A Converse Bound for Cache-Aided Interference Networks

Antonious M. Girgis*, Ozgur Ercetin†, Mohammed Nafie*‡, and Tamer ElBatt§‡
*Wireless Intelligent Networks Center (WINC), Nile University, Cairo, Egypt
† Faculty of Engineering and Natural Sciences, Sabanci University, Turkey.
‡ Electronics and Communications Engineering Dept., Faculty of Engineering, Cairo University, Egypt.
§ Computer Science and Engineering Dept., The American University in Cairo, Egypt.

***Abstract*—In this paper, an interference network with arbitrary number of transmitters and receivers is studied, where each transmitter is equipped with a finite size cache. We obtain an information-theoretic lower bound on both the peak normalized delivery time (NDT), and the *expected* NDT of cache-aided interference networks with uniform content popularity. For the peak NDT, we show that our lower bound is strictly tighter than the bound in the literature for small cache sizes. Moreover, we show that the feasibility region on the expected NDT is bigger than that of the peak NDT.**

I. Introduction

The exponential growth of on-demand video streaming causes an inevitable burden on wireless networks during the peak hours. Caching is a promising solution to alleviate this problem by pushing the popular data content into cache memories at edge nodes during the off-peak hours, where the network resources are under-utilized. Hence, in the peak hours when the network is congested, caches can be exploited to serve the receivers requests with a significant improvement in the system performance. The rule of caching in interference networks is studied in [1]–[10]. In [1], the degrees of freedom (DoF) of a 3×3 interference network with cache-equipped transmitters was studied. In [4], the normalized delivery time (NDT) which defines the delivery latency is introduced as a performance metric to study the fog radio access network (F-RAN) with two transmitters and two receivers. The work in [6]–[10] studied interference networks and F-RANs with caches at both transmitters and receivers.

A. Contribution

In this work, we study a cache-aided interference network with arbitrary number of transmitters and receivers. In contrast to the prior works, we study the information theoretic limits of both the peak NDT and the *expected* NDT under uniform content popularity, where we derive a lower bound on both the expected NDT and the peak NDT for uncoded placement. To the best of the authors knowledge, this paper is the first work discussing the *expected* NDT, since all previous works only study the peak NDT for the worst-case demand. Perhaps, the closest to our work is [3], where the authors derive a lower bound on the peak NDT for uncoded placement schemes. In this paper, we provide a tighter bound for the small cache sizes, i.e., when the cache at each transmitter can store at most the half of the library. Moreover, we show that the feasibility region of the expected NDT is bigger than the feasibility region of the peak NDT. Hence, the achievable schemes designed for the peak NDT should be improved to work with the general demands in which receiver demands are not distinct.

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690893 and a grant from the Egyptian National Telecommunications Regulatory Authority (NTRA).

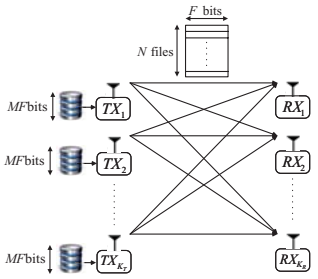


Fig. 1: Cache-aided interference network with K_T transmitters and K_R receivers.

II. System Model

We consider an interference network of K_T transmitters connected to K_R receivers over a Gaussian channel as depicted in Figure 1. There is a content library of N files, $\mathcal{W} \triangleq \{W_1, \cdots, W_N\}$, each of size F bits. Each receiver can randomly and independently request a file from the library according to uniform distribution $\{p_i = \frac{1}{N}\}$ for $i \in [N]$. Each transmitter TX_i, $i \in [K_T]$, has a local cache memory Z_i of size MF bits, where $\mu = M/N$ refers to normalized cache size. The system operates in two separate phases, a *placement phase* and a *delivery phase*. In the placement phase, the transmitters have access to the content library $\mathcal{W}$, and hence, each transmitter fills its cache memory as an arbitrary function of the content library $\mathcal{W}$ under its cache size constraint. We maintain that the caching functions are designed without any prior knowledge of the future receivers demands and the channel coefficients between transmitters and receivers.

