary to guarantee that the entire library $\mathcal{F}$ is available across the caches of all ENs, and hence fronthaul resources may not be used for delivery. Note that, for any request vector, users need to download equal fractions of the requested file from all ENs. This set-up is also known as an X-channel [33]. We first introduce a delivery strategy based on a new interference alignment scheme for an F-RAN with $M = K = 2$. A more scalable strategy based on compress-and-forward is then introduced for any number of ENs and users.

Note that both strategies rely on cooperation between the users. Therefore, due to the latency overhead caused by D2D

communication, they are beneficial only for sufficiently high D2D rate r_D. For low D2D rate, schemes such as cache-aided zero-forcing, cache-aided EN coordination, and cloud-aided soft-transfer, which will be discussed in Section IV-A and do not require receiver cooperation, can achieve lower NDT.

A. Interference Alignment for $M = K = 2$

For the case of $M = 2$ ENs and $K = 2$ users, we present a delivery scheme that integrates D2D communication in the Real Interference Alignment (RIA) scheme introduced in [33]. Our main interest in this scheme stems from its optimality, which will be proved in Section V.

Proposition 1: For a D2D-aided F-RAN with $M = 2$ ENs, each with a fractional cache size $\mu = 1/2$, $K = 2$ users, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$, the minimum NDT under serial delivery is upper bounded as $\delta^*(\mu = 1/M, r_F, r_D) \leq \delta_{\text{D2D-RIA}}$, where

$$\delta_{\text{D2D-RIA}} \triangleq 1 + \frac{1}{2r_D}. \tag{14}$$

Proposition 1 was proved in the conference paper [25] by the authors by leveraging layered transmission, RIA, D2D cooperation, and successive cancellation decoding at the receivers. While referring to [25] for details, we sketch here the main features of the scheme by comparing it to the original RIA scheme introduced in [33] for an X-channel model without D2D cooperation. In RIA, each EN applies layered transmission with two layers by transmitting

$$x_1 = h_{22}a_1 + h_{12}a_2 \quad \text{and} \quad x_2 = h_{21}b_1 + h_{11}b_2, \tag{15}$$

where symbols a_1, a_2, b_1, and b_2 are chosen from a discrete constellation. Each layer is coded using random coding with rate R. Layers a_1 and b_1 are intended for user 1, whereas a_2 and b_2 are intended for user 2. Note that the precoders in (15) are based on perfect knowledge of the CSI at the ENs. The signals (1) received by the two users are hence given as

$$y_1 = h_{11}h_{22}a_1 + h_{12}h_{21}b_1 + h_{11}h_{12}(a_2 + b_2) + z_1,$$
$$y_2 = h_{11}h_{22}b_2 + h_{12}h_{21}a_2 + h_{21}h_{22}(a_1 + b_1) + z_2. \tag{16}$$

As shown in [33], user 1 is able to decode the signal $\tilde{y}_1 \triangleq y_1 - z_1$ from y_1, in the high-SNR regime, if the rate is selected as $R = \log(P)/3$. Next, user 1, which has perfect CSI, searches for a set of symbols $\{a_1, b_1, a_2 + b_2\}$ that generates $\tilde{y}_1$. Since the ENs use a discrete constellation and the channel coefficients are drawn i.i.d. from a continuous distribution, almost surely, this set is unique. This implies that user 1 can decode the desired layers a_1 and b_1 once it has decoded $\tilde{y}_1$. Similarly, user 2 can decode layers a_2 and b_2. Note that the RIA scheme requires $T_E = 3L/(2\log(P))$ channel uses in order to satisfy the users' demands, since each layer consists of $L/2$ bits and is transmitted at a rate of $\log(P)/3$ bits per channel use. It follows that RIA without D2D cooperation achieves an NDT of $3/2$.

In order to leverage D2D cooperation, in the proposed scheme, the ENs apply layered transmission with n_d layers, where n_d is odd. The transmitted signals are hence given as

$$x_1 = \sum_{i=1}^{n_d} g_{1,i}a_i \quad \text{and} \quad x_2 = \sum_{i=1}^{n_d} g_{2,i}b_i, \tag{17}$$

where precoder gains $\{g_{m,i}\}$, with $m \in [2]$ and $i \in [n_d]$, are selected to satisfy $h_{11}g_{1,i} = h_{12}g_{2,i-1}$ and $h_{22}g_{2,i} = h_{21}g_{1,i-1}$. Consider, as an example, the case with $n_d = 3$ layers. The signals (1) received by the two users are hence given as

$$y_1 = h_{11}g_{1,1}a_1 + h_{12}g_{2,3}b_3$$
$$+ h_{11}g_{1,2}(a_2 + b_1) + h_{11}g_{1,3}(a_3 + b_2) + z_1,$$
$$y_2 = h_{22}g_{2,1}b_1 + h_{21}g_{1,3}a_3$$
$$+ h_{22}g_{2,2}(b_2 + a_1) + h_{22}g_{2,3}(b_3 + a_2) + z_2. \tag{18}$$

In a manner similar to the RIA scheme [33], it can be shown that user 1 is able to decode the signal $\tilde{y}_1 = y_1 - z_1$ from y_1, in the high-SNR regime, if each layer is coded with rate $R = \log(P)/4$. Then, user 1 searches for the unique set $\mathcal{R}_1 \triangleq \{a_1, b_3, a_2 + b_1, a_3 + b_2\}$ of symbols that generates $\tilde{y}_1$. The uniqueness of this set is determined by the same arguments used for the RIA scheme. Likewise, user 2 is able to identify the set $\mathcal{R}_2 \triangleq \{b_1, a_3, b_2 + a_1, b_3 + a_2\}$.

In order to decode the desired layers, user 1 transmits the message $v_1 = a_2 + b_1 \in \mathcal{R}_1$ to user 2 over the D2D links, whereas user 2 transmits $v_2 = b_2 + a_1 \in \mathcal{R}_2$ to user 1. User 1 is thus able to decode $\{a_1, b_2, a_3\}$ by means of successive cancellation decoding from $\{\mathcal{R}_1, v_2\}$. To this end, it starts by decoding $a_1 \in \mathcal{R}_1$; then, it uses a_1 together with $v_2 = b_2 + a_1$ to decode b_2; next, it uses b_2 and $(a_3 + b_2) \in \mathcal{R}_1$ to decode a_3. Similarly, user 2 decodes $\{b_1, a_2, b_3\}$ from $\{\mathcal{R}_2, v_1\}$. For general number of layers n_d, as detailed in [25], each user is able to decode $(n_d + 1)/2$ layers from one EN and $(n_d - 1)/2$ layers from the other, where each layer is coded with rate $R = \log(P)/(n_d + 1)$.

The scheme requires $T_E = (n_d + 1)L/((n_d - 1)\log(P))$ downlink channel uses since each EN conveys $L/2$ bits to each user over $(n_d - 1)/2$ layers, which are transmitted at a rate of $\log(P)/(n_d + 1)$ bits per channel use. Unlike RIA, there is an additional latency overhead of $T_D = L/(2r_D \log(P))$ due to sharing $(n_d - 1)/2$ layers over each D2D link. Therefore, assuming an arbitrarily large number of layers at the ENs, the NDT (14) is obtained.

Note that, with respect to RIA, the proposed scheme is able to reduce the edge NDT δ_E (9) from $3/2$ to 1 at the cost of increasing the D2D NDT δ_D from 0 to $1/(2r_D)$. This demonstrates the capability of this scheme to trade-off between the D2D NDT and the edge and fronthaul NDTs, as enabled by D2D communication.

B. Compress-and-Forward D2D Transmission

The scheme discussed above appears to be cumbersome to generalize beyond the case $M = K = 2$. Furthermore, at a practical level, this approach is mostly of theoretical interest since the performance of RIA is known to degrade catastrophically when CSI at the transmitters is imperfect [35]. Therefore, here we present an achievable scheme that applies to all values of M and K and requires only CSI at the receivers. The scheme is based on Compress-and-Forward (CF) D2D communication, and its near-optimality properties will also be discussed in Section V.

Proposition 2: 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 under serial delivery is upper bounded as $\delta^*(\mu = 1/M, r_F, r_D) \leq \delta_{\text{D2D-CF}}$, where the NDT

$$\delta_{\text{D2D-CF}} \triangleq \frac{K}{\min\{M, K\}} \left(1 + \frac{1}{r_D}\right) \qquad (19)$$

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

The NDT (19) is achieved by the following scheme. Consider first the case $M \geq K$. At any time, K out of the M ENs transmit simultaneously, each transmitting a fraction of the requested file to one of the K users. As a result, the ENs' transmissions interfere at each user. After downlink transmission, each user compresses and forwards its received signal to all other users over the D2D links. After D2D communication, each user collects the K received signals, namely the signal that was directly received over the downlink channel and the compressed versions that were shared by the other users. Based on these signals, each user carries out ZF equalization in order to recover the desired signal with no interference from other signals.

