This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

Fundamental limits of Memory-Latency Tradeoff in Fog Radio Access Networks under Arbitrary Demands

Antonious M. Girgis, Ozgur Ercetin, Mohammed Nafie, and Tamer ElBatt

Abstract

We consider a Fog Radio Access Network (F-RAN) with multiple of transmitters and receivers, where each transmitter is connected to the cloud via a fronthaul link. Each network node has a finite cache, where it fills its cache with portions of the library files in the off-peak hours. In the delivery phase, receivers request each library files according to an arbitrary popularity distribution. The cloud and the transmitters are responsible for satisfying the requests. This paper aims to design content placement and coded delivery schemes for minimizing both the *expected* normalized delivery time (NDT) and the peak NDT which measures the transmission latency. We propose achievable transmission policies, and derive an information-theoretic bound on the expected NDT under uniform popularity distribution. Analytical results show that the proposed scheme is within a gap of 2.58 from the derived bound for both the *expected* NDT under uniform popularity distribution and the peak NDT. Next, we investigate the *expected* NDT under an arbitrary popularity distribution for an F-RAN with transmitter-side caches only. The achievable and information-theoretic bounds on the expected NDT are derived, where we analytically prove that our proposed scheme is optimal within a gap of 2 independent of the popularity distribution.

I. Introduction

The explosive growth of the global data traffic has been pushing the wireless systems to a major paradigm shift from a voice-centric to a content-centric architecture, where the multimedia content, such as YouTube and Netflix videos, and data offloading have formed more than half of the overall traffic demand [1]. Caching is considered as one of the most effective techniques to cope with this increasing traffic load. It refers to exploiting inexpensive memories available

Antonious M. Girgis is with Wireless Intelligent Networks Center (WINC), Nile University, Cairo, Egypt.

Ozgur Ercetin is with Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey.

Mohammed Nafie is with Wireless Intelligent Networks Center (WINC), Nile University, Cairo, Egypt and the Dept. of EECE, Faculty of Engineering, Cairo University, Giza, Egypt.

Tamer ElBatt is with the Depart. of Computer Science and Engineering, American University, Cairo, Egypt and the Dept. of EECE, Faculty of Engineering, Cairo University, Giza, Egypt.

at the network edge, such as macro cell, small cell base stations, and user devices, to maximally utilize the scarce wireless network resources. In general, caching systems are modeled to operate in two phases namely the placement phase and the delivery phase. The placement phase occurs in the off-peak times, whereby the network is not congested. Hence, the network resources are utilized to fill the caches of network edge devices with a portion of contents, where the cache memory size represents the only constraint in this phase. The delivery phase occurs during the peak-times in which the wireless network is congested. Thus, the caches at the network edge can be leveraged to reduce the stress on the wireless network by serving portions of the requested content without accessing the network. In [2], the authors proposed the concept of coded caching that achieves a significant reduction of the peak transmission rate for an error-free broadcast channel with caches equipped at the receivers. The main idea of coded caching is to store different contents at each receiver which generates multicast coding opportunities in the delivery phase, such that a single coded transmission becomes useful for multiple receivers. The problem setup in [2] seeks to minimize the peak transmission rate for the worst-case demand over all possible receivers demands. However, in real systems, different contents might have different popularities, where several receivers might request the same content. Hence, the work in [3]–[5] studied the expected transmission rate for non-uniform popularity distribution instead of the peak transmission rate. In [6], the authors provided a full characterization of both the expected rate for uniform popularity distribution and the peak rate of the cache-aided, noiseless broadcast channel with uncoded placement.

Recently, the coded caching concept was studied in interference networks [7]–[13] and in Fog Radio Access Networks (F-RANs) [14]–[21]. Maddah-ali and Niesen in [7] studied a 3×3 cache-aided interference network, where they showed that increasing the caches at transmitters leads to increasing the degrees of freedom (DoF) of the network by allowing cooperation between transmitters. In [8], the authors introduced the normalized delivery time (NDT) which is proportional to the reciprocal DoF as a performance metric for cache-aided interference networks, where the NDT measures the worst case delivery time. In addition, this work developed an information-theoretic bound on the NDT for uncoded cache placement schemes. The authors in [9] proposed a new converse bound on the peak NDT which is tighter than the bound introduced in [8] for small cache sizes, and a converse bound on the expected NDT for uniform popularity distribution was presented. The work in [10]–[12] studied a general interference network with caches equipped not only at the transmitters but also at the receivers. In [13],

the authors characterized the NDT of a 3×3 multi-input-multi-output (MIMO) interference network, where both transmitters and receivers have caches. In [14], a 2×2 F-RAN comprising a central cloud connected to the transmitters via fronthaul links was studied, where each transmitter has an isolated cache memory. The interference network is considered as a special case of F-RAN, in which the fronthaul capacity is zero. In the cache-aided interference networks, the total transmitter-side caches should be larger enough in order to store all the library contents; which is not the case in the F-RANs. Furthermore, F-RAN has an advantage over the interference networks of leveraging the fronthaul links to enable the central processing at the cloud. In [15], the authors extended the work in [14] for arbitrary number of transmitters and receivers. In [16], the authors studied an F-RAN with transmitters equipped with caches, with each transmitter having multiple antennas. F-RAN with caches at both transmitters and receivers was studied in [17]–[20]. Decentralized coded caching was studied in [17] for an F-RAN with two transmitters, where each network node randomly stores some bits from each content independently of each other. It was shown that the proposed decentralized scheme is order optimal for some special cases. The extension for arbitrary number of transmitters and wireless fronthaul link was studied in [18], where the authors showed that their proposed scheme is within a constant factor from the information theoretic bound. In [19], a delivery scheme was proposed for minimizing the NDT of the F-RAN, where the authors designed a centralized placement at transmitters and a decentralized placement at receivers. In [20], the authors developed a delivery scheme for a partially connected F-RAN. In [21], the authors studied the online-caching problem for an F-RAN, in which the contents are updated over time.

In this work, we study F-RANs with caches at both transmitter and receiver sides, where each transmitter is connected to the cloud server via a dedicated fronthaul link with finite capacity. In contrast to prior work in the literature, our goal is to characterize not only the peak NDT of the worst-case receiver demands, but also, the *expected* NDT under an arbitrary popularity distribution. Our work focuses on uncoded placement schemes, in which each network node stores uncoded fragments of the data contents. To the best of our knowledge, this is the first work characterizing the *expected* NDT for F-RANs or interference networks, since all previous works, except our preliminary work in [9], consider only the peak NDT. We summarize our contributions in this work as follows:

- An information-theoretic bound on the expected NDT of the F-RAN under uniform popularity distribution is derived for uncoded placement. Our derived bound follows a similar argument

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

as in [6], where we split the set of all demands into groups, and then we bound each group separately. The novelty of our work is to consider the case of multiple cache-aided transmitters in the presence of fronthaul links. Moreover, we bound the fronthaul-latency and edge-latency separately by applying the cut-set argument.

• We present an achievable delivery scheme for minimizing the expected NDT of the F-RAN. The proposed scheme is a generalization of the delivery scheme in [11] by taking into account the redundancy of receivers demands, and by proposing the coded fronthaul transmission policy for all possible transmitter-cache size. In F-RAN, the fronthaul links can be used to deliver the requested contents to multiple receivers, i.e., increasing the cooperation between transmitters. Although this strategy decreases the edge-latency, the fronthaul-latency would increase, especially for small fronthaul capacity. Thus, there is a trade-off between the fronthaul transmission and edge transmission. In our proposed scheme, the fronthaul links are used only for delivering the bits of the requested contents that are not stored at any transmitter. Furthermore, in the fronhaul transmission policy, we design coded messages for each transmitter in order to reduce the fronthaul-latency.

• We implement a rigorous analysis to bound the gap between the derived lower bound and the achievable bound on the expected NDT for F-RAN under uniform popularity distribution. It is proven that the multiplicative gap is within a constant factor of 2.58 independent of all system parameters. Moreover, we extend the results for the peak NDT. The best characterization of the peak NDT of the F-RAN and the cache-aided interference network is 13.5 for arbitrary placement schemes in [11] and 12 for uncoded placement schemes in [12]. Thus, our analysis improves the previous gap of the peak NDT for uncoded cache placement schemes. Our results show that the fronthaul links are mandatory for small transmitter-cache sizes. However, when the total transmitter-caches can store all library files, using the fornthaul links and/or increasing the transmitter-cache sizes are not necessary for order optimality. This result extends the previous observation about the transmitter-cache sizes in [11].

• The F-RAN with only transmitter-side caches is studied under an arbitrary popularity distribution. One of the most intriguing questions is to determine what to cache at each transmitter for a given popularity distribution. It is shown that caching the most popular files at each transmitter is *not* order optimal, as well as, caching different fractions from each content at each transmitter is *not* order optimal. To this end, we propose a new cache placement policy, in which the library content is divided into two groups according to the popularity distribution

and the total transmitters cache sizes. The transmitters store fractions of the contents from the first group of the most popular contents, while the contents of the second group are not cached at any transmitter. It is proven that our proposed strategy is order optimal, where the expected NDT for F-RAN is characterized within a multiplicative gap of 2 independent of all system parameters and the popularity distribution.

The remainder of the paper is organized as follows. In Section II, the system model and the problem formulation are introduced. Section III summarizes the main results and contributions of our work. An information-theoretic bound on the expected NDT under uniform popularity distribution is derived in Section IV. The achievable scheme for an F-RAN with caches at both transmitters and receivers is proposed in Section V. The expected NDT for an F-RAN with caches at transmitters is characterized in Section VI under an arbitrary popularity distribution. Finally, we conclude the paper in Section VII.

Notations: Let $[K]$ define the set $\{1, \cdots, K\}$ of integers. Let $(x)^+ = \max\{x, 0\}$. The set $[0:1]$ define the real numbers between zero and one. We define $|\mathcal{S}|$ as the cardinality of the set $\mathcal{S}$. For a set $\mathcal{S} = \{S_1, \cdots, S_K\}$, we define a subset $\mathcal{S}_{[k]}$ of the first k elements in the set $\mathcal{S}$.

II. System Model

We consider a Fog-RAN network illustrated in Figure 1 that is comprising of a cloud server that has a library of N files, $\mathcal{W} \triangleq \{W_1, \ldots, W_N\}$, each of size F bits, where each file $W_n \in \mathcal{W}$ is chosen independently and uniformly from $\left[2^F\right]$ at random. A set of K_T transmitters serves a set of K_R receivers over a $K_T \times K_R$ time varying wireless interference channel. Each transmitter TX_i is connected to the cloud via a dedicated fronthaul link of capacity C_F bits per symbol, where the symbol refers to a channel use of the downlink wireless channel. Each transmitter TX_i, $i \in [K_T]$, has a cache memory $\mathcal{V}_i$ of size $M_T F$ bits, where we refer to $\mu_T = M_T/N$ as the normalized transmitter-cache size. Moreover, each receiver RX_j, $i \in [K_R]$, has a cache memory $\mathcal{Z}_j$ of size $M_R F$ bits, where we refer to $\mu_R = M_R/N$ as the normalized receiver-cache size. The system operates in two separate phases, a placement phase and a delivery phase. In the placement phase, the transmitters and receivers have access to the content library $\mathcal{W}$, and hence, each transmitter and each receiver fills its cache memory as an arbitrary function of the content library $\mathcal{W}$ under its cache size constraint without any prior knowledge of the future receiver demands and channel coefficients between the transmitters (TXs) and the receivers (RXs).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

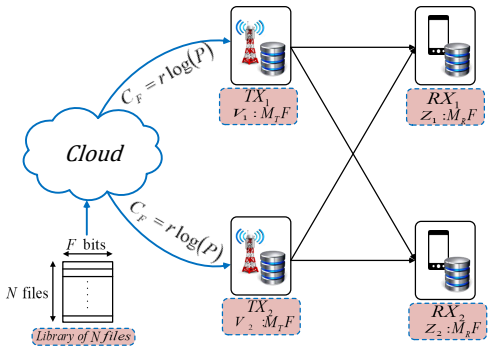

Fig. 1: A Fog Radio Access Network (F-RAN) with $K_T = 2$ TXs and $K_R = 2$ RXs.