In the delivery phase, receiver RX_j requests a file W_{d_j} out of the N files of the library. We consider $\mathbf{d} = [d_1, \cdots, d_{K_R}] \in [N]^{K_R}$ as the vector of receivers demands. The transmitters are informed with receivers demands . Thus, transmitter TX_i, $i \in [K_T]$, responds to the user demands by sending a codeword $\mathbf{x}_i \triangleq (x_i(t))_{t=1}^{T}$ of block length T over the interference chan-

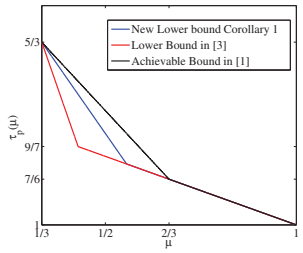


Fig. 2: The peak NDT for a cache-aided interference network with $K_T = 3$ transmitters and $K_R = 3$ receivers.

nel, where $x_i(t) \in \mathbb{C}$ is the transmitted signal of transmitter TX_i at time $t \in [T]$. We impose an average transmit power constraint over the channel input $\frac{1}{T}||\mathbf{x}_i|| \leq P$. In this phase, each transmitter has only access to its own cache contents, therefore, the codeword $\mathbf{x}_i$ of transmitter TX_i is determined by an encoding function in the receivers demands $\mathbf{d}$, the cache contents Z_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 the received signal $\mathbf{y}_j \triangleq (y_j(t))_{t=1}^{T}$ given by

$$y_j(t) = \sum_{i=1}^{K_T} h_{ji} x_i(t) + n(t) \tag{1}$$

where $y_j(t) \in \mathbb{C}$ is the received signal by receiver RX_j at time $t \in [T]$, and $n(t)$ denotes the additive white Gaussian noise at receiver RX_j. $h_{ji} \in \mathbb{C}$ represents the channel gain between transmitter TX_i and receiver RX_j. Let $S(\mathbf{d})$ be a function returning the number of distinct files in the demand $\mathbf{d}$. For a given demand $\mathbf{d}$, the system performance can be characterized by the normalized delivery time (NDT) defined as [4].

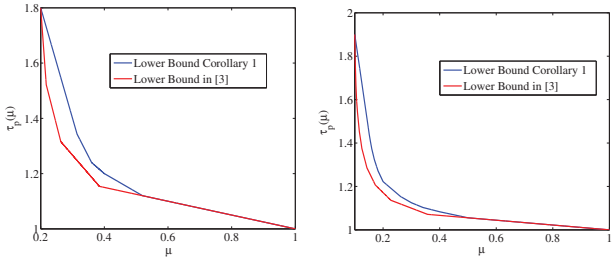
$$\tau(\mu, \mathbf{d}) = \lim_{P\to\infty} \lim_{F\to\infty} \frac{T(\mu, P, \mathbf{d})}{F/\log(P)}, \tag{2}$$

where $T(\mu, P, \mathbf{d})$ denotes the time needed to send the all requested files such that each receiver can decode its requested file with probability one as $F \to \infty$. The NDT refers to the delivery latency with respect to an interference-free baseline system at the high SNR regime. Furthermore, we define $\overline{\tau}(\mu) = \mathbb{E}_{\mathbf{d}}[\tau(\mu, \mathbf{d})]$ as the expected NDT, where the expectation is over the random demand $\mathbf{d}$.

Our objective in this work is to derive an information theoretic lower bound on the expected NDT as a function of the normalized cache size μ for cache-aided interference networks. We point out that the transmitter cache size must satisfy $K_T\mu \geq 1$ to maintain that every bit of the library content is stored at least at one cache of the network. Moreover, if the cache size increases the library size $\mu > 1$, each transmitter is able to cache all the library files and the remaining cache memory would not be used. Therefore, we are interested in the normalized cache size $\frac{1}{K_T} \leq \mu \leq 1$.

III. Main Results

In this section, we first present our main result of this paper which gives a lower bound on the expected NDT for cache-aided networks. Then, for a special case when each receiver requests a distinct file, we compare our results with the cut-set based lower bound in [3, Theorem 1].



(a) $K_T = 5$ and $K_R = 5$. (b) $K_T = 10$ and $K_R = 10$.

Fig. 3: Comparison between our bound in Corollary 1 and the bound in [3] for the peak NDT.

