Cache-Assisted Broadcast-Relay Wireless Networks: A Delivery-Time Cache-Memory Tradeoff

Jaber Kakar, Member, IEEE, Alaa Alameer, Member, IEEE, Anas Chaaban, Senior Member, IEEE, Aydin Sezgin, Senior Member, IEEE, and Arogyaswami Paulraj, Fellow, IEEE

Abstract

An emerging trend of next generation communication systems is to provide network edges with additional capabilities such as storage resources in the form of caches to reduce file delivery latency. To investigate this aspect, we study the fundamental limits of a cache-aided broadcast-relay wireless network consisting of one central base station, $M$ cache-equipped transceivers and $K$ receivers from a latency-centric perspective. We use the normalized delivery time (NDT) to capture the per-bit latency for the worst-case file request pattern at high signal-to-noise ratios (SNR), normalized with respect to a reference interference-free system with unlimited transceiver cache capabilities. The objective is to design the schemes for cache placement and file delivery in order to minimize the NDT. To this end, we establish a novel converse (for arbitrary $M$ and $K$) and two types of achievability schemes applicable to both time-variant and invariant channels. The first scheme is a general one-shot scheme for any $M$ and $K$ that synergistically exploits both multicasting (coded) caching and distributed zero-forcing opportunities. Apart from the obvious advantage of low signaling complexity, we show that the proposed one-shot scheme (i) attains gains attributed to both individual and collective transceiver caches (ii) is NDT-optimal for various parameter settings, particularly at higher cache sizes. The second scheme, on the other hand, designs beamformers to facilitate both subspace interference alignment and zero-forcing at lower cache sizes. Exploiting both schemes, we are able to characterize for various special cases of $M$ and $K$ which satisfy $K + M \leq 4$ the optimal tradeoff between cache storage and latency. The tradeoff illustrates that the NDT is the preferred choice to capture the latency of a system rather than the commonly used sum degrees-of-freedom (DoF). In fact, our optimal tradeoff refutes the popular belief that increasing cache sizes translates to increasing the achievable sum DoF. As such, we identify and discuss cases where increasing cache sizes decreases both the delivery time and the achievable DoF.
Index Terms

Caching, interference alignment, degrees-of-freedom, latency, delivery time.

I. INTRODUCTION

In the last decade, mobile usage in wireless networks has shifted from being connection-centric driven (e.g., phone calls) to content-centric (e.g., HD video) behaviors [1]. In this context, integrating content caching in heterogeneous networks (HetNet) represents a viable solution for highly content-centric next generation (5G) mobile networks. Specifically, caching the most popular contents in HetNet edge nodes, e.g., eNBs and relays, alleviates backhaul traffic, reduces latency and ameliorates quality of service of mobile users. For example, edge caches were used to balance backhaul costs against transmission power costs which results in optimal sparse beamforming solutions [2]. Further, in [3] caching was deployed to minimize the weighted average latency subject to proportional fairness and ergodic resource allocation constraints. In these previous works, the cache was placed only at the base stations. However, it is expected that future networks will be heterogeneous in nature, vastly deploying relay nodes (RN) (e.g., fixed RNs in LTE-A [4] or mobile RNs in form of drones [5], [6]) endowed with content cache capabilities. In this work, we assume that RNs not only provide files to mobile users but also have their own requests from the central base station in a HetNet scenario.

A simplistic HetNet modeling this aspect is shown in Fig. 1. In this model, $M$ RNs act as cache-aided transceivers. Thus, aspects of both transmitter and receiver caching in RNs are captured through this network model enabling a low delivery time of requested files by $M$ RNs and $K$ user equipments (UE). Delivery time refers to the timing overhead required to satisfy all file demands of requesting nodes in the network. Such type of model is of importance from an online cache update perspective in which RNs refresh their cached contents while simultaneously satisfying the UEs file demands in collaboration with the donor eNB (DeNB). In this work, we are interested in studying the fundamental delivery-time cache-memory tradeoff of this particular network.

In prior work, it was shown that both receiver (Rx) and transmitter (Tx) caching can offer significant latency reduction. Rx caching was first studied in [7] for a shared error-free broadcast channel with one server and multiple cache-enabled receivers. The authors show that their coded

\[^1\text{We use the words \textit{delivery time} and \textit{latency} interchangeably.}\]
Fig. 1: A transceiver cache-aided HetNet consisting of one DeNB, $M$ RNs and $K$ UEs. These nodes are connected through the wireless links $f_i, g_k$ and $h_{ij}, i = 1, \ldots, M, j = 1, \ldots, K$. Each RN is equipped with a finite size cache.

caching approach of popular content exploits multicast opportunities and consequently reduces latency. Coded caching has received considerable attention for various related settings of [7]. This includes the rate-memory tradeoff under uncoded cache placement [8], [9], decentralized caching under homogeneous [10] and heterogeneous cache sizes [11], caching with distinct file sizes [12], online caching [13], caching with non-uniform demands [14] and multiple requests [15] amongst others. Further, coded caching concepts have been applied to device-to-device networks with [16] and without secure delivery [17], multi-server networks [18] and hierarchical networks [19].

On the other hand, the impact of Tx caching on latency has mainly been investigated by analyzing the inverse degrees-of-freedom (DoF) metric of Gaussian interference networks [20]. To this end, the authors of [21] developed an interference alignment scheme characterizing the inverse DoF as a function of the cache storage size for a 3-user Gaussian interference network. The caches are prefetched to allow transmitter cooperation so that interference coordination techniques are applicable. In [22], the authors introduce the normalized delivery time (NDT) as a performance metric which is proportional to the inverse DoF. The first lower bounds on the NDT as a delivery time metric were derived in the same paper for a cache-aided interference channel
with an arbitrary number of edge nodes and users. With these bounds, the optimality of schemes presented in [21] for certain regimes of cache sizes was shown under uncoded prefetching of the cached content. These concepts have been recently applied to Fog radio access networks (F-RAN) that consist of a centralized cloud server, cache-assisted edge nodes and mobile users. The NDT of F-RANs has been first fully characterized for two edge nodes and two mobile users [23]. Later on for the setting of arbitrary number of edge nodes and receivers a constant factor characterization of 2 has been established in [24]. The effect of channel strength and fading on the delivery time of partially connected F-RANs has been investigated in [25], [26], [27] on the basis of the binary fading model [28] and the linear deterministic model [29].

Recently, the effect of Tx-Rx caching on the delivery time of interference networks was studied in two new lines of research. The first being, where distinct nodes, i.e., transmitters and receivers are equipped with caches. For this setting, the authors in [30] establish one-shot linear delivery schemes, which avoid channel extension, and show their optimality within a factor of 2 of the lower bounds on the delivery time. In [31], an architecture that separates physical and network layers is proposed and shown to be approximately DoF-optimal for Tx-Rx cache enabled interference networks. In [32], achievability schemes on cooperative X-multicast channels are used to show the multiplicative NDT-optimality of interference channels with caches at both ends of the network. The second line of research focuses on the latency-influence of Tx-Rx caching at a single node, i.e., transceiver caching. Such type of caching is of importance from an online cache update perspective. The first paper studying transceiver caching from a channel strength point of view is [33]. However, in this paper the authors only characterize a single RN and UE setting. Our paper examines the second line of research but as opposed to [33] with an arbitrary number of RNs and UEs. This examination is of interest to understand the compatibility of multicasting (observable in Rx-caching) and interference coordination techniques (observable in Tx-caching) in cache-aided transceiver networks.

In this paper, we study the fundamental limits on the delivery time for a transceiver cache-aided HetNet consisting of one donor eNB (DeNB), $M$ RNs and $K$ UEs. The RNs are equipped with a cache memory of $\mu N_L$ bits with $N_L$ being the entire library size composed of $N$ files with file size $L$. We measure the performance through the latency-centric metric normalized delivery time per bit (NDT) (cf. formal definition of NDT in Eq. (5) in Section II). This metric, first introduced in [22], indicates the worst-case per-bit latency incurred in the wireless network with respect to a reference interference-free system without cache capacity restrictions in the
high signal-to-noise ratio (SNR) regime. The main contributions of this paper are as follows:

- In Section IV, we develop a novel class of information theoretic lower bounds on the NDT under the assumption of perfect channel state information (CSI) and uncoded prefetching of the cached content.
- We show that the optimal schemes for the extreme cases of no caching and full caching are DeNB broadcasting and joint DeNB-RN zero-forcing beamforming, respectively.
- In Section V, we propose a generalized one-shot scheme that integrates multicasting schemes used in Rx caching with ZF beamforming typically deployed in the context of Tx caching. Further, we evaluate its performance with respect to the lower bounds and show its effectiveness at higher cache capacities. Despite the low complexity of the one-shot scheme, we identify regimes as a function \((\mu, K, M)\) for which it is indeed NDT-optimal. Moreover, we show that caching more than \(\frac{[M-1/2]}{M}\) fractions of a file attains (at most) a multiplicative gap of \(\frac{8}{3}\) with respect to the optimal NDT.
- In another scheme, we design precoders that synergistically interlace subspace interference alignment and zero-forcing. This design exploits spatially correlated file fractions through balancing zero-forcing opportunities and alignment opportunities. The scheme is NDT-optimal at lower cache sizes for both time-variant and invariant channels requiring finite signal dimensions (e.g., time).
- With the existence of both schemes, we are able to completely characterize the latency-memory tradeoff for the settings of (a) \(M = 1\) RNs and \(K \in \{1, 2, 3\}\) UEs, (b) \(M = 2\) RNs and \(K \in \{1, 2\}\) UEs and (c) \(M = 3\) RNs and \(K = 1\) UEs.
- Along with our results, we discuss the relationship between achievable (sum) DoF and NDT. To this end, we assess the results from both a rate (e.g., DoF), and latency (e.g., NDT) perspective. In particular, our optimal latency-memory tradeoff for \(K + M \leq 4\) refutes the popular belief that increasing cache sizes translates to increasing the achievable sum DoF. In fact, there are cases where an increase in the cache size decreases the delivery time but also the achievable DoF.

**Notation:** For any two integers \(a\) and \(b\) with \(a \leq b\), we define \([a : b] \triangleq \{a, a+1,\ldots,b\}\) and we denote \([1,b]\) simply as \([b]\). We use \(a_{t_1}^{t_2}\) and \(A_{t_1}^{t_2}\) with \(t_1 \leq t_2\) to refer to a vector \(a[t]\) and a matrix \(A[t]\) concatenated across \(t \in [t_1 : t_2]\). When \(t_1 = 1\), we simply write \(a^t\) and \(A^t\), respectively. The superscript \((\cdot)^\dagger\) represents the transpose of a matrix. We use \(\otimes\) to denote the
II. SYSTEM MODEL AND LATENCY METRIC

In this section, we first outline the system model of the cache-assisted broadcast-relay wireless network illustrated in Fig. 1. Then, we introduce the normalized delivery time per bit (NDT) metric, along with its operational meaning to provide additional context on the adopted model and performance metric.

The network under study consists of $M$ causal full-duplex RNs and a donor eNB (DeNB) which serves $K$ UEs with its desired content over a shared wireless channel. Simultaneously, each RN also requests information from the DeNB. At every transmission interval, we assume that RNs and UEs request one file each from the set $\mathcal{W}$ of $N$ popular files, whose elements are all of $L$ bits in size. The transmission interval terminates when the requested files have been delivered. The system model, notation and main assumptions for a single transmission interval are summarized as follows:

- Let $\mathcal{W} = \{W_1, \ldots, W_N\}$ denote the library of popular files, where each file $W_n$ is of size $L$ bits. Each file $W_n$ is chosen uniformly at random from $[2^L]$, where $N = 2^L$. UEs and RNs request files $W_{d_u}$, $\forall u \in [K]$, and $W_{d_r}$, $\forall r \in [K + 1 : K + M]$, from the library $\mathcal{W}$, respectively. The demand vector $d = (d_1, \ldots, d_{K+M}) \in [N]^{K+M}$ denotes the request pattern of RNs and UEs. This vector is shared among all nodes.
- The RNs are endowed with a cache capable of storing $\mu N L$ bits, where $\mu \in [0, 1]$ corresponds to the fractional cache size. It denotes how much content can be stored at each RN relative to the size of the entire library $\mathcal{W}$.
- The DeNB has access to all $N$ popular files of $\mathcal{W}$.
- Global CSI for the single-antenna setting at time instant $t$ is summarized by the channel vectors $f[t] = \{f_{m}[t]\}_{m=1}^{M} \in \mathbb{C}^{M}$ and $g[t] = \{g_{k}[t]\}_{k=1}^{K} \in \mathbb{C}^{K}$ and the channel matrix $H[t] = \{h_{km}[t]\}_{k=1,m=1}^{K \times M} \in \mathbb{C}^{K \times M}$. Here, $f_{m}$ and $g_{k}$ represent the complex channel coefficients from DeNB to RN$_{m}$ and UE$_{k}$, respectively, while $h_{km}$ is the channel from RN$_{m}$ to UE$_{k}$. We assume that all channel coefficients are drawn i.i.d. from a continuous random distribution.

Communication over the wireless channel occurs in two consecutive phases, (a) placement phase followed by (b) delivery phase. These are detailed next, along with the key performance metric termed as normalized delivery time per bit (NDT).
**a) Placement phase:** During this phase, each RN caches content from the library $\mathcal{W}$ by requesting this content from the DeNB using the caching function defined next.

*Definition 1.* (Caching function) $\text{RN}_m$, $\forall m = 1, \ldots, M$, maps each file $W_n \in \mathcal{W}$ to its local *file cache content* as

$$S_{m,n} = \phi_{m,n}(W_n), \quad \forall n = 1, \ldots, N,$$

where $\phi_{m,n}(\cdot)$ is the caching function. All $S_{m,n}$ are concatenated to form the total cache content

$$S_m = (S_{m,1}, S_{m,2}, \ldots, S_{m,N})$$

at $\text{RN}_m$.

Hereby, we assume symmetry in caching, i.e., each file $W_n, \forall n \in [N]$, is cached with at most $\mu L$ number of bits.\(^2\) In consequence, the entropy $H(S_{m,n})$ of each component $S_{m,n}, n = 1, \ldots, N$, is upper bounded by $\mu^{NL/N} = \mu L$. The definition of the caching function presumes that every file $W_i$ is subjected to individual caching functions. Thus, permissible caching policies allow for intra-file coding but avoid coding across files known as inter-file coding. Moreover, the caching policy is typically kept fixed over long transmission intervals. Thus, it is indifferent to the UEs request pattern and of channel realizations.

**b) Delivery phase:** In this phase, a transmission policy at DeNB and all RNs is applied to satisfy the given requests $d$ under the current channel realizations $f, g$ and $H$. Throughout the remaining definitions, we denote the number of channel uses required to satisfy all file demands by $T$. This time depends on the demand vector $d$ and the channel realizations $f, g$ and $H$, i.e.,

$$T = T(d, f, g, H).$$

In the sequel, we exploit the lowercase subscripts $s, r$ and $u$ for notations concerning DeNB, RNs and UEs, respectively.

*Definition 2.* (Encoding functions) The DeNB encoding function at time instant $t \in [T]$

$$\psi_s^r : [2^{NL}] \times [N]^{M+K} \times \mathbb{C}^{Mt} \times \mathbb{C}^{Kt} \times \mathbb{C}^{Kt \times M} \to \mathbb{C}$$

\(^2\)For instance, in cases, where the files are requested in a non-uniform fashion \[34\], asymmetric caching across files is of relevance.
determines the DeNBs transmission signal \( x_s[t] = \psi_s^{[t]}(W, d, f^T, g^T, H^T) \) subjected to an average power constraint of \( P \). The encoding function of the causal full-duplex RN\(_m\) at time instant \( t \in [T] \) is defined by

\[
\psi_r^{[r]} : [2^\mu NL] \times \mathbb{C}^{d-1} \times [N]^M \times \mathbb{C}^{M_1} \times \mathbb{C}^{K_1} \times \mathbb{C}^{K_1 \times M} \to \mathbb{C},
\]

which determines the symbol \( x_{r,m}[t] = \psi_r^{[r]}(S_m, y_r^{[t-1]}, d, f^T, g^T, H^T) \) while satisfying the average power constraint given by the parameter \( P \).

Hereby, the symbols \( x_s[t] \) and \( x_{r,m}[t] \) are transmitted over \( t \in [T] \) channel uses. For any time instant \( t \), \( \psi_r^{[r]} \) accounts for the simultaneous reception and transmission through incoming and outgoing wireless links at RN\(_m\). To be specific, at the \( t \)-th channel use the encoding function \( \psi_r^{[r]} \) maps the cached content \( S_m \), the received signal \( y_r^{[t-1]} \) (see Eq. (2)), the demand vector \( d \) and global CSI to the symbol \( x_{r,m}[t] \).

After transmission, the received signals at UE\(_k\) is given by

\[
y_{u,k}[t] = g_k[t] x_s[t] + \sum_{m=1}^M h_{km}[t] x_{r,m}[t] + z_{u,k}[t], \forall t \in [T],
\]

where \( z_{u,k}[t] \) denotes complex i.i.d. Gaussian noise of zero mean and unit power. The received signal at RN\(_m\) is given by

\[
y_{r,m}[t] = f_m[t] x_s[t] + z_{r,m}[t], \forall t \in [T],
\]

where \( z_{r,m}[t] \) is additive zero mean, unit-power i.i.d. Gaussian noise. The desired files are decoded using the following functions.

**Definition 3.** (Decoding functions) The decoding operation at UE\(_k\) follows the mapping

\[
\eta_{u,k} : \mathbb{C}^T \times \mathbb{C}^{K_1 \times M_1} \times \mathbb{C}^{K_1} \times \mathbb{C}^{K_1 \times M} \to [2^L]
\]

to provide an estimate \( \hat{W}_{d_k} = \eta_{u,k}(y_{u,k}^T, d, f^T, g^T, H^T) \) of the requested file \( W_{d_k} \). In contrast to decoding at UE\(_k\), all RNs explicitly leverage their cached content according to

\[
\eta_{r,m} : \mathbb{C}^T \times [2^\mu NL] \times [N]^M \times \mathbb{C}^{M_1} \times \mathbb{C}^{K_1} \times \mathbb{C}^{K_1 \times M} \to [2^L]
\]

to generate \( \hat{W}_{d_r} = \eta_{r,m}(y_{r,m}^T, S_m, d, f^T, g^T, H^T) \) as an estimate of the requested file \( W_{d_r} \).

The reliability measure of a jointly proposed placement and delivery scheme is governed by its worst-case error probability defined as

\[
P_e = \max_{d \in [N]^{K+M}} \max_{j \in [K+M]} \mathbb{P}(\hat{W}_{d_j} \neq W_{d_j}) \quad (3)
\]
which is taken over error probabilities of \( M \) RNs and \( K \) UEs for all possible demands. A proper choice of caching, encoding and decoding functions that satisfy the reliability condition; that is, the worst-case error probability \( P_e \) approaches 0 as \( T \to \infty \), is called a feasible policy. For strictly positive rates \( T \to \infty \) is congruent with \( L \to \infty \).\(^3\) Now we are ready to define the delivery time per bit and its normalized version.

**Definition 4.** (Delivery time per bit\(^{[22]}\)) The delivery time per bit (DTB) is defined as

\[
\Delta(\mu, P) = \max_{d \in [N]^{K+M}} \limsup_{L \to \infty} \frac{\mathbb{E}[T(d, f, g, H)]}{L},
\]

where the expectation is over the channel realizations \( f, g \) and \( H \).

In the definition above, \( T \) represents the delivery time \(^{[35]}\). The normalization of the expected delivery time by the file size \( L \) gives insight about the per-bit latency. In this context, the DTB measures the per-bit latency, i.e., the latency incurred per-bit when transmitting the requested files through the wireless channel, within a single transmission interval for the worst-case request pattern of RNs and UEs as \( L \to \infty \). The DTB depends on the fractional cache size \( \mu \) and the power level \( P \).

In analogy to the degrees-of-freedom metric \(^{[36]}\), the normalized delivery time per bit (NDT) is a high-SNR metric that relates the DTB to that of a point-to-point reference system.