In the delivery phase, the receiver demands are revealed in which the receiver RX_j requests a file $W_{d_j} \in \mathcal{W}$, where each receiver demand is applied independently on the other receivers according to popularity distribution $\mathcal{P} = \{p_1, \cdots, p_N\}$ for $\sum_{n=1}^{N} p_n = 1$. In other words, receiver RX_j requests file W_n with probability p_n for all $n \in [N]$. Without loss of generality, the files are sorted such that $p_1 \geq p_2 \geq \cdots \geq p_N$. Let $\mathbf{d} = [d_1, \ldots, d_{K_R}] \in \mathcal{D}$ be the vector of receiver demands, where $\mathcal{D} \triangleq [N]^{K_R}$ denotes the set of all possible demands. Hence, probability of demand vector $\mathbf{d} \in \mathcal{D}$ is given by

$$P(\mathbf{d}) = \prod_{j=1}^{K_R} p_{d_j}.$$

Let $S(\mathbf{d})$ be a function returning the number of distinct files in the demand $\mathbf{d}$. Since, $\mathbf{d}$ is a random vector, $S(\mathbf{d})$ is a random variable that takes values from the set $\mathcal{S} = \{1, \cdots, \min\{N, K_R\}\}$ such that

$$\Pr(S(\mathbf{d}) = s) = \sum_{\substack{\mathbf{d} \in \mathcal{D}: \\ S(\mathbf{d}) = s}} P(\mathbf{d}).$$

At the beginning of the delivery phase, the cloud and the transmitters are aware of all receiver demands . Thus, the cloud maps the library contents $\mathcal{W}$, the receiver demand $\mathbf{d}$, and the channel between TXs and RXs, to a fronthaul message $\mathbf{U}_i \triangleq [U_i(t)]_{t=1}^{T_F}$ of block length T_F that is sent to transmitter TX_i over the fronthaul link. Since the fronthaul link has a limited capacity of C_F bits ber symbol, the size each message $\mathbf{U}_i$, $i \in [K_T]$, cannot exceed $T_F C_F$ bits. Moreover, transmitter TX_i, $i \in [K_T]$, responds to receiver demands by sending a codeword $\mathbf{X}_i \triangleq [X_i(t)]_{t=1}^{T_E}$ of block length T_E over the interference channel with an average power constraint P, where $X_i(t) \in \mathbb{C}$ is the transmitted signal of transmitter TX_i at time $t \in [T_E]$[1]. The transmitted codeword $\mathbf{X}_i$ is an

[1]Note that the fronthaul-transmission latency T_F and the edge-transmission latency T_E are functions of the receiver demand $\mathbf{d}$, however, we drop the index $\mathbf{d}$ to simplify the notations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

encoding function of the receiver demands $\mathbf{d}$, the fronthaul message $\mathbf{U}_i$, the cache contents $\mathcal{V}_i$, and the channel coefficients between TXs and RXs. Afterwards, each receiver RX_j implements a decoding function to estimate the requested file $\hat{W}_{d_j}$ from its cache contents $\mathcal{Z}_j$ and the received signal $\mathbf{Y}_j \triangleq [Y_j(t)]_{t=1}^{T_E}$ which is given by

$$Y_j(t) = \sum_{i=1}^{K_T} h_{ji}(t) X_i(t) + N_j(t), \tag{1}$$

where $Y_j(t) \in \mathbb{C}$ is the received signal by receiver RX_j at time $t \in [T]$, $N_j(t)$ denote the additive white Gaussian noise at receiver RX_j at time $t \in [T_E]$, and $h_{ji}(t) \in \mathbb{C}$ represents the channel gain between transmitter TX_i and receiver RX_j at time t. The channel coefficients $\{h_{ji}(t)\}$ are assumed to be drawn independently and identically distributed (i.i.d.) from a continuous distribution. For a given demand $\mathbf{d} \in \mathcal{D}$, the probability of error for a coding scheme (caching, cloud encoding, transmitter encoding and receiver decoding functions) is given by

$$\mathsf{Pe}(\mathbf{d}) = \max_{j \in [K_R]} \mathbb{P}\left(\hat{W}_{d_j} \neq W_{d_j}\right), \tag{2}$$

which is the maximum error probability over all receivers. We say that a coding scheme is feasible if and only if for each demand $\mathbf{d} \in \mathcal{D}$, the probability of error $\mathsf{Pe}(\mathbf{d}) \to \infty$ when $F \to \infty$. We point out that the end-to-end latency for an F-RAN is given by $T = T_E + T_F$ that depends on two transmission delays: the *fronthaul delay* T_F and the *edge delay* T_E. The fronthaul delay T_F represents the transmission time to deliver the cloud messages to transmitters, while the edge delay T_E represents the transmission time to deliver the messages of transmitters to receivers. In the following, we define the normalized delivery time (NDT) as the performance metric of the network which was first discussed in [14].

Definition 1. For a given demand $\mathbf{d} \in \mathcal{D}$, the Normalized Delivery Time (NDT) for any feasible coding scheme with normalized cache size μ_T, μ_R, and a fronthaul capacity $C_F = r \log(P)$ bits is defined as

$$\tau(\mathbf{d}, \mu_T, \mu_R, r) = \lim_{P \to \infty} \lim_{F \to \infty} \frac{T(\mathbf{d})}{F/\log(P)}, \tag{3}$$

where $T(\mathbf{d}) = T_F(\mathbf{d}) + T_E(\mathbf{d})$ is the end-to-end-latency for satisfying the demand vector $\mathbf{d}$, and r measures the multiplexing gain of the fronthaul links. We point out that the NDT for a given demand $\mathbf{d}$ in (3) represents the transmission delay to serve the receiver demands $\mathbf{d}$ normalized with respect to the interference-free baseline system with transmission rate $\log(P)$ at the high SNR regime. The optimal NDT for a given tuple $(\mathbf{d}, \mu_T, \mu_R, r)$ is defined as

$$\tau^* \triangleq \inf\{\tau : \tau(\mathbf{d}, \mu_T, \mu_R, r) \text{ is feasible}\}. \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

Moreover, we define the minimum expected NDT over all possible demands $\mathbf{d} \in \mathcal{D}$ for a given tuple (μ_T, μ_R, r) by

$$\overline{\tau}^* (\mu_T, \mu_R, r) = \mathbb{E}_{\mathbf{d}} \left\{ \tau^* (\mathbf{d}, \mu_T, \mu_R, r) \right\}, \tag{5}$$

where the expectation is with respect to the random demand $\mathbf{d} \in \mathcal{D}$. Similarly, we define the optimal peak NDT for the worst-case demands as follows

$$\tau^*_{\text{peak}} (\mu_T, \mu_R, r) = \max_{\mathbf{d} \in \mathcal{D}} \tau^* (\mathbf{d}, \mu_T, \mu_R, r), \tag{6}$$

which defines the peak delivery time normalized with respect to the interference-free baseline system. In this work, our objective is to characterize both the expected NDT and the peak NDT as a function of the normalized transmitter-cache size μ_T, the normalized receiver-cache size μ_R, and the fronthaul link multiplexing gain r for uncoded placement schemes.

III. Main Results

In this paper, our main results are divided into two parts. First, we characterize the expected NDT and the peak NDT for a $K_T \times K_R$ F-RAN with caches equipped at both transmitters and receivers, where each receiver requests file W_n, $n \in [N]$, with probability $p_n = 1/N$, i.e., the library files have uniform popularity distribution. Then, we characterize the expected NDT for a $K_T \times K_R$ F-RAN with only transmitter-side caches under an arbitrary popularity distribution.

Theorem 1. *For a $K_T \times K_R$ F-RAN with a normalized transmitter-cache size $\mu_T \in [0:1]$, a normalized receiver-cache size $\mu_R = t/K_R$ for $t \in \{0, \cdots, K_R\}$, and a fronthaul multiplexing gain r, the optimal expected NDT with uniform popularity distribution is bounded by*

$$\begin{aligned} &\overline{\tau}^* (\mu_T, \mu_R, r) \geq \mathbb{E}_{S(\mathbf{d})} \left\{ \overline{\tau}^*_F (S(\mathbf{d}), \mu_T, \mu_R, r) + \overline{\tau}_E (S(\mathbf{d}), \mu_T, \mu_R, r) \right\}, \\ &\overline{\tau}_F (S(\mathbf{d}), \mu_T, \mu_R, r) = \frac{C_t (1 - K_T \mu_T)^+}{r K_T}, \\ &\overline{\tau}^*_E (S(\mathbf{d}), \mu_T, \mu_R, r) = \max \left\{ \frac{C_t}{\min\{K_T, S(\mathbf{d})\}}, (1 - \mu_R) \right\} \end{aligned} \tag{7}$$

for any uncoded cache placement scheme, where $C_t = \frac{\binom{K_R}{t+1} - \binom{K_R - S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}$. Furthermore, the optimal expected NDT for general $\mu_R \in [0:1]$ is bounded by the lower convex envelope of the corner points $\mu_R = t/K_R$ for $t \in \{0, \cdots, K_R\}$.

Proof. The proof is presented in Section IV. ■

This theorem provides a lower bound on the expected NDT under uniform popularity distribution for uncoded placement schemes. To prove Theorem 1, we first divide the demand set $\mathcal{D}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

into groups inspired from the converse bound in [6], where each group of demands has the same number of distinct files $S(\mathbf{d})$. Then, we derive a lower bound on both the fronthaul and edge NDTs separately for each group of demands for uncoded placement schemes at both transmitters and receivers. Our bound is mainly based on the cut-set argument. After that, we take the average over all demands, and optimize the derived bound over all possible uncoded placement schemes to obtain the lower bound in Theorem 1.

Theorem 2. *For a $K_T \times K_R$ F-RAN with a normalized transmitter-cache size $\mu_T \in [0:1]$, a normalized receiver-cache size $\mu_R = t/K_R$ for $t \in \{0, \cdots, K_R\}$, and a fronthaul multiplexing gain r, the achievable expected NDT with uniform popularity distribution is given by*

$$\begin{aligned}
\overline{\tau}(\mu_T, \mu_R, r) &= \mathbb{E}_{S(\mathbf{d})}\left\{\overline{\tau}_F\left(S(\mathbf{d}), \mu_T, \mu_R, r\right) + \overline{\tau}_E\left(S(\mathbf{d}), \mu_T, \mu_R, r\right)\right\}, \\
\overline{\tau}_F\left(S(\mathbf{d}), \mu_T, \mu_R, r\right) &= \frac{\min\{S(\mathbf{d}), \frac{K_R}{t+1}\}}{rK_T}\left(1-\mu_R\right)\left(1-K_T\mu_T\right)^+, \\
\overline{\tau}_E\left(S(\mathbf{d}), \mu_T, \mu_R, r\right) &= \frac{K_T + \min\{S(\mathbf{d}), \frac{K_R}{t+1}\} - 1}{K_T}\left(1-\mu_R\right).
\end{aligned} \tag{8}$$

Furthermore, for general $\mu_R \in [0:1]$, the expected NDT is obtained from the lower convex envelope of the corner points $\mu_R = t/K_R$ for $t \in \{0, \cdots, K_R\}$.

Proof. The proof is presented in Section V. ■

In contrast to prior work in the literature that seeks to design transmission schemes for minimizing the peak NDT, Theorem 2 provides a new delivery scheme for F-RANs with caches at both transmitters and receivers to minimize the expected NDT for uniform popularity distribution, where we take into consideration the redundancy of the receiver requests. In the achievable scheme, we propose two possible methods of transmission, and then we choose the method that has the lower NDT. In the first method, we extend the transmission scheme introduced in [11] for arbitrary transmitter-cache sizes. The novel part in our derivation is designing the fronthaul transmission policy to transmit different coded, multicast messages to every transmitter in order to reduce the fronthaul latency. In the second method, we exploit the redundancy of the receiver requests by grouping the receivers that request the same file and dealing the problem as a compound X-channel defined in [22]. The proposed delivery scheme depends mainly on the value of the transmitter-cache size. When the total transmitters caches cannot store the whole library bits, i.e., $\mu_T < 1/K_T$, the fronthaul links are exploited to deliver the requested bits

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

that are not available at the transmitters. Meanwhile the delivery scheme does not exploit the fronthaul links when $\mu_T \geq 1/K_T$, even at high fronthaul capacity.

Theorem 3. *For a $K_T \times K_R$ F-RAN with a normalized transmitter-cache size $\mu_T \in [0:1]$, a normalized receiver-cache size $\mu_R = [0:1]$, and a fronthaul multiplexing gain r, the multiplicative gap between the achievable expected NDT in Theorem 2 and the lower bound in Theorem 1 is bounded by*

$$\frac{\overline{\tau}\left(\mu_T, \mu_R, r\right)}{\overline{\tau}^*\left(\mu_T, \mu_R, r\right)} \leq 2.58 \tag{9}$$

independent of all system parameters.

Proof. The proof is presented in Appendix A. ■

This theorem characterizes the expected NDT of an F-RAN with caches at transmitters and receivers under uniform popularity distribution within a constant factor of 2.58. In the detailed proof of this theorem, we can see that this gap is reduced to 1.58 for an F-RAN with a single transmitter $K_T = 1$. The results on the expected NDT can be extended to the peak NDT as in the following corollary.

Corollary 1. For a $K_T \times K_R$ F-RAN with a normalized transmitter-cache size $\mu_T \in [0:1]$, a normalized receiver-cache size $\mu_R = [0:1]$, and a fronthaul multiplexing gain r, the multiplicative gap between the achievable peak NDT (for worst-case demand) and the lower bound is bounded by

$$\frac{\tau_{\text{peak}}\left(\mu_T, \mu_R, r\right)}{\tau^*_{\text{peak}}\left(\mu_T, \mu_R, r\right)} \leq 2.58 \tag{10}$$

independent of all system parameters.

Proof. The proof is a special case of the above theorems, where the achievable scheme is obtained from Theorem 2 by setting $S\left(\mathbf{d}\right) = \min\{N, K_R\}$. The lower bound is obtained from Theorem 1 for $S\left(\mathbf{d}\right) = \min\{N, K_R\}$. In the proof of Theorem 3, we show that the gap between the achievable scheme in Theorem 2 and the lower bound in Theorem 1 is less than 2.58 for each value of $S\left(\mathbf{d}\right) \in \mathcal{S}$. Hence, the proof is completed. ■

To the best of our knowledge, Theorem 3 provides the first characterization of the expected NDT for F-RANs, where the work in the literature consider only the peak NDT. Furthermore, Corollary 1 provides the best characterization of the peak NDT for F-RANs. The best characterization for cache-aided interference networks (F-RAN with $r = 0$) known in the literature has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

a multiplicative gap of 12 for $K_T \leq K_R$ and 2 for $K_T \geq K_R$ for uncoded placement schemes in [12, Corollary 2], and has a multiplicative gap of 13.5 for arbitrary placement schemes in [11].