Theorem 1. *For a $K_T \times K_R$ cache-aided interference network with a library of N files, normalized cache size $\mu \in \left[\frac{1}{K_T} : 1\right]$ at each transmitter, and a parameter $t = K_T\mu$, the expected NDT under uniform popularity distribution is lower bounded as*

$$\overline{\tau}(\mu) \geq \mathbb{E}\left[\max_{\mathcal{F}} Conv\left(\frac{t\binom{K_T}{t} + (S(\mathbf{d}) - \sigma)\binom{\sigma-1}{t-1}}{t\binom{K_T}{t}}\right)\right] \tag{3}$$

where $\mathcal{F} \triangleq \{1 \leq \sigma \leq \min\{K_T, S(\mathbf{d})\}\}$, and the expectation is over the random demand $\mathbf{d}$. $Conv(f(t))$ denotes the lower convex envelope of the integer points $[(t, f(t)) : t \in \{1, \cdots, K_T\}]$.

To the best of our knowledge, this theorem gives the first converse bound on the *expected* NDT under uniform popularity distribution for cache-aided interference networks, where the lower bound in [3, Theorem 1] is applied to the peak NDT only wherein each receiver requests a different file. To prove Theorem 1, we first derive a lower bound on the NDT for a given demand $\mathbf{d}$, and uncoded placement scheme. The derived lower bound is mainly based on genie-aided, cut-set arguments. Then, we optimize the derived bound over all possible uncoded placement schemes to get the minimum NDT for a given demand $\mathbf{d}$. Finally, by taking the expectation over all demands $\mathbf{d} \in [N]^{K_R}$, we obtain the lower bound in Theorem 1. The full proof of Theorem 1 is presented in Section IV. We can directly derive a lower bound on the peak NDT from Theorem 1 as in the following corollary.

Corollary 1. For a general $K_T \times K_R$ cache-aided interference network with a library of $N \geq K_R$ files, normalized cache size $\mu \in \left[\frac{1}{K_T} : 1\right]$ at each transmitter, a parameter $t = K_T\mu$, and each receiver requests a distinct file, the NDT is lower bounded as

$$\tau_{\text{p}}(\mu) \geq \max_{1\leq\sigma\leq\min\{K_T,K_R\}} \text{Conv}\left(\frac{t\binom{K_T}{t} + (K_R - \sigma)\binom{\sigma-1}{t-1}}{t\binom{K_T}{t}}\right) \tag{4}$$

The proof is straightforward obtained from Theorem 1 by setting the number of distinct demands $S(\mathbf{d}) = K_R$. Now,

we compare our result in Corollary 1 with the lower bound in [3]. In Figure 2, we plot Maddah-Ali-Neisen (MN) scheme in [1], the lower bound derived in [3], and our proposed lower bound in Corollary 1 for a cache-aided interference network with $K_T = 3$ transmitters and $K_R = 3$ receivers. We can see that our bound is tighter than the bound in [3] for $\mu \leq 0.5$, where the multiplicative gap between the MN scheme and our lower bound is reduced to 1.091. In Figures 3a and 3b, we compare between our bound in Corollary 1 and the bound in [3] with different number of transmitters and receivers. It is shown that our bound is tighter when the normalized cache size $\mu \leq 0.5$, while our bound coincides with the bound in [3] for large cache sizes when $\mu \geq 0.5$.

In Figure 4, we plot the lower bound on the expected NDT in Theorem 1 and the lower bound on the peak NDT in [3] for a cache-aided interference network with $K_T = 5$ transmitters, $K_R = 20$ receivers, and a library of $N = 100$ file. The expected NDT works differently from the peak NDT. In the peak NDT, each receiver requests a different file. Therefore, at each time, there will be K_R different files required to be delivered, while in the expected NDT, there is a redundancy in the receivers requests, i.e., there is a chance that different receivers request the same file. Hence, it is expected that the NDT would be reduced. To see this consider a simple example of a single transmitter. For an extreme case when all receivers request the same file, the transmitter can broadcast this file in a single time slot to all receivers, i.e., $\tau = 1$. While in the worst case, it is required K_R time slots to send the different K_R files, i.e., $\tau = K_R$. This interprets why our bound on the expected NDT is less than the bound on the peak NDT in [3]. Moreover, this observation indicates that the feasibility region on the expected NDT is bigger than the feasibility region on the peak NDT, and hence, it is expected that the achievable schemes for the worst case demand might be no longer order optimal in general.