To elaborate, consider, for example, the case where the first K ENs are active. After D2D cooperation, the signals $\mathbf{v} = [v_1, \ldots, v_K]^T$ available at user $k \in [K]$ can be expressed as $\mathbf{v}_k = \mathbf{H}_K \mathbf{x} + \mathbf{z} + \mathbf{q}_k$, where $\mathbf{x} \triangleq [x_1, \ldots, x_K]^T$ represents the transmitted signals, $\mathbf{H}_K$ is the channel matrix such that $(\mathbf{H}_K)_{i,j} = h_{ij}$, $\mathbf{z} \triangleq [z_1, \ldots, z_K]^T$ represents the white Gaussian noise, and $\mathbf{q}_k \triangleq [q_1, \ldots, q_K]^T$ represents the compression noise vector. We have $q_k = 0$ since user k receives y_k directly over the downlink channel (1). The channel coefficients are drawn i.i.d. from a continuous distribution; therefore, almost surely, matrix $\mathbf{H}_K$ is invertible. Hence, each user can apply ZF equalization, i.e., multiply the received signals by $\mathbf{H}_K^{-1}$, to get $\mathbf{H}_K^{-1} \mathbf{v}_k = \mathbf{x} + \mathbf{H}_K^{-1}(\mathbf{z} + \mathbf{q}_k)$. Note that, after ZF equalization, the ENs' transmissions no longer cause interference. Therefore, the achievable rate is determined by the power of the additive noise $\mathbf{H}_K^{-1}(\mathbf{z} + \mathbf{q}_k)$. As shown in [11, App. II-A], by compressing with a rate equal to $\log(P)$ bits per downlink symbol, we can guarantee that the SNR after compression scales linearly with P. Thus, in the high-SNR regime, each EN is able to transmit with a rate of $R \approx \log(P)$ bits/channel use.

To satisfy the users' demands, each EN must convey L/M bits to each user. To this end, we cluster the ENs into all possible $\binom{M}{K}$ subsets of K ENs, and schedule each cluster into distinct time intervals of duration $T_E / \binom{M}{K}$. Since each EN participates in $\binom{M-1}{K-1}$ clusters, and the total number of bits transmitted by each EN is KL/M, then the duration of each interval is given as

$$\frac{T_E}{\binom{M}{K}} = \frac{KL/M}{\binom{M-1}{K-1} R} = \frac{L}{\binom{M}{K} \log(P)}. \qquad (20)$$

Therefore, the number of downlink channel uses is $T_E = L/\log(P)$, and hence the proposed scheme achieves an ideal edge NDT of $\delta_E = 1$. Since, for each downlink channel use, each user transmits $\log(P)$ bits over the D2D link, a latency

overhead of $T_D = T_E \log(P)/C_D = T_E/r_D$ is added to the delivery time, and hence the total NDT is (19).

For the complementary case in which $M < K$, all ENs are active. If K is a multiple of M, then the users are partitioned into K/M disjoint clusters of M users. For each cluster, ZF equalization requires $L/\log(P)$ downlink channel uses in order to satisfy the demands of the users in the cluster. Therefore, the edge delivery time is $T_E = KL/(M\log(P))$, and hence the total NDT is (19). For the more general case in which K/M may not be an integer, the same edge delivery time can be achieved by clustering the users into all possible $\binom{K}{M}$ subsets of M users, and, for each cluster, setting an interval of duration

$$\frac{T_E}{\binom{K}{M}} = \frac{L}{\binom{K-1}{M-1} \log(P)}. \qquad (21)$$

IV. BOUNDS ON THE MINIMUM NDT FOR SERIAL DELIVERY

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

A. Upper Bounds and Achievable Strategy

In the previous section, we presented schemes for the special case in which the fractional cache size is $\mu = 1/M$. To obtain a policy that applies for any value of fractional cache size μ, we combine, via file-splitting and cache-sharing, the D2D-based CF scheme (Proposition 2) with the best-known general strategies for an F-RAN model with no D2D cooperation. These strategies are described next for reference, followed by a review of file-splitting and cache-sharing.

1) Cache-Aided ZF [11, Lemma 2]: Cache-aided ZF precoding requires that all ENs cache the entire library of files, and hence it only applies for $\mu = 1$. Full caching allows the ENs to cooperate by applying ZF-beamforming, whereby the precoding matrix equals the inverse of the channel matrix. This generates $\min\{M, K\}$ interference-free links to the users. Therefore, in the high-SNR regime, this scheme achieves a sum-rate of $\min\{M, K\} \log(P)$, and hence an NDT of $\delta_{\text{ZF}} \triangleq K/\min\{M, K\}$.

2) Cache-Aided EN Coordination [11, Lemma 3]: The RIA scheme discussed in Section III-A can be applied to arbitrary number of ENs and users. Each EN transmits M layers, and each layer is coded using random coding with rate $R \approx \log(P)/(M+K-1)$ bits per symbol. The layers are precoded such that, at each user, the desired layers can be decoded. The scheme hence achieves an NDT of $\delta_{\text{IA}} \triangleq (M + K - 1)/M$.

3) Cloud-Aided Soft-Transfer [11, Proposition 3]: In this scheme, ZF precoding is carried out at the cloud, which has access to the entire library of files. The resulting encoded signals are then compressed with a resolution of $\log(P)$ bits per downlink baseband sample and conveyed to the ENs over the fronthaul links. Similar to the CF-based scheme (Proposition 2), it can be shown that the effective SNR in the downlink scales proportionally to the power P, and that this scheme achieves an NDT of $\delta_{\text{ST}} \triangleq K/\min\{M, K\} + K/(Mr_F)$,

where the latency overhead of $\delta_F = K/(Mr_F)$ is due to transmission over the fronthaul links.

The described delivery techniques are combined by means of file-splitting and cache-sharing [11, Lemma 1]. That is, all files are split in the same way into a number of fragments, and each fragment is delivered by using a different policy.

To formulate the main result, we define the threshold values

$$r_F^{\text{th}} \triangleq \frac{K(M-1)}{M(\min\{M,K\}-1)}, \tag{22}$$

and

$$r_D^{\text{th}} \triangleq \max\left\{ \frac{\max\{M,K\}}{\min\{M,K\}-1}, \frac{M^2 r_F}{(M-1)\min\{M,K\}} \right\}. \tag{23}$$

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 minimum NDT under serial delivery 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

$$\begin{aligned} &\delta_{\text{ach}}(\mu, r_F, r_D) \\ &= (M+K-1)\mu \\ &\quad + (1-\mu M)\left[\frac{K}{\min\{M,K\}} + \frac{K}{Mr_F} \right]. \end{aligned} \tag{24}$$

- 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

$$\begin{aligned} &\delta_{\text{ach}}(\mu, r_F, r_D) \\ &= \frac{K}{\min\{M,K\}}\left(\frac{\mu M - 1}{M-1} \right) + (1-\mu)\frac{M+K-1}{M-1}. \end{aligned} \tag{25}$$

- 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

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{K}{\min\{M,K\}} + \frac{(1-\mu)K}{Mr_F}. \tag{26}$$

- 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

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{(1+\mu M/r_D)K}{\min\{M,K\}} + \frac{(1-\mu M)K}{Mr_F}. \tag{27}$$

- 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

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{K}{\min\{M,K\}}\left(1 + \frac{(1-\mu)M}{(M-1)r_D} \right). \tag{28}$$

Proof: For the first three regimes, i.e., for low D2D rate $r_D \leq r_D^{\text{th}}$, the NDTs in (24)–(26) are achieved by applying the strategy of [11, Proposition 4], which does not require D2D resources. For the other regimes, please see Appendix A. ∎

For the special case of $M = 2$ ENs and $K = 2$ users, the following NDT is achieved by using the D2D-enhanced RIA scheme of Proposition 1.

Proposition 4: For a 2×2 D2D-aided F-RAN with a fractional cache size $\mu \in [0,1]$, a library of $N \geq 2$ files, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$, the minimum NDT under serial delivery is upper bounded as $\delta^*(\mu, r_F, r_D) \leq \delta_{2\times 2}(\mu, r_F, r_D)$, where

$$\delta_{2\times 2}(\mu, r_F, r_D) \tag{29}$$

$$\triangleq \begin{cases} \max\left\{ 1 + \mu + \dfrac{1-2\mu}{r_F}, \right. \\ \qquad \left. 2 - \mu \right\} & \text{for } 0 \leq r_F, r_D \leq 1, \\[2mm] 1 + \dfrac{1-\mu}{r_F}, & \text{for } r_F \geq \max\{1, r_D\}, \\[2mm] \max\left\{ 1 + \dfrac{\mu}{r_D} + \dfrac{1-2\mu}{r_F}, \right. \\ \qquad \left. 1 + \dfrac{1-\mu}{r_D} \right\} & \text{for } r_D > \max\{1, r_F\}. \end{cases}$$

Proof: Follows from Proposition 3 by replacing the D2D threshold r_D^{th} in (23) with $r_D^{\text{th}} = \max\{1, r_F\}$, and, for D2D rate $r_D > r_D^{\text{th}}$, by applying the D2D scheme of Proposition 1 instead of the CF-based scheme. ∎

B. Lower Bound

A general lower bound on the minimum NDT is given in Proposition 5. Following [11], 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 5: 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 under serial delivery 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

$$\text{minimize} \quad \delta_F + \delta_E + \delta_D \tag{30a}$$
$$\text{subject to} \quad l\delta_E + (M-l)r_F\delta_F + g(l)r_D\delta_D$$
$$\geq K - (M-l)(K-l)\mu, \tag{30b}$$
$$\delta_E \geq \frac{K}{\min\{M,K\}}, \tag{30c}$$
$$\delta_F \geq 0, \ \delta_D \geq 0, \tag{30d}$$

where (30b) is a family of constraints with $l = 0, 1, \ldots, \min\{M, K\}$, and

$$g(l) \triangleq \begin{cases} 0, & \text{for } l = 0, \\ K-1, & \text{for } l = 1, \\ K, & \text{for } l = 2, \ldots, \min\{M, K\}. \end{cases} \tag{31}$$

Proof: See Appendix B. ∎

Note that, without D2D communication, i.e., $r_D = 0$, the linear program (30) is identical to that of [11, Proposition 1]. For $r_D > 0$, the additional term $g(l)r_D\delta_D$ in (30b) reflects the novel trade-off between the D2D NDT δ_D and the edge and fronthaul NDTs δ_E and δ_F, respectively.