In the achievable scheme introduced in Section V, the fronthaul links do not contribute to the transmission when the transmitters can store all the library files, i.e., $\mu_T \geq 1/K_T$. Furthermore, we neglect the gain that can be obtained from the cooperation between transmitters when the transmitter-cache sizes increase. Theorem 3 shows that this achievable bound is within a constant factor of 2.58 from the minimum expected NDT independent of all values of system parameters. Hence, using the fronthaul links when $\mu_T \geq 1/K_T$ and/or increasing the transmitter-cache sizes more than $\mu_T \geq 1/K_T$ can at most improve the expected NDT within a constant factor from our proposed scheme. For cache-aided interference networks, i.e., $r = 0$, the authors in [11] have shown that increasing the transmitter-cache size can just improve the peak NDT at most within a constant factor. Therefore, our insights regarding the transmitter caches coincide with the authors' insights in [11]. In addition, our insights are more general, since we consider the expected NDT (not only the peak NDT) for F-RANs, and we also show that using the fronthaul links when $\mu_T \geq 1/K_T$ is not necessary for order optimality.

In general, transmitters cooperation might improve the performance of F-RANs for *some special cases* of system parameters. For instance, it is shown in [15, Lemma 2] that exploiting the cooperation between transmitters achieves the optimal peak NDT when $\mu_T = 1$ and $\mu_R = 0$. Furthermore, in [12], the authors developed a transmission scheme using zero-forcing and interference alignment to obtain the optimal peak NDT when $K_R\mu_R + K_T\mu_T \geq K_R$. However, our focus in this work is mainly on the expected NDT. When we take into account the redundancy of the receivers requests, the transmitters cooperation is useful in some demands, while it cannot provide any additional gains to the NDT for other demands. For example, when each receiver requests a different file, the optimal peak NDT is obtained by exploiting the transmitters cooperation. However, for demands $\mathbf{d} \in \mathcal{D}$ such that each group of more than K_T receivers requests a single file[2], the transmission problem becomes a *Compound broadcast channel* [22], in which each group of receivers is treated as a single receiver with more than K_T channel states. From [22, Theorem 2], we can verify that the achievable scheme in Section V is optimal *for these demands*. Hence, the cooperation between transmitters does not provide any benefits

[2]We assume that $K_T << K_R$, and each transmitter-cache can store all the library $\mu_T = 1$, while there are no caches at the receivers $\mu_R = 0$.

for such demands. Now, we focus on the F-RANs with caches at transmitters only, i.e., there is no caches at receivers $\mu_R = 0$. Moreover, we consider an arbitrary popularity distribution $\mathcal{P}$ not only uniform distribution as in the previous theorems.

Theorem 4. *For a $K_T \times K_R$ F-RAN with a normalized transmitter-cache size $\mu_T \in [0:1]$, a fronthaul multiplexing gain r, the optimal expected NDT under popularity distribution $\mathcal{P}$ achieves*

$$\frac{1}{2}\overline{\tau}\left(\mu_T, r\right) \leq \overline{\tau}^{*}\left(\mu_T, r\right) \leq \overline{\tau}\left(\mu_T, r\right), \tag{11}$$

where $\overline{\tau}\left(\mu_T, r\right)$ denotes the achievable expected NDT given by

$$\begin{aligned} \overline{\tau}\left(\mu_T, r\right) &= \overline{\tau}_F\left(\mu_T, r\right) + \overline{\tau}_E\left(\mu_T, r\right), \\ \overline{\tau}_F\left(\mu_T, r\right) &= \frac{\sum_{n=L+1}^{N}\left(1-\left(1-p_n\right)^{K_R}\right)}{rK_T}, \\ \overline{\tau}_E\left(\mu_T, r\right) &= \mathbb{E}_{\mathbf{d}}\left\{\frac{K_T + S\left(\mathbf{d}\right) - 1}{K_T}\right\}, \end{aligned} \tag{12}$$

where $L = \min\{N, K_T M_T\}$.

Proof. The proof is presented in Section VI. ■

In the achievable bound on the expected NDT, the library files are divided into two groups depending on the aggregate transmitter-cache sizes $K_T M_T$. The first group contains the most $L = \min\{N, K_T M_T\}$ popular files, while the second group contains the remaining least popular files. Each transmitter stores a different fractions of each file belonging to the first group, while the files of the second group are not cached at the caches of the transmitters. Theorem 4 shows that this caching strategy is approximately optimal within a multiplicative gap of 2 from the derived lower bound. When we deal with non-uniform popularity distribution of the library files, there is a debate about the best placement and delivery polices in order to minimize the expected NDT. One of the possible placement schemes is to store the most popular files at each transmitter, i.e., each transmitter stores the first M_T files from the library. This policy is called the Highest-Popularity-First(HPF) scheme. The HPF scheme boosts the cooperation between transmitters to deliver the requests coming to the most popular M_T files, and hence, this policy reduces the NDT of delivering these files. Meanwhile, the requests coming to the least $N - M_T$ popular files that are not available at the transmitters should first be delivered to the transmitters through the fronthaul link increasing the fronthaul NDT. The second possible placement scheme

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

called Highest-Content-Disparity (HCD) aims to store different contents from each file in the library at each transmitter. The HCD policy aims to store as much as possible of the library contents at the transmitters caches in order to reduce the number of bits delivered through the fronthaul links. However, this placement strategy does not take the popularity distribution of the files into consideration. Although the discussed policies might achieve better performance than our proposed scheme in Theorem 4 at certain values of the systems parameters, these policies are not order optimal as we can see in the following example.

Example 1. Consider an F-RAN with $K_T = 2$ transmitters, $K_R = 2$ receivers, and a multiplexing gain of the fronthaul link r. The library contains $N = 3$ files, $\mathcal{W} = \{A, B, C\}$, with popularity distribution $\mathcal{P} = \left\{\frac{4}{9}, \frac{4}{9}, \frac{1}{9}\right\}$.

First, assume that each transmitter has a cache of size $M_T = 2$ files. In the HPC, policy both transmitters TX_1 and TX_2 store the two most popular files A and B, while file C is not stored at none of them, $\mathcal{V}_1 = \mathcal{V}_2 = (A, B)$. Hence, the expected NDT for the HPF policy is bounded by

$$\overline{\tau}_{\text{HPF}} \geq \frac{1 - \left(1 - \frac{1}{9}\right)^2}{2r} + 1 = \frac{17/81}{2r} + 1, \tag{13}$$

In the fronthaul transmission, half of file C is delivered from each transmitter if and only if one of the receivers requests it. In our proposed policy introduced in Section VI-A as well as the HCD policy, each file is split into two smaller subfiles, e.g., $A = (A_1, A_2)$, and transmitter TX_1 stores the first subfile, and transmitter TX_2 stores the second subfile. Thus, $\mathcal{V}_1 = (A_1, B_1, C_1)$ and $\mathcal{V}_2 = (A_2, B_2, C_2)$. Hence, every bit of the library is already stored at one of the transmitters. By using the proposed delivery scheme in Theorem 4, the expected NDT is given by

$$\overline{\tau}_{\text{Proposed}} = \mathbb{E}_{\mathbf{d}} \left\{ \frac{K_T + S(\mathbf{d}) - 1}{K_T} \right\} = \frac{105}{81}. \tag{14}$$

By comparing the expected NDT of our proposed policy and the HPF policy, we can see that the HPF has a lower expected NDT for $r \geq 17/48$. However, our proposed scheme is better for small multiplexing gain $r < 17/48$. Furthermore, we can verify that the multiplicative gap between the expected NDT of the HPF and the lower bound derived in Section VI-B is a function of the reciprocal multiplexing gain r. Hence, this gap increases to infinity as the multiplexing gain goes to zero. **As a result, HPF is not order optimal for all system parameters**.

Second, we assume that each transmitter has a cache of size $M_T = 1$ file. In the HCD, each transmitter stores different fraction of $1/3F$ bits from each file in the library. Hence, each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

file $W_n \in \mathcal{W}$ is split into three subfiles, e.g., $A = (A_1, A_2, A_3)$. The transmitter TX_1 stores $\mathcal{V}_1 = (A_1, B_1, C_1)$, and transmitter TX_2 stores $\mathcal{V}_2 = (A_2, B_2, C_2)$, while the remaining third subfile (A_3, B_3, C_3) from each file is not stored at any transmitter. Hence, the expected NDT of the HCD is bounded by

$$\overline{\tau}_{\text{HCD}} \geq \frac{\mathbb{E}_{\mathbf{d}}\left\{S\left(\mathbf{d}\right)\right\}}{6r} + 1 = \frac{129/81}{6r} + 1, \tag{15}$$

The half of the third subfile from each requested file is delivered to each transmitter through the fronthaul link[3]. In the proposed strategy from Theorem 4, each file is split into two smaller subfiles, for example file A is split to $A = (A_1, A_2)$. Transmitter TX_1 stores $\mathcal{V}_1 = (A_1, B_1)$ and transmitter TX_2 stores $\mathcal{V}_2 = (A_2, B_2)$, while file C is not cached at any of them. From Theorem 4, the expected NDT is given by

$$\begin{aligned}\overline{\tau}_{\text{Proposed}} &= \frac{\sum_{n=L+1}^{N}\left(1-(1-p_n)^{K_R}\right)}{K_T r} + \mathbb{E}_{\mathbf{d}}\left\{\frac{K_T + S\left(\mathbf{d}\right) - 1}{K_T}\right\} \\ &= \frac{17/81}{2r} + \frac{105}{81},\end{aligned} \tag{16}$$

where $L = \min\{N, K_T M_T\} = 2$. When we compare the performance of the proposed scheme with that of the HCD scheme, it can be seen that the HCD scheme is better when the multiplexing gain $r \geq 13/24$, while the proposed scheme outperforms the HCD scheme for $r < 13/24$. Moreover, similar to the HPF scheme, the gap between the HCD scheme and the lower bound derived in Section VI-B is a function of the reciprocal multiplexing gain r. Hence, **not only the HPF scheme but also the HCD scheme is not order optimal for all system parameters**.

This example explains that the HPF and HCD schemes can be considered suboptimal policies, since their performance is not guaranteed for all system parameters. While, our proposed scheme is robust against all system parameters and the popularity distribution.

IV. Converse Bound on the optimal expected NDT

This section establishes a lower bound on the expected NDT under uniform popularity distribution for uncoded cache placement in Theorem 1. The proof is based on the cut-set argument. Let $\mathcal{V} = [\mathcal{V}_1, \cdots, \mathcal{V}_{K_T}]$ and $\mathcal{Z} = [\mathcal{Z}_1, \cdots, \mathcal{Z}_{K_R}]$ denote the placement scheme at

[3]We point out that equations (13) and (15) do not give the exact performance of the HPF policy and the HCD policy; however, they provide a lower bound on the expected NDT of these policies. In other words the performance of the HPC and the HCD might be worse than given by these equations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

transmitters and receivers, respectively. Furthermore, let $\tau_F(\mathbf{d},\mathcal{V},\mathcal{Z})$ and $\tau_E(\mathbf{d},\mathcal{V},\mathcal{Z})$ denote the NDT of fronthaul and edge transmissions for given demand $\mathbf{d}$ and placement schemes at transmitters and receivers. The expected NDT is obtained by

$$\begin{aligned}\overline{\tau}^*(\mu_T,\mu_R,r) &= \min_{\mathcal{V},\mathcal{Z}} \mathbb{E}_{\mathbf{d}}\left\{\tau(\mathbf{d},\mathcal{V},\mathcal{Z})\right\} \\ &\overset{(a)}{=} \min_{\mathcal{V},\mathcal{Z}} \mathbb{E}_{S(\mathbf{d})}\left\{\mathbb{E}_{\mathbf{d}|S(\mathbf{d})}\left\{\tau(\mathbf{d},\mathcal{V},\mathcal{Z})\,|S(\mathbf{d})=s\right\}\right\} \\ &\overset{(b)}{\geq} \mathbb{E}_{S(\mathbf{d})}\left\{\min_{\mathcal{V},\mathcal{Z}} \mathbb{E}_{\mathbf{d}|S(\mathbf{d})}\left\{\tau(\mathbf{d},\mathcal{V},\mathcal{Z})\,|S(\mathbf{d})=s\right\}\right\}.\end{aligned} \tag{17}$$