**Definition 5.** (Normalized delivery time\(^{[22]}\)) The NDT is defined as

\[
\delta(\mu) = \lim_{P \to \infty} \frac{\Delta(\mu, P)}{1/\log(P)}.
\]

The minimum NDT \( \delta^*(\mu) \) is the infimum of \( \delta(\mu) \) over all feasible policies.

The NDT compares the delivery time per bit achieved by the feasible coding scheme for the worst-case demand scenario to that of a baseline interference-free system in the high SNR regime. The feasible scheme, on the one hand, allows for reliable transmission of one file of \( L \) bits to each Rx on average in \( \mathbb{E}[T(f, g, H)] \) channel uses, i.e., 1 bit in \( \mathbb{E}[T(f, g, H)]/L \) channel uses. The baseline system (e.g., a point-to-point channel), on the other hand, can transmit \( \log(P) \) bits to a single Rx in one channel use, i.e., 1 bit in \( 1/\log(P) \) channel uses in the worst case. Therefore, the resulting NDT \( \delta(\mu) \) indicates that the worst-case delivery time for one bit of the

\(^3\)This is due to the fact that \( N = 2^L \) files are chosen uniformly at random from the index set \([2^T]^R \) with \( R \) being the rate. In consequence, \( L = TR \) which shows the equivalence of \( T \to \infty \) and \( L \to \infty \).
cache-aided network at fractional cache size $\mu$ is $\delta(\mu)$ times larger than the time needed by the baseline system.

From [26, Lemma 1], it readily follows that the NDT is a convex function in $\mu$. This means that a cache-aided network shown in Fig. 1 operating at fractional cache size $\mu = \alpha \mu_1 + (1 - \alpha) \mu_2$ for any $\alpha \in [0, 1]$ achieves less (or equal) NDT than the convex combination $a\delta(\mu_1) + (1 - a)\delta(\mu_2)$ through applying known feasible schemes applicable at fractional cache sizes $\mu_1$ and $\mu_2$ on distinct $\alpha$ and $1 - \alpha$-fractions of the files, respectively. This strategy is known as memory sharing.

III. MAIN RESULTS

In this section, we state our main results on the minimum NDT for the cache-enabled broadcast-relay wireless network of Fig. 1 for $M$ RNs and $K$ UEs. Hereby, our main results are presented in Theorems 1–3. They include, respectively, a novel lower bound, an upper bound (achievability) on the NDT and a complete NDT-tradeoff characterization for $K + M \leq 4$. Further, we formulate multiple corollaries that evaluate the performance of the scheme presented in Theorem 2 with respect to the lower bound of Theorem 1 in terms of a multiplicative gap given by

$$\frac{\delta_{\text{ach}}(\mu)}{\delta_{\text{LB}}(\mu)}$$

Hereby, $\delta_{\text{ach}}(\mu)$ and $\delta_{\text{LB}}(\mu)$ denote, respectively, an upper and lower bound on the NDT.

**Theorem 1 (Lower bound on NDT).** For the transceiver cache-aided network with one DeNB, $M$ RNs each endowed with a cache of fractional cache size $\mu \in [0, 1]$, $K$ UEs and a file library of $N \geq K + M$ files, the optimal NDT is lower bounded under perfect CSI at all nodes by

$$\delta^*(\mu) \geq \max \left\{ 1, \max_{\ell \in \mathbb{I}^{\left[1:M]\sbar}} \delta_{\text{LB}}(\mu, \ell, s) \right\},$$

where $\bar{s} = M + 1 - s$ and

$$\delta_{\text{LB}}(\mu, \ell, s) = \frac{K + \ell - \mu(\bar{s}(K - s + \frac{(\bar{s}-1)}{2}) + \frac{\ell}{2}(\ell + 1))}{s}.$$  

**Proof.** The proof of Theorem 1 will be given in Section IV. To provide some insight into the lower bound presented in Theorem 1 however, we outline a short sketch of the proof. Particularly, we summarize the ideas when deriving the two terms in Eq. (6).

First, we find the bound $\delta_{\text{LB}}(\mu, \ell, s)$ by exploiting the following main observation in the high SNR regime (where noise becomes negligible). That is, given the channel outputs of any $s$ UEs (e.g., of UE$_1$, UE$_2$, . . . , UE$_s$ denoted by $y_{u_{[1:s]}}^T$), in addition to the cached content of $\ell$ RNs (e.g,
cached contents of RN$_1$, RN$_2$, . . . , RN$_\ell$ represented by $S_{[1:\ell]}$ such that $s+\ell \geq M+1$ enables the decoding of all $K$ files requested by the UEs as well as $\ell$ files desired by the RNs. This is due to the fact that with this information set, all $M+1$ transmit signals consisting of the DeNB signal $x_s$ and the RNs transmit signals $x_{r,m}$, $\forall m \in [M]$, can be reproduced. This in turn, allows the reconstruction of the following channel outputs: On the one hand, the remaining $K-s$ channel outputs of the UEs and on the other hand $\ell$ outputs of the RNs. With the availability of $K$ UE channel outputs as well $\ell$ RN channel outputs and cached contents, $K+\ell$ files in total become decodable.

Second, the unity lower bound follows from the fact that the NDT is bounded from below by the performance of the reference interference-free system with an NDT of 1. The maximum over these two lower bounds concludes the proof of Theorem [1]. □

Before establishing the achievability at fractional cache sizes in the range $\mu \in (0,1)$, we consider two special corner points at fractional cache sizes $\mu = 0$ and $\mu = 1$ for arbitrary $M$ and $K$. These are the cases where the RN has either zero-cache ($\mu = 0$) or full-cache ($\mu = 1$) capabilities. In the following lemma, we expound the optimal NDT for these two points.

**Lemma 1.** For the transceiver cache-aided network with one DeNB, $M$ RNs each endowed with a cache of fractional cache size $\mu$, $K$ UEs and a file library of $N \geq M+K$ files, the optimal NDT is

$$\delta^*(\mu) = K + M \text{ for } \mu = 0, \quad (8)$$

achievable via DeNB broadcasting to $M$ RNs and $K$ UEs, and

$$\delta^*(\mu) = \max \left\{ 1, \frac{K}{M+1} \right\} \text{ for } \mu = 1, \quad (9)$$

achievable via zero-forcing beamforming for an $(M+1,K)$ MISO broadcast channel.

**Proof.** For the proof, it suffices to find a cache transmission policy that matches the lower bound in Thoerem [1] for $\mu = 0$ and $\mu = 1$, respectively. On the one hand, if $\mu = 0$, we note that $\delta_{LB}(0,M,1) = K + M$. On the other hand, if $\mu = 1$, we observe that $\delta_{LB}(1,0,M+1) = K/(M+1)$ if $M+1 \leq K$ and $\delta_{LB}(1,\ell,s) < 1$ if $M+1 > K$. Next, we consider the achievability at $\mu = 0$ and $\mu = 1$. For these two fractional cache sizes, the network in Fig. [1] reduces to a SISO broadcast

---

$^4$In MISO broadcast channels, we use the notation, $(a,b)$ for integers $a$ and $b$ to denote a broadcast channel with $a$ transmit antennas and $b$ single antenna receivers.
channel (BC) with $K + M$ users for $\mu = 0$ and an $(M + 1, K)$ MISO broadcast channel for $\mu = 1$. The approximate per-user rate (neglecting $o(\log(P))$ bits) for these two channels are known to be $\frac{1}{(K+M)} \log(P)$ [37] (achievable through unicasting each user’s message) for $\mu = 0$ and $\frac{1}{K} \min\{M + 1, K\} \log(P)$ (achievable through zero-forcing beamforming) [38] for $\mu = 1$, respectively. Equivalently, each user needs the reciprocal per-user rate of signaling dimensions (e.g., channel uses in time) to retrieve one desired bit reliably. Thus, the approximate DTB becomes, respectively, $\frac{(K+M)}{\log(P)}$ and $\frac{K}{\min\{M+1,K\} \log(P)}$. Normalizing the delivery time per bit by the point-to-point reference DTB $\frac{1}{\log(P)}$ generates the NDTs $K + M$ and $\max\{1, K/(M+1)\}$. This establishes the NDT-optimality at these fractional cache sizes. □

**Remark 1.** From Lemma 1, we infer that the caching problem for the system illustrated in Fig. 1 establishes the behavior of the network in terms of delivery time between the two extremes – SISO BC with $K + M$ users and an $(M + 1, K)$ MISO BC. This analysis will reveal what kind of schemes other than simple unicasting and zero-forcing will be optimal for $0 < \mu < 1$.

**Remark 2 (Optimality of Memory Sharing of Zero-cache and Full-cache Schemes).** Memory sharing of optimal schemes at extreme points of zero-cache ($\mu = 0$) and full-cache ($\mu = 1$) may be optimal. Or in other words, the optimal NDT of intermediate points at fractional cache sizes $0 < \mu < 1$ may be achievable through successively time-sharing between unicasting and zero-forcing on $(1 - \mu)$ and $\mu$ fractions of the files, respectively. This implies that treating uncached and cached file fractions independently by applying two separate delivery schemes – unicasting and zero-forcing – can be delivery time optimal. However, as shown in Fig. 2, this only happens
for cases where \( M = 1 \) and \( K \leq 2 \). In these cases, the lower bound \( \delta_{LB}(\mu, 1, 1) \) coincides with the achievable NDT.

Intuitively, the observation of Remark 2 makes sense because of the following two reasons. On the one hand, when \( M = 1 \), there are (with respect to \((1 - \mu)\) fractions of uncached information on each file) (i) no multicasting opportunities on the DeNB-RN broadcast channel and (ii) no zero-forcing opportunities on the RN-UE channel. On the other hand, when \( K \leq 2 \) joint DeNB-RN zero-forcing beamforming guarantees (with respect to the cached fractions of the files all \( K \) UEs desire) that every UE receives in each channel use desired information on its requested file since \( K \leq M + 1 = 2 \). Thus, in conclusion, for these cases applying unicasting and joint DeNB-RN beamforming successively on \((1 - \mu)\) and \( \mu \) file fractions is NDT-optimal. However, this is in general not true for arbitrary instances of \( K \) and \( M \) as we shall see next.

So far we have discussed cases where successively applying unicasting and joint DeNB-RN beamforming on uncached and cached file fractions is NDT-optimal at fractional cache sizes \( 0 < \mu < 1 \). However, in general, (with the exception of \( M = 1, K \leq 2 \)) delivery schemes that treat uncached and cached file fractions independently are suboptimal for arbitrary \( K \) and \( M \). To this end, we propose a general one-shot (OS) scheme that treats uncached and cached file fractions jointly by exploiting multicasting and zero-forcing opportunities. Our proposed one-shot scheme is with the exception of the full-duplex requirement at the RNs simple in implementation because all receiving nodes are able to decode their desired symbols on a single channel use basis. In other words, these schemes explicitly preclude symbol decoding over multiple channel uses. The next theorem specifies the achievable NDT for these one-shot schemes.

**Theorem 2 (Achievable One-Shot NDT).** For \( N \geq K + M \) files, \( K \) UEs and \( M \) RNs each with a cache of (fractional) size \( \mu \in \{0, 1/M, 2/M, 3/M, \ldots, (M-1)/M, 1\} \),

\[
\delta_{OS}(\mu) \triangleq \max \left\{ \delta_{MAN}(\mu), \frac{K + \delta_{MAN}(\mu) \mathbf{1}_{K > \mu M}}{\min\{K, 1 + \mu M\}} \right\}
\]

(10)
is achievable, where \( \mathbf{1}_{K > \mu M} \) is the indicator function and \( \delta_{MAN}(\mu) \) is the achievable Maddah-Ali Niesen (MAN) NDT given by

\[
\delta_{MAN}(\mu) = M \cdot (1 - \mu) \cdot \frac{1}{1 + \mu M}
\]
such that \( \delta^*(\mu) \leq \delta_{OS}(\mu) \). For arbitrary \( \mu \in [0, 1] \), the lower convex envelope of these points is achievable.
Fig. 3: Illustration of the proposed one-shot scheme with $M = 4$ RNs, $K = 2$ UEs and $\mu M = 2$ for the worst-case demand scenario. On the one hand, in each channel use of the first phase [cf. (a)], the MAN scheme is used on the SISO DeNB-RN broadcast channel to convey desired symbols of (any combination of) $1 + \mu M$ RNs. In the worst-case scenario, where UEs request other files, these symbols represent interference which are zero-forced through cooperative DeNB-RN interference cancelation at (any combination) of $\min\{K, \mu M\}$ UEs. Simultaneously, the scheme exploits RN caches by providing the same UEs with their desired symbols. After $T_1$ channel uses, of the first phase, the demand of the RNs is satisfied. On the other hand, the second phase [cf. (b)] is devoted to communicate, if necessary, the remaining file symbols of the UEs by applying cooperative DeNB-RN zero-forcing beamforming.

Proof. Details on the scheme are presented in Section V. Nevertheless, we use this paragraph to highlight the main idea of the scheme. Recall that the file length is denoted by $L$. Assuming Gaussian signaling, the file is comprised of $L' = L/\log(P)$ symbols. The scheme we develop (potentially) consists of two phases requiring $T_1$ and $T_2$ channel uses, respectively, to send uncached $(1 - \mu)L'$ Gaussian symbols (each symbol carrying approximately $\log(P)$ bits) to each
RN and also $L'$ symbols to each UE.

In every channel use of the first phase depicted in Fig. 3a, beamforming facilitates the integration of the MAN scheme [7] with zero-forcing beamforming to (i) pre-cancel interference caused by applying the MAN scheme on the DeNB-RN broadcast channel at the UEs and (ii) convey desired symbols to UEs. Precisely, the MAN scheme is applied on the DeNB-RN broadcast channel to provide each RN in a subset $S_R \subset [M]$ with $|S_R| = 1 + \mu M$ RNs with a desired symbol. Simultaneously, the full-duplex capabilities at the RNs are exploited by conveying to each UE in the subset $S_U \subset [K]$ with $|S_U| = \min\{K, \mu M\}$ UEs in total with a desired symbol by zero-forcing the contribution of all interfering symbols that $|S_R| = 1 + \mu M$ RNs in $S_R$ desire. Recall that the first phase consumes $T_1$ channel uses. We show that $T_1$ channel uses suffice in sending each RN $m$, $\forall m \in [M]$, the remaining $(1 - \mu)L'$ symbols of its requested file. Simultaneously, in $T_1$ channel uses each UE $k$, $\forall k \in [K]$, receives $\tilde{L}$ symbols of its desired file, with $\tilde{L}$ being proportional to $|S_U| = \min\{K, \mu M\}$. Thus, we may encounter cases where it is either feasible or infeasible to communicate all $L'$ symbols of each requested file to the respective UEs in $T_1$ channel uses ($\tilde{L} \geq L'$ or $\tilde{L} < L'$).

Only in the case of missing symbols ($\tilde{L} < L'$) that all $K$ UEs still require after $T_1$ channel uses, additional $T_2 > 0$ channel uses are required in phase two to deliver the remaining desired symbols as shown in Fig. 3b. To this end, in every channel use cooperative DeNB-RN zero-forcing beamforming is deployed to send one symbol in total to $\psi' = \min\{K, 1 + \mu M\}$ UEs. The decoding at the RNs and UEs does not involve symbol decoding over multiple channel uses. Instead, decoding occurs on a one-shot, or single channel use, basis. In conclusion, the achievable NDT becomes either $\frac{T_1}{L'}$ if $T_2 = 0$ or $\frac{T_1 + T_2}{L'}$ if $T_2 > 0$ with $L'$ being the number of symbols per file. □
### Region Name Definition

| Region Name | Definition | Channel limitation | Achievable NDT |
|-------------|------------|--------------------|----------------|
| Region A    | $K \leq \mu M < M < \frac{1}{1-2\mu}, \mu \leq \frac{1}{2}$  
$K \leq \mu M \leq M, M > \frac{1}{1-2\mu}, \mu > \frac{1}{2}$ | ✓ | $\delta_{\text{OS}}^{(A)}(\mu) = 1$ |
| Region B    | $K \leq \mu M, \frac{1}{1-2\mu} \leq M, \mu \leq \frac{1}{2}$ | ✓ | $\delta_{\text{OS}}^{(B,E)}(\mu) = \delta_{\text{MAN}}(\mu)$ |
| Region E    | $\mu M < K \leq \mu M \cdot \delta_{\text{MAN}}(\mu) \leq M$ | – | $\delta_{\text{OS}}^{(C,D)}(\mu) = \frac{K+\delta_{\text{MAN}}(\mu)}{1+\mu M}$ |
| Region C    | $\mu M < M < K$ | – | ✓ |
| Region D    | $\mu M \cdot \max \left\{ 1, \delta_{\text{MAN}}(\mu) \right\} < K \leq M$ | – | ✓ |

**TABLE I:** Definition of $(\mu, K, M)$ region triplets and their achievable one-shot NDT. The achievable one-shot NDT in Region A coincides with the lower bound and is thus NDT-optimal.

The delivery time of the proposed one-shot scheme is devoted to both RNs and UEs. It is intuitive to expect cases where the delivery of requested files by the UEs may take longer than the delivery of uncached file fractions by the RNs. For instance, we expect that for $M \ll K$, irrespective of the fractional cache size, the file delivery to UEs through the interference channel represents the channel limitation from a delivery time perspective. However, finding the exact areas (including the transition) as a function of $\mu, K$ and $M$ where either RN or UE file delivery through broadcast or interference channel represents the bottleneck from a latency perspective for one-shot schemes is of interest. To this end, we conclude from Theorem 2 that the functional behavior of the achievable one-shot NDT changes for different region triplets $(\mu, K, M)$ as follows. Specifically, when neglecting the discretization of the fractional cache size $\mu$ to values $\{0, 1/M, 2/M, 3/M, \ldots, (M-1)/M, 1\}$, Table II specifies how the one-shot NDT expression (10) simplifies for the given region triplets. The regions of Table II are illustrated in Fig. 4 for constant $K$ ($K = 2$). We state two interesting observations on these regions in the following remarks.

**Remark 3.** Interestingly, when $M \geq 2K + 1$, we see that for $\mu \geq 1/M$, the achievable one-shot NDT does not depend on $K$, i.e., the number of UEs. Instead, the NDT is solely dependent on the number of RNs $M$. Consequently, our one-shot NDT behaves identical to the achievable NDT of the MAN scheme. This is due to the fact that the delivery time is governed by the delivery of uncached file fragments to the RNs through the broadcast channel. Then, this problem reduces to the initial receiver-based, single server coded caching problem of Maddah-Ali and Niesen.
Fig. 4: 2D \((\mu, M)\)-plot of all Regions A, B, C, D and E for constant \(K (K = 2)\). The labels on the graph indicate the functional relationship at the borders of neighboring regions. The discrete points illustrate the fractional cache sizes \(\mu \in \left\{ 0, \frac{1}{M}, \ldots, \frac{M-1}{M}, 1 \right\} \) for which the achievable one-shot NDT expression \(\delta_{OS}(\mu)\) in Eq. (10) actually hold. The annotations to the regions specify the main characteristics of the respective region. The channel limitations specify which channel – broadcast or interference channel – is characteristic for the delivery time overhead. The RN standalone frontier, where \(\mu M = K\) holds, represents scenarios for which all \(K\) UEs can be served by any subset of \(\mu M\) RNs without the need of the DeNB.

In this setting, we recall that the MAN NDT for \(M\) receivers consists of the local and global caching gains (captured by the factors \(1 - \mu\) and \(\frac{1}{1+\mu M}\) in \(\delta_{MAN}(\mu)\), respectively) \([7]\). In conclusion, we refer, respectively, to the attainable NDT as the DeNB-to-RN) (broadcast-limited NDT and cases where \(M/K > 2\) as cases of high \(M/K\) (with respect to the number of UEs).

Remark 4. As opposed to the previous remark, we observe that in the one-shot scheme the interference channel to the UEs functions as the bottleneck from a delivery time perspective as long as \(M < 2K + 1\). We call these instances, all with respect to the number of UEs, as cases of moderate \(M/K\) when \(1 \leq M/K \leq 2\) and low \(M/K\) when \(M/K < 1\). We then name the attainable
NDT of $\frac{K+\delta_{\text{MAN}}(\mu)}{1+\mu M}$ the interference-limited NDT.