IV. Proof of Theorem 1

In this section, we present the detailed proof of Theorem 1. Let $\tau(\mathbf{d}, \mu, \mathcal{Z})$ denote the NDT for a given demand $\mathbf{d}$ and placement scheme $\mathcal{Z} \triangleq \{Z_1, \cdots, Z_{K_T}\}$. Then, the expected NDT can be bounded by

$$\begin{aligned}\overline{\tau}(\mu) &= \min_{\mathcal{Z}} \mathbb{E}_{\mathbf{d}}\left[\tau(\mathbf{d}, \mu, \mathcal{Z})\right] \\ &\overset{(a)}{=} \min_{\mathcal{Z}} \mathbb{E}_{S(\mathbf{d})}\left[\mathbb{E}_{\mathbf{d}|S(\mathbf{d})}\left[\tau(S(\mathbf{d}), \mu, \mathcal{Z})\right]\right] \\ &\overset{(b)}{\geq} \mathbb{E}_{S(\mathbf{d})}\left[\min_{\mathcal{Z}} \mathbb{E}_{\mathbf{d}|S(\mathbf{d})}\left[\tau(S(\mathbf{d}), \mu, \mathcal{Z})\right]\right]\end{aligned} \tag{5}$$

where in step (a), we first take the expectation over demands on condition that $S(\mathbf{d}) = s$, i.e., the number of distinct files in demand $\mathbf{d}$ is equal to s. Then, we take the expectation over all values of s. Notice that $\mathbf{d}$ is a random vector, and hence, $S(\mathbf{d})$ is a random variable taking values from $\{1, \cdots, \min\{K_R, N\}\}$. Thus, we divide the demands $\mathbf{d} \in [N]^{K_R}$ into categories $\{\mathcal{D}_s\}$, where $\mathcal{D}_s$ is the set of demands satisfying that $S(\mathbf{d}) = s$, i.e., the demands that have

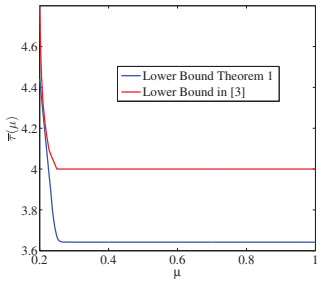


Fig. 4: Converse bound on the expected NDT for a cache-aided interference network with $K_T = 5$ transmitters, $K_R = 20$ receivers, and a library of $N = 100$.

exactly s distinct files. In step (b), we bound the expected NDT by designing the placement scheme to minimize individually the NDT for each demand category instead of designing the placement scheme to minimize the expected NDT.

To obtain the result in Theorem 1, we derive a lower bound on the NDT for demand category $\mathcal{D}_s$ by using cut-set and genie-aided arguments. Then, we run an optimization problem to find the tight cut over all possible cuts, and to minimize the NDT over all possible uncoded placement schemes. Finally, we take the expectation with respect to $S(\mathbf{d})$.

For a given demand $\mathbf{d} \in \mathcal{D}_s$, let $\mathcal{R}$ be an arbitrary set of $S(\mathbf{d})$ receivers, in which each receiver requests a different file. Let $\mathcal{S}_t$ be a set of transmitters with cardinality σ, and $\mathcal{S}_r$ be a set of receivers with cardinality σ, where $1 \leq \sigma \leq s$. We define $\overline{\mathcal{S}}_t = [K_T] \backslash \mathcal{S}_t$, and $\overline{\mathcal{S}}_r = \mathcal{R} \backslash \mathcal{S}_r$. The cache contents of set $\mathcal{S}_t$ of transmitters is defined by $\mathcal{Z}_{\mathcal{S}_t} \triangleq \{Z_i\}_{i \in \mathcal{S}_t}$. Moreover, we define the following disjoint set of bits

$$\begin{aligned}\mathcal{W}_{\mathcal{S}_t} &\triangleq \left\{B_{d_j,i} : B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}, j \in \mathcal{R}\right\} \\ \mathcal{W}_{\mathcal{S}_r} &\triangleq \left\{B_{d_j,i} : B_{d_j,i} \in \mathcal{Z}_{\overline{\mathcal{S}}_t}, j \in \mathcal{S}_r\right\} \\ \overline{\mathcal{W}} &\triangleq \left\{B_{d_j,i} : B_{d_j,i} \in \mathcal{Z}_{\overline{\mathcal{S}}_t}, j \in \overline{\mathcal{S}}_r\right\}\end{aligned} \tag{6}$$