Step (a) follows from the conditional expectation, where we take the expectation with respect to $\mathbf{d}$ conditioned that $S(\mathbf{d})=s$, then, we take the expectation over all values of s. Step (b) follows from the fact that minimization of the weighted sum of non-negative terms is not lower than the weighted sum of minimizing each term individually. Thus, we split all possible demands $\mathbf{d}\in\mathcal{D}$ into categories $\{\mathcal{D}_s\}$, where $\mathcal{D}_s$ contains all the demands $\mathbf{d}$ that have exactly s number of distinct files. Hence, we bound the NDT for each category $\mathcal{D}_s$ in order to obtain the lower bound on the expected NDT. For a set $\mathcal{S}$, we define $\prod_{\mathcal{S}}$ as a set of $|\mathcal{S}|!$ permutations of $\mathcal{S}$. We say that set $\mathcal{B}=\{\mathcal{B}_1,\cdots,\mathcal{B}_s\}$ is a partition of the set $[K_R]$ into s blocks, when $\{\mathcal{B}_i\subseteq[K_R]: i\in[s]\}$ are nonempty, disjoint sets with $\bigcup_{i=1}^{s}\mathcal{B}_i \triangleq [K_R]$. Let $\mathcal{S}$ be a set containing all possible partition sets $\mathcal{B}$, and hence, $|\mathcal{S}|=\text{Str}(K_R,s)$, where $\text{Str}(K_R,s)$ is the Stirling number of the second kind. Furthermore, we define set $\mathcal{S}_O=\bigcup_{\mathcal{B}}\prod_{\mathcal{B}}$ containing permutations of all possible partition sets $\mathcal{B}$, where $|\mathcal{S}_O|=s!\text{Str}(K_R,s)$. As a result, the total number of demands in $\mathcal{D}_s$ can be calculated by

$$|\mathcal{D}_s| = s!\binom{N}{s}\ \text{Str}(K_R,s), \tag{18}$$

where the term $s!\binom{N}{s}$ counts the number of ways to make an ordered selection of distinct s files from the library of N files. The second term $\text{Str}(K_R,s)$ counts the number of ways to distribute K_R distinct RXs into s distinct files, such that each file is requested at least by one RX [4]. Instead of averaging over the set $\mathcal{D}_s$, we will average over a larger set $\tilde{\mathcal{D}}_s \triangleq \bigcup_{\mathcal{W}_s\in\prod_{\mathcal{F}\subseteq[N]}}\ \bigcup_{\pi\in\mathcal{S}_O}\mathbf{d}(\pi,\mathcal{W}_s)$, where each demand $\mathbf{d}\in\mathcal{D}_s$ is repeated $s!$ times in the set $\tilde{\mathcal{D}}_s$. Let $\pi=(\mathcal{B}_1,\cdots,\mathcal{B}_s)$. For demand $\mathbf{d}(\pi,\mathcal{W}_s)$, receivers in set $\mathcal{B}_j$ request the j-th file in the ordered set $\mathcal{W}_s$. For a given demand $\mathbf{d}(\pi,\mathcal{W}_s)$, we consider the set $\mathcal{R}=\{R_1,\cdots,R_s\}$ of s receivers, in which receiver R_j is picked

[4]This counting problem is the same as distributing K_R distinguishable balls into s distinguishable boxes, such that each box contains at least one ball (See [23, Ch.2]).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

uniformly at random from the set $\mathcal{B}_j$ in which the ordered set $\mathcal{W}_s = \{W_{d_{R_1}}, \cdots, W_{d_{R_s}}\}$. The reason behind using permutations of partition sets and picking receivers uniformly at random is to make a symmetry of selecting an arbitrary receiver in position $i \in [s]$. In following, we bound the NDT of the fronthaul transmission and the edge transmission.

A. Bound on The NDT of The Fronthaul Transmission

The main idea is that the transmission signals $\mathbf{X}_{[K_T]}$ can be constructed from the cache contents at the K_T transmitters ($\mathcal{V}$), and the fronthaul messages received by K_T transmitters ($\mathbf{U}_{[K_T]}$). Therefore, the s distinct files $\mathcal{W}_s$ can be decoded from the cache contents $\mathcal{V}$, the fronthaul messages $\mathbf{U}_{[K_T]}$, and the cache contents at the set $\mathcal{R}$ of s receivers $\mathcal{Z}_{\mathcal{R}}$. Thus, by using Fano's inequality,

$$H\left(\mathcal{W}_s|\mathbf{U}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}}\right) \leq sF\epsilon_F, \tag{19}$$

where $\epsilon_F \to 0$ as $F \to \infty$. Now, consider the following bound

$$\begin{aligned}
H\left(\mathcal{W}_s\right) &\overset{(a)}{=} H\left(\mathcal{W}_s \middle| \tilde{\mathcal{W}}_s\right) \\
&\overset{(b)}{=} I\left(\mathcal{W}_s; \mathbf{U}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}} \middle| \tilde{\mathcal{W}}_s\right) + H\left(\mathcal{W}_s \middle| \tilde{\mathcal{W}}_s, \mathbf{U}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}}\right) \\
&\overset{(c)}{\leq} H\left(\mathbf{U}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}} \middle| \tilde{\mathcal{W}}_s\right) - H\left(\mathbf{U}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}} | \mathcal{W}\right) + sF\epsilon_F \\
&= H\left(\mathbf{U}_{[K_T]} \middle| \tilde{\mathcal{W}}_s\right) + H\left(\mathcal{V} \middle| \tilde{\mathcal{W}}_s, \mathbf{U}_{[K_T]}\right) + H\left(\mathcal{Z}_{\mathcal{R}} \middle| \tilde{\mathcal{W}}_s, \mathbf{U}_{[K_T]}, \mathcal{V}\right) + sF\epsilon_F \\
&\overset{(d)}{\leq} H\left(\mathbf{U}_{[K_T]}\right) + H\left(\mathcal{V} \middle| \tilde{\mathcal{W}}_s\right) + \sum_{i=1}^{s} H\left(\mathcal{Z}_{R_i} \middle| \tilde{\mathcal{W}}_s, \mathbf{X}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}_{[i-1]}}\right) + sF\epsilon_F \\
&\overset{(e)}{=} T_F K_T r \log\left(P\right) + H\left(\mathcal{V} \middle| \tilde{\mathcal{W}}_s\right) + \sum_{i=1}^{s} H\left(\mathcal{Z}_{R_i} \middle| \tilde{\mathcal{W}}_s, \mathbf{X}_{[K_T]}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}_{[i-1]}}\right) + sF\epsilon_F \\
&\overset{(f)}{\leq} T_F K_T r \log\left(P\right) + H\left(\mathcal{V} \middle| \tilde{\mathcal{W}}_s\right) + \sum_{i=1}^{s} H\left(\mathcal{Z}_{R_i} \middle| \tilde{\mathcal{W}}_s, \mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}_{[1:i-1]}}\right) + F\epsilon_F,
\end{aligned} \tag{20}$$

where $\tilde{\mathcal{W}}_s = \mathcal{W} \setminus \mathcal{W}_s$. Step (a) follows from the fact that the files are independent. (b) follows from the chain rule. (c) follows from the Fano's inequality in (19). Step (d) follows from the chain rule and the fact that conditioning reduces entropy. Moreover, step (d) follows from the fact that $\mathbf{X}_{[K_T]}$ is a function of $\mathcal{V}$ and $\mathbf{U}_{[K_T]}$. Step (e) follows from the bound on the capacity of the fronthaul link over T_F symbol transmission in the high SNR regime. Step (f) follows from the fact that $W_{d_{R_i}}$ can be decoded from the transmitted signals $\mathbf{X}_{[K_T]}$ and the cache contents $\mathcal{Z}_{R_i}$ of the receiver R_i, where $\mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}} = \{W_{d_{R_1}}, \cdots, W_{d_{R_{i-1}}}\}$. By taking the limit $P \to \infty$ and dividing both terms by F, we get

$$\tau_F^*\left(\mathbf{d}, \mu_T, \mu_R, r\right) \geq \frac{1}{rK_TF}\left(H\left(\mathcal{W}_s\right) - H\left(\mathcal{V} \middle| \tilde{\mathcal{W}}_s\right) - \sum_{i=1}^{s} H\left(\mathcal{Z}_{R_i} \middle| \tilde{\mathcal{W}}_s, \mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}_{[1:i-1]}}\right)\right). \tag{21}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

Since, we consider uncoded cache placement, the term $H\left(\mathcal{V}\left|\tilde{\mathcal{W}}_s\right.\right)$ denotes the bits of files $\mathcal{W}_s$ stored at the transmitters caches $\mathcal{V}$. Moreover, the term $H\left(\mathcal{Z}_{R_i}\left|\tilde{\mathcal{W}}_s, \mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}}, \mathcal{V}, \mathcal{Z}_{\mathcal{R}_{[1:i-1]}}\right.\right)$ denotes the bits of files $\{W_{d_{R_i}}, \cdots, W_{d_{R_s}}\}$ stored at the receiver R_i cache ($\mathcal{Z}_{R_i}$) and not stored at the transmitters caches $\mathcal{V}$. Hence, (21) is rewritten by

$$\tau_F^*\left(\mathbf{d}, \mu_T, \mu_R, r\right) \geq \frac{1}{rK_TF}\left(\sum_{i=1}^{s}\sum_{j=1}^{F}\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{Z}_{\mathcal{R}_{[i]}}\right)\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{V}\right)\right), \tag{22}$$

where $B_{d_{R_i},j}$ denotes the j-th bit of the file $W_{d_{R_i}}$. Let $\mathcal{K}_{d_{R_i},j}$ denotes the set of receivers that stores the j-th bit of the file $W_{d_{R_i},j}$, and hence, $\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{Z}_{\mathcal{R}_{[i]}}\right) = \mathbf{1}\left(\mathcal{K}_{d_{R_i},j} \cap \mathcal{R}_{[i]} = \phi\right)$. By taking the average over all demands $\mathbf{d} \in \tilde{\mathcal{D}}_s$, we get

$$\begin{aligned}
\overline{\tau}_F^*\left(S\left(\mathbf{d}\right), \mu_T, \mu_R, r\right) &= \frac{1}{|\tilde{\mathcal{D}}_s|}\sum_{\mathcal{W}_s \in \prod_{\mathcal{F}\subseteq[N]}}\sum_{\pi\in\mathcal{S}_O}\overline{\tau}_F^*\left(\mathbf{d}\left(\pi, \mathcal{W}_s\right), \mu_T, \mu_R, r\right) \\
&\geq \frac{1}{rK_TF}\left(\frac{1}{|\tilde{\mathcal{D}}_s|}\sum_{\mathcal{W}_s \in \prod_{\mathcal{F}\subseteq[N]}}\sum_{\pi\in\mathcal{S}_O}\sum_{i=1}^{s}\sum_{j=1}^{F}\mathbf{1}\left(\mathcal{K}_{d_{R_i},j} \cap \mathcal{R}_{[i]} = \phi\right)\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{V}\right)\right) \\
&= \frac{1}{rK_TF}\left(\frac{1}{s!\binom{N}{s}s!\ \mathrm{Str}\left(K_R, s\right)}\sum_{\mathcal{W}_s \in \prod_{\mathcal{F}\subseteq[N]}}\sum_{\pi\in\mathcal{S}_O}\sum_{i=1}^{s}\sum_{j=1}^{F}\mathbf{1}\left(\mathcal{K}_{d_{R_i},j} \cap \mathcal{R}_{[i]} = \phi\right)\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{V}\right)\right),
\end{aligned} \tag{23}$$

where $\overline{\tau}_F^*\left(S\left(\mathbf{d}\right), \mu_T, \mu_R, r\right)$ denotes the expected fronthaul NDT over all demands $\mathbf{d} \in \mathcal{D}_s$. By changing the order of summation, we calculate the term $\frac{1}{s!\binom{N}{s}}\sum_{\mathcal{W}_s}\prod_{\mathcal{F}\subseteq[N]}\mathbf{1}\left(\mathcal{K}_{d_{R_i},j} \cap \mathcal{R}_{[i]} = \phi\right)\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{V}\right)$ as follows. Note that a file $W_n \in \mathcal{W}$ appears $(s-1)!\binom{N-1}{s-1}$ times at the i-the order in the ordered set $\mathcal{W}_s$. Thus, we have

$$\frac{1}{s!\binom{N}{s}}\sum_{\mathcal{W}_s}\prod_{\mathcal{F}\subseteq[N]}\mathbf{1}\left(\mathcal{K}_{d_{R_i},j} \cap \mathcal{R}_{[i]} = \phi\right)\mathbf{1}\left(B_{d_{R_i},j} \notin \mathcal{V}\right) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left(\mathcal{K}_{n,j} \cap \mathcal{R}_{[i]} = \phi\right)\mathbf{1}\left(B_{n,j} \notin \mathcal{V}\right). \tag{24}$$

Hence, the inequality (23) can be evaluated by using (24)

$$\overline{\tau}_F^*\left(S\left(\mathbf{d}\right), \mu_T, \mu_R, r\right) \geq \frac{1}{rK_TF}\left(\frac{1}{Ns!\ \mathrm{Str}\left(K_R, s\right)}\sum_{\pi\in\mathcal{S}_O}\sum_{i=1}^{s}\sum_{n=1}^{N}\sum_{j=1}^{F}\mathbf{1}\left(\mathcal{K}_{n,j} \cap \mathcal{R}_{[i]} = \phi\right)\mathbf{1}\left(B_{n,j} \notin \mathcal{V}\right)\right). \tag{25}$$