V. CHARACTERIZATION OF THE MINIMUM NDT FOR SERIAL DELIVERY

In this section, based on the lower and upper bounds presented in Section IV, we discuss the optimality properties of the D2D-based strategies.

A. 2×2 D2D-Aided F-RAN

For the case of $M = 2$ ENs and $K = 2$ users, as detailed in the following proposition, the D2D-based strategy of Proposition 4 is optimal.

Proposition 6: The minimum NDT for the 2×2 F-RAN system with number of files $N \geq 2$, a fractional cache size $\mu \in [0,1]$, a fronthaul rate $r_F \geq 0$, and a D2D rate $r_D \geq 0$ is given as $\delta^*(\mu, r_F, r_D) = \delta_{2\times 2}(\mu, r_F, r_D)$.

Proof: See Appendix C. ∎

Proposition 6 can be used 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 $\delta_{2\times 2}(\mu, r_F, r_D)$ (29) is identical to the minimum NDT without D2D links derived in [11, Corollary 3]. Therefore, D2D communication provides a latency reduction only when we have $r_D > \max\{1, r_F\}$.

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, when $r_D > \max\{1, r_F\}$, D2D communication reduces the minimum NDT for all values $0 < \mu < 1$. Furthermore, when $\mu > 1/2$, 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.

B. General D2D-Aided F-RAN

For arbitrary number of ENs and users, we start with the main result in the following proposition, which shows that the achievable CF-based strategy of Proposition 3 is optimal to within a multiplicative factor of two.

Proposition 7: 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 Proposition 3 achieves the minimum NDT under serial delivery to within a factor of two, i.e.,

$$\frac{\delta_{\text{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \leq 2. \tag{32}$$

Proof: See Appendix D. ∎

The key result in Proposition 7 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. 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 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 gap identified in (32) is generally not zero, the next corollary states that CF is close to optimal for sufficiently high D2D rate.

Corollary 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 $r_D \geq \max\{r_D^{\text{th}}, 1/\epsilon\}$ with r_D^{th} in (23) and $\epsilon > 0$, the achievable strategy of Proposition 3 is close to optimal in the sense that we have

$$\frac{\delta_{\text{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \leq 1 + \epsilon. \tag{33}$$

Proof: Corollary 1 follows directly from the proof of Proposition 7 (Appendix D) since, for $r_D \geq r_D^{\text{th}}$, we have $\delta_{\text{ach}}(\mu, r_F, r_D)/\delta^*(\mu, r_F, r_D) \leq 1 + 1/r_D$ (cf. (64) and (67)). ∎

Corollary 1 is illustrated in Fig. 4. 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 the D2D rate r_D increases, the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ is seen to approach the lower bound $\delta_{\text{lb}}(\mu, r_F, r_D)$. For instance, 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. Note that, for low D2D rate, Fig. 4 suggests that the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ and the lower bound $\delta_{\text{lb}}(\mu, r_F, r_D)$ do not depend on r_D. Intuitively, this is because, under serial delivery, D2D communication with low D2D rate is not beneficial due to the significant latency overhead associated with D2D communication.

Fig. 4 also shows that, for high D2D rate, the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ is much lower than the minimum NDT for $r_D = 0$, i.e., when no receiver cooperation is allowed. To further highlight the gains that can be achieved with sufficiently high D2D rate, in Fig. 5, we plot the achievable NDT $\delta_{\text{ach}}(\mu, r_F, r_D)$ as a function of the fractional cache size μ, for fronthaul rate $r_F = 1$, D2D rate $r_D = 10$, and $M = K = 5$ ENs and users, as well as the lower and upper bounds on the minimum NDT for the case of $r_D = 0$. For all fractional cache size $0 < \mu < 1$, the strategy presented in Proposition 3, which uses CF-based D2D communication, improves upon the minimum NDT without receiver cooperation. Moreover,
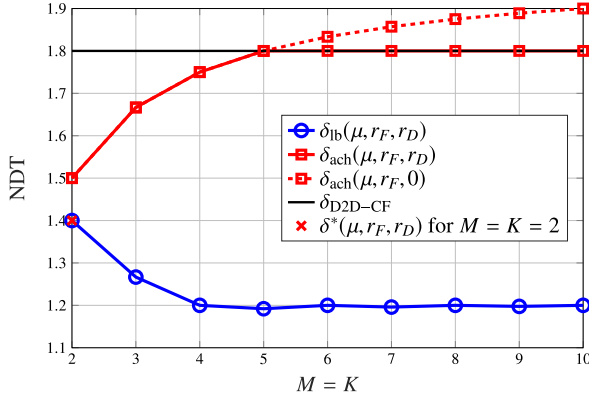
Fig. 3.   Lower and upper bounds on the minimum NDT as a function of the number of ENs and users $M = K$ for $r_F = 1$, $\mu = 1/M$, and $r_D = 1.25$ or $r_D = 0$.
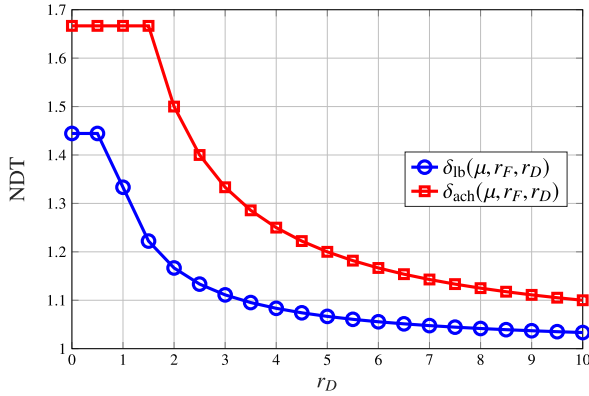

Fig. 4.   Lower and upper bounds on the minimum NDT as a function of r_D for $r_F = 1$, $M = K = 3$, and $\mu = 1/3$.

the gain is maximal for $\mu = 1/M$, i.e., when no cache-aided EN cooperation is possible.

VI. PIPELINED DELIVERY

In this section, we study the D2D-aided F-RAN model with pipelined delivery as defined in Section II-C. We proceed in a manner similar to serial delivery by first deriving lower and upper bounds on the minimum NDT, and then discussing the optimality of CF-based D2D delivery.

A. Lower Bound on the Minimum NDT

A lower bound on the minimum NDT for an $M \times K$ D2D-aided F-RAN under pipelined delivery policies is given in Corollary 2. The lower bound is derived by following the same arguments as in Proposition 5, with the caveat that, under pipelined delivery policies, fronthaul, edge, and D2D transmissions occur simultaneously rather than sequentially.

Corollary 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 under pipelined delivery is lower bounded as $\delta_P^*(\mu, r_F, r_D) \geq \delta_{P,lb}(\mu, r_F, r_D)$, where

$$\delta_{P,lb}(\mu, r_F, r_D)$$
$$= \max \left\{ \frac{K}{\min\{M, K\}}, \max_{l=0,\ldots,\min\{M,K\}} \frac{K - (M - l)(K - l)\mu}{l + (M - l)r_F + g(l)r_D} \right\}, \quad (34)$$
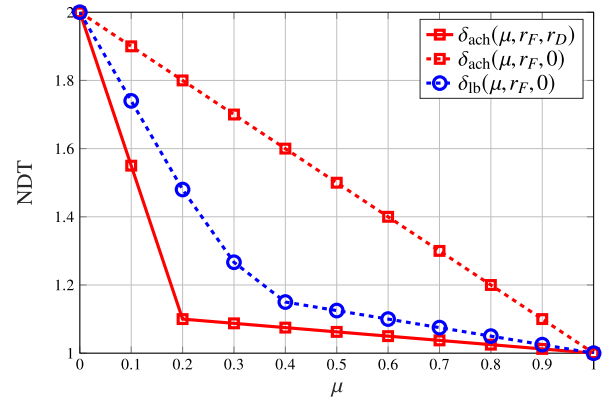
and $g(l)$ is defined in (31).

Fig. 5.   Lower and upper bounds on the minimum NDT as a function of μ for $r_F = 1$, $r_D = 10$, and $M = K = 5$.

Proof: Follows from the proof of Proposition 5 (Appendix B) with the following difference. For pipelined delivery policies, vectors $\mathbf{u}_m$, $\mathbf{x}_m$, $\mathbf{y}_k$, $\mathbf{z}_k$, and $\mathbf{v}_k$, which represents fronthaul message sent to EN $m \in [M]$, output codeword transmitted by EN m, signal received by user $k \in [K]$ on the shared wireless channel, white Gaussian noise at user k, and D2D message transmitted by user k, respectively, have T entries, where T is the delivery latency. ∎

B. Upper Bound on the Minimum NDT

To upper bound the minimum NDT, we consider a strategy that converts the CF-based serial transmission policies discussed in Section IV-A into a pipelined delivery policy by means of block-Markov encoding and per-block file splitting. The approach is a generalization of the method presented in [11, Sec.VI-B] for an F-RAN with no D2D links. To elaborate, fix a serial delivery policy with its fronthaul, edge, and D2D transmission strategy. As illustrated in Fig. 6, in order to convert this strategy into one that leverages pipelining, every file in the library is split into B blocks of size L/B bits each, and every TI is divided into $B + 2$ slots. In each slot $b \in [B + 2]$, the CP uses the fronthaul links to deliver the bth block of the requested files using the fronthaul transmission strategy of the selected serial policy. At the same time, the ENs, having received the fronthaul message for the $(b-1)$th block in the previous slot, apply the edge transmission strategy of the serial policy to deliver the $(b-1)$th block of the requested files to the users; and the users apply the corresponding conferencing scheme to cooperate in the decoding of the $(b-2)$th block of the requested files.