where $B_{d_j,i}$ denotes the ith bits in file W_{d_j}, for all $i \in [F]$. Observe that each bit of the library should be stored at least at one of the transmitter caches. Hence, if $B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}$, then $B_{d_j,i} \in \mathcal{Z}_{\mathcal{S}_t}$. The set $\mathcal{W}_{\mathcal{S}_t}$ contains the bits of files $\{W_{d_j}\}_{j \in \mathcal{R}}$ that are stored exclusively at the caches of transmitters $\mathcal{S}_t$, while the set $\mathcal{W}_{\mathcal{S}_r}$ contains the bits of files $\{W_{d_j}\}_{j \in \mathcal{S}_r}$ that are available at transmitters $\overline{\mathcal{S}}_t$. We can easily verify that $\mathcal{W}_{\mathcal{S}_t} \bigcup \mathcal{W}_{\mathcal{S}_r}$ has all the bits of files $\{W_{d_j}\}_{j \in \mathcal{S}_r}$ in addition to the bits of files $\{W_{d_j}\}_{j \in \overline{\mathcal{S}}_r}$ that are exclusively stored at transmitters $\mathcal{S}_t$.

Assume that a genie provides the receivers in set $\mathcal{S}_r$ with the bits in set $\overline{\mathcal{W}}$, and provides the receivers in set $\overline{\mathcal{S}}_r$ with bits in set $\mathcal{W}_{\mathcal{S}_r} \bigcup \overline{\mathcal{W}}$. We prove that the set $\mathcal{S}_r$ of σ receivers can decode all bits $\mathcal{W}_{\mathcal{S}_t} \bigcup \mathcal{W}_{\mathcal{S}_r}$ using their received signal and the genie-aided information. Consider the receivers in set $\mathcal{S}_r$ can fully cooperate between each others. We present the received signals of $\mathcal{S}_r$ and $\overline{\mathcal{S}}_r$ receivers as follows:

$$\begin{aligned}\mathbf{Y}_{\mathcal{S}_r} &= \mathbf{H}_{\mathcal{S}_r}^{\mathcal{S}_t} \mathbf{X}_{\mathcal{S}_t} + \mathbf{H}_{\mathcal{S}_r}^{\overline{\mathcal{S}}_t} \mathbf{X}_{\overline{\mathcal{S}}_t} + \mathbf{Z}_{\mathcal{S}_r}, \\ \mathbf{Y}_{\overline{\mathcal{S}}_r} &= \mathbf{H}_{\overline{\mathcal{S}}_r}^{\mathcal{S}_t} \mathbf{X}_{\mathcal{S}_t} + \mathbf{H}_{\overline{\mathcal{S}}_r}^{\overline{\mathcal{S}}_t} \mathbf{X}_{\overline{\mathcal{S}}_t} + \mathbf{Z}_{\overline{\mathcal{S}}_r}.\end{aligned} \tag{7}$$

where $\mathbf{Y}_{\mathcal{K}_r}$ is a $|\mathcal{K}_r| \times 1$ concatenated vector of the received signals of receivers in set $\mathcal{K}_r$, and $\mathbf{X}_{\mathcal{K}_t}$ is a $|\mathcal{K}_t| \times 1$ concatenated vector of the transmitted signals of transmitters in set $\mathcal{K}_t$. Furthermore, $\mathbf{H}_{\mathcal{K}_r}^{\mathcal{K}_t} = [h_{ji}]_{j\in\mathcal{K}_r}^{i\in\mathcal{K}_t}$ is a $|\mathcal{K}_r| \times |\mathcal{K}_t|$ channel matrix between transmitters in set $\mathcal{K}_t$ and receivers in set $\mathcal{K}_r$. For any coding scheme, receivers in $\mathcal{S}_r$ should be able to decode the bits $\mathcal{W}_{\mathcal{S}_r}$. Therefore, receivers in $\mathcal{S}_r$ can compute $\mathbf{X}_{\overline{\mathcal{S}}_t} = \{x_i\}_{i\in\overline{\mathcal{S}}_t}$ and subtract it from the received signal using the decoded bits $\mathcal{W}_{\mathcal{S}_r}$ and the genie-aided information $\overline{\mathcal{W}}$, where the encoding function of the transmitters are as follows

$$x_i = f_i\left(B_{d_j,l} : j \in \mathcal{R}, B_{d_j,l} \in Z_i\right). \tag{8}$$