In the following four corollaries, we state the relations of the NDT lower bound from Theorem 1 and the one-shot scheme upper bound of Theorem 2 of the aforementioned regions for discretized $\mu$.

**Corollary 1 (One-Shot NDT Optimality).** The one-shot scheme is optimal (i.e., it coincides with the lower bound) achieving the minimum NDT given by $\delta^*(\mu) = 1$, when the triplet $(\mu, K, M)$ satisfies any of the following conditions:

1a) $K < M, \mu \geq \frac{K}{M}, \mu M \geq \left\lceil \frac{M-1}{2} \right\rceil, \mu \leq \frac{1}{2}$; or
1b) $K \leq M, \mu \geq \frac{K}{M}$ and $\mu > \frac{1}{2}$.

**Proof.** The proof follows from the fact that all discretized $\mu$ values inside Region A attain an NDT $\delta_{\text{OS}}^A(\mu) = 1$ that matches the lower bound. □

**Remark 5.** With this corollary, we make the observation that at high $M/K$ a fractional cache size of approximately $1/2$ is sufficient of achieving the lowest attainable NDT of 1. In other words, this shows that when $M/K > 2$ prefetching half of each file (from the entire library of files) and applying the one-shot scheme is delivery time optimal. Caching more than that will not reduce the delivery time any further.

The remaining three corollaries state the multiplicative gap of the one-shot scheme with respect to the lower bound.

**Corollary 2 (Gap of NDT for High $M/K$ $(M/K > 2)$).** The multiplicative gap between the one-shot scheme presented in Section V and the lower bound on the NDT for $M \geq 2K + 1$ is upper bounded by

1a) $\frac{M-1}{2}$ for $\mu \leq \frac{1}{M}$,
1b) $\frac{M-\theta}{1+\theta}$ for any $\mu \in \left[\mu'(\theta), \left\lceil \frac{(M-1)/2}{M} \right\rceil \right]$, where $\mu'(\theta) = \left\lceil \frac{\theta}{M} \right\rceil$ and $\theta \in \left[1, \frac{M-3}{2} \right]$.

**Proof.** The proof of this corollary is given in the appendix. Specifically, details of the proof can be found in subsections A-A and A-C. □

**Corollary 3 (Gap of NDT for Moderate $M/K$ $(1 \leq M/K \leq 2)$).** The multiplicative gap between the one-shot scheme presented in Section V and the lower bound on the NDT at moderate $M/K$ is
1) 1 for $K = 1$ and within (or upper bounded by)

2a) $\frac{K}{2} + \frac{M-1}{4}$ for $\mu \leq \frac{1}{M}$ and $K \geq 2$,

2b) $\frac{dK}{M+1} + \frac{d(d-1)}{M+1}$ for any $\mu \in \left[\mu'(\kappa_d), \frac{K}{M}\right]$, where $\mu'(\kappa_d) = \frac{\lfloor \kappa_d \rfloor}{M}$, $\kappa_d = \frac{M+1-d}{d}$ and $d \in \left[\frac{M+1}{K}, \frac{M+1}{2}\right]$ for $K \geq 2$.

Proof. The proof of this corollary is given in subsection A-B.

\[\square\]

Corollary 4 (Gap of NDT for Low $M/K$ ($M/K < 1$)). The multiplicative gap between the one-shot scheme presented in Section V and the optimal NDT at low $M/K$ is

1) 1 for $(K, M) = (1, 2)$ and upper bounded by

2) $1 + \frac{1}{K} \left(\frac{K}{2} - \frac{3}{4}\right)$ for $M = 1, K > 2$,

3a) $1 + \left(\frac{K}{2} + \frac{M-5}{4}\right) \cdot \frac{M}{K} \cdot \frac{(K+M+1)}{(K+M-1)}$ for $\mu \leq \frac{1}{M}$ and $M \geq 2$,

3b) $d + \frac{d(d-1)}{K}$ for any $\mu \in \left[\mu'(\kappa_d), 1\right]$, where $\mu'(\kappa_d) = \frac{\lfloor \kappa_d \rfloor}{M}$, $\kappa_d = \frac{M+1-d}{d}$ and $d \in \left[\frac{M+1}{M}, \frac{M+1}{2}\right]$ for $M \geq 2$.

Proof. The proof of this corollary is given in subsection A-B of the appendix.

\[\square\]

We use the last three corollaries to show that a fractional cache size of approximately $1/2$ generates a constant multiplicative gap less than 3 for arbitrary $K$ and $M$. This is stated in the following corollary.

Corollary 5 (Constant Gap). For fractional cache sizes $\mu \geq \mu_C = \frac{\lceil(M-1/2\rceil)}{M}$ the multiplicative gap between the one-shot scheme presented in Section V and the optimal NDT is constant versus $\mu$ and is bounded by a factor of $8/3$.

Proof. At high $M/K$, the NDT optimality at $\mu_C$ is shown in Corollary 1. For this case, the gap is per definition of optimality 1. Next, at moderate $M/K$, the gap is, respectively, for $K = 1$ and $(K, M) = (2, 2)$ 1 and $\frac{5}{4}$ as stated by Corollary 3 (points 1 and 2b) with $d = \frac{3}{2}$. Also at moderate $M/K$, when $K \geq 2, M \geq 3$, we use point 2b) with $d = 2$ to find the multiplicative gap

$$\frac{dK}{M+1} + \frac{d(d-1)}{M+1} \bigg|_{d=2} = \frac{2(K+1)}{M+1} \leq 2$$

for fractional cache sizes $\mu \geq \mu'(\kappa_2) = \mu_C = \frac{\lceil(M-1/2\rceil)}{M}$. Finally, at low $M/K$, the gap for the special cases $(K, M) = (2, 1)$ and $M = 1, K > 2$ are according to 1) and 2) of Corollary 4. 1 and

$$1 + \frac{1}{K} \left(\frac{K}{2} - \frac{2}{K}\right) \leq \frac{3}{2},$$
Combining all cases, we conclude that the gap is bounded from above by \( \frac{8}{3} \). \(\square\)

In conclusion, this corollary shows that one-shot schemes at fractional cache sizes of \( \frac{1}{2} \left( \frac{1}{2} \geq \frac{\lfloor (M-1)/2 \rfloor}{M} \right) \) are optimal within a constant (with respect to the lower bound) multiplicative gap of optimality. The main disadvantage of this scheme becomes visible when considering regions C and D. For these regions, the first and second phase provide (per channel use) only a subset of \( \mu M \) and \( \min\{K, 1 + \mu M\} \) UEs with their desired symbols while the remaining UEs observe interference. Interference alignment provides the opportunity of alleviating the effect of undesired symbols. To this end, we establish achievability schemes that involve a novel beamforming design that facilitates (i) multicasting opportunities (when \( M \geq 2 \)) on the DeNB-RN broadcast channel, (ii) (joint) zero-forcing opportunities and (iii) subspace interference alignment. The notion of subspace alignment was first introduced in [39].

The idea is to align interferences into a multi-dimensional subspace instead of a single dimension [40]. Through this beamforming design and using the insights from Lemma 1 and Theorem 2 we are able to establish the complete NDT-tradeoff for \( K + M \leq 4 \). The following theorem specifies this tradeoff.

**Theorem 3 (Optimal NDT Tradeoff)**. The optimal NDT tradeoff for the transceiver cache-aided network with one DeNB, \( M \) RNs each endowed with a cache of fractional cache size \( \mu \in [0, 1] \), \( K \) UEs satisfying \( K + M \leq 4 \) with \( N \geq K + M \) number of files, is given as

\[
\delta^*(\mu) = \max \left\{ 0, K + M - \mu M(K + M - 1), \frac{K + M - \mu (M^2 + (K - 3)(M - 1))}{2}, \frac{K + M - 1 - \mu (M^2 + (K - 3)(M - 1) - M)}{2} \right\}.
\]

**Proof.** The lower bound on the NDT for this setting readily follows from Theorem 1. Concretely, the last three terms inside the max-expression of (11) correspond to \( \delta_{LB}(\mu, M, 1) \), \( \delta_{LB}(\mu, M, 2) \) and \( \delta_{LB}(\mu, M - 1, 2) \), respectively. The outer bound (achievability) on the NDT is presented in Section VI. Shortly, we establish the achievability for (at most) four corner points at zero-cache and full-cache fractional cache sizes \( \mu = 0, \mu = 1 \) as well as intermediate cache sizes \( \mu' \) in the interval \( 0 < \mu' < 1 \). The first two corner points are achievable through DeNB broadcasting and cooperative DeNB-RN zero-forcing beamforming (cf. Lemma 1). In the achievability scheme at intermediate
fractional cache size $\mu = 1/M$, on the other hand, the one-shot scheme of Theorem 2 is optimal for $(K, M) \in \{(1, 2), (2, 2), (1, 3)\}$. This is shown in Figs. 5b and 5c for the cases $(K, M) \in \{(1, 2), (2, 2)\}$. For the cases $(K, M) \in \{(3, 1), (2, 2)\}$, respectively, we establish the optimal NDT at fractional cache sizes $4/5$ and $4/9$ (cf. Figs. 5a and 5c) through optimized precoding design that synergistically integrates subspace alignment with zero-forcing beamforming consuming finite channel uses.

In the following remarks, we discuss the results of Theorem 3 in further detail. In the
Fig. 6: Achievable sum DoF as a function of $\mu$ for (a) $(K, M) = (3, 1)$, (b) $(K, M) = (1, 2)$ and (c) $(K, M) = (2, 2)$. The dashed line shows the DoF of a time-sharing based unicasting-zero-forcing scheme.

discussion, we assume that RNs and UEs all request distinct files. This represents the worst-case demand scenario.

Remark 6 (Subspace Interference Alignment). It is of interest to discuss why aligning all interferences into one dimension is not feasible when $\mu < \frac{1}{M}$. In cases, where $\mu < \frac{1}{M}$ the collection of all $M$ caches cannot hold the entire library of files. Thus, under the placement strategy when each RN caches $\mu L'$ ($L' = L/\log(P)$) independent Gaussian symbols (i.e., no
overlaps in file chunks being cached) of a file, we observe that \((1 - \mu M)L\) symbols are only available at the DeNB. In the following, we argue why multiple subspace dimensions have to be allocated for the interference from both RN and UE perspective.

First, we consider the delivery of file content to all \(M\) RNs. We observe that each RN is interested in \((1 - \mu)L'\) uncached symbols of its requested file; out of which \((1 - \mu M)L'\) symbols are only available at the DeNB. In the worst-case scenario, this amounts to \(M(1 - \mu M)L\) symbols in total since there are \(M\) distinct files that \(M\) RNs request. These symbols have to be broadcast to the RNs. However, they represent interference to all \(K\) UEs. Aligning these symbols at all \(K\) UEs to a single signaling dimension is not feasible because alignment at the UEs would make these symbols indistinguishable at all \(M\) RNs. To preclude this, all \(M(1 - \mu M)L'\) symbols have to be aligned at distinct subspaces.

Second, subspace alignment is also necessary when focusing on the delivery of \(K\) distinct files with one single file being desired by one UE. A similar line of argument as in the previous paragraph suggests that \(K(1 - \mu M)L'\) symbols are only available at the DeNB (and not at the \(M\) RNs) and have to be conveyed from the DeNB to \(K\) UEs. For the sake of reliable decodability at the UEs, all these symbols have to be distinguishable from each other. Since only \((K - 1)(1 - \mu M)L'\) of those symbols represent interference at a single UE, the interference dimension increases by \((K - 1)(1 - \mu M)L'\).

In conclusion, at least \((K + M - 1)(1 - \mu M)L'\) interference dimensions are required. For the cases \((K, M) \in \{(3, 1), (2, 2)\}\), at fractional cache sizes \(\frac{4}{5}\) \((L' = 5)\) and \(\frac{4}{9}\) \((L' = 9)\), the number of interference dimensions is in agreement with \((K + M - 1)(1 - \mu M)L' = 3\).

*Remark 7 (Feasibility for Constant Channels)*. For completeness, we would like to emphasize that the complete NDT tradeoff of Theorem 3 is applicable to constant channels as well. This is due the fact that both the one-shot scheme as well as the alignment scheme are feasible for time-invariant channels. In particular, the synergistic beamforming design is feasible for constant channels under the umbrella of real interference alignment [41], [42]. Whether a two-phase precoding design with constant channels (similar to previous work on relay-aided X-channels [43]) attains close-to-optimal performance is an interesting extension to work on. However, it is beyond the scope of this paper.

*Remark 8 (Inverse Sum DoF vs. NDT)*. From the optimal NDT tradeoff of Theorem 3, we may
compute the resulting achievable (sum) DoF as follows

\[
\text{DoF} \geq K \cdot \frac{1}{\delta^*(\mu)} \cdot \text{per-UE DoF} + M \cdot \frac{(1 - \mu)}{\delta^*(\mu)} \cdot \text{per-RN DoF}.
\]

The achievable DoF for \((K, M) \in \{(3, 1), (1, 2), (2, 2)\}\) is shown in Fig. 6. When comparing NDT and DoF (Fig. 5 vs. Fig. 6), we clearly see that an increase in sum rate (measured by the DoF) is not necessarily equivalent to a decrease in delivery time (measured by the NDT). In other words, we observe that the NDT metric is not necessarily proportional to the inverse of the sum DoF for fractional cache sizes exceeding \(\tilde{\mu} = \frac{4}{5}\) \((\tilde{\mu} = \frac{1}{2})\) when \((K, M) = (3, 1) ((K, M) \in \{(1, 2), (2, 2)\})\). In fact, interestingly there are cases when both NDT and achievable DoF decrease for increasing \(\mu\).

At first glance, this may seem counterintuitive. However, a closer look reveals that an increasing \(\mu\) has the advantage of elevating the per-UE DoF at the cost of a declining per-RN DoF. In consequence, the overall achievable sum DoF may drop, but the increase in the per-UE DoF comes along with a decay in NDT. Ultimately, this observation suggests that the inverse sum DoF can be a misleading metric with respect to the delivery time. This is mainly due to the fact that the DoF metric loses information on the per-user DoF for asymmetric rate allocation scenarios.

IV. LOWER BOUND (CONVERSE) OF THE MINIMUM NDT

In this section, we present the proof of the lower bound of the minimum NDT in Theorem 1. The method of the proof extends on the approaches of [22] and [33].

In the following, we expound the key idea of the proof. To this end, we introduce the following worst-case considerations. First, we presume that all \(K\) UEs request distinct files \(W_{d_j} (d_j \neq d_{\ell}, j, \ell \in [K], j \neq \ell)\). Second, for the sake of notational simplicity and without loss of generality, we set the requested files by the \(K\) UEs to \(W_{[1:K]} = (W_1, W_2, \ldots, W_K)\), or shortly \(d[1 : K] = [K]\). Third, in our proof we assume that there at least \(K + M\) distinct files \(W_{d_k}\) available, i.e., \(N \geq K + M\). Under this UE demand pattern and given channel realizations \(f, g\) and \(H\), we obtain a lower bound on the delivery time \(T = T(d, f, g, H)\), and therefore ultimately on the NDT, of any feasible scheme. The UE request pattern is one possible (out of \(\binom{N}{K} K!\) requests) worst-case scenario due to the following line of argument. Redundancy in the UE request pattern can, if anything, only decrease the NDT further. This is due to the fact that the NDT of a smaller network (with respect to the number of UEs) is in general a lower bound on bigger networks. For instance, when we
consider redundancy in the file request pattern at the UE side with respect to one single file, we make the following observation. The transceiver cache-aided network under study with $M$ RNs and $K$ UEs with $\tilde{K}, \tilde{K} \leq K$, UEs having a redundant request (i.e., requesting the same file) at the UE side behaves similarly (with respect to the NDT performance) to a smaller network with $M$ RNs but only $K - \tilde{K} + 1$ UEs.

The key idea in establishing the lower bound on the NDT is that $K + \ell$ unique files, comprising of all $K$ files $W_{[1:K]}$ requested by the UEs and $\ell$ files desired by at least a subset of $\ell$ RNs (out of $M$ RNs), e.g., $W_{[K+1;K+\ell]}$ can be retrieved in the high SNR regime from

- $s$ output signals of the UEs, e.g., $y^T_{u,[1:s]}$ for $1 \leq s \leq \min\{M + 1, K\}$, and
- $\ell$ cached contents of $\ell$ RNs, e.g., $S_{[1:\ell]}$, where $\tilde{s} \leq \ell \leq M$ and $\tilde{s} = M + 1 - s$.

We note that since $s + \ell \geq M + 1$ holds, we are able to reconstruct all $M + 1$ transmit signals $(x_s[t]$ and $x_{r,m}[t], m \in [M])$ at all $T$ time instants of the delivery phase within bounded noise. This is shown in Fig. 7. We would like to emphasize to the reader that through the choice of $\ell$, the bounds account for distinct RN request patterns. We start the converse as follows:

\[
(K + \ell)L = H(W_{[1:K+\ell]}|W_{[K+\ell+1:N]}) \\
= I(W_{[1:K+\ell]}; y^T_{u,[1:s]}|W_{[K+\ell+1:N]}), S_{[1:\ell]}), H(W_{[1:K+\ell]}|y^T_{u,[1:s]}|W_{[K+\ell+1:N]}), S_{[1:\ell]}, W_{[K+\ell+1:N]}).
\] (12)

In what follows, we upper bound each summand in Eq. (12) individually. Using the chain rule of mutual information, the first term in (12) can be rewritten as shown below

\[
I(W_{[1:K+\ell]}; y^T_{u,[1:s]}|W_{[K+\ell+1:N]}), S_{[1:\ell]}), H(W_{[1:K+\ell]}|y^T_{u,[1:s]}|W_{[K+\ell+1:N]}), S_{[1:\ell]}, W_{[K+\ell+1:N]}).
\] (13)

Next, we use the non-negativity of mutual information ($I(W_{[1:K+\ell]}; W_{[1:s]}|W_{[K+\ell+1:N]}), y^T_{u,[1:s]}|S_{[1:\ell]}))$ followed by the chain rule of mutual information to upper bound (13) by

\[
I(W_{[1:K+\ell]}; y^T_{u,[1:s]}|W_{[K+\ell+1:N]}), S_{[1:\ell]}), H(W_{[1:K+\ell]}; W_{[1:s]}|W_{[K+\ell+1:N]}), y^T_{u,[1:s]}).
\] (14)

\footnote{For ease of notation, we assume that the $\ell$ files retrieved by the RNs are $W_{[K+1;K+\ell]}$.}
Fig. 7: Illustration of the proof of converse. The top part of the figure shows how $K + \ell$ unique files become (reliably) decodable in the high SNR regime when a receiver is aware of the information subset $\{y_u^T, S_{[1:T]}\}$ with $s + \ell \geq M + 1$. 
Rewriting the three mutual information terms of (14) by their respective differential or discrete entropy terms and bounding them results in the following chain of inequalities.

\[
\begin{align*}
&h(y^T_{u,[1:s]}|W_{[K+\ell+1:N]}) - h(y^T_{u,[1:s]}|W_{[1:N]}) + H(W_{[1:s]}|W_{[K+\ell+1:N]}, y^T_{u,[1:s]}) \\
&- H(W_{[1:s]}|W_{[1:N]}, y^T_{u,[1:s]} + H(S_{[1:s]}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}) - H(S_{[1:s]}|W_{[1:N]}, y^T_{u,[1:s]}) \\
&\leq (a) h(y^T_{u,[1:s]}) - h(z^T_{u,[1:s]}) + L\varepsilon_L + H(S_{[1:s]}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}) \\
&\leq (b) sT \log(2\pi e(cP + 1)) - h(z^T_{u,[1:s]}) + L\varepsilon_L + \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}) \\
&\leq (c) sT \log(cP + 1) + L\varepsilon_L + \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}) \\
&\leq (d) sT \log(cP + 1) + L\varepsilon_L + \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N]) + \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}) \\
&\leq (e) sT \log(cP + 1) + L\varepsilon_L + \bar{s}(K + \ell - s)\mu L + \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}) \\
&\leq (f) sT \log(cP + 1) + L\varepsilon_L + \bar{s}(K + \ell - s)\mu L \\
&+ \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}, x_{u,[1:s]}{1:1}, z^T_{r,[1:s]}{1:1}) \\
&\leq (g) sT \log(cP + 1) + L\varepsilon_L + \bar{s}(K + \ell - s)\mu L + T\varepsilon_P \log(P) \\
&+ \sum_{i=1}^{\ell} H(S_{i}|W_{[1:s]}\cup[K+\ell+1:N], y^T_{u,[1:s]}, S_{[1:s]}_{1:1}, x_{u,[1:s]}{1:1}, z^T_{r,[1:s]}{1:1}) \\
&\leq (d) sT \log(cP + 1) + L\varepsilon_L + \bar{s}(K + \ell - s)\mu L + T\varepsilon_P \log(P) + \sum_{i=1}^{\ell} H(S_{i,j})
\end{align*}
\]
where the steps in (15) are explained as follows:

- Step (a) follows from the fact that dropping the conditioning on the first term does not increase entropy. Further, we apply Fano’s inequality to the third term. Hereby, \( \epsilon \) is a function, independent of \( P \) which vanishes in the limit as \( L \to \infty \). The fourth and sixth term are zero, respectively, because \( W_{1:x} \subseteq W_{1:N} \) and \( S_{1} \) is a deterministic function of \( W_{1:N} \).