For a serial delivery scheme that achieves fronthaul, edge, and D2D transmission durations T_F, T_E, and T_D, respectively, the block-Markov approach, with arbitrarily large number of blocks B, achieves the pipelined NDT

$$\delta_{P,ach}(\mu, r_F, r_D)$$
$$= \lim_{B \to \infty} \lim_{P \to \infty} \lim_{L \to \infty} \frac{B + 2}{B} \cdot \frac{\max\{T_F, T_E, T_D\}}{L/\log(P)}$$
$$= \max\{\delta_F, \delta_E, \delta_D\}, \quad (35)$$

where δ_F, δ_E, and δ_D are the fronthaul, edge, and D2D NDTs of the serial transmission scheme as defined in (9). Moreover,
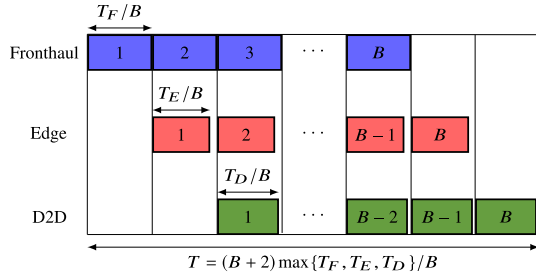
Fig. 6. Pipelining via block-Markov encoding.

for two serial transmission schemes, one that achieves NDTs $\delta_F^{(1)}$, $\delta_E^{(1)}$, and $\delta_D^{(1)}$, whereas the other achieves NDTs $\delta_F^{(2)}$, $\delta_E^{(2)}$, and $\delta_D^{(2)}$, and for some $\alpha \in [0, 1]$, the following pipelined NDT is achievable [11, Sec. VI-B]

$$\delta_{\text{P,ach}}(\mu, r_F, r_D) = \max \left\{ \begin{array}{l} \alpha\delta_F^{(1)} + (1-\alpha)\delta_F^{(2)}, \\ \alpha\delta_E^{(1)} + (1-\alpha)\delta_E^{(2)}, \\ \alpha\delta_D^{(1)} + (1-\alpha)\delta_D^{(2)} \end{array} \right\}. \quad (36)$$

Proposition 8: For an $M \times K$ D2D-aided F-RAN with a fractional cache size $\mu \in [0, 1]$, 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 under pipelined delivery is upper bounded as $\delta^*(\mu, r_F, r_D) \leq \delta_{\text{P,ach}}(\mu, r_F, r_D)$, where the achievable NDT $\delta_{\text{P,ach}}(\mu, r_F, r_D)$ is given for two distinct regimes of operation as follows:

- High fronthaul rate ($r_F \geq \min\{M, K\}/M$):

$$\delta_{\text{P,ach}}(\mu, r_F, r_D) = \frac{K}{\min\{M, K\}}. \quad (37)$$

- Low fronthaul rate ($r_F < \min\{M, K\}/M$):

$$\delta_{\text{P,ach}}(\mu, r_F, r_D)$$
$$= \begin{cases} \dfrac{(1 - M\mu)K}{Mr_F} & \text{for } \mu \in [0, \mu_1], \\ \dfrac{(1 - M\mu_1)K}{Mr_F} \cdot \dfrac{\mu_2 - \mu}{\mu_2 - \mu_1} & \\ \quad + \dfrac{K}{\min\{M, K\}} \cdot \dfrac{\mu - \mu_1}{\mu_2 - \mu_1} & \text{for } \mu \in (\mu_1, \mu_2), \\ \dfrac{K}{\min\{M, K\}} & \text{for } \mu \in [\mu_2, 1], \end{cases} \quad (38)$$

where we have defined

$$\mu_1 \triangleq \frac{K - \max\{M, K\}r_F}{KM + Mr_F[\min\{M, K\} - 1]}, \quad (39)$$

and

$$\mu_2 \triangleq \max \left\{ 1 - \frac{Mr_F}{\min\{M, K\}} - \frac{M - 1}{M}r_D, \right.$$
$$\left. \frac{1}{M} - \frac{r_F}{\min\{M, K\}} \right\}. \quad (40)$$

Proof: See Appendix E. ∎

C. Characterization of the Minimum NDT

In the following propositions we discuss the optimality of the D2D CF-based strategy under pipelined delivery. First, we prove that the multiplicative suboptimality factor of two, identified in Proposition 7, applies also to pipelined delivery policies.

Proposition 9: For a D2D-aided F-RAN with M ENs, K users, a library of $N \geq K$ files, a fronthaul rate $r_F < \min\{M, K\}/M$, and a D2D rate $r_D < 1 - Mr_F/\min\{M, K\}$, the strategy of Proposition 8 achieves the minimum NDT under pipelined delivery to within a factor of two, i.e.,

$$\frac{\delta_{\text{P,ach}}(\mu, r_F, r_D)}{\delta_{\text{P}}^*(\mu, r_F, r_D)} \leq 2, \quad \forall \mu \in [\mu_1, \mu_2] \quad (41)$$

Proof: See Appendix F. ∎

Next, we show that the achievable strategy of Proposition 8 is optimal for the high fronthaul regime with $r_F \geq \min\{M, K\}/M$; for the high D2D regime with $r_D \geq 1 - Mr_F/\min\{M, K\}$; for the low cache regime with $\mu \in [0, \mu_1]$; and for the high cache regime with $\mu \in [\mu_2, 1]$.

Proposition 10: 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 characterized for three distinct regimes of operation as follows:

- High fronthaul rate ($r_F \geq \min\{M, K\}/M$):

$$\delta_{\text{P}}^*(\mu, r_F, r_D) = \frac{K}{\min\{M, K\}}. \quad (42)$$

- Low fronthaul rate and high D2D rate ($r_F < \min\{M, K\}/M$ and $r_D \geq 1 - Mr_F/\min\{M, K\}$):

$$\delta_{\text{P}}^*(\mu, r_F, r_D) = \max \left\{ \frac{(1 - M\mu)K}{Mr_F}, \frac{K}{\min\{M, K\}} \right\}. \quad (43)$$

- Low fronthaul and D2D rates ($r_F < \min\{M, K\}/M$ and $r_D < 1 - Mr_F/\min\{M, K\}$):

$$\delta_{\text{P}}^*(\mu, r_F, r_D) = \begin{cases} \dfrac{(1 - M\mu)K}{Mr_F} & \text{for } \mu \in [0, \mu_1], \\ \dfrac{K}{\min\{M, K\}} & \text{for } \mu \in [\mu_2, 1], \end{cases} \quad (44)$$

where μ_1 and μ_2 are defined in (39) and (40), respectively.

Proof: See Appendix G. ∎

In the pipelined case, as seen in Fig. 6, the latency is dictated by the largest among fronthaul, D2D, and edge NDTs. Therefore, whenever the fronthaul rate is large enough to enable ZF precoding on the wireless channel without causing a bottleneck, the minimum NDT can be achieved without using D2D communication. However, for low fronthaul rate and low cache capacity, cooperation via CF-based ZF equalization allows the delivery latency to be reduced by alleviating fronthaul load without increasing the edge NDT.

Comparing the results for serial and pipelined delivery policies, we observe that both the achievable NDT in Proposition 3 and the lower bound in Proposition 5 are strictly decreasing

functions of r_D for all $r_D \geq r_D^{\text{th}}$, and hence the minimum NDT under serial delivery is strictly decreasing as well (cf. Fig. 4). In contrast, under pipelined delivery, the minimum NDT (43) for large r_D is a constant function of r_D. This is because, when $r_D \geq 1 - Mr_F/\min\{M, K\}$, the duration of the D2D transmission in each slot of the optimal block-Markov strategy is smaller than the fronthaul or edge transmissions, and hence increasing the D2D rate further does no reduce the minimum NDT.

The role of D2D cooperation in improving the delivery latency under pipelined delivery policies is further illustrated in Fig. 7, where we plot the lower and upper bounds on the minimum NDT as a function of the fractional cache size μ for an F-RAN with $M = 10$ ENs, $K = 10$ users, and a fixed fronthaul rate $r_F = 0.4$. For small cache capacities satisfying $\mu \leq \mu_1$, D2D communication cannot reduce the minimum NDT because, in this regime, the total delivery time is dictated by fronthaul communication, which is required to deliver a large part of the requested files. In addition, for $\mu \geq 1 - Mr_F/\min\{M, K\}$, the cache capacity is large enough to support delivery via cache-aided ZF with a fronthaul overhead that does not affect the achievability of the ideal NDT of one. However, for $\mu_1 < \mu < 1 - Mr_F/\min\{M, K\}$, a D2D-based scheme provides a latency reduction. For example, as depicted in Fig. 7, for $r_D \geq 1 - Mr_F/\min\{M, K\}$, an ideal NDT of one can be achieved with a fractional cache size M times smaller than is required when no D2D communication is allowed ($r_D = 0$).