Similarly, receivers in set $\overline{\mathcal{S}}_r$ can compute $\mathbf{X}_{\overline{\mathcal{S}}_t}$ and subtract it from the received signal using the genie-aided information $\mathcal{W}_{\mathcal{S}_r} \bigcup \overline{\mathcal{W}}$. As a result, we can rewrite the received signals of receivers in $\mathcal{S}_r$ and $\overline{\mathcal{S}}_r$ as

$$\begin{aligned} \tilde{\mathbf{Y}}_{\mathcal{S}_r} &= \mathbf{H}_{\mathcal{S}_r}^{\mathcal{S}_t}\mathbf{X}_{\mathcal{S}_t} + \mathbf{Z}_{\mathcal{S}_r}, \\ \tilde{\mathbf{Y}}_{\overline{\mathcal{S}}_r} &= \mathbf{H}_{\overline{\mathcal{S}}_r}^{\mathcal{S}_t}\mathbf{X}_{\mathcal{S}_t} + \mathbf{Z}_{\overline{\mathcal{S}}_r}. \end{aligned} \tag{9}$$

where receivers $j \in \overline{\mathcal{S}}_r$ are able to decode their bits $\{B_{d_j,i} : B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}, j \in \overline{\mathcal{S}}_r\}$ from the received signal vector $\tilde{\mathbf{Y}}_{\overline{\mathcal{S}}_r}$, and receivers $j \in \mathcal{S}_r$ are able to decode their intended bits $\{B_{d_j,i} : B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}, j \in \mathcal{S}_r\}$ from the received signal vector $\tilde{\mathbf{Y}}_{\mathcal{S}_r}$. Notice that the $\sigma \times \sigma$ submatrix channel $\mathbf{H}_{\mathcal{S}_r}^{\mathcal{S}_t}$ is invertable almost surely. Thus, by reducing noise at receivers $\mathcal{S}_r$ and multiplying the constructed signal $\tilde{\mathbf{Y}}_{\mathcal{S}_r}$ at receivers $\mathcal{S}_r$ by $\mathbf{H}_{\overline{\mathcal{S}}_r}^{\mathcal{S}_t}\left(\mathbf{H}_{\mathcal{S}_r}^{\mathcal{S}_t}\right)^{-1}$, we have

$$\tilde{\mathbf{Y}}'_{\mathcal{S}_r} = \mathbf{H}_{\overline{\mathcal{S}}_r}^{\mathcal{S}_t}\mathbf{X}_{\mathcal{S}_t} + \tilde{\mathbf{Z}}_{\mathcal{S}_r}, \tag{10}$$

which is a degraded version of $\tilde{\mathbf{Y}}_{\overline{\mathcal{S}}_r}$, where $\tilde{\mathbf{Z}}_{\mathcal{S}_r}$ represents the reduced noise vector at receivers $\mathcal{S}_r$. Therefore, receivers in set $\mathcal{S}_r$ can decode all messages $\mathcal{W}_{\mathcal{S}_t}$. Thus, by using Fano's inequality, we have

$$H\left(\mathcal{W}_{\mathcal{S}_t}|\mathbf{Y}_{\mathcal{S}_r}, \overline{\mathcal{W}}\right) \le H\left(\mathcal{W}_{\mathcal{S}_t}|\mathbf{Y}_{\overline{\mathcal{S}}_r}, \overline{\mathcal{W}}, \mathcal{W}_{\mathcal{S}_r}\right) \le |\mathcal{W}_{\mathcal{S}_t}|T\epsilon. \tag{11}$$