Consider the term $\frac{1}{s!\ \mathrm{Str}(K_R,s)}\sum_{\pi\in\mathcal{S}_O}\mathbf{1}\left(\mathcal{K}_{n,j} \cap \mathcal{R}_{[i]} = \phi\right)$. We take the average over all possible ordered partition sets $\pi = (\mathcal{B}_1, \cdots, \mathcal{B}_s)$, and receiver $\mathcal{R}_l$ is picked uniformly at random from the set $\mathcal{B}_l$. Therefore, this term refers to the probability of choosing i receivers from K_R receivers, and none of them belongs to $\mathcal{K}_{n,j}$. Hence, this term is equal

$$\frac{1}{s!\ \mathrm{Str}\left(K_R, s\right)}\sum_{\pi\in\mathcal{S}_O}\mathbf{1}\left(\mathcal{K}_{n,j} \cap \mathcal{R}_{[i]} = \phi\right) = \frac{\binom{K_R - |\mathcal{K}_{n,j}|}{i}}{\binom{K_R}{i}}. \tag{26}$$

Let $a_{t,0}$ denotes the fraction of bits from the library that are exclusively stored at t receivers (i.e., when $|\mathcal{K}_{n,j}| = t$) and none of the transmitters. Moreover, let $a_{t,1}$ denotes the fraction of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

bits from the library that exclusively stored at t receivers (i.e., when $|\mathcal{K}_{l,j}| = t$) and at least one of the transmitters. Thus, we have

$$\frac{1}{NF}\sum_{n=1}^{N}\sum_{j=1}^{F}\frac{\binom{K_R-|\mathcal{K}_{n,j}|}{i}}{\binom{K_R}{i}}\mathbf{1}\left(B_{n,j}\notin\mathcal{V}\right)=\sum_{t=0}^{K_R}a_{t,0}\frac{\binom{K_R-t}{i}}{\binom{K_R}{i}}=\sum_{t=0}^{K_R}a_{t,0}\frac{\binom{K_R-i}{t}}{\binom{K_R}{t}},\tag{27}$$

where we use the equality $\binom{K_R-t}{i}/\binom{K_R}{i}=\binom{K_R-i}{t}/\binom{K_R}{t}$. Now , we follow similar steps as in [6] to get

$$\sum_{i=1}^{s}\sum_{t=0}^{K_R}a_{t,0}\frac{\binom{K_R-i}{t}}{\binom{K_R}{t}}=\sum_{t=0}^{K_R}a_{t,0}\frac{\binom{K_R}{t+1}-\binom{K_R-s}{t+1}}{\binom{K_R}{t}},\tag{28}$$

where we use the equality [6, Eq (37)]. Substituting from (28) into (25), then we obtain

$$\overline{\tau}_F^*\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)\geq\frac{1}{rK_T}\sum_{t=0}^{K_R}a_{t,0}\frac{\binom{K_R}{t+1}-\binom{K_R-S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}\tag{29}$$

Now, we minimize both sides in (29) over all possible uncoded placement at transmitters $\mathcal{V}$ and at receivers $\mathcal{Z}$ such that

$$\sum_{t=0}^{K_R}a_{t,0}+a_{t,1}=1\tag{30}$$

$$\sum_{t=0}^{K_R}a_{t,0}=1-K_T\mu_T\tag{31}$$

$$\sum_{t=0}^{K_R}t\left(a_{t,0}+a_{t,1}\right)=K_R\mu_R,\tag{32}$$

where the constraint (30) comes from the total number of bits in the library. (31) is to maintain the total cache sizes at the transmitters[5], while (32) is to maintain the total cache sizes at receivers. Let $a_{t,0}=\alpha_{t,0}\left(1-K_T\mu_T\right)$. Thus, the inequality (29) is equal to

$$\overline{\tau}_F^*\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)\geq\frac{1}{rK_T}\sum_{t=0}^{K_R}\alpha_{t,0}\frac{\binom{K_R}{t+1}-\binom{K_R-S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}\left(1-K_T\mu_T\right)^+,\tag{33}$$

where the constraint (31) is equal to $\sum_{t=0}^{K_R}\alpha_{t,0}=1$. Furthermore, we consider the fact that $\overline{\tau}_F^*\left(.,.,.,.\right)\geq 0$. Now, we want to minimize the right hand side of (33) over all possible uncoded cache placement schemes, such that $\sum_{t=0}^{K_R}\alpha_{t,0}=1$. Since the right hand side in (33) is convex combination of the corner points $\left[\left(t,C_t\left(1-K_T\mu_T\right)^+/rK_T\right):t\in\{0,\cdots,K_R\}\right]$, the fronthaul NDT $\overline{\tau}_F^*\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)$ is bounded by the lower convex envelope of these points, where $C_t=\frac{\binom{K_R}{t+1}-\binom{K_R-S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}$.

[5]Note that the constraint (31) should be $\sum_{t=0}^{K_R}a_{t,0}\geq 1-K_T\mu_T$, since the transmitters can store the same contents. However, setting this inequality with equality does not violate the lower bound, since the variables $\{a_{t,0}\}$ are multiplied with non-negative values in the left hand side in (29).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

B. Bound on The NDT of The Edge Transmission

We consider two converse bounds on the NDT of edge transmission. In the first bound, we take a cut at the set $\mathcal{R}$ of receivers. While in the second bound, we take a cut at a single receiver.

1) Multiple Receivers Cut: The main idea is that the s distinct files $\mathcal{W}_s$ can be decoded from the received signals of $\mathcal{R}$ receivers $\mathbf{Y}_{\mathcal{R}}$ and the cache contents at the set $\mathcal{R}$ of s receivers $\mathcal{Z}_{\mathcal{R}}$. Thus, by using Fano's inequality,

$$H\left(\mathcal{W}_s|\mathbf{Y}_{\mathcal{R}},\mathcal{Z}_{\mathcal{R}}\right)\leq sF\epsilon_F, \tag{34}$$

where $\epsilon_F\to 0$ as $F\to\infty$. Now, consider the following bound

$$\begin{aligned}
H\left(\mathcal{W}_s\right)&\overset{(a)}{=}H\left(\mathcal{W}_s\left|\tilde{\mathcal{W}}_s\right.\right)\\
&\overset{(b)}{=}I\left(\mathcal{W}_s;\mathbf{Y}_{\mathcal{R}},\mathcal{Z}_{\mathcal{R}}\left|\tilde{\mathcal{W}}_s\right.\right)+H\left(\mathcal{W}_s\left|\tilde{\mathcal{W}}_s,\mathbf{Y}_{\mathcal{R}},\mathcal{Z}_{\mathcal{R}}\right.\right)\\
&\overset{(c)}{\leq}I\left(\mathcal{W}_s;\mathbf{Y}_{\mathcal{R}}\left|\tilde{\mathcal{W}}_s\right.\right)+I\left(\mathcal{W}_s;\mathcal{Z}_{\mathcal{R}}\left|\tilde{\mathcal{W}}_s,\mathbf{Y}_{\mathcal{R}}\right.\right)+sF\epsilon_F\\
&\overset{(d)}{\leq}I\left(\mathbf{X}_{[K_T]};\mathbf{Y}_{\mathcal{R}}\right)+H\left(\mathcal{Z}_{\mathcal{R}}\left|\tilde{\mathcal{W}}_s,\mathbf{Y}_{\mathcal{R}}\right.\right)-H\left(\mathcal{Z}_{\mathcal{R}}\left|\mathcal{W},\mathbf{Y}_{\mathcal{R}}\right.\right)+sF\epsilon_F\\
&\overset{(e)}{=}\min\{K_T,s\}T_E\log\left(P\right)+\sum_{i=1}^{s}H\left(\mathcal{Z}_{R_i}\left|\tilde{\mathcal{W}}_s,\mathbf{Y}_{\mathcal{R}},\mathcal{Z}_{\mathcal{R}_{[i-1]}}\right.\right)+sF\epsilon_F\\
&\overset{(f)}{\leq}\min\{K_T,s\}T_E\log\left(P\right)+\sum_{i=1}^{s}H\left(\mathcal{Z}_{R_i}\left|\tilde{\mathcal{W}}_s,\mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}},\mathcal{Z}_{\mathcal{R}_{[i-1]}}\right.\right)+F\epsilon_F,
\end{aligned} \tag{35}$$

where $\tilde{\mathcal{W}}_s=\mathcal{W}\setminus\mathcal{W}_s$. Step (a) follows from the fact that the files are independent. (b) follows from the chain rule. (c) follows from the Fano's inequality in (19). Step (d) follows from the data processing inequality, where $\mathbf{X}_{[K_T]}$ are functions of $\mathcal{W}_s$, and the fact that the transmitted and received signals are independent of the non-requested files. Step (e) follows from the capacity bound of the $K_t\times s$ broadcast channel over T_E symbol transmissions in the high SNR regime. Step (f) follows from the fact that $W_{d_{R_i}}$ can be decoded from the received signals $\mathbf{Y}_{\mathcal{R}}$ and the cache contents $\mathcal{Z}_{R_i}$ of the receiver R_i, where $\mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}}=\{W_{d_{R_1}},\cdots,W_{d_{R_{i-1}}}\}$. By taking the limit $P\to\infty$ and dividing both terms by F, we get

$$\tau_E^*\left(\mathbf{d},\mu_T,\mu_R,r\right)\geq\frac{1}{\min\{K_T,s\}F}\left(H\left(\mathcal{W}_s\right)-\sum_{i=1}^{s}H\left(\mathcal{Z}_{R_i}\left|\tilde{\mathcal{W}}_s,\mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}},\mathcal{Z}_{\mathcal{R}_{[i-1]}}\right.\right)\right). \tag{36}$$

For uncoded cache placement, the term $H\left(\mathcal{Z}_{R_i}\left|\tilde{\mathcal{W}}_s,\mathcal{W}_{\mathbf{d}_{\mathcal{R}_{[i-1]}}},\mathcal{Z}_{\mathcal{R}_{[1:i-1]}}\right.\right)$ denotes the bits of files $\{W_{d_{R_i}},\cdots,W_{d_{R_s}}\}$ that are available at the receiver R_i cache ($\mathcal{Z}_{R_i}$). Hence, (36) is rewritten by

$$\tau_E^*\left(\mathbf{d},\mu_T,\mu_R,r\right)\geq\frac{1}{\min\{K_T,s\}F}\left(\sum_{i=1}^{s}\sum_{j=1}^{F}\mathbf{1}\left(B_{d_{R_i},j}\notin\mathcal{Z}_{\mathcal{R}_{[i]}}\right)\right), \tag{37}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

where $B_{d_{R_i},j}$ denotes the j-th bit of the file $W_{d_{R_i}}$. By following similar steps as in Subsection IV-A, we obtain

$$\overline{\tau}_E^* \left(S\left(\mathbf{d}\right), \mu_T, \mu_R, r \right) \geq \frac{1}{\min\{K_T, S\left(\mathbf{d}\right)\}} \sum_{t=0}^{K_R} a_t \frac{\binom{K_R}{t+1} - \binom{K_R - S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}, \tag{38}$$

where $a_t = a_{t,0} + a_{t,1}$. Now, we minimize both sides in (38) over all possible uncoded placement at transmitters $\mathcal{V}$ and at receivers $\mathcal{Z}$ such that

$$\sum_{t=0}^{K_R} a_t = 1 \tag{39}$$

$$\sum_{t=0}^{K_R} t a_t = K_R \mu_R, \tag{40}$$

where the constraint (39) comes from the total number of bits in the library, while (40) is to maintain the total cache sizes at receivers. Thus, $\overline{\tau}_E^* \left(S\left(\mathbf{d}\right), \mu_T, \mu_R, r \right)$ is bounded by the lower convex envelope of the point $[(t, C_t / \min\{K_T, S\left(\mathbf{d}\right)\}) : t \in \{0, \cdots, K_R\}]$, where $C_t = \frac{\binom{K_R}{t+1} - \binom{K_R - S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}$.