VII. CONCLUSION

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, and under both serial and pipelined delivery policies. 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. Furthermore, we characterized the minimum NDT for the case of two ENs and users, and it was demonstrated that D2D communication can alleviate the load on the network infrastructure by reducing the traffic on the fronthaul links.

Among related open problems we mention the design of robust delivery strategies that cope with cases in which: some of the D2D links may be in outage; the D2D communication takes place over an interference channel; CSI at the ENs and cloud may be imperfect; inter-file coding is allowed; and security constraints are imposed on the ENs [36]. Another related problem is the characterization of the optimal D2D strategy under linear precoding and hard-transfer constraints where each EN is equipped with multiple antennas [22].

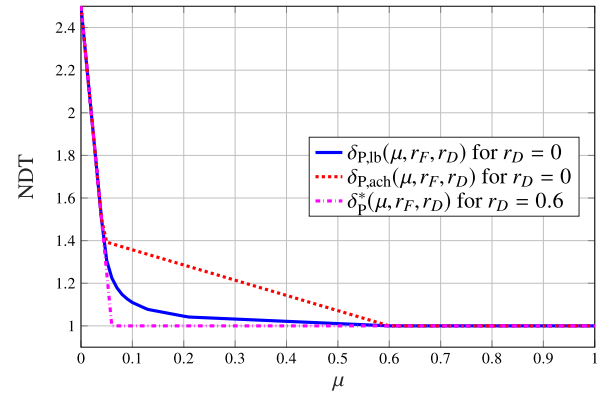Finally, we remark that D2D communication is not only useful for improving content delivery latency, but can be

Fig. 7. Lower and upper bounds on the minimum NDT as a function of μ for $r_F = 0.4$ and $M = K = 10$.

used, in a similar fashion, to lower the communication cost of other services such as edge computing, as implemented by cloudlets [37] and ENs [38], [39], for latency-sensitive applications. By sharing their messages over out-of-band high-capacity D2D links, the users can cooperate, perform joint-precoding of the transmitted signals, and achieve a higher spectral efficiency, which translates to a lower communication latency over the shared wireless channel. Characterizing the optimal user-cooperation scheme for edge computing while properly accounting for the latency overhead associated with D2D communication is also an open problem.

APPENDIX

A. Proof of Proposition 3 for High D2D Rate $r_D > r_D^{th}$

For low cache, i.e., for $\mu \leq 1/M$, a fraction μM of each of the requested files is delivered via D2D-based CF, whereas the remaining $(1 - \mu M)$ fraction is delivered via cloud-aided soft-transfer. The cache capacity constraint is satisfied since $\mu M \times 1/M + (1 - \mu M) \times 0 = \mu$, and the overall NDT is

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \mu M \delta_{\text{D2D-CF}} + (1 - \mu M)\delta_{\text{ST}}. \quad (45)$$

For high cache, i.e., for $\mu > 1/M$, a fraction $(1 - \mu)M/(M-1)$ of each of the requested files is delivered via D2D-based CF, whereas the remaining $(\mu M - 1)/(M - 1)$ fraction is delivered via cache-aided ZF. The cache capacity constraint is satisfied since $(1 - \mu)M/(M - 1) \times (1/M) + (\mu M - 1)/(M - 1) \times 1 = \mu$, and the overall NDT is

$$\delta_{\text{ach}}(\mu, r_F, r_D) = \frac{(1-\mu)M}{M-1}\delta_{\text{D2D-CF}} + \frac{\mu M - 1}{M-1}\delta_{\text{ZF}}. \quad (46)$$

B. Proof of Proposition 5

For the proof of Proposition 5, we use the notation introduced in [11, App. I]. Accordingly, for integers $0 \leq a \leq b \leq K$ and $0 \leq c \leq d \leq M$, we define $\mathbf{f}_{[a:b]} \triangleq (f_a, f_{a+1}, \ldots, f_b)$, $\mathbf{s}_{[c:d]} \triangleq (s_c, s_{c+1}, \ldots, s_d)$, $\mathcal{U}_{[c:d]} \triangleq \{\mathbf{u}_c, \mathbf{u}_{c+1}, \ldots, \mathbf{u}_d\}$, as well as the matrix of channel outputs

$$\mathbf{Y}_{[a:b]} \triangleq \begin{bmatrix} y_a[1] & y_a[2] & \cdots & y_a[T_E] \\ y_{a+1}[1] & y_{a+1}[2] & \cdots & y_{a+1}[T_E] \\ \vdots & \vdots & \ddots & \vdots \\ y_b[1] & y_b[2] & \cdots & y_b[T_E] \end{bmatrix}, \quad (47)$$

and similarly for $\mathbf{Z}_{[a:b]}$ and $\mathbf{X}_{[c:d]}$. Furthermore, we define the following sub-matrix of the channel matrix $\mathbf{H}$

$$\mathbf{H}_{[a:b]}^{[c:d]} \triangleq \begin{bmatrix} h_{a,c} & h_{a,c+1} & \cdots & h_{a,d} \\ h_{a+1,c} & h_{a+1,c+1} & \cdots & h_{a+1,d} \\ \vdots & \vdots & \ddots & \vdots \\ h_{b,c} & h_{b,c+1} & \cdots & h_{b,d} \end{bmatrix}. \quad (48)$$

It follows from [11, App. I] that, for $l = 0, 1, \ldots, \min\{M, K\}$,

$$KL = I\left(\mathbf{f}_{[1:K]}; \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]} \,\middle|\, \mathbf{f}_{[K+1:N]}\right) \\ + H\left(\mathbf{f}_{[1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right), \quad (49)$$

and

$$I\left(\mathbf{f}_{[1:K]}; \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]} \,\middle|\, \mathbf{f}_{[K+1:N]}\right) \\ \leq H\left(\mathbf{f}_{[1:l]} \,\middle|\, \mathbf{Y}_{[1:l]}\right) + (M-l)(K-l)\mu L \\ + l T_E \log(\Lambda P + 1) + (M-l) r_F T_F \log(P), \quad (50)$$

where $\Lambda = \max_{k\in[l]} \left[\sum_{m=1}^{M} |h_{km}|^2 + \sum_{m\neq\tilde{m}} h_{km} h_{k\tilde{m}}^*\right]$, and with the abuse of notation $\mathbf{Y}_{[1:0]} = \emptyset$ and $\mathbf{f}_{[1:0]} = \emptyset$.

We bound $H\left(\mathbf{f}_{[1:l]} \,\middle|\, \mathbf{Y}_{[1:l]}\right)$ in (50) as follows

$$H\left(\mathbf{f}_{[1:l]} \,\middle|\, \mathbf{Y}_{[1:l]}\right) \\ = H\left(\mathbf{f}_{[1:l]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{V}_1, \ldots, \mathcal{V}_l\right) \\ \qquad + I\left(\mathbf{f}_{[1:l]}; \mathcal{V}_1, \ldots, \mathcal{V}_l \,\middle|\, \mathbf{Y}_{[1:l]}\right) \\ \leq \sum_{k=1}^{l} H(f_k | \mathbf{y}_k, \mathcal{V}_k) + I\left(\mathbf{f}_{[1:l]}; \mathcal{V}_1, \ldots, \mathcal{V}_l \,\middle|\, \mathbf{Y}_{[1:l]}\right) \\ \overset{(a)}{\leq} l L \epsilon_L + I\left(\mathbf{f}_{[1:l]}; \mathcal{V}_1, \ldots, \mathcal{V}_l \,\middle|\, \mathbf{Y}_{[1:l]}\right) \\ \leq l L \epsilon_L + H\left(\mathcal{V}_1, \ldots, \mathcal{V}_l\right), \quad (51)$$

where $\epsilon_L \geq 0$ is a function of L, independent of P, such that $\epsilon_L \to 0$ as $L \to \infty$; and (a) follows from Fano's inequality. For $l = 0$, we have $\{\mathcal{V}_1, \ldots, \mathcal{V}_l\} = \emptyset$, whereas, for $l = 1$, $\{\mathcal{V}_1, \ldots, \mathcal{V}_l\} = \{\mathbf{v}_2, \ldots, \mathbf{v}_K\}$, and, for $l \geq 2$, $\{\mathcal{V}_1, \ldots, \mathcal{V}_l\} = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$. Hence, $H\left(\mathcal{V}_1, \ldots, \mathcal{V}_l\right) \leq g(l) T_D r_D \log(P)$, where $g(l)$ is defined in (31), and we can further bound $H\left(\mathbf{f}_{[1:l]} \,\middle|\, \mathbf{Y}_{[1:l]}\right)$ as

$$H\left(\mathbf{f}_{[1:l]} \,\middle|\, \mathbf{Y}_{[1:l]}\right) \leq l L \epsilon_L + g(l) T_D r_D \log(P). \quad (52)$$

Next, we bound $H\left(\mathbf{f}_{[1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right)$ in (49) as follows

$$H\left(\mathbf{f}_{[1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right) \\ = H\left(\mathbf{f}_{[1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathbf{Y}_{[l+1:K]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right) \\ \quad + I\left(\mathbf{f}_{[1:K]}; \mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right) \\ \leq H\left(\mathbf{f}_{[1:K]} \,\middle|\, \mathbf{Y}_{[1:K]}\right) \\ \quad + H\left(\mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right) \\ \quad - H\left(\mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[1:N]}\right) \\ \leq KL\epsilon_L + H\left(\mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathbf{X}_{[1:(M-l)]}\right) - H\left(\mathbf{Z}_{[l+1:K]}\right). \\ \hspace{8cm} (53)$$