The applied assumptions (genie-aided information, cooperation between subset of receivers, reducing noise) cannot hurt the coding scheme. Thus, we have

$$\begin{aligned} & H(\mathcal{W}_{\mathcal{S}_t}, \mathcal{W}_{\mathcal{S}_r}) \qquad (12) \\ &= \sum_{j\in\mathcal{S}_r}\sum_{i=1}^{F}\mathbf{1}\left(B_{d_j,i} \in \mathcal{Z}\right) + \sum_{j\in\overline{\mathcal{S}}_r}\sum_{i=1}^{F}\mathbf{1}\left(B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}\right) \\ &\overset{(a)}{=} H\left(\mathcal{W}_{\mathcal{S}_t}, \mathcal{W}_{\mathcal{S}_r}|\overline{\mathcal{W}}\right) \\ &\overset{(b)}{=} I\left(\mathcal{W}_{\mathcal{S}_t}, \mathcal{W}_{\mathcal{S}_r}; \mathbf{Y}_{\mathcal{S}_r}|\overline{\mathcal{W}}\right) + H\left(\mathcal{W}_{\mathcal{S}_t}, \mathcal{W}_{\mathcal{S}_r}|\mathbf{Y}_{\mathcal{S}_r}, \overline{\mathcal{W}}\right) \qquad (13) \\ &\overset{(c)}{\le} I\left(\mathbf{X}_{[K_T]}; \mathbf{Y}_{\mathcal{S}_r}\right) + H\left(\mathcal{W}_{\mathcal{S}_t}, \mathcal{W}_{\mathcal{S}_r}|\mathbf{Y}_{\mathcal{S}_r}, \overline{\mathcal{W}}\right) \\ &\overset{(d)}{\le} T\sigma\log(P) + H\left(\mathcal{W}_{\mathcal{S}_r}|\mathbf{Y}_{\mathcal{S}_r}, \overline{\mathcal{W}}\right) + H\left(\mathcal{W}_{\mathcal{S}_t}|\mathbf{Y}_{\mathcal{S}_r}, \overline{\mathcal{W}}, \mathcal{W}_{\mathcal{S}_r}\right) \\ &\overset{(e)}{\le} T\sigma\log(P) + |\mathcal{S}_r|T\epsilon + |\mathcal{S}_t|T\epsilon \end{aligned}$$

where $\mathbf{1}(.)$ is an indicator function. (a) follows from the fact that the sets of bits are independent. Step (b) follows from the chain rule. Step (c) follows from data processing inequality, where the signal $\mathbf{X}_{[K_T]}$ is a function of $\mathcal{W}_{\mathcal{S}_t} \bigcup \mathcal{W}_{\mathcal{S}_r}$. Step (d) follows from the bound of the degrees of freedom of multiple access channel (MAC) with K_T single-antenna transmitters and a receiver with $|\mathcal{S}_r|$ antennas. Finally, step (e) follows from Fano's inequality. By diving on F, and taking $P \to \infty$ and $\epsilon \to 0$, we get.

$$\frac{1}{F}\left(\sum_{j\in\mathcal{S}_r}\sum_{i=1}^{F}\mathbf{1}\left(B_{d_j,i} \in \mathcal{Z}\right) + \sum_{j\in\overline{\mathcal{S}}_r}\sum_{i=1}^{F}\mathbf{1}\left(B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}\right)\right) \le \sigma\tau\left(\mu, \mathbf{d}, \mathcal{Z}\right). \tag{14}$$

Notice $\mathbf{1}\left(B_{d_j,i} \in \mathcal{Z}\right) = 1$ for any bit in the library, since every bit should be available at least at one of the transmitter caches. Hence, the first term in the left hand side (LHS) is equal to σF. Then, by taking the average of the above inequality over all possible set $\mathcal{S}_r \subset \mathcal{R}$, we have

$$\sigma + \frac{\binom{s-1}{s-\sigma-1}}{F\binom{s}{\sigma}}\left(\sum_{j\in\mathcal{R}}\sum_{i=1}^{F}\mathbf{1}\left(B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}\right)\right) \le \sigma\tau\left(\mu, \mathbf{d}, \mathcal{Z}\right). \tag{15}$$