- Step (b) follows by applying [22, Lemma 1] on the first differential entropy term. Additionally, we use the chain rule to rewrite the fourth term.

- Step (c) is due to the fact that the channel noise \( \hat{z}^{T}_{u,1:x} \) is i.i.d. across time and has a Gaussian distribution with zero mean and unit variance.

- Step (d) follows from the fact that in a discrete entropy of the form \( H(S_i|\hat{W}) \), \( \hat{W} \subseteq W \) with \( W = \{W_1, \ldots, W_N\} \), only the files \( W \setminus \hat{W} \) have non-zero contribution. This is because the cached content at the \( i \)-th RN of the \( j \)-th file is solely a function of \( W_j \), i.e., \( S_{i,j} = \phi_{i,j}(W_j) \).

- Step (e) is since the entropy of each local file cache content is upper bounded according to \( H(S_{i,j}) \leq \mu L \).

- In step (f), we use the fact that in the first channel use \( (t = 1) \), the transmit signal \( x_{r,m}[1] \) at RN \( m \) depends (apart from CSI) solely on the cached content \( S_{m} \). Further, we note that the noise term \( \hat{z}^{T}_{r,1:j-1} \) is independent of all the remaining random variables in the conditional entropy term.

- In step (g), we upper bound \( H(W_{s+1:K+i-1}, \hat{z}^{T}_{u,1:x} | W_{1:x} \cup [K+\ell+1:N], y^{T}_{u,1:x}, S_{1:j-1}, x_{r,1:j-1}[1], z^{T}_{r,1:j-1}) \) by \( T \epsilon + L \epsilon \), where \( \epsilon \) is any function in \( P \) which satisfies \( \lim P \to \infty \epsilon P = 0 \). This is because

\[
\mathcal{h}(z^{T}_{u,1:x} | W_{1:x} \cup [K+\ell+1:N], y^{T}_{u,1:x}, S_{1:j-1}, x_{r,1:j-1}[1], z^{T}_{r,1:j-1}) \leq \mathcal{h}(z^{T}_{u,1:x}) = sT \log(2\pi e)
\]

and

\[
H(W_{s+1:K+i-1} | W_{1:x} \cup [K+\ell+1:N], y^{T}_{u,1:x}, z^{T}_{u,1:x}, S_{1:j-1}, x_{r,1:j-1}[1], z^{T}_{r,1:j-1}) \leq L \epsilon
\]

due to Fano’s inequality. In the nutshell, this bound states that the files \( W_{s+1:K+i-1} \) can be resolved reliably when knowing \( z^{T}_{u,1:x} \) in addition to \( S_{1:j-1} \) and \( y^{T}_{u,1:x} \) (if \( i \geq M+2-s = \bar{s}+1 \)
(cf. Fig. 7). This works since the Markov chain $z^T \in [1:s] \rightarrow (y^T \in [s+1:K], x^T, S_{\ell}, W_{[s+1:K]}) \rightarrow W_{[K+1:K+\ell-1]}$ is applicable to the network.

Now we consider the second term of (12). Using the chain rule of discrete entropies for this term leads to

$$H(W_{[1:s]} | y^T \in [1:s], S_{\ell}, W_{[K+\ell+1:N]}) = H(W_{[s+1:K+\ell]} | y^T \in [1:s], S_{\ell}, W_{[s+1:K+\ell+1:N]}) + H(W_{[s+1:K+\ell]} | y^T \in [1:s], S_{\ell}, W_{[1:s]} \cup [K+\ell+1:N])$$

(16)

Next, we use the aforementioned steps (a), (f) and (g), respectively, to upper bound (16).

$$H(W_{[1:s]} | y^T \in [1:s], S_{\ell}, W_{[K+\ell+1:N]}) \leq LεL + H(W_{[s+1:K+\ell]} | y^T \in [1:s], S_{\ell}, W_{[1:s]} \cup [K+\ell+1:N]) \leq LεL + H(W_{[s+1:K+\ell]} | y^T \in [1:s], S_{\ell}, W_{[1:s]} \cup [K+\ell+1:N])$$

(17)

Combining (15) and (17) in (12), we get the following inequality

$$(K + \ell)L \leq sT \log(P) \left[ 1 + \frac{\log(c + 1/P)}{\log(P)} + \frac{εP}{s} \right] + \left[ \frac{sK - s + (\bar{s} - 1)}{2} \right] \mu L + LεL. \quad (18)$$

Dividing both sides of (18) by $L$, letting $L \to \infty$ and $P \to \infty$, we obtain the following class of lower bounds on the minimum NDT

$$\delta^{\ast} (\mu) \geq \delta_{LB}(\mu, \ell, s) \triangleq \frac{K + \ell - \mu(\bar{s} - s + (\bar{s} - 1)) + \ell(\ell + 1)}{s}. \quad (19)$$

By optimizing the bound $\delta_{LB}(\mu, \ell, s)$ in (19) with respect to $\ell \in [\bar{s} : M]$ and $s \in \min\{M +1, K\}$, we determine one term of the NDT lower bound in (6). Additionally, we recall that the NDT is also bounded from below by the performance of the reference interference-free system which has an NDT of 1. The maximum over these two lower bounds concludes the proof of Theorem 1. We note that the lower bound simplifies to $\delta_{LB}(0, M, 1) = K + M$ for $\mu = 0$, while it reduces to max $\{1, \delta_{LB}(1, 0, M + 1)\} = \max \{1, K/(M+1)\}$ for $\mu = 1$. These NDT lower bounds coincide with the upper bounds one would intuitively assume to be optimal in the worst-case scenario. These are respectively, unicasting $K + M$ files for $\mu = 0$ from the DeNB and zero-forcing beamforming of $K$ files to $K$ UEs from $M + 1$ identical transmitters ($M$ RNs and DeNB) at $\mu = 1$. 
V. NDT-ONE SHOT SCHEME: SYNERGISTIC INTEGRATION OF MULTICASTING AND ZERO-FORCING SCHEMES

We now develop a general one-shot scheme. We refer to a scheme to be one-shot if all receiving nodes are able to decode their desired symbols on a single channel use basis. Such schemes explicitly preclude symbol decoding over multiple channel uses. Hereby, our one-shot scheme uses a combination of the Maddah-Ali Niesen (MAN) scheme \[7\] and ZF to simultaneously convey a subset of RNs and UEs with their desired symbols. The achievability for the extreme cases of zero-cache ($\mu = 0$) and full-cache ($\mu = 1$) are based on one-shot schemes (These are, respectively, unicasting and zero-forcing.) and readily follow from Lemma [1] Thus in the sequel, we consider fractional cache sizes that are strictly larger than zero and strictly less than one. Henceforth, we explicitly assume $M \geq 2$.

To this end, let us consider the achievability at fractional cache sizes $\mu \in \{1/M, 2/M, 3/M, \ldots, (M-1)/M\}$. We split each file $W_n, \forall n \in [N]$, into $\Gamma(M \mu M)$ symbols, where $\Gamma = \binom{K}{\psi}$ and $\psi = \min\{K, \mu M\}$. Symbols of every file are labeled according to $W_n = (\eta_{n,T,U} : T \subset [M], |T| = \mu M, U \subset [K], |U| = \psi)$.

In the placement phase, for each $n \in [N]$, symbol $\eta_{n,T,U}$ is prefetched at RN$_m$ if $m \in T$. Thus, each RN caches a total of $N\Gamma(M-1)/(\mu M)$ symbols. It is easy to see that the cache constraint at each RN is satisfied since

$$\frac{\text{# of cached symbols}}{\text{# of total symbols}} = \frac{N\Gamma(M-1)/(\mu M)}{N\Gamma(M \mu M)} = \mu.$$ 

Next, we describe the delivery phase. Consider for the sake of notational simplicity, the worst-case request vector $$(1, 2, \ldots, K, K + 1, K + 2, \ldots, K + M)$$ for $N \geq K + M$. At each time instant $t$, we focus on the delivery of desired symbols for a subset $S_R \subset [M]$ of $|S_R| = 1 + \mu M$ RNs and $S_U \subset [K]$ of $|S_U| = \psi$ UEs. Hereby, the UEs are served by
a subset of $|S'_R| = \psi$ RNs belonging to the set $S'_R \subset [M]$. To this end, DeNB and RN transmit the following signals:

$$x_k[t] = \sum_{m \in S_R} v_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t] \eta_{K+m,S_R\setminus\{m\}},S_U$$

(20)

$$x_{r,m'}[t] = \begin{cases} 
\sum_{m \in S_R} \beta^{(m')}_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t] \eta_{K+m,S_R\setminus\{m\}},S_U + \sum_{p \in S_U} \beta^{(m')}_{\eta_{p,S_R\setminus\{m\}},S_U}[t] \eta_{p,S_R\setminus\{m\}},S_U & \text{if } m' \in (S_R \cap S'_R) \\
\sum_{m \neq m'} \beta^{(m')}_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t] \eta_{K+m,S_R\setminus\{m\}},S_U & \text{if } m' \in (S_R \setminus S'_R) \\
0 & \text{otherwise}
\end{cases}$$

(21)

In these equations, the precoders for symbol $\eta_{h,T,S_U}$ originating from the DeNB and RN are denoted by $v_{\eta_{h,T,S_U}}$ and $\beta^{(m)}_{\eta_{h,T,S_U}}$, respectively. Decoding at the RNs in $S_R$ follows along the standard MAN manner. That is, each RN $m' \in S_R$ exploits its knowledge of symbols $\eta_{K+m,S_R\setminus\{m\}},S_U$ for all $m \in S_R \setminus \{m'\}$ to recover its desired symbol $\eta_{K+m',S_R\setminus\{m'\}},S_U$. Thus, we shift our focus to the UEs. The received signal at UE $k$ is specified by the equation

$$y_{u,k}[t] = \sum_{m \in S_R} \eta_{K+m,S_R\setminus\{m\}},S_U \left( g_k[t] v_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t] + \sum_{m' \in S_R} h_{km}[t] \beta^{(m')}_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t] \right) +$$

\[ + \sum_{p \in S_U} \sum_{m' \in S'_R} h_{km'}[t] \beta^{(m')}_{\eta_{p,S_R\setminus\{m\}},S_U}[t] \eta_{p,S_R\setminus\{m\}},S_U \]  

(22)

$$= \sum_{m \in S_R} \eta_{K+m,S_R\setminus\{m\}},S_U \left( g_k[t], h_{k,S_R\setminus\{m\}}[t] \right) \left( v_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t], \beta^{(m')}_{\eta_{K+m,S_R\setminus\{m\}},S_U}[t] \right)^\dagger +$$

\[ + \sum_{p \in S_U} h_{k,S_R}[t] \beta^{(m')}_{\eta_{p,S_R\setminus\{m\}},S_U}[t] \eta_{p,S_R\setminus\{m\}},S_U \]  

(23)

This equation can be rewritten in a compact form by making the following definitions.

First, we define the collection of channel coefficients from RNs in the set $\mathcal{W}$ to UE $k$ as the vector

$$\mathbf{h}_{k;W}[t] \doteq \{ h_{km}[t] \}_{m \in \mathcal{W}}, \quad \mathbf{h}_{k;W}[t] \in \mathbb{C}^{||\mathcal{W}||}.$$  

Similarly, channel coefficients from RNs in the set $\mathcal{W}$ to UEs in the set $\mathcal{U}$ are denoted by the channel matrix

$$\mathbf{H}_{\mathcal{U};W}[t] \doteq \{ h_{k;W}[t] \}_{k \in \mathcal{U}}, \quad \mathbf{H}_{\mathcal{U};W}[t] \in \mathbb{C}^{||\mathcal{U}|| \times ||\mathcal{W}||}.$$
Second, we concatenate the precoders of the RNs for symbol $\eta_{n,T,U}$ to the vector
\[ \beta_{\eta_{n,T,U}}^B[t] \triangleq \{ \beta_{\eta_{n,T,U}}^{(m)}[t] \}_{m \in B}, \quad \beta_{\eta_{n,T,U}}^B[t] \in \mathbb{C}^{|B|}. \]

When $B = T$, we simply write $\beta_{\eta_{n,T,U}}[t]$ instead of $\beta_{\eta_{n,T,U}}^B[t]$. These definitions are also applicable to the channel coefficients $g_k[t]$ as well. They allow us to rewrite Eq. (22) to (23).

Recall that $\text{UE}_k$ is only provided with its desired symbol $\eta_{k,S_U} \in S_U$ as long as $k \in S_U$. To this end, all interferences in the concatenated vector $y_{u,S_U}[t] = \{y_{u,k}\}_{k \in S_U}$ given by
\[ y_{u,S_U}[t] = \sum_{m \in S_R} \eta_{K+m,S_R \setminus \{m\},S_U}[t] \left( g_{S_U}[t], H_{S_U,S_R \setminus \{m\},S_U}[t] \right) \left( \nu_{\eta_{K+m,S_R \setminus \{m\},S_U}[t], \beta^\dagger_{\eta_{K+m,S_R \setminus \{m\},S_U}[t]} \right)^\dagger + \]
\[ \underbrace{\sum_{p \in S_U} H_{S_U,S_R'[p],S_U}[t] \beta_{\eta_{p,S_R'[p],S_U}}[t] \eta_{p,S_R'[p],S_U}}_{\psi-1 \text{ interferences and 1 desired component for } \text{UE}_k, k \in S_U} \]

have to be zero-forced. This is equivalent to
\[ (g_{S_U}[t], H_{S_U,S_R \setminus \{m\},S_U}[t]) \tilde{\beta}_{\eta_{K+m,S_R \setminus \{m\},S_U}}[t] = 0_{|S_U|}, \] \[ \forall m \in S_R \quad \quad \tag{25} \]
\[ \forall m \in S_R \text{ and} \]
\[ H_{S_U \setminus \{p\},S_R'[p],S_U}[t] \beta_{\eta_{p,S_R'[p],S_U}}[t] = 0_{|S_U|-1}, \] \[ \forall p \in S_U. \text{ Note that we used} \]
\[ \tilde{\beta}_{\eta_{K+m,S_R \setminus \{m\},S_U}}[t] = \left( \nu_{\eta_{K+m,S_R \setminus \{m\},S_U}[t], \beta^\dagger_{\eta_{K+m,S_R \setminus \{m\},S_U}[t]} \right)^\dagger \]

in Eq. (25) for reasons of compactness. It is easy to see that
\[ \text{rank} \left( (g_{S_U}[t], H_{S_U,S_R \setminus \{m\},S_U}[t]) \right) = \psi, \] \[ \text{rank} \left( H_{S_U \setminus \{p\},S_R'[p],S_U}[t] \right) = \psi - 1, \] \[ \forall m \in S_R, \forall p \in S_U. \text{ The nullspace dimension for these two matrices thus become} \]
\[ |S_R| - \psi = 1 + (\mu M - K)^+, \] \[ |S_R'| - (\psi - 1) = 1 \text{ for } \psi \geq 2. \]
When \( \psi = 1 \), we note that there are no interference terms \( (\psi - 1 = 0) \) in the second sum of Eq. (24), which makes zero-forcing for this component obsolete. We choose the precoding vectors in Eqs. (25) and (26) such that

\[
\tilde{\beta}_{\eta_{K+m,S_R \setminus \{m\}}, S_U} [t] \in \mathcal{N} \left( \left( \mathbf{g}_{S_U} [t], \mathbf{H}_{S_U \setminus \{m\}} [t] \right) \right),
\]

\[
\beta_{\eta_{p,S_R \setminus \{p\}}, S_U} [t] \in \mathcal{N} \left( \mathbf{H}_{S_U \setminus \{p\}} [t] \right),
\]

where \( \mathcal{N}(\mathbf{A}) \) denotes the (right) nullspace of \( \mathbf{A} \). Consequently, all UEs in the subset \( S_U \) will be free from interference. We conclude that in a single channel use \( 1 + \mu M \) RNs in \( S_R \) and \( \psi \) UEs in \( S_U \) were able to decode their desired symbols through a combination of ZF and the MAN scheme.

Assume that we deploy this scheme for \( T_1 = \Gamma \left( \frac{M}{1 + \mu M} \right) \) channel uses. The probability that RN\(_m\) is served at the \( t \)-th channel use, \( t \in [T_1] \), with its desired symbol \( \eta_{K+m,S_R \setminus \{m\}}, S_U \) is determined whether \( m \in S_R \). In \( T_1 \) channel uses this happens \((M-1)/(\mu M)\)-fraction of the time \( T_1 \). Thus, RN\(_m\), \( \forall m \in [M] \), receives uncached symbols \( \eta_{K+m,S_R \setminus \{m\}}, S_U \) of the requested file \( W_{K+m} \). Recall that RN\(_m\) has also

\[
N_{\text{RN}}^c = T_1 \frac{(M-1)}{\mu M} = \Gamma \left( \frac{M - 1}{\mu M} \right)
\]

uncached symbols \( \eta_{K+m,S_R \setminus \{m\}}, S_U \) of the requested file \( W_{K+m} \). Using uncached and cached symbols, RN\(_m\) can reconstruct file \( W_{K+m} \) of size

\[
|W_{K+m}| = N_{\text{RN}}^c + N_{\text{RN}}^c = \Gamma \left( \frac{M}{\mu M} \right)
\]

symbols. With respect to the RNs, we conclude that \( T_1 \) channel uses suffice to allow them to retrieve their desired files \( W_{K+m}, \forall m \in [M] \). We now shift our focus to the \( K \) UEs. In \( \tilde{T}_1 \leq T_1 \) channel uses, where

\[
\tilde{T}_1 = \min \left\{ T_1, \Gamma \frac{K}{\psi} \left( \frac{M}{\mu M} \right) \right\},
\]

\footnote{This fraction can be interpreted as the probability \( p(m \in S_R) \).}
we try to provide each UE with the same number of desired symbols. To this end, UE\(_k\) obtains its desired symbols \(\eta_{k,s_k}^{R_k,U} (\frac{K-1}{\psi})\)-fraction of time duration \(\tilde{T}_1\). Hereby, RN\(_m\), \(\forall m \in [M]\), is used

\[
N_{RN,Tx} = \tilde{T}_1\left(\frac{\psi - 1}{M}\right) = \Gamma \min \left\{\frac{\psi}{M} \left(\frac{M}{1 + \mu M}\right), \frac{K}{\mu M}\right\}
\]

times for transmission\(^7\) of these symbols such that every UE receives

\[
N_{UE} = \tilde{T}_1\left(\frac{K - 1}{\psi}\right) = \min \left\{\left(\frac{M}{\mu M + 1}\right)(K - 1), \Gamma\left(\frac{M}{\mu M}\right)\right\}
\] (30)
symbols. Depending on whether

\[
|W_k| - N_{UE} \geq 0,
\] (31a)

or

\[
\frac{K}{\psi} = \frac{\binom{K}{\psi - 1}}{\binom{K - 1}{\psi - 1}} \leq \frac{\binom{M}{1 + \mu M}}{\binom{M}{\mu M}} = \frac{M - \mu M}{1 + \mu M},
\] (31b)

\(\tilde{T}_1\) channel uses are sufficient or insufficient for the delivery of files \(W_k\) to UE\(_k\), \(\forall k \in [K]\). We state the following conditions (see also Fig. [3]) with respect to the inequalities in (31). First, when \(\psi = K\), \(|W_k| > N_{UE}\) holds if

\[
\begin{cases}
K \leq \mu M < M < \frac{1}{1 - 2\mu}, \mu \leq \frac{1}{2} & \text{(Region A\(_1\))} \\
K \leq \mu M \leq M, M > \frac{1}{1 - 2\mu}, \mu > \frac{1}{2} & \text{(Region A\(_2\))}
\end{cases}
\] (32)

while \(|W_k| = N_{UE}\) is satisfied if

\[
K \leq \mu M, \frac{1}{1 - 2\mu} \leq M, \mu \leq \frac{1}{2} \quad \text{(Region B)}.
\] (33)

We denote the union of regions A\(_1\) and A\(_2\) as A. Second, when \(\psi = \mu M\), \(|W_k| > N_{UE}\) is met if

\[
\begin{cases}
\mu M < M < K \quad \text{(Region C)} \\
\max \left\{\frac{\mu M^2(1 - \mu)}{1 + \mu M}, \mu M\right\} < K \leq M \quad \text{(Region D)}
\end{cases}
\] (34)

whereas on the other hand \(|W_k| = N_{UE}\) if

\[
\mu M < K \leq \frac{\mu M^2(1 - \mu)}{1 + \mu M} \leq M \quad \text{(Region E)}.
\] (35)

\(^7\)If \(N_{RN,Tx}\) is not an integer, file symbols require further fragmentation and rate splitting ought to be applied.
Fig. 8: 2D ($\mu, M$)-plot of all regions described by inequalities (32)–(35) for $K = 2$. Specifically, the regions described by inequalities (32), (33) and (35) are termed Region A, B and E, respectively. The two inequalities in (34) are represented by Region C and D. The discrete points illustrate the fractional cache sizes $\mu = m/M$, $m \in [M]$ for an integer $M$.