By applying [11, Lemma 7], we get

$$H\left(\mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathbf{X}_{[1:(M-l)]}\right) \\ = H\left(\mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathbf{X}_{[1:(M-l)]}, \right. \\ \left. \mathbf{Y}_{[l+1:K]} + \tilde{\mathbf{Z}}_{[l+1:K]} - \mathbf{Z}_{[l+1:K]}\right) \\ \leq H\left(\mathbf{Y}_{[l+1:K]} \,\middle|\, \mathbf{Y}_{[l+1:K]} + \tilde{\mathbf{Z}}_{[l+1:K]} - \mathbf{Z}_{[l+1:K]}\right) \\ \leq H\left(\tilde{\mathbf{Z}}_{[l+1:K]} - \mathbf{Z}_{[l+1:K]}\right), \quad (54)$$

where we define $\tilde{\mathbf{Z}}_{[l+1:K]} \triangleq (\mathbf{H}_2 \cdot \mathbf{H}_1^\dagger) \mathbf{Z}_{[1:l]}$ with $\mathbf{H}_1 \triangleq \mathbf{H}_{[1:l]}^{[(M-l)+1:M]}$ and $\mathbf{H}_2 \triangleq \mathbf{H}_{[l+1:K]}^{[(M-l)+1:M]}$. Matrix $\mathbf{H}_1^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{H}_1$. Therefore,

$$H\left(\mathbf{f}_{[1:K]} \,\middle|\, \mathbf{Y}_{[1:l]}, \mathcal{U}_{[1:(M-l)]}, \mathbf{s}_{[1:(M-l)]}, \mathbf{f}_{[K+1:N]}\right) \\ \leq KL\epsilon_L + T_E \log\det\left(\mathbf{I}_{[K-l]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right), \quad (55)$$

where $\tilde{\mathbf{H}} \triangleq \mathbf{H}_2 \cdot \mathbf{H}_1^\dagger$.

Overall, for $l = 0, 1, \ldots, \min\{M, K\}$, it follows from (49), (50), (52), and (55) that

$$K \leq (M-l)(K-l)\mu + l\frac{T_E}{L}\log(\Lambda P + 1) \\ + (M-l)r_F\frac{T_F}{L}\log(P) + (K+l)\epsilon_L \\ + \frac{T_E}{L}\log\det\left(\mathbf{I}_{[K-l]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right) + g(l)\frac{T_D}{L}r_D\log(P). \\ \hspace{8cm} (56)$$

Now, we take the limit $L \to \infty$ and then $P \to \infty$, and arrive at (30b).

Finally, since the Degrees of Freedom (DoF) of (1) are upper bounded by the DoF of the $M \times K$ MIMO point-to-point channel, i.e., $\min\{M, K\}$ [40], then

$$\delta_E = \lim_{P\to\infty} \lim_{L\to\infty} \frac{T_E}{L/\log(P)} \\ \geq \lim_{P\to\infty} \lim_{L\to\infty} \frac{KL/(\min\{M,K\}\log(P))}{L/\log(P)} \\ = \frac{K}{\min\{M,K\}}, \quad (57)$$

i.e., inequality (30c). Inequalities (30d) follows trivially from the definitions of δ_F and δ_D (9).

C. Proof of Proposition 6

Since the achievability was established by Proposition 4, here we prove the converse, i.e., $\delta^*(\mu, r_F, r_D) \geq \delta_{2\times2}(\mu, r_F, r_D)$. For $M = K = 2$, the constraints (30b), (30c), in Proposition 5 can be written as:

$$\delta_E + r_F\delta_F + r_D\delta_D \geq 2 - \mu, \quad (58)$$

$$\delta_F \geq \frac{1 - 2\mu}{r_F}, \quad (59)$$

$$\delta_E \geq 1. \quad (60)$$

For $0 \leq r_F, r_D \leq 1$, using (58) gives the lower bound $\delta^*(\mu, r_F, r_D) \geq 2 - \mu$. Furthermore, using $(1 - r_F) \times (59) + (58)$ gives the lower bound $\delta^*(\mu, r_F, r_D) \geq 1 + \mu + (1 - 2\mu)/r_F$.

For $r_F \geq \max\{1, r_D\}$, using $[(r_F - 1) \times (60) + (58)]/r_F$ gives the lower bound $\delta^*(\mu, r_F, r_D) \geq 1 + (1 - \mu)/r_F$.

For $r_D > \max\{1, r_F\}$, using $[(r_D - 1) \times (60) + (58)]/r_D$ gives the lower bound $\delta^*(\mu, r_F, r_D) \geq 1 + (1 - \mu)/r_D$. Moreover, using $[(58) + (r_D - r_F) \times (59) + (r_D - 1) \times (60)]/r_D$ gives the lower bound $\delta^*(\mu, r_F, r_D) \geq 1 + \mu/r_D + (1 - 2\mu)/r_F$.

D. Proof of Proposition 7

We prove Proposition 7 by showing that the ratio (32) holds in each of the five regimes described in Proposition 3. First, note that, due to (30c)-(30d), we have

$$\delta^*(\mu, r_F, r_D) \geq \frac{K}{\min\{M, K\}}. \tag{61}$$

Another lower bound on the minimum NDT follows from (30b)-(30d) (with $l = 0$) as

$$\delta^*(\mu, r_F, r_D) \geq \frac{K}{\min\{M, K\}} + \frac{K(1 - M\mu)}{Mr_F}. \tag{62}$$

1) High Cache and High D2D ($\mu > 1/M$ and $r_D > r_D^{th}$): Since $\mu > 1/M$, the achievable NDT (28) satisfies

$$\delta_{\mathrm{ach}}(\mu, r_F, r_D) \leq \frac{K}{\min\{M, K\}}\left(1 + \frac{1}{r_D}\right). \tag{63}$$

Dividing (63) by (61) gives

$$\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \leq 1 + \frac{1}{r_D}. \tag{64}$$

Next, since $r_D > r_D^{th}$, where r_D^{th} is the threshold defined in (23), we have

$$r_D > r_D^{th} \geq \frac{\max\{M, K\}}{\min\{M, K\} - 1} > 1. \tag{65}$$

Thus,

$$\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} < 2. \tag{66}$$

2) Low Cache and High D2D ($\mu \leq 1/M$ and $r_D > r_D^{th}$): Dividing (27) by (62) gives

$$
\begin{aligned}
&\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \\
&\leq 1 + \frac{\max\{M, K\}\mu/r_D}{K/\min\{M, K\} + K(1 - \mu M)/(Mr_F)} \\
&\overset{(a)}{\leq} 1 + \frac{1}{r_D} \\
&\overset{(b)}{<} 2,
\end{aligned} \tag{67}
$$

where (a) follows from $\mu \leq 1/M$, and (b) is due to $r_D > 1$ (65).

3) High Cache, Low Fronthaul, and Low D2D ($\mu > 1/M$, $r_F \leq r_F^{th}$, and $r_D \leq r_D^{th}$): In this regime, we have

$$
\begin{aligned}
&\delta_{\mathrm{ach}}(\mu, r_F, r_D) \\
&= \frac{K}{\min\{M, K\}}\left(\frac{\mu M - 1}{M - 1}\right) + (1 - \mu)\frac{M + K - 1}{M - 1} \\
&\leq \frac{M - 1}{M} \cdot \frac{M + K - 1}{M - 1} \\
&= \frac{M + K - 1}{M},
\end{aligned} \tag{68}
$$

where the inequality follows from $\delta_{\mathrm{ach}}(\mu, r_F, r_D)$ being a monotonically decreasing function of μ and since $\mu \geq 1/M$. Dividing (68) by (61) gives

$$
\begin{aligned}
\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} &\leq \frac{M + K - 1}{M} \cdot \frac{\min\{M, K\}}{K} \\
&= 1 + \frac{\min\{M, K\} - 1}{\max\{M, K\}} \\
&< 2.
\end{aligned} \tag{69}
$$

4) High Fronthaul and Low D2D ($r_F > r_F^{th}$ and $r_D \leq r_D^{th}$): In this regime, we have

$$
\begin{aligned}
\delta_{\mathrm{ach}}(\mu, r_F, r_D) &= \frac{K}{\min\{M, K\}} + \frac{(1 - \mu)K}{Mr_F} \\
&\overset{(a)}{\leq} \frac{K}{\min\{M, K\}} + \frac{K}{Mr_F} \\
&\overset{(b)}{<} \frac{K}{\min\{M, K\}} + \frac{\min\{M, K\} - 1}{M - 1},
\end{aligned} \tag{70}
$$

where (a) follows from $\mu \geq 0$, and (b) follows from $r_F > r_F^{th}$. Dividing (70) by (61) gives

$$\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} < 1 + \frac{\min\{M, K\} - 1}{M - 1} \cdot \frac{\min\{M, K\}}{K} \leq 2. \tag{71}$$

5) Low Cache, Low Fronthaul, and Low D2D ($\mu \leq 1/M$, $r_F \leq r_F^{th}$, and $r_D \leq r_D^{th}$): We first consider the case of $K \leq M$. Dividing (24) by (62) gives

$$
\begin{aligned}
\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} &\leq 1 + \frac{\mu(K - 1)}{1 + K(1 - M\mu)/(Mr_F)} \\
&\overset{(a)}{\leq} 1 + \frac{K - 1}{M} \\
&< 2,
\end{aligned} \tag{72}
$$

where (a) follows from $\mu \leq 1/M$.