2) Single Receiver Cut: Consider the following inequality for a given demand $\mathbf{d} \in \mathcal{D}$ and an arbitrary receiver RX_j.

$$\begin{aligned} H\left(W_{d_j}\right) &= H\left(W_{d_j} \middle| \tilde{\mathcal{W}}_{d_j}\right) \\ &= I\left(W_{d_j}; \mathbf{Y}_j, \mathcal{Z}_j \middle| \tilde{\mathcal{W}}_{d_j}\right) + H\left(W_{d_j} \middle| \tilde{\mathcal{W}}_{d_j}, \mathbf{Y}_j, \mathcal{Z}_j\right) \\ &\overset{(a)}{\leq} I\left(W_{d_j}; \mathbf{Y}_j \middle| \tilde{\mathcal{W}}_{d_j}\right) + I\left(W_{d_j}; \mathcal{Z}_j \middle| \tilde{\mathcal{W}}_{d_j}, \mathbf{Y}_j\right) + F\epsilon_F \\ &\overset{(b)}{\leq} I\left(\mathbf{X}_{[K_T]}; \mathbf{Y}_j\right) + H\left(\mathcal{Z}_j \middle| \tilde{\mathcal{W}}_{d_j}, \mathbf{Y}_j\right) - H\left(\mathcal{Z}_j \middle| \mathcal{W}, \mathbf{Y}_j\right) + F\epsilon_F \\ &\leq T_E \log\left(P\right) + H\left(\mathcal{Z}_j \middle| \tilde{\mathcal{W}}_{d_j}\right), \end{aligned} \tag{41}$$

where (a) follows from Fano's inequality. (b) follows from data processing inequality. By taking $\lim P \to \infty$ and dividing both sides by F, we get

$$\begin{aligned} \tau_E^* \left(\mathbf{d}, \mu_T, \mu_R, r\right) &\geq \frac{1}{F} \left(H\left(W_{d_j}\right) - H\left(\mathcal{Z}_j \middle| \tilde{\mathcal{W}}_{d_j}\right) \right) \\ &= \frac{1}{F} \sum_{l=1}^{F} \mathbf{1}\left(B_{d_j,l} \notin \mathcal{Z}_j\right), \end{aligned} \tag{42}$$

where $H\left(\mathcal{Z}_j \middle| \tilde{\mathcal{W}}_{d_j}\right)$ denotes the number of bits of file W_{d_j} stored at receiver RX_j. By taking the average over all demands $\mathbf{d} \in \mathcal{D}$, the expected NDT for the edge transmission

$$\begin{aligned} \overline{\tau}_E^* \left(\mu_T \mu_R, r\right) &\geq \frac{1}{F N^{K_R}} \sum_{\mathbf{d} \in \mathcal{D}} \sum_{l=1}^{F} \mathbf{1}\left(B_{d_j,l} \notin \mathcal{Z}_j\right) \\ &\overset{(a)}{=} \frac{1}{FN} \sum_{n=1}^{N} \sum_{l=1}^{F} \mathbf{1}\left(B_{n,l} \notin \mathcal{Z}_j\right) = \left(1 - \mu_R\right), \end{aligned} \tag{43}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

where (a) follows from that the file $W_n \in \mathcal{W}$ appears in the j-th order of demand $\mathbf{d} \in \mathcal{D}$ by N^{K_R-1} times. From (38) and (43), we conclude that

$$\overline{\tau}_E^* \left(S\left(\mathbf{d}\right), \mu_T \mu_R, r \right) \geq \max \left\{ \frac{C_t}{\min\{K_T, S\left(\mathbf{d}\right)}, (1-\mu_R) \right\}. \tag{44}$$

This completes the proof of Theorem 1.

V. Achievable Scheme for a General F-RAN

In this section, we present the content placement and the transmission delivery of some corner points of the tuple (μ_T, μ_R). Then, we combine the introduced schemes by using memory-time sharing to obtain the results of Theorem 2. In [11], an achievable scheme is proposed to minimize the peak NDT of the cache-aided interference network which is a special case of the F-RANs in which the fronthaul capacity is zero, i.e., $r = 0$. In the cache-aided interference networks, the total transmitter-cache sizes should satisfy $K_T \mu_T \geq 1$ to guarantee that any bit of the library is stored at least at one of the transmitters. We extend this scheme for all possible values of the transmitter-cache sizes, and for all possible demands (not only the worst-case demands) by proposing the coded fronthaul transmission policy. Throughout this section, we consider an arbitrary demand vector $\mathbf{d} \in \mathcal{D}$ with a total of $S\left(\mathbf{d}\right) = s$ distinct files.

A. Content Placement

Consider a normalized receiver-cache size $\mu_R = t/K_R$ for $t \in \{0, 1, \cdots, K_R\}$, where the achievable scheme for general μ_R is obtained by using memory-time sharing. Each file $W_n \in \mathcal{W}$ is split into $K_T \binom{K_R}{t}$ subfiles with equal size $F/K_T \binom{K_R}{t}$. The subfiles are defined by $W_n = \{W_{n,i,\mathcal{R}}\}$ for $i \in [K_T]$ and $\mathcal{R} \subseteq [K_R]$ with cardinality $|\mathcal{R}| = t$. The receiver RX_j stores the subfiles $W_{n,i,\mathcal{R}}$ for all $n \in [N]$, $i \in [K_T]$, and $j \in \mathcal{R}$. Hence, each receiver stores $K_T \binom{K_R-1}{t-1} N$ subfiles each with size $F/K_T \binom{K_R}{t}$. In other words, the total number of bits stored at each receiver is given by $K_T \binom{K_R-1}{t-1} NF/K_T \binom{K_R}{t} = M_R F$ bits that satisfies the receiver memory constraint. Moreover when $\mu_T \geq 1/K_T$, transmitter TX_i stores subfiles $W_{n,i,\mathcal{R}}$ for all $n \in [N]$ and $\mathcal{R} \subseteq [K_R]$. Thus, each transmitter stores $\binom{K_R}{t} N$ subfiles which is equivalent to $\binom{K_R}{t} NF/K_T \binom{K_R}{t} \leq M_T F$ bits satisfying the transmitter memory constraint. When $\mu_T \leq 1/K_T$, transmitter TX_i stores a fraction μ_T from each subfile $W_{n,i,\mathcal{R}}$ for all $n \in [N]$ and $\mathcal{R} \subseteq [K_R]$, and hence, the total number of bits stored at each transmitter is given by $\mu_T K_T \binom{K_R}{t} NF/K_T \binom{K_R}{t} = M_T F$ bits satisfying the transmitter memory constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

B. Coded Delivery for The F-RAN with Transmitter-Cache size $\mu_T = 0$

Here, we assume that transmitters have no caches, i.e., $\mu_T = 0$. Therefore, the bits requested by receivers should be delivered from the cloud to transmitters via fronthaul links, since there is no direct link between the cloud and the receivers. We introduce two possible methods for transmission delivery in this case.

1) In the first method, we assume that each receiver requests a different file, i.e., files $\{W_{d_j}\}$ are treated as K_R different files. Hence, receiver RX_j desires subfiles $\{W_{d_j,i,\mathcal{R}} : i \in [K_T], j \notin \mathcal{R}\}$, while the remaining subfiles are already available at its cache memory. Given $i \in [K_T]$ and an arbitrary set $\mathcal{S}$ of $t+1$ receivers, each receiver $j \in \mathcal{S}$ wants subfile $W_{d_j,i,\mathcal{S}\setminus\{j\}}$ that is available at receivers $\tilde{j} \in \mathcal{S} \setminus \{j\}$. Therefore, each receiver $j \in \mathcal{S}$ can extract its desired subfile $W_{d_j,i,\mathcal{S}\setminus\{j\}}$ from the coded message

$$W_{i,\mathcal{S}} = \oplus_{j\in\mathcal{S}} W_{d_j,i,\mathcal{S}\setminus\{j\}}, \tag{45}$$

where $\oplus$ denotes the bitwise XOR. Note that each of these coded messages $\{W_{i,\mathcal{S}}\}$ has a length of $F/K_T\binom{K_R}{t}$ bits. We can easily verify that each receiver RX_j, $j \in [K_R]$, can decode its requested file from its cache contents and the received coded messages $\{W_{i,\mathcal{S}} : i \in [K_T], j \in \mathcal{S}\}$, where any subfile $W_{d_j,i,\mathcal{R}}$ which is desired by RX_j can be extracted from the coded message $W_{i,\mathcal{R}\cup\{j\}}$. The delivery of these messages is applied into two hops. First, the cloud sends the coded messages $\{W_{i,\mathcal{S}} : \mathcal{S} \subseteq [K_R]\}$ to transmitter TX_i. Hence, the total number of bits delivered to transmitter TX_i is given by $R_F = \binom{K_R}{t+1} F/K_T\binom{K_R}{t} = K_R(1-\mu_R)F/K_T(t+1)$, where the cloud sends $\binom{K_R}{t+1}$ coded messages, each of size $F/K_T\binom{K_R}{t}$ bits, to each transmitter. Thus, the NDT of the fronthaul link is given by

$$\tau_F(\mathbf{d}, 0, \mu_R, r) = \lim_{F\to\infty}\lim_{P\to\infty} \frac{R_F}{r\log(P)F/\log(P)} = \frac{\frac{K_R}{t+1}}{rK_T}(1-\mu_R). \tag{46}$$

Now, there are $K_T\binom{K_R}{t+1}$ coded messages required to be delivered to receivers in which each transmitter TX_i has message $W_{i,\mathcal{S}}$ for every subset $\mathcal{S} \subseteq [K_R]$ of t_R+1 receivers. This communication problem is called multicast X-channel, where the optimal sum-DoF of this channel is $\mathsf{DoF} = K_T K_R/(t+1)\left(K_T + \frac{K_R}{t+1} - 1\right)$ obtained from [11, Theorem 2]. Moreover the total number of delivered bits is $R_E = K_T\binom{K_R}{t+1}F/K_T\binom{K_R}{t} = K_R(1-\mu_R)/(t+1)F$ bits. Thus, the NDT of the edge transmission is obtained by

$$\tau_E(\mathbf{d}, 0, \mu_R, r) = \lim_{F\to\infty} \frac{R_E}{F\mathsf{DoF}} = \frac{K_T + \frac{K_R}{t+1} - 1}{K_T}(1-\mu_R). \tag{47}$$

2) In the second method, we focus on the $S(\mathbf{d}) = s$ distinct files in the request vector $\mathbf{d}$. Without loss of generality, the s files are denoted by $\{W_1, \cdots, W_s\}$, and the K_R receivers are divided into s groups, where the jth group contains the receivers that request file W_j. Consider $\mu_R = 0$, i.e., the receivers have no caches. First, each file of the s files is divided into K_T smaller subfiles of equal size: $W_j \triangleq \{W_{j,i} : i \in [K_T]\}$ for all $j \in [s]$. Then, the cloud sends subfiles $\{W_{j,i} : j \in [s]\}$ to transmitters TX_i. The number of bits delivered for each transmitter is $R_F = sF/K_T$. Hence the NDT of the fronthaul transmission is obtained by

$$\tau_F(\mathbf{d}, 0, 0, r) = \lim_{F\to\infty} \lim_{P\to\infty} \frac{R_F}{r \log(P) F / \log(P)} = \frac{S(\mathbf{d})}{rK_T}. \tag{48}$$

Now, each transmitter has a dedicated subfile for every group of receivers. Each a group of receivers can be treated as a single receiver with different channel states. Thus, the problem is equivalent to the compound X-channel [22], where the optimal DoF is $K_T S(\mathbf{d})/K_T + S(\mathbf{d}) - 1$ obtained from [22, Theorem 4]. As a result the NDT of the edge transmission is given by

$$\tau_E(\mathbf{d}, 0, 0, r) = \lim_{F\to\infty} \frac{S(\mathbf{d}) F}{F\mathsf{DoF}} = \frac{K_T + S(\mathbf{d}) - 1}{K_T}. \tag{49}$$

We know that the NDT of the fronthaul and edge transmissions are zero when $\mu_R = 1$. Hence, by applying memory-time sharing between $\mu_R = 0$ and $\mu_R = 1$, we get

$$\tau_F(\mathbf{d}, 0, \mu_R, r) = \frac{S(\mathbf{d})}{rK_T}(1 - \mu_R). \tag{50}$$

$$\tau_E(\mathbf{d}, 0, \mu_R, r) = \frac{K_T + S(\mathbf{d}) - 1}{K_T}(1 - \mu_R). \tag{51}$$

In order to obtain the minimum achievable NDT, we choose one of the two introduced schemes that has the lower NDT. Therefore, the edge and fronthaul NDT are given by

$$\tau_F(\mathbf{d}, 0, \mu_R, r) = \frac{\min\{S(\mathbf{d}), \frac{K_R}{t+1}\}}{rK_T}(1 - \mu_R). \tag{52}$$

$$\tau_E(\mathbf{d}, 0, \mu_R, r) = \frac{K_T + \min\{S(\mathbf{d}), \frac{K_R}{t+1}\} - 1}{K_T}(1 - \mu_R). \tag{53}$$

C. Coded Delivery for The F-RAN with Transmitter-Cache size $\mu_T \geq 1/K_T$

In this case, the total transmitter caches can store whole the library files. The proposed achievable scheme is based on delivering the requested bits from the transmitters without the help of the cloud. Thus, the fronthaul NDT is $\tau_F(\mathbf{d}, \mu_T, \mu_R, r) = 0$ for $\mu_T \geq 1/K_T$. We consider two possible methods for transmission delivery similar to Subsection V-B.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

1) We assume that each receiver requests a different file. According to the content placement introduced in Subsection V-A, transmitter TX_i has subfiles $\{W_{d_j,i,\mathcal{R}} : j \in [K_R], \mathcal{R} \subseteq [K_R]\}$. Therefore, each transmitter TX_i can construct the coded messages $\{W_{i,\mathcal{S}} : \mathcal{S} \subseteq [K_R], |\mathcal{S}| = t+1\}$ in (45) from its cache contents. By completing the transmission delivery as the first method of Subsection V-B, the achievable NDT of the edge transmission is obtained by

$$\tau_E(\mathbf{d}, \mu_T, \mu_R, r) = \frac{K_T + \frac{K_R}{t+1} - 1}{K_T}(1 - \mu_R). \tag{54}$$

2) Let us denote the distinct s files in the request demand $\mathbf{d}$ by $\{W_1, \cdots, W_s\}$. For $\mu_R = 0$, each transmitter has already subfiles $\{W_{j,i} : j \in [s]\}$. Thus, the problem is equivalent to the compound X channel as discussed in the second method in Subsection V-B, where each transmitter has a distinct subfile for every group of receivers. As a result the NDT of the edge transmission is given by $\tau_E(\mathbf{d}, \mu_T, 0, r) = \frac{K_T + S(d) - 1}{K_T}$. By using memory-time sharing between $\mu_R = 0$ and $\mu_R = 1$, we get

$$\tau_E(\mathbf{d}, \mu_T, \mu_R, r) = \frac{K_T + S(d) - 1}{K_T}(1 - \mu_R). \tag{55}$$