Thus, for cases where $|W_k| = N_{\text{UE}}$ (under conditions (33) or (35)), $T_1$ channel uses suffice to meet the demands of all $M$ RNs and $K$ UEs. The achievable NDT under these cases results in the NDT of the MAN scheme given by

$$\delta_{\text{MAN}}(\mu) = M \cdot (1 - \mu) \cdot \frac{1}{1 + \mu M}$$

(36)

under conditions (33) or (35) at fractional cache sizes $\mu \in \{1/M, 2/M, \ldots, (M-1)/M\}$. For the cases where $|W_k| > N_{\text{UE}}$ (i.e., (32) or (34)), however, $T_2$ additional channel uses are required to convey each UE with its remaining $|W_k| - N_{\text{UE}}$ desired symbols. Hereby, in each channel use, $\psi' = \min\{K, 1 + \mu M\}$ UEs can be provided with their desired symbols through applying ZF beamforming for a $(\psi', K)$ MISO broadcast channel such that

$$T_2 = \frac{K(|W_k| - N_{\text{UE}})}{\psi'}.$$  

(37)

As opposed to the first block of $T_1$ channel uses, both RNs and DeNB are now involved in providing the UEs with their desired symbols. This is due to the fact that the delivery of the RNs requested files is terminated. The achievable NDT for the cases (32) and (34) are given by

$$\delta(\mu) = \frac{T_1 + T_2}{\Gamma(M_{\psi})}.$$
Simplification of this term leads to the following expression

$$\delta(\mu) = \begin{cases} 
\delta_{\text{MISO-BC}} = 1 & \text{for (32)} \\
\frac{K + \delta_{\text{MAN}}(\mu)}{\psi'} & \text{for (34)}
\end{cases},$$

(38)

where $\delta_{\text{MAN}}(\mu)$ is the NDT expression given in Eq. (36) and $\delta_{\text{MISO-BC}}$ is the NDT of the $(\psi, K)$ MISO BC if $\psi = K$. Combining the two NDT expressions in Eqs. (36) and (38) to a single one generates the achievable one-shot NDT $\delta_{\text{OS}}(\mu)$ of Theorem 2.

VI. NDT-OPTIMAL SCHEMES FOR SPECIAL INSTANCES: INTEGRATION OF SUBSPACE ALIGNMENT, MULTICASTING AND ZERO-FORCING

In this section, we present our novel achievability schemes that combine the well-known schemes – interference alignment, multicasting and zero-forcing – for special instances of $M \in \{1, 2, 3\}$ and $K \in \{1, 2, 3\}$ that satisfy $K + M \leq 4$. For these special instances, these schemes allow us to fully characterize the optimal NDT cache-memory tradeoff for any $\mu \in [0, 1]$.

We start with the case where $M = 1$.

A. Achievability for $M = 1$

The NDT-optimal scheme for $M = 1$ and $K \leq 3$ is presented. Hereby, the following proposition quantifies the achievable NDT.

**Proposition 1.** The achievable NDT of the network under study for $M = 1$ RNs, $K \in \{1, 2, 3\}$ UEs and $\mu \in [0, 1]$ is given by

$$\delta(\mu) = \begin{cases} 
K + 1 - \mu K & \text{if } K \leq 2 \\
\max\left\{ K + 1 - \mu K, \frac{K + 1 - \mu}{2} \right\} & \text{if } K = 3
\end{cases}.$$ 

(39)

In what follows the proof of this proposition is outlined. As shown in Fig. 2 in the case, where $K \leq 2$, the lower bound $\delta_{\text{LB}}(\mu, 1, 1)$ coincides with the achievable NDT. Lemma 1 establishes the achievability for the corner points at $\mu = 0$ and $\mu = 1$. Hereby, due to arguments of memory sharing, intermediate points at fractional cache sizes $0 < \mu < 1$ become achievable through successively time-sharing between unicasting and zero-forcing for $\mu$ and $(1 - \mu)$ file fractions, respectively.

However, for $K = 3$ such type of memory sharing scheme is in fact suboptimal. This is indicated when comparing the achievable NDT of memory sharing between the zero-cache and
Fig. 9: The NDT lower bound for $M = 1$ and $K \geq 3$ is shown by the solid line. For $K = 3$, this line is in fact achievable. The dashed line shows the achievable NDT of a time-sharing based unicasting-zero-forcing scheme.

$\mu_1 = K + \frac{1}{2K - 1}$

Full-cache scheme with the coinciding lower bound for $K = 3$ illustrated in Fig. 9. In fact, we will show that simultaneously combining subspace alignment with zero-forcing through appropriate precoder design allows us to show the achievability of the corner point $\left( \frac{K+1}{2K-1}, \frac{K^2-1}{2K-1} \right) = \left( \frac{4}{3}, \frac{8}{5} \right)$ for $K = 3$. To this end, we describe, respectively, (i) the RN cache placement for $\mu = 4/5$, (ii) the encoding at DeNB and RN ($M = 1$), (iii) and finally the decoding at RN and the three UEs ($K = 3$). We begin with the cache placement.

**RN Cache Placement:** Assume without loss of generality one possible worst-case demand scenario for $N = 4$. That is, the UEs request files $W_1, W_2$ and $W_3$ while the RN is interested in

| File W1 | $\eta_{1,1}$ | $\eta_{1,2}$ | $\eta_{1,3}$ | $\eta_{1,4}$ |
|---------|-------------|-------------|-------------|-------------|
| File W2 | $\eta_{2,1}$ | $\eta_{2,2}$ | $\eta_{2,3}$ | $\eta_{2,4}$ |
| File W3 | $\eta_{3,1}$ | $\eta_{3,2}$ | $\eta_{3,3}$ | $\eta_{3,4}$ |
| File W4 | $\eta_{4,1}$ | $\eta_{4,2}$ | $\eta_{4,3}$ | $\eta_{4,4}$ |

| RN & DeNB | DeNB only |
|-----------|-----------|
| $\eta_{1,5}$ | $\eta_{1,5}$ |
| $\eta_{2,5}$ | $\eta_{2,5}$ |
| $\eta_{3,5}$ | $\eta_{3,5}$ |
| $\eta_{4,5}$ | $\eta_{4,5}$ |

Fig. 10: Requested files by $K = 3$ users and $M = 1$ RNs and the availability exclusively at the DeNB or both at the DeNB and the RN. The highlighted symbols are zero-forcing symbols, i.e., the (spatial) availability at both the DeNB and the RN is used to zero-force these symbols at one of the $K = 3$ UEs.
file W4. According to Fig. 10, all files are broken into parts of 5 symbols each. For the sake of simplicity, it suffices to focus on single file parts. Thus, the remaining discussion focuses on 5 symbols per file denoted by $\eta_{i,1}, \eta_{i,2}, \eta_{i,3}, \eta_{i,4}$ and $\eta_{i,5}$ with respect to the $i$-th file. For notational simplicity, we stack these symbols to the vector $\eta_i = (\eta_{i,1}, \eta_{i,2}, \ldots, \eta_{i,5})^\dagger$. The vector $\eta$ on the other hand, concatenates all $\eta_i, i = 1, \ldots, 4$, to a single column vector, i.e., $\eta = (\eta_1^\dagger, \eta_2^\dagger, \eta_3^\dagger, \eta_4^\dagger)^\dagger$. All symbols in $\eta$ are available at the DeNB. However, as far as the RN is concerned, only the first four symbols of all files are locally available in its cache. Since, the RN is interested in file W4 and it knows $\eta_{4,1}, \eta_{4,2}, \eta_{4,3}$ and $\eta_{4,4}$ already, the only missing symbol it desires is $\eta_{4,5}$. Thus, the transmission policy has to be designed such that all symbols of files W1, W2 and W3 as well as $\eta_{4,5}$ are conveyed jointly or exclusively by DeNB and/or RN. These are in total 16 information symbols. Next, we describe the encoding strategy for both DeNB and RN.

| ZF map |  
|--------|  
| UE1    | $\eta_{2,1}$ $\eta_{2,2}$ $\eta_{3,3}$  
| UE2    | $\eta_{3,1}$ $\eta_{3,2}$ $\eta_{1,3}$  
| UE3    | $\eta_{1,1}$ $\eta_{1,2}$ $\eta_{2,3}$  

Fig. 11: Map that assigns which symbol to zero-force at which UE.

**Encoding at DeNB and RN:** The transmission strategy will exploit the correlation that arises between the availability of shared symbols at RN and DeNB by leveraging zero-forcing (ZF) opportunities while simultaneously facilitating (subspace) interference alignment (IA) at the UEs. This is why our scheme (as shown in Fig. 10) only zero-forces symbols $\eta_{1,1}, \eta_{1,2}, \eta_{1,3}, \eta_{2,1}, \eta_{2,2}, \eta_{2,3}$ and $\eta_{3,1}, \eta_{3,2}, \eta_{3,3}$. Symbols $\eta_{1,4}, \eta_{2,4}$ and $\eta_{3,4}$ are not zero-forced but are instead used to enable alignment amongst others with $\eta_{4,5}$ at the UEs. The map that assigns which symbol is zero-forced at which UE is given in Figure 11. To this end, DeNB and RN form their transmit signals according to

$$x_s[t] = v^\dagger[t]\eta$$
$$x_r[t] = \beta^\dagger[t]C_{RN}\eta$$

IA is facilitated by the fact that the DeNB does not transmit these symbols (even though it knows them). Thus, effectively, the DeNB does not need to be aware of $\eta_{i,4}, i \in [N]$. This is accounted for in Eq. (43) for $N = 4$. 

---

8IA is facilitated by the fact that the DeNB does not transmit these symbols (even though it knows them). Thus, effectively, the DeNB does not need to be aware of $\eta_{i,4}, i \in [N]$. This is accounted for in Eq. (43) for $N = 4$. 

---
∀t ∈ [T] for T = 8, respectively. In these two equations, the vectors ν[t] and β[t] denote the precoding vectors at time instant t for DeNB and RN with respect to all symbols in η, whereas \( C_{RN} \in \mathbb{R}^{20 \times 20} \) is a binary (caching) matrix accounting for the cache placement of the RN. The elements of the precoding vectors ν[t] and β[t] comprise of complex precoding scalars \( ν_{η,i}[t] \) and \( β_{η,i}[t] \) of symbol \( η_{i,j} \). These elements are stacked to \( ν[t] \) and \( β[t] \) which themselves are then concatenated to \( ν[t] \) and \( β[t] \) in the exact same fashion as \( η_i \) in \( η \). In (40) and (41), we have implicitly fixed the following elements of \( ν[t] \) and \( β[t] \) to be

\[
ν_{η,i}[t] = β_{η,i}[t] = 0, \quad j = 1, \ldots, 4 \\
ν_{η,i}[t] = 0, \quad i = 1, \ldots, 4
\]

∀t ∈ [T]. Further, we fix the RN caching matrix \( C_{RN} \) in accordance to Fig. 10 and the definition of \( η \) to

\[
C_{RN} = I_4 \otimes \begin{pmatrix} I_4 & 0 \\ 0 & 0 \end{pmatrix}.
\]

(44)

From Eq. (43), we infer that the elements \( ν_{η,i}[t] \) are 0, i.e., irrelevant in \( ν[t] \), which is equivalent to introduce an effective caching matrix \( C^\text{eff}_{DeNB} \)

\[
C^\text{eff}_{DeNB} = I_4 \otimes \begin{pmatrix} I_3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.
\]

(45)

that ignores the availability of symbols \( η_{i,4}, i = 1, \ldots, 4 \), at the DeNB.\(^9\) In consequence, we may rewrite (40) as follows

\[
x_{u}[t] = ν^\dagger[t]C^\text{eff}_{DeNB}η.
\]

(46)

where we now implicitly assumed that \( ν_{η,i}[t] \neq 0 \). For ease of presentation, we first write the received signal at UEk, \( k \in [K] = [3] \) according to

\[
y_{u,k}[t] = g_k[t]x_{s}[t] + h_{k1}[t]x_{r}[t] + z_{u,k}[t]
\]

\[
\quad = \left( g_k[t]ν^\dagger[t]C^\text{eff}_{DeNB} + h_{k1}[t]β^\dagger[t]C_{RN}\right)η + z_{u,k}[t]
\]

\[
\quad = \sum_{i=1}^{3} \left( g_k[t]ν^\dagger_{i,[1:3]}[t] + h_{k1}[t]β^\dagger_{i,[1:3]}[t] \right)I_3η_i,[1:3] + h_{k1}[t]β_{η,i,[1:3]}[t]η_{i,4}
\]

\[
\quad + \sum_{\ell=1}^{4} g_k[t]ν_{η,\ell}[t]η_{\ell,5} + z_{u,k}[t].
\]

(47)

\(^9\)From this argument, we infer that the effective cache size of the DeNB is \( ρ = \frac{4}{3} \).
In (47), all components for which \( i, \ell \neq k \) represent interference. Further, we observe that the effective channel coefficient of the \( j \)-th ZF symbol of the \( i \)-th file at UE\( k \) corresponds to 
\[
e_{\eta_k+j,k}^{ZF}[t] = g_k[t]v_{\eta_k+j}[t] + h_{k1}[t]\beta_{\eta_k+j}[t] \].
\[
\text{The vectors } v[t] \text{ and } \beta[t] \text{ are chosen such that both ZF and IA at the UEs become feasible. According to the ZF map of Fig. [11] the ZF conditions at UE}_k \text{ become}
\[
e_{\eta_k+1,k}^{ZF}[t] = g_k[t]v_{\eta_k+1}[t] + h_{k1}[t]\beta_{\eta_k+1}[t] = 0, \quad (48a)
\]
\[
e_{\eta_k+2,k}^{ZF}[t] = g_k[t]v_{\eta_k+2}[t] + h_{k1}[t]\beta_{\eta_k+2}[t] = 0, \quad (48b)
\]
\[
e_{\eta_k+3,k}^{ZF}[t] = g_k[t]v_{\eta_k+3}[t] + h_{k1}[t]\beta_{\eta_k+3}[t] = 0. \quad (48c)
\]
For the sake of compact notation, we have used modulo-\( K \) indexing with respect to symbols \( \eta_{k+1,1}, \eta_{k+1,2} \) and \( \eta_{k+2,1} \) in the formulation of the ZF conditions (48). Simultaneously, we design the precoding scalars \( v_{\eta_k,j}[t] \) and \( \beta_{\eta_k,j}[t] \) such that the interference at each UE is aligned into a three-dimensional signal space. (The remaining 5 dimensions are reserved for the 5 symbols of the desired file.) The interference graph in Fig. [12] shows which symbols align with each other at which UE. This graph consists of 3 layers. In the first layer, two symbols, namely \( \eta_{4,5} \) and \( \eta_{1,4}, \eta_{2,4} \) or \( \eta_{3,4} \) align at the three UEs. At layers two and three, on the other hand, three symbols align per UE. Symbols \( \eta_{1,4}, \eta_{2,4} \) and \( \eta_{3,4} \) link layers 1 and 2, while \( \eta_{1,5}, \eta_{2,5} \) and \( \eta_{3,5} \) connect layers 2 and 3. In analogy to the graph in Fig. [12], the alignment conditions at UE\( k \) can be written as 
\[
g_k[t]v_{\eta_4}[t] = h_{k1}[t]\beta_{k+1,4}[t] \quad (49)
\]
for Layer 1,
\[
h_{k1}[t]\beta_{k+2,4}[t] = e_{\eta_k+2,4,k}^{ZF}[t] = g_k[t]v_{\eta_k+1,4}[t] \quad (50)
\]
for Layer 2, and
\[
g_k[t]v_{\eta_k+2,5}[t] = e_{\eta_k+2,5,k}^{ZF}[t] = e_{\eta_k+1,3,k}^{ZF}[t] \quad (51)
\]
for Layer 3. Under the given ZF and IA conditions (cf. (48) and (49)–(51)), the precoders are functions of the channels \( g[t] \) and \( H[t] \). We fix the precoder for symbol \( \eta_{4,5} \) to 
\[
v_{\eta_{4,5}}[t] = j_{13}[t]j_{23}[t]g_1[t]g_2[t]g_3[t]h_{11}[t]h_{21}[t]h_{31}[t], \quad (52)
\]
Fig. 12: Interference alignment graph for the achievability at corner point \((\frac{3}{8}, \frac{3}{8})\) for \(M = 1\) and \(K = 3\). The graph consists of three (subspace) alignment chains. By definition, the first alignment chain is the path from the node \(\eta_{4,5}\) to \(\eta_{3,1}\), the second from \(\eta_{4,5}\) to \(\eta_{1,1}\) and the third from \(\eta_{4,5}\) to \(\eta_{2,1}\).

where

\[
\begin{align*}
  j_{13}[t] &= g_2[t]h_{31}[t] - g_3[t]h_{21}[t], \\
  j_{23}[t] &= g_3[t]h_{11}[t] - g_1[t]h_{31}[t], \\
  j_{33}[t] &= g_1[t]h_{21}[t] - g_2[t]h_{11}[t].
\end{align*}
\] (53a–53c)

The remaining precoders scalars depend on \(v_{\eta_{4,5}}[t]\) and can be computed by using (52) in (48) and (49)–(51). To this end, we compute the precoders along the \(r\)-th alignment chain \([44], r \in [3]\), of the graph (the first chain in Fig. 12 for instance being the entire path from node \(\eta_{4,5}\) to \(\eta_{3,1}\)) as a function of \(v_{\eta_{4,5}}[t]\). The resulting precoders under modulo-\(K\) indexing then become:

\[
\begin{align*}
  \beta_{\eta_{4,4}}[t] &= v_{\eta_{4,5}}[t] \cdot \frac{g_{r+2}[t]}{h_{r+2,1}[t]}, \\
  p_{\eta_{r,2}}[t] &\triangleq \left( \frac{v_{\eta_{r,2}}[t]}{\beta_{\eta_{r,2}}[t]} \right) = v_{\eta_{4,5}}[t] \cdot \frac{g_{r+2}[t]h_{r+1,1}[t]}{j_{r,3}[t]} \left( \frac{1}{-\frac{g_{r+2}[t]}{h_{r+2,1}[t]}} \right), \\
  v_{\eta_{r+2,5}}[t] &= v_{\eta_{4,5}}[t] \cdot \frac{g_{r+2}[t]h_{r+1,1}[t]}{h_{r+2,1}[t]g_{r+1}[t]}, \\
  p_{\eta_{r+1,3}}[t] &\triangleq \left( \frac{v_{\eta_{r+1,3}}[t]}{\beta_{\eta_{r+1,3}}[t]} \right) = v_{\eta_{4,5}}[t] \cdot \frac{g_{r+2}[t]h_{r+1,1}[t]g_{r}[t]}{g_{r+1}[t]j_{r+1,3}[t]} \left( \frac{-1}{\frac{g_{r+2}[t]}{h_{r+2,1}[t]}} \right).
\end{align*}
\] (54a–54d)
Recall that the scheme spans over $T$ channel uses. Thus, due to the time-variant nature of the wireless channel, the UEs and the RN have 8 (noise-corrupted) linear independent observations.