Next, for $M < K$ and $r_F \geq 1$, the achievable NDT (24) satisfies

$$\delta_{\mathrm{ach}}(\mu, r_F, r_D) \overset{(a)}{\leq} \frac{K}{M} + \frac{K}{Mr_F} \overset{(b)}{\leq} \frac{2K}{M}, \tag{73}$$

where (a) follows from $\delta_{\mathrm{ach}}(\mu, r_F, r_D)$ being a monotonically decreasing function of μ and since $\mu > 0$, whereas (b) is due to $r_F \geq 1$. Dividing (73) by (61) gives

$$\frac{\delta_{\mathrm{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)} \leq \frac{2K}{M} \cdot \frac{M}{K} = 2. \tag{74}$$

Finally, we consider the case of $M < K$ and $r_F < 1$. Let the integer l^* be defined as $l^* \triangleq \lceil M/2 \rceil$. Note that since we

consider a case with $r_F < 1$, then we have the following inequality

$$(M - l^*)r_F \leq l^*. \tag{75}$$

We further divide the case of $M < K$ and $r_F < 1$ into two regimes: $r_D \leq l^*/g(l^*)$ and $l^*/g(l^*) < r_D \leq r_D^{\text{th}}$. For $r_D \leq l^*/g(l^*)$, it follows from (30b) with $l = l^*$ that

$$l^*\delta_E + (M - l^*)r_F\delta_F + l^*\delta_D \geq K - (M - l^*)(K - l^*)\mu. \tag{76}$$

Furthermore, we have

$$[l^* - (M - l^*)r_F]\delta_F \geq [l^* - (M - l^*)r_F] \cdot \frac{K(1 - M\mu)}{Mr_F} \tag{77}$$

due to (30b) (with $l = 0$) and (75). By adding (76) and (77), and dividing by l^*, we get the following lower bound on the minimum NDT

$$\delta^*(\mu, r_F, r_D) \geq \frac{K(1 - M\mu)}{M}\left(1 + \frac{1}{r_F}\right) + (M + K - l^*)\mu. \tag{78}$$

Dividing (24) by (78) gives

$$\frac{\delta_{\text{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)}$$
$$\leq 1 + \frac{(l^* - 1)\mu}{K(1 - M\mu)\cdot(1 + 1/r_F)/M + (M + K - l^*)\mu}$$
$$\leq 1 + \frac{l^* - 1}{M + K - l^*}$$
$$\overset{(a)}{<} 1 + \frac{M/2}{M/2 + K - 1}$$
$$< 2, \tag{79}$$

where (a) follows from $l^* < M/2 + 1$.

Now, for $l^*/g(l^*) < r_D \leq r_D^{\text{th}}$, we have $g(l^*)r_D > l^* \geq (M - l^*)r_F$. Thus, (30b) (for $l = 0$) and (30c) imply

$$(g(l^*)r_D - (M - l^*)r_F)\delta_F$$
$$\geq (g(l^*)r_D - (M - l^*)r_F) \cdot \frac{K(1 - M\mu)}{Mr_F} \tag{80}$$

and

$$(g(l^*)r_D - l^*)\delta_E \geq (g(l^*)r_D - l^*) \cdot \frac{K}{M}, \tag{81}$$

respectively. By adding (80) and (81) to (30b) (with $l = l^*$), and dividing by $g(l^*)r_D$, we get the following lower bound on the minimum NDT

$$\delta^*(\mu, r_F, r_D)$$
$$\geq \frac{K(1 - M\mu)}{M}\left(1 + \frac{1}{r_F}\right) + K\mu + \frac{(M - l^*)l^*\mu}{g(l^*)r_D}$$
$$\overset{(a)}{\geq} \frac{K(1 - M\mu)}{M}\left(1 + \frac{1}{r_F}\right) + K\mu + \frac{(M - l^*)(M - 1)l^*\mu}{g(l^*)K}, \tag{82}$$

where (a) follows from $r_D \leq r_D^{\text{th}} = K/(M-1)$. Dividing (24) by (82) gives

$$\frac{\delta_{\text{ach}}(\mu, r_F, r_D)}{\delta^*(\mu, r_F, r_D)}$$
$$\leq 1 + \frac{(M - 1)\mu[1 - (M - l^*)l^*/(g(l^*)K)]}{K\mu + \frac{K(1 - M\mu)\cdot(1 + 1/r_F)}{M} + \frac{(M - l^*)(M - 1)l^*\mu}{g(l^*)K}}$$
$$\leq 1 + \frac{(M - 1)\mu}{K\mu}$$
$$< 2. \tag{83}$$

E. Proof of Proposition 8

For high fronthaul rate, $r_F \geq \min\{M, K\}/M$, we apply block-Markov encoding with cloud-aided soft-transfer [11, Proposition 3]; the resulting NDT is

$$\delta_{\text{P}}(\mu, r_F, r_D) = \max\left\{\frac{K}{Mr_F}, \frac{K}{\min\{M, K\}}, 0\right\}$$
$$= \frac{K}{\min\{M, K\}}. \tag{84}$$

Note that, in this regime, no caching and D2D resources are required.

Next, we consider low fronthaul rate, i.e., $r_F < \min\{M, K\}/M$. For $\mu \in [0, \mu_1]$, where μ_1 is defined in (39), no D2D communication is utilized. As in [11, Proposition 9], we apply the following per-block file-splitting with block-Markov encoding: Part $(1 - M\mu)$ of each requested file is delivered using cloud-aided soft-transfer [11, Proposition 3]; and part $M\mu$ of each requested file is delivered using cache-aided EN coordination [11, Lemma 3]. The cache capacity constraint is satisfied since $\mu \leq \mu_1 \leq 1/M$. This achieves the NDT

$$\delta_{\text{P}}(\mu, r_F, r_D)$$
$$= \max\left\{\frac{(1 - M\mu)K}{Mr_F}, \frac{(1 - M\mu)K}{\min\{M, K\}} + \frac{M\mu(M + K - 1)}{M}, 0\right\}$$
$$= \frac{(1 - M\mu)K}{Mr_F}, \tag{85}$$

where the last equality follows from $\mu \leq \mu_1$.

For $\mu \in [\mu_2, 1]$, where μ_2 is defined in (40), we apply the following per-block file-splitting with block-Markov encoding: Part $\alpha_1 \triangleq \min\{Mr_F/\min\{M, K\}, 1\}$ of each requested file is delivered using cloud-aided soft-transfer; part $\alpha_2 \triangleq \min\{r_D, 1 - \alpha_1\}$ of each requested file is delivered using D2D-based compress-and-forward (Proposition 2); and part $(1 - \alpha_1 - \alpha_2)$ of each requested file is delivered using cache-aided ZF [11, Lemma 2]. The cache capacity constraint is satisfied since $1 - \alpha_1 - \alpha_2 + \alpha_2/M = \mu_2 \leq \mu$. This achieves the NDT

$$\delta_{\text{P}}(\mu, r_F, r_D)$$
$$= \max\left\{\frac{\alpha_1 K}{Mr_F}, \frac{K}{\min\{M, K\}}, \frac{\alpha_2 K}{r_D \min\{M, K\}}\right\}$$
$$= \frac{K}{\min\{M, K\}}. \tag{86}$$

Finally, for $\mu \in [\mu_1, \mu_2]$, we apply file-splitting and cache-sharing [11, Lemma 1] between the policies for the corner points $\mu = \mu_1$ and $\mu = \mu_2$. This achieves the NDT

$$
\begin{aligned}
&\delta_{\mathrm{P}}(\mu, r_F, r_D) \\
&= \frac{\mu_2 - \mu}{\mu_2 - \mu_1} \cdot \frac{(1 - M\mu_1)K}{Mr_F} + \frac{\mu - \mu_1}{\mu_2 - \mu_1} \cdot \frac{K}{\min\{M, K\}}. \quad (87)
\end{aligned}
$$

F. Proof of Proposition 9

In [11, App. VIII-C], it was proved that, without D2D communication,

$$
\frac{\delta_{\mathrm{P,ach}}(\mu, r_F, r_D = 0)}{K/\min\{M, K\}} \leq 2, \quad \forall \mu \in [\mu_1, \tilde{\mu}_2], \quad (88)
$$

where $\tilde{\mu}_2 \triangleq 1 - Mr_F/\min\{M, K\}$. Thus, for all $\mu \in [\mu_1, \mu_2]$, the ratio between the achievable NDT and the minimum NDT under pipelined delivery is upper bounded as

$$
\begin{aligned}
\frac{\delta_{\mathrm{P,ach}}(\mu, r_F, r_D)}{\delta_{\mathrm{P}}^*(\mu, r_F, r_D)} &\overset{(a)}{\leq} \frac{\delta_{\mathrm{P,ach}}(\mu, r_F, r_D)}{K/\min\{M, K\}} \\
&\overset{(b)}{\leq} \frac{\delta_{\mathrm{P,ach}}(\mu, r_F, r_D = 0)}{K/\min\{M, K\}} \\
&\overset{(c)}{\leq} 2, \quad (89)
\end{aligned}
$$

where (a) follows from Corollary 2; (b) holds since D2D cooperation does not increase the achievable NDT of Proposition 8; and (c) follows from (88) and since $\tilde{\mu}_2 \geq \mu_2$.