Finally, we choose the minimum NDT among the two introduced schemes to obtain

$$\tau_E(\mathbf{d}, \mu_T, \mu_R, r) = \frac{K_T + \min\{S(d), \frac{K_R}{t+1}\} - 1}{K_T}(1 - \mu_R). \tag{56}$$

D. Coded Delivery for The F-RAN with Transmitter-Cache size $\mu_T < 1/K_T$

In this case, we apply memory-sharing between corner points of the transmitter-cache size $\mu_T = 0$ and $\mu_T = 1/K_T$. According to the content placement in Subsection V-A, each transmitter stores a fraction μ_T of each file when $\mu_T < 1/K_T$. Therefore a total fraction of $K_T\mu_T$ bits which are available at the transmitters is delivered using the achievable scheme of Subsection V-C for $\mu_T = 1/K_T$. While the remaining fraction of $(1 - K_T\mu_T)$ bits that are not stored at any transmitter is delivered by using the achievable scheme of subsection V-B for $\mu_T = 0$. Thus, the NDT of the fronthaul and edge transmissions for $\mu_T < 1/K_T$ are given by

$$\tau_F(\mathbf{d}, \mu_T, \mu_R, r) = \frac{K_T + \min\{S(d), \frac{K_R}{t+1}\} - 1}{K_T}(1 - \mu_R)(1 - K_T\mu_T)^+. \tag{57}$$

$$\tau_E(\mathbf{d}, \mu_T, \mu_R, r) = \frac{K_T + \min\{S(d), \frac{K_R}{t+1}\} - 1}{K_T}(1 - \mu_R). \tag{58}$$

It is observed that demands $\mathbf{d} \in \mathcal{D}$ having the exact number $S(\mathbf{d})$ of distinct files have the same achievable NDT. Moreover, since the achievable NDT is convex in μ_T, μ_R, we can exchange the order of expectation and memory-time sharing. This completes the proof of Theorem 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

VI. The Expected NDT of The F-RAN with Transmitter-Side Caches

In this section, we prove Theorem 4 by proposing a lower (converse) and upper (achievable) bound on the expected NDT of an F-RAN with caches at transmitter-side only under an arbitrary popularity distribution $\mathcal{P}$. Afterwards, we show that the multiplicative gap between the lower and upper bounds is bounded by 2 independent of all system parameters and the popularity distribution.

A. Achievable Scheme

For any popularity distribution $\mathcal{P}$, the library files are partitioned into two groups. The first group $\mathcal{W}_p \triangleq \{W_1, \cdots, W_L\}$, $L = \min\{N, K_T M_T\}$, contains the first L most popular files in the library. The second group $\mathcal{W}_{\text{un}} \triangleq \{W_{L+1}, \cdots, W_N\}$ contains the remaining $N - L$ unpopular files. We point out that the partitioning strategy depends mainly on the total cache sizes at all transmitters, and does not depend on the popularity distribution $\mathcal{P}$.

1) Content Placement: In the placement phase, each file is split into K_T smaller subfiles: $W_n \triangleq \{W_{n,i} : i \in [K_T]\}$ for all $n \in [N]$. Each transmitter TX_i stores the subfiles $\{W_{n,i} : W_n \in \mathcal{W}_p\}$ which represent the i-th subfile of the most L popular files. Thus, each transmitter stores L subfiles, where each subfile has a size of F/K_T bits. Therefore, the number of bits stored at each transmitter is equal $LF/K_T \leq M_T F$ for $L = \min\{N, K_T \mu_T N\}$ that satisfies the transmitter-cache size constraint.

2) Delivery Scheme: Consider an arbitrary demand vector $\mathbf{d} \in \mathcal{D}$ with $S(\mathbf{d})$ distinct files. Without loss of generality, we define $S(\mathbf{d})$ distinct files with $\mathcal{W}_{\mathbf{d}}$

• Fronthaul transmission: For files $W_n \in \mathcal{W}_{\mathbf{d}}$ and $W_n \in \mathcal{W}_p$ which are the files requested from the popular group, each transmitter TX_i, $i \in [K_T]$, has subfile $W_{n,i}$. Hence, these file can be transmitted without the help of the cloud. On the other hand, the files $W_n \in \mathcal{W}_{\mathbf{d}}$ and $W_n \in \mathcal{W}_{\text{un}}$, which are the files requested from the unpopular group, are not available at the transmitters. Therefore, the cloud should deliver these files to the transmitters. The cloud sends subfile $W_{n,i}$ for $W_n \in \mathcal{W}_{\mathbf{d}}$ and $W_n \in \mathcal{W}_{\text{un}}$ to transmitter TX_i. As a result the fronthaul time is given by $T_F(\mathbf{d}, \mu_T, r) = \sum_{W_n \in \mathcal{W}_{\text{un}}} \mathbf{1}(W_n \in \mathcal{W}_{\mathbf{d}}) F/K_T r \log(P)$, and the fronthaul NDT is given by

$$\tau_F(\mathbf{d}, \mu_T, r) = \lim_{P\to\infty} \lim_{F\to\infty} \frac{T_F(\mathbf{d}, \mu_T, r)}{F/\log(P)} = \frac{\sum\limits_{W_n \in \mathcal{W}_{\text{un}}} \mathbf{1}(W_n \in \mathcal{W}_{\mathbf{d}})}{r K_T}. \tag{59}$$

By taking the expectation over all demands $\mathbf{d} \in \mathcal{D}$, we get

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

$$\begin{aligned}\overline{\tau}_F\left(\mu_T, r\right) &= \sum_{\mathbf{d}\in\mathcal{D}} \sum_{W_n\in\mathcal{W}_{\text{un}}} \mathbf{1}\left(W_n \in \mathcal{W}_{\mathbf{d}}\right) P\left(\mathbf{d}\right)/rK_T \\ &\overset{(a)}{=} \sum_{W_n\in\mathcal{W}_{\text{un}}} \sum_{\mathbf{d}\in\mathcal{D}} \mathbf{1}\left(W_n \in \mathcal{W}_{\mathbf{d}}\right) P\left(\mathbf{d}\right)/rK_T \\ &= \sum_{W_n\in\mathcal{W}_{\text{un}}} \sum_{\mathbf{d}\in\mathcal{D}} \left(1-\mathbf{1}\left(W_n \notin \mathcal{W}_{\mathbf{d}}\right)\right) P\left(\mathbf{d}\right)/rK_T \\ &\overset{(b)}{=} \sum_{W_n\in\mathcal{W}_{\text{un}}} \left(1-\sum_{\mathbf{d}\in\mathcal{D}} \mathbf{1}\left(W_n \notin \mathcal{W}_{\mathbf{d}}\right) P\left(\mathbf{d}\right)\right)/rK_T \\ &\overset{(c)}{=} \sum_{n=L+1}^{N} \left(1-\left(1-p_n\right)^{K_R}\right)/rK_T,\end{aligned} \tag{60}$$

where (a) follows from exchanging the order of summations. Step (b) follows from the fact that $\sum_{\mathbf{d}\in\mathcal{D}} P\left(\mathbf{d}\right) = 1$. Step (c) follows from that $\sum_{\mathbf{d}\in\mathcal{D}} \mathbf{1}\left(w_n \notin \mathcal{W}_{\mathbf{d}}\right) P\left(\mathbf{d}\right)$ is the probability that none of the receivers requests file W_n.

• Edge transmission: We divide the receivers into $S\left(\mathbf{d}\right)$ groups, where each group of receivers requests one of the filez $W_n \in \mathcal{W}_{\mathbf{d}}$. After the fronthaul transmission, transmitter TX_i has subfiles $W_{n,i}$ for all $W_n \in \mathcal{W}_{\mathbf{d}}$. Hence, the problem is equivalent to the compound X-channel, where the optimal DoF is $K_T S\left(\mathbf{d}\right)/K_T + S\left(\mathbf{d}\right) - 1$ obtained from [22, Theorem 4]. As a result the NDT of the edge transmission is given by

$$\tau_E\left(\mathbf{d}, \mu_T, r\right) = \lim_{F\to\infty} \frac{S\left(\mathbf{d}\right)F}{F\mathsf{DoF}} = \frac{K_T + S\left(\mathbf{d}\right) - 1}{K_T}. \tag{61}$$

By taking the expectation over all demands $\mathbf{d} \in \mathcal{D}$, we get

$$\overline{\tau}_E\left(\mu_T, r\right) = \mathbb{E}_{\mathbf{d}}\left\{\frac{K_T + S\left(\mathbf{d}\right) - 1}{K_T}\right\}. \tag{62}$$

B. Converse Bound

Now, we derive the converse bound of the expected NDT, where we bound the NDT of the fronthaul and edge transmissions, separately.

1) Bound on The NDT of The Fronthaul Transmission: For a given demand $\mathbf{d} \in \mathcal{D}$, consider the following inequality.

$$\begin{aligned}\tau_F^*\left(\mathbf{d}, \mu_T, r\right) &\overset{(a)}{\geq} \frac{1}{rK_TF}\left(H\left(\mathcal{W}_{\mathbf{d}}\right) - H\left(\mathcal{V}\left|\overline{\mathcal{W}}_{\mathbf{d}}\right.\right)\right), \\ &\overset{(b)}{=} \frac{1}{rK_TF}\left(\sum_{W_n\in\mathcal{W}_{\mathbf{d}}} \sum_{j=1}^{F} \mathbf{1}\left(B_{n,j} \notin \mathcal{V}\right)\right),\end{aligned} \tag{63}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

where $\mathcal{W}_{\mathbf{d}}$ denotes the $S(\mathbf{d})$ distinct files in the demand $\mathbf{d}$, and $B_{n,j}$ is the j-th bit of file W_n. Step (a) is obtained from (22) by setting $\mu_R = 0$. Step (b) follows from considering an arbitrary uncoded placement scheme. By taking expectations over all demands $\mathbf{d} \in \mathcal{D}$, we get

$$\begin{aligned}
\overline{\tau}_F^*(\mu_T, r) &\geq \frac{1}{rK_TF} \sum_{\mathbf{d}\in\mathcal{D}} \left(\sum_{W_n \in \mathcal{W}_{\mathbf{d}}} \sum_{j=1}^{F} \mathbf{1}(B_{n,j} \notin \mathcal{V}) \right) P(\mathbf{d}) \\
&= \frac{1}{rK_TF} \left(\sum_{\mathbf{d}\in\mathcal{D}} \sum_{n=1}^{N} \sum_{j=1}^{F} \mathbf{1}(B_{n,j} \notin \mathcal{V}) \mathbf{1}(W_n \in \mathcal{W}_{\mathbf{d}}) P(\mathbf{d}) \right) \\
&\overset{(a)}{=} \frac{1}{rK_TF} \left(\sum_{n=1}^{N} \sum_{j=1}^{F} \mathbf{1}(B_{n,j} \notin \mathcal{V}) \sum_{\mathbf{d}\in\mathcal{D}} \mathbf{1}(W_n \in \mathcal{W}_{\mathbf{d}}) P(\mathbf{d}) \right) \\
&\overset{(b)}{=} \frac{1}{rK_TF} \left(\sum_{n=1}^{N} \sum_{j=1}^{F} \mathbf{1}(B_{n,j} \notin \mathcal{V}) \sum_{\mathbf{d}\in\mathcal{D}} (1 - \mathbf{1}(W_n \notin \mathcal{W}_{\mathbf{d}})) P(\mathbf{d}) \right) \\
&= \frac{1}{rK_TF} \left(\sum_{n=1}^{N} \sum_{j=1}^{F} \mathbf{1}(B_{n,j} \notin \mathcal{V}) \left(1 - (1-p_n)^{K_R}\right) \right),
\end{aligned} \tag{64}$$

where step (a) is obtained by exchanging the summation orders. Step (b) is obtained from the fact that $\sum_{\mathbf{d}\in\mathcal{D}} \mathbf{1}(W_n \notin \mathcal{W}_{\mathbf{d}}) P(\mathbf{d})$ denotes the probability that none of the K_R receivers requests file W_n. Note that $\left(1 - (1-p_n)^{K_R}\right)$ is non-increasing function in n, since $p_n \geq p_{n+1}$. By minimizing both sides over all possible placement strategies $\mathcal{V}$, we obtain

$$\overline{\tau}_F^*(\mu_T, r) \geq \frac{1}{rKT} \left(\sum_{n=L+1}^{N} \left(1 - (1-p_n)^{K_R}\right) \right), \tag{65}$$

where $L = \min\{N, K_TM_T\}$.