At the RN, on the other hand, the knowledge of $\eta_{i,j}, i, j \in [3]$, as side information prefetched in its cache is exploited to cancel the contribution of these components. Thus, at the $t$-th channel use, the RN observes

$$y_{r,1}^t = f_t^t \left( v_{\eta_{4,5}}^t \eta_{4,5} + \sum_{i=1}^{3} v_{\eta_{i,5}}^t [t] \eta_{i,5} \right) + z_{r,1}^t. \quad (59)$$

Recall that the scheme spans over $T = 8$ channel uses.
\{y_{u,k}[I]\}_{t=1}^{8} \text{ and } \{y'_{r,1}[I]\}_{t=1}^{8} \text{ (in the field of reals } \mathbb{R} \text{ or complex } \mathbb{C} \text{) according to (56) and (59), respectively. In consequence, UE}_k, k \in [3], \text{ on the one hand, is able to decode its } 5 \text{ desired symbols}

- \eta_{k,1}, \eta_{k,2}, \eta_{k,3}, \eta_{k,4}, \eta_{k,5}

and 3 aligned interfering symbols

- \eta_{4,5} + \eta_{k+1,4},
- \eta_{k+2,4} + \eta_{k+2,2} + \eta_{k+1,5}
- \text{and } \eta_{k+2,5} + \eta_{k+1,3} + \eta_{k+2,1},

while the RN, on the other hand, decodes its desired symbol \eta_{4,5} \text{ and } 3 \text{ interfering (but not aligned) symbols } \eta_{1,5}, \eta_{2,5} \text{ and } \eta_{3,5}. \text{ In fact, as far as the RN is concerned, it actually only requires } 4 \text{ channel uses to allow for the decoding of the aforementioned symbols. The UEs are the reason why the scheme spans over } 8, \text{ and not } 4, \text{ channel uses. Consequently, the achievable NDT becomes } \frac{8}{5}. \text{ Next, we present the NDT-optimal schemes for which } M = 2 \text{ and } K \leq 2.

B. Achievability for } M = 2

In the following, we will outline the delivery time optimal schemes for } M = 2 \text{ and } K \leq 2. \text{ The following proposition states the achievable NDT.}

**Proposition 2.** The achievable NDT of the network under study for } M = 2 \text{ RNs, } K \in \{1, 2\} \text{ UEs and } \mu \in [0, 1] \text{ is given by}

\[
\delta(\mu) = \begin{cases} 
\max \left\{1, 3 - 4\mu\right\} & \text{if } K = 1 \\
\max \left\{4 - 6\mu, \frac{4 - 3\mu}{2}, \frac{3 - \mu}{2}\right\} & \text{if } K = 2
\end{cases}
\]  

(60)

Now, we will present the proof of this proposition. Fig. 5 shows that the achievable NDT in Proposition 2 coincides with the lower bound. Due to Lemma 1 and arguments of memory sharing, we will only establish the achievability of the corner point(s) \(\left(\frac{1}{2}, 1\right)\) for } K = 1 \text{ and } \(\left(\frac{4}{5}, \frac{5}{2}\right), \left(\frac{1}{2}, \frac{5}{3}\right)\) for } K = 2, \text{ respectively. In this context, the NDT-optimal schemes for corner points } \(\left(\frac{1}{2}, 1\right)\) \text{ and } \(\left(\frac{1}{2}, \frac{5}{3}\right)\) \text{ are based on the one-shot}^{10} \text{ reception of desired symbols, whereas at corner point } \(\left(\frac{4}{5}, \frac{5}{2}\right)\) \text{ symbol decoding over multiple channel uses (similarly to the achievability for } (K, M) = (3, 1) \text{ in subsection VI-A) is required. We start the achievability for the latter corner point } \(\left(\frac{4}{5}, \frac{5}{2}\right)\).

\(^{10}\text{One-shot in this context is equivalent to requiring a single channel use for conveying desired symbols to the receiving nodes.}\)
\begin{center}
\begin{tabular}{|c|c|c|c|c|c|}
\hline
 & RN$_1$ & DeNB & RN$_2$ & DeNB & DeNB only \\
\hline
File W$_1$ & $\eta_{1,1}$ & $\eta_{1,2}$ & $\eta_{1,3}$ & $\eta_{1,4}$ & $\eta_{1,5}$ & $\eta_{1,6}$ & $\eta_{1,7}$ & $\eta_{1,8}$ & $\eta_{1,9}$ \\
File W$_2$ & $\eta_{2,1}$ & $\eta_{2,2}$ & $\eta_{2,3}$ & $\eta_{2,4}$ & $\eta_{2,5}$ & $\eta_{2,6}$ & $\eta_{2,7}$ & $\eta_{2,8}$ & $\eta_{2,9}$ \\
File W$_3$ & $\eta_{3,1}$ & $\eta_{3,2}$ & $\eta_{3,3}$ & $\eta_{3,4}$ & $\eta_{3,5}$ & $\eta_{3,6}$ & $\eta_{3,7}$ & $\eta_{3,8}$ & $\eta_{3,9}$ \\
File W$_4$ & $\eta_{4,1}$ & $\eta_{4,2}$ & $\eta_{4,3}$ & $\eta_{4,4}$ & $\eta_{4,5}$ & $\eta_{4,6}$ & $\eta_{4,7}$ & $\eta_{4,8}$ & $\eta_{4,9}$ \\
\hline
\end{tabular}
\end{center}

ZF symbols

Fig. 13: Requested files by $K = 2$ users and $M = 2$ RNs and the availability illustrated by the symbols transmitted from the DeNB only or from both at the DeNB and one of the two RNs RN$_1$ and RN$_2$.

**RNs Cache Placement:** Again, we assume without loss of generality $N = 4$ and one possible worst-case demand scenario in which the UEs request files W$_1$, W$_2$ while RN$_1$ and RN$_2$ are interested in files W$_3$ and W$_4$, respectively. As shown in Fig. 13, all files are comprised of 9 symbols, i.e., the $i$-th file is composed of symbols $\eta_i = (\eta_{i,1}, \eta_{i,2}, \ldots, \eta_{i,9})^\dagger$. Similarly to the previously described scheme, $\eta = (\eta_1^\dagger, \eta_2^\dagger, \eta_3^\dagger, \eta_4^\dagger)$ represent the column-wise concatenation of all $N = 4$ files. All symbols in $\eta$ are available at the DeNB. In contrast, RN$_1$ and RN$_2$ prefetch symbols $\eta_{i,j}$, $j = 1, \ldots, 4$ and $\eta_{i,j}$, $j = 5, \ldots, 8$, respectively in its cache.$^{11}$ For the sake of compact notation, these cached symbols are denoted by $\eta_{i,[1:4]}$ and $\eta_{i,[5:8]}$. Since, RN$_1$ (RN$_2$) is, interested in file W$_3$ (W$_4$) and it knows a-priori $\eta_{3,[1:4]}$ ($\eta_{4,[1:4]}$), the missing symbols RN$_1$ (RN$_2$) desires are $\eta_{3,[5:9]}$ ($\eta_{4,[5:8]}$) and $\eta_{4,9}$. In consequence, we design the transmission policy such that DeNB, RN$_1$ and RN$_2$ convey all receiving nodes with their desired symbols of files W$_1$ and W$_2$ as well as $\eta_{3,[5:9]}$, $\eta_{4,[1:4]}$ and $\eta_{4,9}$. These are in total 28 information symbols. Next, we consider the encoding strategy for DeNB and RNs.

\begin{center}
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
 & ZF map \\
\hline
UE$_1$ & $\eta_{2,1}$ & $\eta_{2,2}$ & $\eta_{2,5}$ & $\eta_{2,6}$ & $\eta_{3,5}$ & $\eta_{3,6}$ & $\eta_{4,1}$ & $\eta_{4,2}$ \\
UE$_2$ & $\eta_{1,1}$ & $\eta_{1,2}$ & $\eta_{1,5}$ & $\eta_{1,6}$ & $\eta_{3,7}$ & $\eta_{3,8}$ & $\eta_{4,3}$ & $\eta_{4,4}$ \\
\hline
\end{tabular}
\end{center}

Fig. 14: Map that assigns which symbol to zero-force at which UE.

$^{11}$We note that at $\mu = 4/9$, the joint cache content of the two RNs cannot contain the entire library of popular files.
Encoding at DeNB and RNs: Similarly to the previously described scheme, we use the spatial correlation of shared symbols between RN$_1$ or RN$_2$ with the DeNB to facilitate ZF and IA at the UEs. However, leveraging all ZF opportunities at the UEs has one main drawback from the RNs perspective – namely, it maximizes the number of interfering symbols imposed on RN$_1$ and RN$_2$. To limit this number, we do not send symbols $\eta_i[3:4]$ and $\eta_i[7:8]$, $i = 1, 2$, through the DeNB (at the cost of missing these symbols’ ZF opportunities) but rather through RN$_1$ or RN$_2$ directly to the UEs. All in all, this constitutes the transmit signals

$$x_s[t] = v^\dagger[t]\eta$$

$$x_{r,m}[t] = \beta^{(m)}[t]C_{rn,m}\eta,$$

\forall m \in [2] \text{ and } \forall t \in [T] \text{ for } T = 12, \text{ respectively. In the equation above, the notation for the precoding scalars is identical to the one utilized in the previous achievability scheme with the slight difference that } \beta^{(m)}[t] \text{ is the precoding vector at the } m\text{-th RN with respect to symbols } \eta. \text{ In (61) and (62), the following precoding subvectors in } v[t] \text{ and } \beta^{(m)}[t] \text{ are set to}

$$v_{i,[3:4]}[t] = v_{i,[7:8]}[t] = 0, \quad i = 1, 2$$

$$v_{3,[1:4]}[t] = \beta^{(1)}_{3,[1:4]}[t] = 0,$$

$$v_{4,[5:8]}[t] = \beta^{(2)}_{4,[5:8]}[t] = 0,$$

\forall t \in [T]. \text{ The binary caching matrices } C_{rn,1}, C_{rn,2} \in \mathbb{F}_2^{36 \times 36} \text{ at the two RNs can be represented by the following Kronecker products}

$$C_{rn,1} = I_4 \otimes \begin{pmatrix} I_4 & 0_{4 \times 5} \\ 0_{5 \times 4} & 0_{5 \times 5} \end{pmatrix},$$

$$C_{rn,2} = I_4 \otimes \begin{pmatrix} 0_{4 \times 4} & 0_{4 \times 4} & 0 \\ 0_{4 \times 4} & I_4 & 0 \\ 0^t & 0^t & 0^t \end{pmatrix}.$$ 

Similarly to the previously described scheme, zero elements in $v[t]$ (as stated in Eqs. (63)–(65)) can be incorporated to the effective DeNB caching matrix\textsuperscript{12}

$$C_{eff,DeNB} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes C_1 \otimes C_2 \otimes C_3$$

\textsuperscript{12}The structure of this matrix suggests that effectively the fractional cache size of the DeNB is $\mu = \frac{5}{9}$.\n
All components in (70) for which \( \eta \) are at UE

Further, we use modulo-\( K \) are at UE

such that we can rewrite the DeNB transmit signal in Eq. (61) by

\[
x_k[t] = \nu^\dagger[t]C_{\text{DeNB}}^{\text{eff}} \eta.
\]

Under these transmit signals, the observation at the \( k \)-th UE becomes

\[
y_{uk}[t] = g_k[t]x_k[t] + h_{k1}[t]x_{r,1}[t] + h_{k2}[t]x_{r,2}[t] + z_{uk}[t]
\]

\[
= \left( g_k[t]\nu^\dagger[t]C_{\text{DeNB}}^{\text{eff}} + h_{k1}[t]\beta^{(1)}[t]C_{\text{RN}1} + h_{k2}[t]\beta^{(2)}[t]C_{\text{RN}2} \right) \eta + z_{uk}[t]
\]

\[
= \sum_{i=1}^{2} \left( g_k[t]\nu^\dagger[i,1;2][t] + h_{k1}[t]\beta^{(1)}[i,1;2][t] \right) \eta_{i,1;2} + \left( g_k[t]\nu^\dagger[5,6;5,6][t] + h_{k2}[t]\beta^{(2)}[5,6;5,6][t] \right) \eta_{5,6}
\]

\[
+ h_{k1}[t]\beta^{(1)}[i,3;4][t] \eta_{i,3;4} + h_{k2}[t]\beta^{(2)}[i,3;4][t] \eta_{i,3;4} + g_k[t]\nu_{i,0}[t] \eta_{0,0} + z_{uk}[t]
\]

\[
+ \sum_{\ell=3}^{4} g_k[t]\nu_{\ell,9}[t] \eta_{\ell,9} + z_{uk}[t].
\]

All components in (70) for which \( i, \ell \neq k \) represent interference. The coefficient vector \( e_{\eta_{i,[a:b];[k,m]}}^{\text{ZF}}[t] \) denotes the effective channel coefficients of \( (b-a+1) \) ZF symbols \( \eta_{i,[a:b]} \) at UE_\( k \). If \( a = b \), the vector becomes a scalar and we simply write it as \( e_{\eta_{i,a,[k,m]}}^{\text{ZF}} \). Further, we use modulo-\( K \) indexing with respect to symbols \( \eta_{k+1,\ell}, \forall \ell \in [9] \) and their respective beamformers. These notations help us to state the ZF conditions (cf. ZF map in Fig. 14) which are at UE_\( k \):

\[
e_{\eta_{k+1,[1;2];[k,1]}}^{\text{ZF}}[t] \triangleq g_k[t]\nu_{k+1,[1;2]}[t] + h_{k1}[t]\beta^{(1)}_{k+1,[1;2]}[t] = 0,
\]

\[
e_{\eta_{k+1,[5;6];[k,2]}}^{\text{ZF}}[t] \triangleq g_k[t]\nu_{k+1,[5;6]}[t] + h_{k2}[t]\beta^{(2)}_{k+1,[5;6]}[t] = 0,
\]

\[
e_{\eta_{3,[2k+1;2k+2];[k,2]}}^{\text{ZF}}[t] \triangleq g_k[t]\nu_{3,[2k+1;2k+2]}[t] + h_{k2}[t]\beta^{(2)}_{3,[2k+1;2k+2]}[t] = 0,
\]

\[
e_{\eta_{4,[2k-1;2k];[1,1]}}^{\text{ZF}}[t] \triangleq g_k[t]\nu_{4,[2k-1;2k]}[t] + h_{k1}[t]\beta^{(1)}_{4,[2k-1;2k]}[t] = 0.
\]
In addition to the ZF conditions, we also impose additional IA conditions to the precoding design (as shown in Fig. [15], which are for UE_k:

\[ g_k[t]v_{\eta_{k-3}}[t] = h_k[1]h_{k+1}[t] = h_k[2]h_{k+2}[t], \]

\[ g_k[t]v_{\eta_{k+2}}[t] = e_{ZF}^{\eta_{k-3}}[k, 1] = h_k[1]h_{k-1}[2] = e_{ZF}^{\eta_{k-2}}[k, 2], \]

\[ g_k[t]v_{\eta_{k+1}}[t] = e_{ZF}^{\eta_{k-2}}[k, 1] = h_k[1]h_{k+1}[2] = e_{ZF}^{\eta_{k+1}}[k, 2]. \]

\[ \eta_{k+1, j} \quad \eta_{k+1, 4+j}, \quad j = 1, 2 \] (in total 8), are only subjected to ZF conditions (71a) and (71b) but not to any alignment conditions. This is due to the fact these are ZF symbols that are canceled at one of the UEs while being desired at the other UE. In consequence, we do not need to align these symbols at any UE. However, as \( \eta_{k+1, 1} \) and \( \eta_{k+1, 2} \) have identical ZF conditions (similarly for the pair of symbols \( \eta_{k+1, 5}, \eta_{k+1, 6} \)), we need to distinguish these symbols at UE_k for the sake of reliable decodability. To achieve this, we introduce additional complex random factors \( c_{\eta_{k+1, j}}[t] \) and \( c_{\eta_{k+1, 4+j}}[t] \), \( j = 1, 2 \), when fixing the precoding scalars \( v_{\eta_{k+1}}[t], \beta_{\eta_{k+1}, j}[t] \) and \( v_{\eta_{k+1, 4+j}}[t], \beta_{\eta_{k+1, 4+j}}[t] \). These random factors are \( a-priori \) known by all transmitting and receiving nodes.

Now we provide the solution of the system of equations comprising of ZF and IA conditions. For this purpose, we determine the precoding scalars in all 6 alignment chains as a function of \( v_{\eta_{1, 3}}[t], v_{\eta_{2, 9}}[t], v_{\eta_{3, 9}}[t] \) or \( v_{\eta_{4, 9}}[t] \). Hereby, we fix these scalars to

\[ v_{\eta_{k+1}}[t] = l_{13}[t]l_{23}[t]h_{k+1}[t], \]

\[ v_{\eta_{k-3}}[t] = l_{13}[t]l_{23}[t]h_{k-1}[t]h_k[2]h_{k-1}[2], \]

where

\[ l_{13}[t] = g_1[t]h_{21}[t] - g_2[t]h_{11}[t], \]

\[ l_{23}[t] = g_2[t]h_{12}[t] - g_1[t]h_{22}[t]. \]

As a consequence, the relevant precoders for \( j = 1, 2 \) become

\[ p_{\eta_{k+1}}[t] = \begin{pmatrix} v_{\eta_{k+1}, j}[t] \\ \beta_{\eta_{k+1}, j}[t] \end{pmatrix} = c_{\eta_{k+1}, j}[t] \begin{pmatrix} h_{k+1}[t] \\ -g_k[t] \end{pmatrix}, \]

\[ \beta_{\eta_{k+1}, j}[t] = v_{\eta_{k-3}}[t]g_k[t]h_{k+1}[t], \]

\[ \beta_{\eta_{k+1}, 4+j}[t] = v_{\eta_{k+1}}[t]g_k[t]h_{k+1}[t]. \]
Fig. 15: Interference alignment graph from the UE perspective for the achievability at corner point \((\frac{4}{9}, \frac{4}{9})\) for \(M = 2\) and \(K = 2\). The graph consists of six alignment chains and 20 nodes (i.e., symbols). The remaining 8 symbols \((\eta_{k+1,j}, \eta_{k+1,4+j}, j, k = 1,2)\) do not appear in the alignment graph as these symbols are zero-forced at the unwanted UE (according to the ZF conditions (71a) and (71b)) while they are desired by the other UE.

\[
p_{\eta_{k+1,4+j}}[r] \triangleq \left( \frac{v_{\eta_{k+1,4+j}}[r]}{\beta_{\eta_{k+1,4+j}}[r]} \right) = c_{\eta_{k+1,4+j}}[r] \begin{pmatrix} -h_{k2}[r] \\ g_k[r] \end{pmatrix}, \tag{75d}
\]

\[
\beta_{\eta_{k+1,4+j}}^{(2)}[r] = v_{\eta_{5-k,5}}[r] \frac{g_k[r]}{h_{k2}[r]}, \tag{75e}
\]

\[
\beta_{\eta_{k+1,k}}^{(2)}[r] = v_{\eta_{k+2,5}}[r] \frac{g_k[r]}{h_{k2}[r]}, \tag{75f}
\]

\[
p_{\eta_{3,9-2k}}[r] \triangleq \left( \frac{v_{\eta_{3,9-2k}}[r]}{\beta_{\eta_{3,9-2k}}^{(2)}} \right) = (-1)^{k+1} v_{\eta_{k+1,9}}[r] \frac{g_k[r]}{l_{23}[r]} \begin{pmatrix} -h_{k+1,2}[r] \\ g_{k+1}[r] \end{pmatrix}, \tag{75g}
\]

\[
p_{\eta_{3,10-2k}}[r] \triangleq \left( \frac{v_{\eta_{3,10-2k}}[r]}{\beta_{\eta_{3,10-2k}}^{(2)}} \right) = (-1)^{k+1} v_{\eta_{k+2,5}}[r] \frac{g_k[r]}{l_{23}[r]} \begin{pmatrix} -h_{k+1,2}[r] \\ g_{k+1}[r] \end{pmatrix}, \tag{75h}
\]

\[
p_{\eta_{4,5-2k}}[r] \triangleq \left( \frac{v_{\eta_{4,5-2k}}[r]}{\beta_{\eta_{4,5-2k}}^{(1)}} \right) = (-1)^{k+1} v_{\eta_{k+1,9}}[r] \frac{g_k[r]}{l_{13}[r]} \begin{pmatrix} h_{k+1,1}[r] \\ -g_{k+1}[r] \end{pmatrix}, \tag{75i}
\]
Hereby, the concatenated beamforming vectors in (75a), (75d) and (75g)–(75j) are perpendicular to

\[
\mathbf{p}_{\eta_{k+1,j}}^\perp [t] = \mathbf{h}_{k1}[t] \triangleq \begin{pmatrix} g_k[t] \\ h_{k1}[t] \end{pmatrix},
\]

(76a)

\[
\mathbf{p}_{\eta_{k+1,4+j}}^\perp [t] = \mathbf{h}_{k2}[t] \triangleq \begin{pmatrix} g_k[t] \\ h_{k2}[t] \end{pmatrix},
\]

(76b)

\[
\mathbf{p}_{\eta_{k+1,10-2k}}^\perp [t] = \mathbf{h}_{k+1,2}[t] \triangleq \begin{pmatrix} g_{k+1}[t] \\ h_{k+1,2}[t] \end{pmatrix},
\]

(76c)

\[
\mathbf{p}_{\eta_{k+1,6-2k}}^\perp [t] = \mathbf{h}_{k+1,1}[t] \triangleq \begin{pmatrix} g_{k+1}[t] \\ h_{k+1,1}[t] \end{pmatrix}.
\]

(76d)

This observation concurs with the four ZF conditions of Eq. (71). The decoding at RNs and UEs is described next.