G. Proof of Proposition 10

The lower bound of Corollary 2 can be relaxed by considering only $l = 0$, i.e.,

$$
\delta_{\mathrm{P}}^*(\mu, r_F, r_D) \geq \max\left\{ \frac{(1 - M\mu)K}{Mr_F}, \frac{K}{\min\{M, K\}} \right\}. \quad (90)
$$

For high fronthaul rate and for low fronthaul rate with fractional cache capacity μ that satisfies $\mu \in [0, \mu_1]$ or $\mu \in [\mu_2, 1]$, the lower bound (90) coincides with the achievable NDT of Proposition 8 ((37) and (38)), hence the minimum NDT is given by (42) and (44), respectively. Next, for low fronthaul rate $r_F < \min\{M, K\}/M$ and high D2D rate $r_D \geq 1 - Mr_F/\min\{M, K\}$, the strategy of Proposition 8 achieves an NDT of $\delta_{\mathrm{P,ach}}(\mu, r_F, r_D) = K/\min\{M, K\}$ for all $\mu \geq \mu_2 = 1/M - r_F/\min\{M, K\}$; and an NDT of $\delta_{\mathrm{P,ach}}(\mu, r_F, r_D) = K/(Mr_F)$ for cloud-only F-RAN, i.e., for $\mu = 0$ (see (38)). For $\mu \in (0, \mu_2)$ we apply file-splitting and cache-sharing between the policies for $\mu = 0$ and $\mu = \mu_2$. This achieves the NDT (43), which equals the lower bound (90), and hence optimal.

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, 3rd Quart., 2018.

[2] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098–3130, 4th Quart., 2018.

[3] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[4] 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, Aug. 2016.

[5] 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.

[6] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003–5009, 2016.

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

[8] 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, 4th Quart., 2014.

[9] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 90–96, Feb. 2014.

[10] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2133–2168, 3rd Quart., 2018.

[11] 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.

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

[13] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5359–5380, Jul. 2018.

[14] 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.

[15] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2016, pp. 320–325.

[16] 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.

[17] S.-H. Park, O. Simeone, W. Lee, and S. Shamai (Shitz), "Coded multicast fronthauling and edge caching for multi-connectivity transmission in fog radio access networks," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2017, pp. 1–5.

[18] J. Kakar, A. Chaaban, A. Sezgin, and A. Paulraj, "Delivery time minimization in cache-assisted broadcast–relay wireless networks with imperfect CSI," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

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

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

[21] K. Wan, D. Tuninetti, M. Ji, and G. Caire, "On the fundamental limits of fog-RAN cache-aided networks with downlink and sidelink communications," Nov. 2018, *arXiv:1811.05498*. [Online]. Available: https://arxiv.org/abs/1811.05498

[22] J. Zhang and O. Simeone, "Fundamental limits of cloud and cache-aided interference management with multi-antenna edge nodes," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5197–5214, Aug. 2019.

[23] S. P. Shariatpanahi, J. Zhang, O. Simeone, B. H. Khalaj, and M.-A. Maddah-Ali, "Cloud-aided interference management with cache-enabled edge nodes and users," Jan. 2019, *arXiv:1901.06698*. [Online]. Available: https://arxiv.org/abs/1901.06698

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

[25] R. Karasik, O. Simeone, and S. Shamai (Shitz), "Fundamental latency limits for D2D-aided content delivery in fog wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 2461–2465.

[26] R. Karasik, O. Simeone, and S. Shamai (Shitz), "Information-theoretic analysis of D2D-aided pipelined content delivery in fog-RAN," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Lisbon, Portugal, Aug. 2018, pp. 1–5.

[27] R. Karasik, O. Simeone, and S. Shamai (Shitz), "Latency limits for content delivery in a fog-RAN with D2D communication," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Aug. 2018, pp. 1–5.

[28] 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, Mar. 2017.

[29] Y. Lan, X. Wang, D. Wang, Z. Liu, and Y. Zhang, "Task caching, offloading, and resource allocation in D2D-aided fog computing networks," *IEEE Access*, vol. 7, pp. 104876–104891, 2019.

[30] 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, Sep. 2009.

[31] 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.

[32] B. Kananian, M. A. Maddah-Ali, and B. Hossein Khalaj, "K-User interference channels with backhaul cooperation: DoF vs. backhaul load trade-Off," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7253–7267, Nov. 2019.

[33] 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.

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

[35] A. G. Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling conjectures on the collapse of degrees of freedom under finite precision CSIT," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5603–5618, Oct. 2016.

[36] M. Zeyde, O. Simeone, and S. Shamai (Shitz), "Confidential communication in C-RAN systems with infrastructure sharing," in *Proc. IEEE Int. Conf. Sci. Electr. Eng. Israel (ICSEE)*, Dec. 2018, pp. 1–5.

[37] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM–based cloudlets in mobile computing," *IEEE Pervasive Comput.*, vol. 8, no. 4, pp. 14–23, Oct. 2009.

[38] Y.-Y. Shih, W.-H. Chung, A.-C. Pang, T.-C. Chiu, and H.-Y. Wei, "Enabling low–latency applications in fog–radio access networks," *IEEE Netw.*, vol. 31, no. 1, pp. 52–58, Jan. 2017.

[39] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi–access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[40] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecomm.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

Roy Karasik (Student Member, IEEE) received the B.Sc. *(summa cum laude)* and M.Sc. (Hons.) degrees in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2011 and 2014, respectively, where he is currently pursuing the Ph.D. degree. From 2013 to 2017, he worked as an Engineer in the Israeli Defense Forces. His research interests include information theory and wireless communications.

Osvaldo Simeone (Fellow, IEEE) received the M.Sc. (Hons.) and Ph.D. degrees in information engineering from the Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively. From 2006 to 2017, he was a Faculty Member with the Electrical and Computer Engineering (ECE) Department, New Jersey Institute of Technology (NJIT), where he was with the Center for Wireless Information Processing (CWiP). He is currently a Professor of information engineering with the Centre for Telecommunications Research, Department of Engineering, King's College London, where he directs the King's Communications, Learning, and Information Processing (KCLIP) Laboratory. He is a coauthor of two monographs, an edited book published by Cambridge University Press, and more than 100 research journal articles. His research has been supported by U.S. NSF, ERC, Vienna Science and Technology Fund, and by a number of industrial collaborations. His research interests include wireless communications, information theory, optimization, and machine learning. He is a Fellow of the IET. He was a co-recipient of the 2019 IEEE Communication Society Best Tutorial Paper Award, the 2018 IEEE Signal Processing Best Paper Award, the 2017 JCN Best Paper Award, the 2015 IEEE Communication Society Best Tutorial Paper Award, and of the Best Paper Awards of IEEE SPAWC 2007 and IEEE WRECOM 2007. He was awarded a Consolidator Grant by the European Research Council (ERC) in 2016. He currently serves the Editorial Board for the IEEE Signal Processing Magazine, and he is a Distinguished Lecturer of the IEEE Information Theory Society.

Shlomo Shamai (Shitz) (Life Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Technion–Israel Institute of Technology, in 1975, 1981 and 1986 respectively. During 1975–1985 he was with the Communications Research Labs, in the capacity of a Senior Research Engineer. Since 1986 he is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, where he is now a Technion Distinguished Professor, and holds the William Fondiller Chair of Telecommunications. His research interests encompasses a wide spectrum of topics in information theory and statistical communications.

Dr. Shamai (Shitz) is an URSI Fellow, a member of the Israeli Academy of Sciences and Humanities and a foreign member of the US National Academy of Engineering. He is the recipient of the 2011 Claude E. Shannon Award, the 2014 Rothschild Prize in Mathematics/Computer Sciences and Engineering and the 2017 IEEE Richard W. Hamming Medal. He is a co-recipient of the 2018 Third Bell Labs Prize for Shaping the Future of Information and Communications Technology. He has been awarded the 1999 van der Pol Gold Medal of the Union Radio Scientifique Internationale (URSI), and is a co-recipient of the 2000 IEEE Donald G. Fink Prize Paper Award, the 2003, and the 2004 joint IT/COM societies paper award, the 2007 IEEE Information Theory Society Paper Award, the 2009 and 2015 European Commission FP7, Network of Excellence in Wireless COMmunications (NEWCOM++, NEWCOM#) Best Paper Awards, the 2010 Thomson Reuters Award for International Excellence in Scientific Research, the 2014 EURASIP Best Paper Award (for the EURASIP Journal on Wireless Communications and Networking), the 2015 IEEE Communications Society Best Tutorial Paper Award and the 2018 IEEE Signal Processing Best Paper Award. Dr. Shamai (Shitz) is listed as a Highly Cited Researcher (Computer Science) for the years 2013/4/5/6/7/8. He is also the recipient of 1985 Alon Grant for distinguished young scientists and the 2000 Technion Henry Taub Prize for Excellence in Research. He has served as Associate Editor for the Shannon Theory of the IEEE Transactions on Information Theory, and has also served twice on the Board of Governors of the Information Theory Society. He has also served on the Executive Editorial Board of the IEEE Transactions on Information Theory and on the IEEE Information Theory Society Nominations and Appointments Committee, and on the IEEE Information Theory Society, Shannon Award Committee.