2) Bound on The NDT of The Edge Transmission: For a given demand $\mathbf{d} \in \mathcal{D}$, consider the following inequality.

$$\tau_E^*(\mathbf{d}, \mu_T, r) \overset{(a)}{\geq} \frac{H(\mathcal{W}_{\mathbf{d}})}{\min\{K_T, S(\mathbf{d})\}F} = \frac{S(\mathbf{d})}{\min\{K_T, S(\mathbf{d})\}}, \tag{66}$$

where step (a) is obtained from (36) by setting $\mu_R = 0$. By taking the expectation over all demands $\mathbf{d} \in \mathcal{D}$, we obtain

$$\overline{\tau}_E^*(\mu_T, r) \geq \mathbb{E}_{\mathbf{d}} \left\{ \frac{S(\mathbf{d})}{\min\{K_T, S(\mathbf{d})\}} \right\}. \tag{67}$$

C. Multiplicative Gap

Here, we bound the gap between the achievable scheme introduced in Section VI-A and the converse bound introduced in Section VI-B. The gap is obtained by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

$$
\begin{aligned}
G &= \frac{\overline{\tau}\left(\mu_T, r\right)}{\overline{\tau}^*\left(\mu_T, r\right)} \\
&= \frac{\overline{\tau}_F\left(\mu_T, r\right) + \overline{\tau}_E\left(\mu_T, r\right)}{\overline{\tau}_F^*\left(\mu_T, r\right) + \overline{\tau}_E^*\left(\mu_T, r\right)} \\
&= \frac{\sum_{n=L+1}^{N}\left(1-\left(1-p_n\right)^{K_R}\right)/rK_T + \mathbb{E}_{\mathbf{d}}\left\{\frac{K_T+S(\mathbf{d})-1}{K_T}\right\}}{\sum_{n=L+1}^{N}\left(1-\left(1-p_n\right)^{K_R}\right)/rK_T + \mathbb{E}_{\mathbf{d}}\left\{\frac{S(\mathbf{d})}{\min\{K_T,S(\mathbf{d})\}}\right\}} \\
&\overset{(a)}{\leq} \frac{\sum_{n=L+1}^{N}\left(1-\left(1-p_n\right)^{K_R}\right)/rK_T + \mathbb{E}_{\mathbf{d}}\left\{2\frac{S(\mathbf{d})}{\min\{K_T,S(\mathbf{d})\}}\right\}}{\sum_{n=L+1}^{N}\left(1-\left(1-p_n\right)^{K_R}\right)/rK_T + \mathbb{E}_{\mathbf{d}}\left\{\frac{S(\mathbf{d})}{\min\{K_T,S(\mathbf{d})\}}\right\}} \leq 2,
\end{aligned}
\tag{68}
$$

where step (a) is obtained from that

$$
\frac{\frac{K_T+S(\mathbf{d})-1}{K_T}}{\frac{S(\mathbf{d})}{\min\{K_T,S(\mathbf{d})\}}} = \frac{K_T + S\left(\mathbf{d}\right) - 1}{\max\{K_T, S\left(\mathbf{d}\right)\}} \leq 2.
$$

This completes the proof of Theorem 4.

VII. Conclusion

In this paper, a Fog Radio Access Network (F-RAN) with caches at both transmitters and receivers has been studied from an information-theoretic perspective. Unlike previous work, we have characterized, both, the peak normalized delivery time (NDT) and the expected NDT under uniform popularity distribution within a constant factor independent of all system parameters. To prove these results, we have proposed a new achievable scheme which takes into account the redundancy of the receivers demands. Furthermore, we have developed a lower bound of the expected NDT under the assumptions of uncoded placement and uniform popularity. Although we have characterized the expected NDT for F-RANs under uniform popularity distribution within a constant gap from the lower bound, it is still an open problem to characterize the expected NDT under arbitrary popularity distribution with caches at receivers. Although the proposed achievable scheme can be generalized for arbitrary popularity distribution, the problem remains challenging because it is hard to find a tight lower bound even in the case of a single transmitter [5]. Another important direction to extend our work is considering the case of coded placement, in which the transmitters and receivers are not restricted to store uncoded fragments of the library files. Comparing to previous work in the literature, we have improved the multiplicative gap of the peak NDT for uncoded placement schemes. After that, we turn our attention to F-RANs with caches at transmitters only under an arbitrary popularity distribution. We have proposed a novel cache placement strategy that divides the library contents into two groups of the most popular contents and the least popular contents, where the transmitters caches contents from only the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

most popular group. It has been proven that our proposed scheme is order optimal for all system parameters and for any popularity distribution.

APPENDIX A
PROOF OF THEOREM 3

In this appendix, we bound the multiplicative gap of the expected NDT with uniform popularity distribution.

Remark 1. Let X be a random variable taking values from the set $x \in \mathcal{X}$, and $f_1(X)$, $f_2(X)$ are two different functions of the random variable X. The relation between the two functions is $f_1(x)/f_2(x) \leq c$ for every realization $x \in \mathcal{X}$, where c is constant. Hence, we have

$$\frac{E_X\{f_1(X)\}}{E_X\{f_2(X)\}} \leq \frac{E_X\{cf_2(X)\}}{E_X\{f_2(X)\}} \leq c. \tag{69}$$

For $\mu_R = t/K_R$, $t \in \{0, \cdots, K_R\}$, the achievable expected NDT is obtained from Theorem 2

$$\overline{\tau}(\mu_T, \mu_R, r) = \mathbb{E}_{S(\mathbf{d})}\left\{\left(\frac{\min\{S(\mathbf{d}), \frac{K_R}{t+1}\}}{K_T}\left(\frac{(1-K_T\mu_T)^+}{r}+1\right)+\frac{K_T-1}{K_T}\right)\left(1-\frac{t}{K_R}\right)\right\}. \tag{70}$$

In addition, the lower bound on the expected NDT is obtained from Theorem 1

$$\begin{aligned}
&\overline{\tau}^*(\mu_T, \mu_R, r) \geq \mathbb{E}_{S(\mathbf{d})}\left\{\overline{\tau}_F(S(\mathbf{d}), \mu_T, \mu_R, r) + \overline{\tau}_E(S(\mathbf{d}), \mu_T, \mu_R, r)\right\},\\
&\overline{\tau}_F^*(S(\mathbf{d}), \mu_T, \mu_R, r) = \frac{C_t(1-K_T\mu_T)^+}{rK_T},\\
&\overline{\tau}_E^*(S(\mathbf{d}), \mu_T, \mu_R, r) = \max\left\{\frac{C_t}{\min\{K_T, S(\mathbf{d})\}}, (1-\mu_R)\right\},
\end{aligned} \tag{71}$$

where $C_t = \frac{\binom{K_R}{t+1} - \binom{K_R - S(\mathbf{d})}{t+1}}{\binom{K_R}{t}}$. While for general μ_R both the achievable expected NDT and the lower bound are obtained from the lower convex envelope of the integer points of t. Therefore, it is just required to bound the corner points $t \in \{0, \cdots, K_R\}$. From Remark 1, it is sufficient to bound the gap between the achievable and lower bound NDT for a given value of the random variable $S(\mathbf{d}) = s$. Consider the following inequality.

$$\begin{aligned}
C_t &= \frac{\binom{K_R}{t+1} - \binom{K_R - s}{t+1}}{\binom{K_R}{t}}\\
&= \frac{K_R - t}{t+1} - \frac{1}{t+1}\frac{(K_R-t)!\,(K_R-s)!}{K_R!\,(K_R-t-s-1)!}\\
&= \frac{K_R - t}{t+1}\left[1 - \frac{(K_R-t-1)(K_R-t-2)\ldots(K_R-t-s)}{K_R(K_R-1)\ldots(K_R-s+1)}\right]\\
&= \frac{K_R - t}{t+1}\left[1 - \prod_{i=0}^{s-1}\left(1-\frac{t+1}{K_R-i}\right)\right]\\
&\geq \frac{K_R - t}{t+1}\left[1 - \left(1-\frac{t+1}{K_R}\right)^s\right]\\
&= \frac{K_R}{t+1}\left[1 - \left(1-\frac{t+1}{K_R}\right)^s\right]\left(1-\frac{t}{K_R}\right).
\end{aligned} \tag{72}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2019.2919028, IEEE Transactions on Wireless Communications

The gap between the achievable expected NDT and the lower bound for a given $S(\mathbf{d})$ is obtained

$$\frac{\overline{\tau}\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)}{\overline{\tau}^*\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)}=G_1+G_2 \tag{73}$$

where G_2 is given by

$$G_2=\frac{\frac{K_T-1}{K_T}\left(1-\frac{t}{K_R}\right)}{\overline{\tau}^*\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)}\leq 1. \tag{74}$$

We can easily verify that $G_2=0$ when $K_T=1$. The second gap G_1 is given by

$$\begin{aligned}
G_1&=\frac{\frac{\min\{S(\mathbf{d}),\frac{K_R}{t+1}\}}{K_T}\left(\frac{(1-K_T\mu_T)^+}{r}+1\right)\left(1-\frac{t}{K_R}\right)}{\overline{\tau}^*\left(S\left(\mathbf{d}\right),\mu_T,\mu_R,r\right)}\\
&\leq\frac{\frac{\min\{S(\mathbf{d}),\frac{K_R}{t+1}\}}{K_T}\left(\frac{(1-K_T\mu_T)^+}{r}+1\right)\left(1-\frac{t}{K_R}\right)}{\frac{C_t}{K_T}\left(\frac{(1-K_T\mu_T)^+}{r}+\frac{K_T}{\min\{K_T,S(\mathbf{d})\}}\right)}\\
&\leq\frac{\min\{S\left(\mathbf{d}\right),\frac{K_R}{t+1}\}\left(1-\frac{t}{K_R}\right)}{C_t}\\
&\overset{(a)}{\leq}\frac{\min\{S\left(\mathbf{d}\right),\frac{K_R}{t+1}\}}{\frac{K_R}{t+1}\left[1-\left(1-\frac{t+1}{K_R}\right)^{S(\mathbf{d})}\right]},
\end{aligned} \tag{75}$$

where (a) follows from substituting the bound on C_t from (72). When $S\left(\mathbf{d}\right)\leq\frac{K_R}{t+1}$, let $S\left(\mathbf{d}\right)\left(t+1\right)/K_R=x$, and hence $0\leq x\leq 1$. We obtain

$$\begin{aligned}
\frac{S\left(\mathbf{d}\right)}{\frac{K_R}{t+1}\left[1-\left(1-\frac{t+1}{K_R}\right)^{S(\mathbf{d})}\right]}&=\frac{1}{\frac{1-(1-x/S(\mathbf{d}))^{S(\mathbf{d})}}{x}}\\
&\overset{(a)}{\leq}\frac{1}{\frac{1-e^{-x}}{x}}\overset{(b)}{\leq}\frac{1}{1-e^{-1}}\simeq 1.58,
\end{aligned} \tag{76}$$

where (a) follows from $(1-x)^y\leq e^{-xy}$ for all $0<x\leq 1$. Step (b) follows from that the function $\left(1-e^{-x}\right)/x$ is an increasing function in the interval $0\leq x\leq 1$. When $S\left(\mathbf{d}\right)>\frac{K_R}{t+1}$, let $\frac{K_R}{t+1}=x>0$. Then, we obtain

$$\frac{\frac{K_R}{t+1}}{\frac{K_R}{t+1}\left[1-\left(1-\frac{t+1}{K_R}\right)^{S(\mathbf{d})}\right]}\overset{(a)}{\leq}\frac{1}{1-\left(1-1/x\right)^x}\overset{(b)}{\leq}\frac{1}{1-e^{-1}}\simeq 1.58, \tag{77}$$

where (a) follows from $S\left(\mathbf{d}\right)>x$. Step (b) follows from $\left(1-1/x\right)^x\leq e^{-1}$ for all $x>0$. As a result, $G_1\leq 1.58$. Hence, the proof is completed.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast white paper," Mar. 2017. [Online]. Available: http://goo.gl/OL6mYY

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

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

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

[5] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. in ITA*, Feb. 2015, pp. 98–107.

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

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

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

[9] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, "A converse bound for cache-aided interference networks," in *to be published in the proceeding of the 52nd Asilomar Conference on Signals, Systems, and Computers, CA*. IEEE, Oct. 2018.

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

[11] J. Hachem, U. Niesen, and S. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Transactions on Information Theory*, 2018.

[12] 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.

[13] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided mimo interference networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5061–5076, 2017.

[14] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc.in IEEE ISIT*, 2016, pp. 2029–2033.

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

[16] J. Zhang and O. Simeone, "Fundamental limits of cloud and cache-aided interference management with multi-antenna base stations," in *ISIT*, June 2018, pp. 1425–1429.

[17] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, "Decentralized coded caching in wireless networks: Trade-off between storage and latency," in *Proc. in IEEE ISIT*, June 2017, pp. 2443–2447.

[18] F. Xu and M. Tao, "Fundamental limits of decentralized caching in fog-rans with wireless fronthaul," in *ISIT*, June 2018, pp. 1430–1434.

[19] J. S. P. Roig, F. Tosato, and D. Guenduez, "Storage-latency trade-off in cache-aided fog radio access networks," in *International Conference on Communications (ICC)*, May 2018, pp. 1–6.

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

[21] S. M. Azimi, O. Simeone, A. Sengupta, and R. Tandon, "Online edge caching and wireless delivery in fog-aided networks with dynamic content popularity," *IEEE Journal on Selected Areas in Communications*, 2018.

[22] M. A. Maddah-Ali, "On the degrees of freedom of the compound miso broadcast channels with finite states," in *Proc. in IEEE ISIT*, 2010, pp. 2273–2277.

[23] M. Bona, *Introduction to enumerative combinatorics*. McGraw-Hill Science/Engineering/Math, 2007.