**Decoding at the RNs and the UEs:** In the following, we will formulate the received signals of the RNs and the UEs. Without loss of generality, we will focus on the received signals of UE\(_1\) and RN\(_1\).

\[
y_{u,1}[t] = D_1(\eta_{1,[1:9]}) + I_1(\eta_{2,9} + \eta_{3,4} + \eta_{2,4} + \eta_{3,7} + \eta_{3,9} + \eta_{4,4} + \eta_{2,8} + \eta_{3,8} + \eta_{4,9} + \eta_{2,3} + \eta_{2,7}) + z_{u,1}[t]
\]

(77)

In this equation, \(D_1\) denotes a linear combination of UE\(_1\)’s 9 desired symbols \(\eta_{1,[1:9]}\). \(I_1\), on the other hand, is a linear combination of 3 aligned interference symbols\(^{13}\). In explicit form, these two linear combinations are given by

\[
D_1(\eta_{1,[1:9]}) = \left( g_1[t]v_{1,[1:2]}^\dagger [t] + h_{11}[t]\beta_{1,[1:2]}^{(1)}[t] \right)\mathbf{l}_1\eta_{1,[1:2]} + h_{11}[t]\beta_{1,[3:4]}^{(1)}[t]\mathbf{l}_2\eta_{1,[3:4]}
\]

+ \left( g_1[t]v_{1,[5:6]}^\dagger [t] + h_{12}[t]\beta_{1,[5:6]}^{(2)}[t] \right)\mathbf{l}_2\eta_{1,[5:6]} + h_{12}[t]\beta_{1,[7:8]}^{(2)}[t]\mathbf{l}_2\eta_{1,[7:8]}
\]

+ \left( g_1[t]v_{\eta_{1,9}}[t] \right)\eta_{1,9}
\]

(78)

\(^{13}\)These symbols are specified in the argument of \(I_1\).
and

$$I_1 = g_1[r] \nu_{q,9}[r](\eta_2 + \eta_4 + \eta_7) + g_1[r] \nu_{q,9}[r](\eta_3 + \eta_4 + \eta_8 + \eta_3)$$

$$+ g_1[r] \nu_{q,9}[r](\eta_4 + \eta_2 + \eta_7).$$

(79)

On the other hand, the received signal of RN$_1$ after canceling known interfering components (by exploiting the cached content) can be written as follows:

$$y'_{r,1}[r] = f_1[r] \left( \nu_{q,9}^\dagger I_4 \eta_3 + \nu_{q,9} I_3 \eta_3 + \sum_{i=1}^{2} (\nu_{q,9}^\dagger I_2 \eta_i) \eta_i \right)$$

$$+ \nu_{q,9}[r] \eta_{q,9} + z_{r,1}[r].$$

(80)

When choosing the precoding scalars $\nu_{q,9}[r]$, $i = 1, \ldots, 4$ independently (e.g., as in Eq. (73a)) in Eqs. (77) and (80) over $T = 12$ channel uses for a time-varying wireless channel, both RN$_1$ and UE$_1$ make in total 12 noise-corrupted independent observations $\{y_{u,1}[r]\}_{i=1}^{12}$ and $\{y'_{r,1}[r]\}_{i=1}^{12}$ as a function of their desired symbols and (possibly aligned) interfering symbols. This facilitates that UE$_1$, on the one hand, is capable of decoding its 9 desired symbols

- $\eta_1 = (\eta_{1,1}, \eta_{1,2}, \ldots, \eta_{1,9})$

and 3 distinct sums of interfering symbols in the form

- $\eta_2 + \eta_4 + \eta_7$,
- $\eta_3 + \eta_4 + \eta_8 + \eta_3$,
- $\eta_4 + \eta_2 + \eta_7$.

RN$_1$, on the other hand, decodes its 5 remaining uncached desired symbols ($\eta_3, \eta_6, \ldots, \eta_9$) and 7 interfering symbols ($\eta_1, \eta_6, \eta_2, \eta_5, \eta_1, \eta_9, \eta_9$ and $\eta_4$). Due to symmetry, the observation of UE$_2$ and RN$_2$ are similar to UE$_1$ and RN$_1$, respectively.

As opposed to the previous scheme, both RNs and UEs require $T = 12$ channel uses to determine their desired symbols. Since every file consists of 9 symbols, the achievable NDT thus corresponds to $12/9 = \frac{4}{3}$.

Now, we establish the achievability for corner points $\left(\frac{1}{2}, 1\right)$ (when $M = 2$, $K = 1$ as shown in Fig. 16) and $\left(\frac{1}{2}, \frac{5}{4}\right)$ (when $M = 2$, $K = 2$ as shown in Fig. 17), respectively. These schemes are one-shot schemes that combine the Maddah-Ali Niesen (MAN) scheme with cooperative zero-forcing beamforming among DeNB and RNs. The generalized scheme is outlined in detail in Section [V]. Thus, in the following, we will only briefly outline the achievability of these two corner points. We start with the cache placement at the RNs.
RNs cache placement: At corner points \( \left( \frac{1}{2}, 1 \right) \) and \( \left( \frac{1}{2}, \frac{5}{4} \right) \), we assume that each file \( W_i \) comprises of \( L'_1 = 2 \) and \( L'_2 = 4 \) symbols, respectively. Hereby, RN\(_m\), \( m = 1, 2 \), caches the \( m\)-th block of \( \frac{L'_1}{2} \left( \frac{L'_2}{2} \right) \) symbols. For instance, at corner point \( \left( \frac{1}{2}, \frac{5}{4} \right) \), RN\(_2\) prefetches the last \( \frac{L'_2}{2} = 2 \) symbols (\( \eta_{i,3} \) and \( \eta_{i,4} \)) of each file \( W_i \) (cf. \( S_2 \) in Fig. 17).

![Diagram](image)

Fig. 16: One-shot scheme at corner point \( \left( \frac{1}{2}, 1 \right) \) for \( M = 2 \) RNs, \( K = 1 \) UEs and \( N = 3 \) files. Each file \( W_i, i = 1, 2, 3 \), is comprised of two symbols \( \eta_{i,1} \) and \( \eta_{i,2} \). The figure shows, respectively, the DeNB transmit signal \( x_s[t] \), the \( m\)-th RNs transmit signal \( x_{r,m}[t] \), the received signals of UE\(_1\) \( y_{u,1}[t] \) and RN\(_m\) \( y'_{r,m}[t] \) (when RN\(_m\) exploits its cached content \( S_m \) as side information) that leads to an NDT of 1 after \( T = 2 \) channel uses.

Encoders at DeNB and RNs: For both corner points, we use the first \( T_1 \)\(^{14}\) channel uses to multicast all symbols desired by RN\(_1\) and RN\(_2\) over the DeNB-RN broadcast channel. Simultaneously, we exploit the side information at the RNs caches to zero-force undesired symbols at one of the UEs and provide this particular UE with one of its desired symbols.

\(^{14}\)In the schemes of corner points \( \left( \frac{1}{2}, 1 \right) \) and \( \left( \frac{1}{2}, \frac{5}{4} \right) \), \( T_1 \) corresponds to 1 and 2 channel uses, respectively.
For example, at \( t = 1 \) for corner point \( \left( \frac{1}{2}, \frac{5}{3} \right) \), we multicast the symbols \( \eta_{3,3} \) and \( \eta_{4,1} \) at the DeNB in an MAN scheme manner according to:

\[
x_s[1] = \nu_{\eta_{3,3}}[1] \eta_{3,3} + \nu_{\eta_{4,1}}[1] \eta_{4,1}
\]

This transmit signal allows each RN to retrieve its desired symbol by exploiting its knowledge of the channel and of the undesired symbol in \( x_s[1] \) through its cache (e.g., RN\(_1\) is interested in retrieving \( \nu_{\eta_{3,3}}[1] \eta_{3,3} \)).
in symbol $\eta_{3,3}$ and has $\eta_{4,1}$ stored in its cache). However, neither UE$_1$ nor UE$_2$ are interested in the symbols of $x_d[1]$ ($\eta_{3,3}, \eta_{4,1}$). Thus, we do not only leverage the knowledge of RN$_1$ in $\eta_{4,1}$ and of RN$_2$ in $\eta_{3,3}$ from a receiver-caching perspective, but instead we also capitalize on this cognizance from a transmitter-caching perspective by sending

$$x_{r,1}[1] = \beta_{\eta_{4,1}}[1] \eta_{4,1} + \eta_{1,1} \tag{82a}$$

$$x_{r,2}[1] = \beta_{\eta_{3,3}}[1] \eta_{3,3} \tag{82b}$$

from RN$_1$ and RN$_2$, respectively, to overcome this issue. Hereby, we choose the following concatenated vectors to satisfy:

$$\begin{pmatrix} v_{\eta_{3,3}}[1] \\ \beta_{\eta_{3,3}}[1] \end{pmatrix} \perp \begin{pmatrix} g_1[1] \\ h_{12}[1] \end{pmatrix}, \tag{83a}$$

$$\begin{pmatrix} v_{\eta_{4,1}}[1] \\ \beta_{\eta_{4,1}}[1] \end{pmatrix} \perp \begin{pmatrix} g_1[1] \\ h_{11}[1] \end{pmatrix}. \tag{83b}$$

This facilitates that UE$_1$ is free from interference due to undesired symbols $\eta_{3,3}$ and $\eta_{4,1}$. Instead, it receives a noise-corrupted signal that depends on its desired symbol $\eta_{1,1}$. UE$_2$, on the other hand, observes a signal that depends solely on undesired symbols $\eta_{1,1}, \eta_{3,3}$ and $\eta_{4,1}$ denoted by $I(\eta_{1,1}, \eta_{3,3}, \eta_{4,1})$ in Fig. 17. In the remaining $T_2$ channel uses, cooperative DeNB-RN zero-forcing is applied for the case when $M = K = 2$ and simple RN unicasting when $M = 2, K = 1$.

**Decoding at the RNs and the UEs:** As opposed to all prior schemes, RNs and UEs can decode each desired symbol on a single-channel use basis, i.e., symbol decoding over multiple channel uses is not required in attaining the optimal NDT. In consequence, channel diversity over multiple channel uses is not required.

Specifically, in the proposed scheme for corner point $\left(\frac{1}{2}, 1\right)$, UE$_1$ decodes its $L'_1 = 2$ desired symbols $\eta_{1,1}$ and $\eta_{1,2}$ in $T_1 + T_2 = 2$ channel uses, namely, 1 and 2 (cf. Fig. 16), respectively. The RNs, on the contrary, only need the first channel use ($T_1 = 1$) to retrieve $(1 - \mu)L'_1 = 1$ desired symbols. In conclusion, this conforms to an NDT of $\frac{T_1 + T_2}{L'_1} = 1$.

In the other scheme, as shown in Fig. 17, UE$_1$ (UE$_2$) decodes its $L'_2 = 4$ desired symbols $\eta_{1,1}, \eta_{1,2}, \eta_{1,3}$ and $\eta_{1,4} (\eta_{2,1}, \eta_{2,2}, \eta_{2,3}$ and $\eta_{2,4})$ in 4 out of $T_1 + T_2 = 5$ channel uses, namely, 1, 2, 3 and 4 (4, 5, 2 and 3), respectively. Unlike the UEs, each RN only needs the first $T_1 = 2$ channel uses to retrieve its $(1 - \mu)L'_2 = 2$ desired symbols. Ultimately, the NDT becomes $\frac{T_1 + T_2}{L'_2} = \frac{5}{4}$.

\footnote{In the schemes of corner points $\left(\frac{1}{2}, 1\right)$ and $\left(\frac{1}{2}, \frac{5}{4}\right)$, $T_2$ corresponds to 1 and 3 channel uses, respectively.}
C. Achievability for $M = 3$

The optimal delivery-time cache-memory tradeoff for $M = 3$ and $K = 1$ is presented. Hereby, the following proposition quantifies the achievable NDT.

Proposition 3. The achievable NDT of the network under study for $M = 3$ RNs, $K = 1$ UEs and $\mu \in [0, 1]$ is given by

$$\delta(\mu) = \max \left\{ 1, 4 - 9\mu \right\}. \quad (84)$$

Similarly, to the achievability of corner points when $M = 2$, it suffices to establish a scheme that attains an NDT of 1 at $\mu = \frac{1}{3}$. The NDT-optimal scheme that establishes this result is the one-shot scheme of Section V. Since the main idea of this scheme has already been illustrated through explicit examples for corner points $(\frac{1}{2}, 1)$ when $(K, M) = (1, 2)$ and $(\frac{1}{2}, \frac{5}{4})$ when $(K, M) = (2, 2)$, we omit further details for the sake of brevity. The interested reader can reconstruct the explicit scheme for $(K, M) = (1, 3)$ from the generalized scheme of Section V for Region B.

VII. Directions for Future Work

In this section, we discuss some of the open problems and directions for future work on the topic of cache-aided broadcast-relay wireless networks. In particular, we focus on aspects that are left open in this paper.

A. Imperfect CSI

An interesting aspect is the influence of imprecise CSI at transmitters (CSIT) on the minimum NDT. In detail, we may consider the system model of Fig. 1 in which at channel use $t$, DeNB and/or RNs may have access to imprecise CSI of $\{f(t), g(t), H(t)\}$. Specifically, we distinguish the following CSIT settings:

1) Delayed CSIT: At channel use $t$, DeNB and/or RNs are aware of the previous $t - 1$ time instants CSI $\{f^{t-1}, g^{t-1}, H^{t-1}\}$.

2) Mixed CSIT: In addition to delayed CSIT, DeNB and/or RNs have at time instant $t$ access to current CSI estimates $\{\hat{f}(t), \hat{g}(t), \hat{H}(t)\}$ of some quality $\alpha \in [0, 1]$ [45, Chapter 4]. The extreme cases of $\alpha = 0$ and $\alpha = 1$ represent, respectively, the cases of delayed CSIT only and perfect current quality CSIT.

While there is plenty of degrees-of-freedom studies on the impact of delayed and mixed CSIT on interference networks (cf. survey paper [45]), the interplay of caching and imperfect CSI
is with the exception of [46] far less understood. The authors in [46] identify DoF gains both due to current CSIT and coded caching for an MISO broadcast channel with Rx caching. When focusing on the extreme cases of zero-cache ($\mu = 0$) and full cache ($\mu = 1$), we can determine the achievable NDT for mixed CSIT

$$\delta_{\text{ach}}(\mu, \alpha) = \begin{cases} K + M & \text{for } \mu = 0 \\
\frac{\sum_{i=1}^{K} \frac{1}{(1-\alpha) + \alpha \sum_{i=1}^{K} \frac{1}{i}}} \text{for } \mu = 1, K \leq M + 1
\end{cases}$$

based on results for the MISO BC [47], [48]. The case of delayed CSIT only is given in above equation when $\alpha = 0$. It is of interest to understand how cache placement and file delivery needs to be adjusted for the aforementioned CSIT models when $0 < \mu < 1$.

**B. Partial Connectivity**

It is of interest to understand the implications of partial connectivity (with respect to the RN-UE links) on the NDT for cache-assisted broadcast-relay networks. In particular, one question of interest is whether the cache placement has to account for the topology. In addition, how does optimal file splitting may look like.

In recent works, caching has been, amongst others, applied to combination networks with receiver caches [49], partially-connected interference channels with transmitter caches under complete file placement [50] and interference networks with Tx and Rx-caching [51]. However, the impact of partial connectivity on cache-aided channels is unknown at large, especially, with respect to transceiver cache-aided networks.

**C. Proof-of-Concept Implementation**

A practical proof-of-concept implementation allows to verify to what extent the theoretically postulated delivery times in this paper are achievable. Further, implementation issues such as large subpacketization levels [52], practicality of centralized cache placement [10] and self-interference cancelation in full-duplex communication [53] have to be handled.

**VIII. Conclusion**

In this paper, we have studied the fundamental information-theoretic limits on the delivery time of a transceiver cache-aided broadcast relay network consisting of a central base station (DeNB), $M$ cache-endowed relay nodes (RNs) and $K$ mobile users (UEs). We used the normalized delivery
time (NDT) as our performance metric. The NDT measures the worst-case delivery time per bit with respect to an interference-free system in the high SNR regime. We established a converse result for the NDT of a general broadcast relay network with arbitrary number of RNs and UEs. Next, we presented two achievability schemes which exploit the RNs caches and its full-duplex capability in collaboration with the DeNB. The first scheme is a one-shot achievability scheme that synergistically interlaces zero-forcing (ZF) and multicasting strategies proposed in the framework of coded caching. The second scheme, on the other hand, integrates subspace interference alignment (IA) with zero-forcing through carefully designing ZF and IA maps. With these results, we were able to characterize the optimal NDT-cache-memory tradeoff for $K + M \leq 4$. In addition, we identify NDT-optimal regimes of the proposed one-shot scheme and also show that caching more than a fraction of $\left\lceil \frac{(M-1)/2}{M} \right\rceil$ files attains a constant gap of $\frac{8}{3}$ with respect to the optimal NDT. We discussed the inadequacy of the inverse sum DoF metric in capturing the delivery time of cache-assisted broadcast relay networks. Finally, we presented directions for future work.

APPENDIX A

MAXIMUM MULTIPLICATIVE GAP FOR ONE-SHOT SCHEME

In this section, we present the maximum multiplicative gap of the achievable one-shot scheme presented in Section V. We remind the reader about the $(\mu, K, M)$ regions presented in Fig. 8 where the achievable NDTs vary. To study the gap with respect to lower bounds on the NDT, we consider the pair of regions (B,E) with achievable one-shot NDT

$$\delta_{OS}^{(B,E)}(\mu) = \delta_{MAN}(\mu) = M \cdot (1 - \mu) \cdot \frac{1}{1 + \mu M},$$

(85)

as well the pair of regions (C,D) with achievable one-shot NDT

$$\delta_{OS}^{(C,D)}(\mu) = \frac{K + \delta_{MAN}(\mu)}{1 + \mu M},$$

(86)

separately. Finally, we consider the transitional region of regime pairs (D,E) for fractional cache sizes $\mu \in \left[0, \frac{1}{M}\right]$. 
A. (B,E) Region Pair

For this region pair, the one-shot NDT corresponds to (85). We focus on fractional cache sizes \( \mu \) inside the interval \( \mu \in \left[ \mu'(\theta), \frac{[M-1/2]}{M} \right] \), where \( \theta \in \left[ 1, \frac{M-3}{2} \right] \) adjusts the left endpoint \( \mu'(\theta) \) of the interval according to

\[
\mu'(\theta) = \left\lceil \frac{\theta}{M} \right\rceil. \tag{87}
\]

The NDT \( \delta_{OS}^{(B,E)}(\mu) \) is decreasing in \( \mu \), which is why we upper bound it (in the interval specified above) by

\[
\delta_{OS}^{(B,E)}(\mu) \leq \delta_{OS}^{(B,E)}(\frac{\theta}{M}) = \frac{M - \theta}{1 + \theta}. \tag{88}
\]

From Theorem 1, we infer that

\[
\delta^*(\mu) \geq 1. \tag{89}
\]

Consequently, the multiplicative gap becomes

\[
\frac{\delta_{OS}^{(B,E)}(\mu)}{\delta^*(\mu)} \leq \frac{M - \theta}{1 + \theta}. \tag{90}
\]

B. (C,D) Region Pair

In these two regions, the achievable NDT is given by (86). We consider three different intervals of fractional cache sizes: (i) the first being \( \mu \in [0, \bar{\mu}] \), (ii) the second being \( \mu \in \left( \bar{\mu}, \frac{1}{M} \right) \) (iii) and the third being \( \mu \in \left[ \mu'(\kappa_d), \min \left\{ 1, \frac{K}{M} \right\} \right] \). In the following, we treat these three cases individually.

Case (i): For this case, the right endpoint of the interval equates to

\[
\bar{\mu} = \min \left\{ \frac{1}{M}, \frac{K + M + 1}{(M + 1)(K + M - 1)} \right\}. \tag{91}
\]

The achievable NDT in the interval \( \left[ 0, \frac{1}{M} \right] \), which subsumes the interval of case (i), is attainable through memory sharing of schemes at corner point \( (0, K + M) \) and the one-shot scheme at \( \mu = \frac{1}{M} \) with corresponding NDT

\[
\delta_{OS}^{(C,D)}(\mu = \frac{1}{M}) = \begin{cases} 
1 & \text{if } K = 1 \\
\max \left\{ 1, \frac{K}{M+1} \right\} & \text{if } M = 1 \\
\frac{K + \delta_{MAN}^{(\mu = \frac{1}{M})}}{2} = \frac{K}{2} + \frac{M-1}{4} & \text{if } K, M \geq 2
\end{cases}
\]
This results in the achievable NDT
\[
\delta_{\text{OS}}^{(C,D)}(\mu) = \begin{cases} 
K + M - \mu M(K + M - 1) & \text{if } K = 1 \\
K + M - \mu M\left( K + M - \max\left\{1, \frac{K}{M+1}\right\} \right) & \text{if } M = 1 \\
K + M - \mu M\left( \frac{K}{2} + \frac{3M+1}{4} \right) & \text{if } K, M \geq 2 
\end{cases}
\] (92)
in the interval \([0, \frac{1}{M}]\). By substituting \(\ell = M\) and \(s = 1\) in Eq. \((6)\) of Theorem \(1\) we get
\[
\delta^*(\mu) \geq \delta_{\text{LB}}(\mu, M, 1) = K + M - \mu M(K + M - 1).
\] (93)

We now combine (92) and (93) to upper bound the multiplicative gap according to:
\[
\frac{\delta_{\text{OS}}^{(C,D)}(\mu)}{\delta^*(\mu)} \leq 1
\] (94)

if \(K = 1\),
\[
\frac{\delta_{\text{OS}}^{(C,D)}(\mu)}{\delta^*(\mu)} \leq \frac{K + M - \mu M\left( K + M - \max\left\{1, \frac{K}{M+1}\right\} \right)}{K + M - \mu M(K + M - 1)} \\
\overset{(a)}{=} 1 + \frac{\mu M\left( \frac{K}{M+1} - 1 \right)^+}{K + M - \mu M(K + M - 1)} \\
\overset{(b)}{\leq} 1 + \frac{\mu M\left( \frac{K}{M+1} - 1 \right)^+}{\max\left\{1, \frac{K}{M+1}\right\}} = 1 + \frac{M}{K} \cdot \frac{K + M + 1}{(K + M - 1)} \left( \frac{K}{M+1} - 1 \right)^+ \\
= \begin{cases} 
1 & \text{if } K \leq 2 \\
1 + \frac{K}{2} \cdot \left( \frac{K}{2} - \frac{2}{K} \right) & \text{if } K > 2 
\end{cases}
\] (95)

if \(M = 1\) and
\[
\frac{\delta_{\text{OS}}^{(C,D)}(\mu)}{\delta^*(\mu)} \leq \frac{K + M - \mu M\left( \frac{K}{2} + \frac{3M+1}{4} \right)}{K + M - \mu M(K + M - 1)} \\
\overset{(a)}{=} 1 + \frac{\mu M\left( \frac{K}{2} + \frac{M-5}{4} \right)}{K + M - \mu M(K + M - 1)} \\
\overset{(b)}{\leq} 1 + \frac{\mu M\left( \frac{K}{2} + \frac{M-5}{4} \right)}{\max\left\{1, \frac{K}{M+1}\right\}} = 1 + \left( \frac{K}{2} + \frac{M-5}{4} \right) \cdot \min\left\{1, \frac{M}{K} \cdot \frac{K + M + 1}{(K + M - 1)} \right\}
\] (96)
if $K, M \geq 2$. Hereby, step $(a)$ follows by addition and subtraction of the denominator $\delta_{LB}(\mu, M, 1)$ from the numerator and step $(b)$ from the fact that the rational function is maximized by one of the endpoints of the domain of fractional cache sizes (in this specific case, $\mu^* = \tilde{\mu}$ in the domain $0 \leq \mu \leq \tilde{\mu}$).

**Case (ii):** Now we examine fractional cache sizes $\mu \in \left(\tilde{\mu}, \frac{1}{M}\right]$. From the definition of $\tilde{\mu}$, we determine that $\tilde{\mu} = \frac{1}{M}$ if $K \leq M$ and $\tilde{\mu} < \frac{1}{M}$ if $K > M$. Consequently, it suffices to restrict the focus for this case to $K > M$. However, we may still use the expressions on the achievable NDT of Eq. (92). In addition, we deduce from Theorem 1 that for $K > M$

$$\delta^*(\mu) \geq \frac{K}{M + 1}. \quad (97)$$

Using (92) and (97), we get for $K > M$ and $M \geq 2$

$$\frac{\delta^{(C,D)}_{OS}(\mu)}{\delta^*(\mu)} \leq \frac{K + M - \mu M \left(\frac{K}{2} + \frac{3M+1}{4}\right)}{\frac{K}{M+1}} \quad (b) \leq \frac{K + M - \tilde{\mu} M \left(\frac{K}{2} + \frac{3M+1}{4}\right)}{\frac{K}{M+1}}$$

$$= 1 + \left(\frac{K}{2} + \frac{M - 5}{4}\right) \cdot \frac{K}{K+1} \cdot \frac{(K+M+1)}{(K+M-1)} \quad (98)$$

**Case (iii):** Consider all fractional cache sizes $\mu \in \left[\mu'(k_d), \min\left\{1, \frac{K}{M}\right\}\right]$, where we choose $\mu'(k_d)$ according to

$$\mu'(k_d) = \left\lceil \frac{k_d}{M} \right\rceil \quad (99)$$

and the numerator $k_d$ in (99) parametrized by $d$ to

$$k_d = \frac{M + 1 - d}{d}, \quad \forall d \in \left[\frac{M+1}{\min\{K,M\}}, \frac{M+1}{2}\right]. \quad (100)$$

Due to the decreasing monotony of $\delta^{(C,D)}_{OS}(\mu)$, we may upper bound the achievable NDT for any $\mu$ in the interval $\left[\mu'(k_d), \min\left\{1, \frac{K}{M}\right\}\right]$ by

$$\delta^{(C,D)}_{OS}(\mu) \leq \delta^{(C,D)}_{OS}\left(\frac{k_d}{M}\right) = \frac{K(1 + k_d) + M - k_d}{(1 + k_d)^2}. \quad (101)$$

The lower bound of Theorem 1 can be further bounded from below as follows:

$$\delta^*(\mu) \geq \max\left\{1, \max_{s \in [\min\{M+1,K\}]} \delta_{LB}(\mu, s)\right\} \geq \max\left\{1, \frac{K}{M + 1}\right\} \quad (102)$$
With (101) and (102), we are able to upper bound the multiplicative gap.

\[
\frac{\delta_{\text{OS}}^{(C,D)}(\mu)}{\delta^*(\mu)} \leq \frac{\delta_{\text{OS}}^{(C,D)}(\frac{k_d}{M})}{\max\left\{1, \frac{\kappa_d}{M+1}\right\}} = \frac{K(M+1)(1 + \kappa_d) + (M - \kappa_d)(M + 1)}{\max\{M + 1, K\}(1 + \kappa_d)^2}
\]  

(103)

When substituting (100) in (103), we have

\[
\frac{k_d}{d^2} + \frac{(M+1)^2(d-1)}{d^2} = \frac{dK}{\max\{M + 1, K\}} + \frac{d(d - 1)}{\max\{M + 1, K\}}
\]

Thus, we conclude

\[
\frac{\delta_{\text{OS}}^{(C,D)}(\mu)}{\delta^*(\mu)} \leq \frac{dK}{\max\{M + 1, K\}} + \frac{d(d - 1)}{\max\{M + 1, K\}}.
\]

(104)

**C. Transition (D,E) Region Pair for \( \mu \leq \frac{1}{M} \)**

At this transitional region, the one-shot NDT for arbitrary fractional cache sizes \( \mu < \frac{1}{M} \) is achievable through memory sharing of the broadcasting scheme at corner point \((0, K + M)\) and the OS-scheme at \( \mu = \frac{1}{M} \) with NDT

\[
\delta_{\text{MAN}}\left(\mu = \frac{1}{M}\right) = \frac{M - 1}{2}.
\]

This results in the following memory-sharing based transitional NDT for the region pair (D,E):

\[
\delta_{\text{OS}}^{(D,E)}(\mu) = K + M - \mu M\left(K + \frac{M + 1}{2}\right)
\]

(105)

We substitute \( \ell = M \) and \( s = 1 \) in Eq. (6) of Theorem 1 to get

\[
\delta^*(\mu) \geq \delta_{\text{LB}}(\mu, M, 1) = K + M - \mu M(K + M - 1).
\]

(106)

Now, from (105) and (106), we have

\[
\frac{\delta_{\text{OS}}^{(D,E)}(\mu)}{\delta^*(\mu)} \leq \frac{K + M - \mu M\left(K + \frac{M + 1}{2}\right)}{K + M - \mu M(K + M - 1)}
\]

\[
= \frac{\mu M\left(M-3\right)}{2} = \frac{1}{2} \leq 1 + \frac{M - 3}{2} = \frac{M - 1}{2},
\]

(107)

where the reasoning for the steps \((a)\) and \((b)\) are in agreement with the aforementioned steps of the other region pairs.
REFERENCES

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

[2] M. Tao, E. Chen, H. Zhou, and W. Yu, “Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6118–6131, Sept 2016.

[3] S. Rezvani, N. Mokari, M. R. Javan, and E. A. Jorswieck, “Fairness and Transmission-Aware Caching and Delivery Policies in OFDMA-Based HetNets,” 2017. [Online]. Available: http://arxiv.org/abs/1711.02776

[4] 3GPP, “Relay architectures for E-UTRA (LTE-Advanced),” Tech. Rep. 36.806, 2011.

[5] J. Kakar, “UAV Communications: Spectral Requirements, MAV and SUAV Channel Modeling, OFDM Waveform Parameters, Performance and Spectrum Management,” Master’s thesis, Virginia Tech, Blacksburg, USA, 2015.

[6] J. Kakar and V. Marojevic, “Waveform and Spectrum Management for Unmanned Aerial Systems Beyond 2025,” 2017. [Online]. Available: http://arxiv.org/abs/1708.01664

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

[8] K. Wan, D. Tuninetti, and P. Piantanida, “On the optimality of uncoded cache placement,” in *ITW*, Sept 2016, pp. 161–165.

[9] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “The exact rate-memory tradeoff for caching with uncoded prefetching,” in *ISIT*, June 2017, pp. 1613–1617.

[10] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[11] M. M. Amiri, Q. Yang, and D. Gündüz, “Decentralized caching and coded delivery with distinct cache capacities,” *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4657–4669, Nov 2017.

[12] J. Zhang, X. Lin, C. C. Wang, and X. Wang, “Coded caching for files with distinct file sizes,” in *ISIT*, June 2015, pp. 1686–1690.

[13] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, “Online coded caching,” *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 836–845, April 2016.

[14] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, “On the average performance of caching and coded multicasting with random demands,” in *ISWCS*, Aug 2014, pp. 922–926.

[15] Y. P. Wei and S. Ulukus, “Coded caching with multiple file requests,” in *Allerton*, Oct 2017, pp. 437–442.

[16] Z. H. Awan and A. Sezgin, “Fundamental limits of caching in D2D networks with secure delivery,” in *ICC*, June 2015, pp. 464–469.

[17] M. Ji, G. Caire, and A. F. Molisch, “Fundamental Limits of Caching in Wireless D2D Networks,” *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb 2016.

[18] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, “Multi-server coded caching,” *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec 2016.

[19] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, “Hierarchical coded caching,” *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3212–3229, June 2016.

[20] S. Gharekhloo and A. Sezgin, “Latency-Limited Broadcast Channel with Cache-Equipped Helpers,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4192–4203, July 2017.

[21] M. A. Maddah-Ali and U. Niesen, “Cache-aided interference channels,” in *ISIT*, June 2015, pp. 809–813.

[22] A. Sengupta, R. Tandon, and O. Simeone, “Cache aided wireless networks: Tradeoffs between storage and latency,” in *CISS*, March 2016, pp. 320–325.
[23] R. Tandon and O. Simeone, “Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in Fog Radio Access Networks,” in ISIT, July 2016, pp. 2029–2033.

[24] 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.

[25] S. M. Azimi, O. Simeone, and R. Tandon, “Fundamental Limits on Latency in Small-Cell Caching Systems: An Information-Theoretic Analysis,” in GLOBECOM, Dec 2016, pp. 1–6.

[26] J. Kakar, S. Gherekhloo, and A. Sezgin, “Fundamental Limits on Delivery Time in Cloud- and Cache-Aided Heterogeneous Networks,” 2017. [Online]. Available: http://arxiv.org/abs/1706.07627

[27] J. Kakar, S. Gherekhloo, Z. H. Awan, and A. Sezgin, “Fundamental limits on latency in cloud- and cache-aided HetNets,” in ICC, May 2017, pp. 1–6.

[28] A. Vahid, M. A. Maddah-Ali, and A. S. Avestimehr, “Capacity Results for Binary Fading Interference Channels with Delayed CSIT,” IEEE Trans. Inf. Theory, vol. 60, no. 10, pp. 6093–6130, Oct 2014.

[29] A. S. Avestimehr, S. N. Diggavi, and D. N. C. Tse, “Wireless network information flow: A deterministic approach,” IEEE Trans. Inf. Theory, vol. 57, no. 4, pp. 1872–1905, April 2011.

[30] 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.

[31] J. Hachem, U. Niesen, and S. N. Diggavi, “Degrees of Freedom of Cache-Aided Wireless Interference Networks,” CoRR, vol. abs/1606.03175, 2016.

[32] 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.

[33] J. Kakar, S. Gherekhloo, and A. Sezgin, “Fundamental limits on latency in transceiver cache-aided HetNets,” in ISIT, June 2017, pp. 2955–2959.

[34] U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands,” IEEE Trans. Inf. Theory, vol. 63, no. 2, pp. 1146–1158, Feb 2017.

[35] Y. Liu and E. Erkip, “Completion time in broadcast channel and interference channel,” in Annual Allerton Conference on Communication, Control, and Computing, Sept 2011, pp. 1694–1701.

[36] R. H. Etkin, D. N. C. Tse, and H. Wang, “Gaussian Interference Channel Capacity to Within One Bit,” IEEE Trans. Inf. Theory, vol. 54, no. 12, pp. 5534–5562, Dec 2008.

[37] S. A. Jafar and M. J. Fakhereddin, “Degrees of Freedom for the MIMO Interference Channel,” IEEE Trans. Inf. Theory, vol. 53, no. 7, pp. 2637–2642, July 2007.

[38] H. Weingarten, Y. Steinberg, and S. S. Shamai, “The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel,” IEEE Trans. Inf. Theory, vol. 52, no. 9, pp. 3936–3964, Sept 2006.

[39] C. Suh and D. Tse, “Interference Alignment for Cellular Networks,” in Annual Allerton Conference on Communication, Control, and Computing, Sept 2008, pp. 1037–1044.

[40] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, “Communication Over MIMO X Channels: Interference Alignment, Decomposition, and Performance Analysis,” IEEE Trans. Inf. Theory, vol. 54, no. 8, pp. 3457–3470, Aug 2008.

[41] 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.

[42] M. A. Maddah-Ali, “On the degrees of freedom of the compound MISO broadcast channels with finite states,” in ISIT, June 2010, pp. 2273–2277.
[43] D. Frank, K. Ochs, and A. Sezgin, “A systematic approach for interference alignment in CSIT-less relay-aided X-networks,” in WCNC, April 2014, pp. 1126–1131.

[44] C. Wang, T. Gou, and S. A. Jafar, “Subspace alignment chains and the degrees of freedom of the three-user mimo interference channel,” IEEE Trans. Inf. Theory, vol. 60, no. 5, pp. 2432–2479, May 2014.

[45] J. Kakar and A. Sezgin, “A Survey on Robust Interference Management in Wireless Networks,” Entropy, vol. 19, no. 7, 2017.

[46] J. Zhang and P. Elia, “Fundamental limits of cache-aided wireless bc: Interplay of coded-caching and csit feedback,” IEEE Trans. Inf. Theory, vol. 63, no. 5, pp. 3142–3160, May 2017.

[47] M. A. Maddah-Ali and D. Tse, “Completely Stale Transmitter Channel State Information is Still Very Useful,” IEEE Trans. Inf. Theory, vol. 58, no. 7, pp. 4418–4431, July 2012.

[48] P. de Kerret, D. Gesbert, J. Zhang, and P. Elia, “Optimal Sum-DoF of the K-user MISO BC with Current and Delayed Feedback,” CoRR, vol. abs/1604.01653, 2016.

[49] A. A. Zewail and A. Yener, “Coded caching for combination networks with cache-aided relays,” in ISIT, June 2017, pp. 2433–2437.

[50] X. Yi and G. Caire, “Topological coded caching,” in ISIT, July 2016, pp. 2039–2043.

[51] F. Xu and M. Tao, “Cache-aided interference management in partially connected wireless networks,” CoRR, vol. abs/1708.09117, 2017.

[52] K. Shanmugam, A. M. Tulino, and A. G. Dimakis, “Coded caching with linear subpacketization is possible using ruzsa-szemeredi graphs,” in ISIT, June 2017, pp. 1237–1241.

[53] H. Vogt, K. Ramm, and A. Sezgin, “Practical Secret-Key Generation by Full-Duplex Nodes with Residual Self-Interference,” in WSA, March 2016, pp. 1–5.