where every indicator $\mathbf{1}\left(B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}\right)$ in the second term in the LHS is counted $\binom{s-1}{s-\sigma-1}$ times. Now, we follow similar steps as in [11] to average the above inequality over all possible demands $d \in \mathcal{D}_s$, and all possible transmitter sets. Let $\mathcal{K}_{d_j,i}$ denote the set of transmitters that exclusively store the i-th bit of the file W_{d_j}. Thus, $\mathbf{1}\left(B_{d_j,i} \notin \mathcal{Z}_{\overline{\mathcal{S}}_t}\right) = \mathbf{1}\left(\mathcal{K}_{d_j,i} \bigcap \overline{\mathcal{S}}_t = \phi\right)$. By taking the average of all possible set $\mathcal{S}_t \subset [K_T]$, the second term in the LHF is equal

$$\frac{s-\sigma}{Fs}\left(\sum_{j\in\mathcal{R}}\sum_{i=1}^{F}\frac{\sum_{\mathcal{S}_t\subset[K_T]}\mathbf{1}\left(\mathcal{K}_{d_j,i} \bigcap \overline{\mathcal{S}}_t = \phi\right)}{\binom{K_T}{\sigma}}\right). \tag{16}$$

where $\frac{\binom{s-1}{s-\sigma-1}}{\binom{s}{\sigma}} = \frac{s-\sigma}{s}$, and we exchange the order of summations. The term $\frac{1}{\binom{K_T}{\sigma}}\sum_{\mathcal{S}_t\subset[K_T]}\mathbf{1}\left(\mathcal{K}_{d_j,i} \bigcap \overline{\mathcal{S}}_t = \phi\right)$ is equal to the probability of selecting $K_T - \sigma$ transmitters uniformally at random, and none of them belongs to $\mathcal{K}_{d_j,i}$. Hence, this term can be computed as follows[1]

$$\frac{1}{\binom{K_T}{\sigma}}\sum_{\mathcal{S}_t\subset[K_T]}\mathbf{1}\left(\mathcal{K}_{d_j,i} \bigcap \overline{\mathcal{S}}_t = \phi\right) = \frac{\binom{K_T-|\mathcal{K}_{d_j,i}|}{K_T-\sigma}}{\binom{K_T}{K_T-\sigma}}. \tag{17}$$

Let a_{n,d_j} denote the number of bits of file W_{d_j} that are stored exclusively at n transmitters, and hence, $|\mathcal{K}_{d_j,i}| = n$ for a fraction $a_{n,d_j}/F$. By taking the average (17) over all bits of file W_{d_j}, we obtain

$$\frac{1}{F}\sum_{i=1}^{F}\frac{\binom{K_T-|\mathcal{K}_{d_j,i}|}{K_T-\sigma}}{\binom{K_T}{K_T-\sigma}} = \sum_{n=1}^{K_T}\frac{a_{n,d_j}}{F}\frac{\binom{K_T-n}{K_T-\sigma}}{\binom{K_T}{K_T-\sigma}} = \sum_{n=1}^{K_T}\frac{a_{n,d_j}}{F}\frac{\binom{\sigma}{n}}{\binom{K_T}{n}} \tag{18}$$

[1]We assume that $\binom{n}{k} = 0$ if $n < k$.

where we use the equality $\binom{K-n}{l}/\binom{K}{l} = \binom{K-l}{n}/\binom{K}{n}$. Substituting from (18) into (16), then we obtain

$$1 + \frac{s-\sigma}{\sigma s} \sum_{j\in\mathcal{R}} \sum_{n=1}^{K_T} \frac{a_{n,d_j}}{F} \frac{\binom{\sigma}{n}}{\binom{K_T}{n}} \leq \tau\left(\mu, \mathbf{d}, \mathcal{Z}\right). \tag{19}$$

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

$$1 + \frac{s-\sigma}{\sigma s} \frac{1}{|\mathcal{D}_s|} \sum_{\mathbf{d}\in\mathcal{D}_s} \sum_{j\in\mathcal{R}} \sum_{n=1}^{K_T} \frac{a_{n,d_j}}{F} \frac{\binom{\sigma}{n}}{\binom{K_T}{n}} \leq \overline{\tau}\left(\mu, s\left(\mathbf{d}\right), \mathcal{Z}\right). \tag{20}$$

It is easy to verify that demands $\mathbf{d} \in \mathcal{D}_s$ are uniformally distributed, since $\mathbf{d} \in [N]^{K_R}$ is a random vector with uniform distribution. Moreover, for file W_j, $j \in [N]$, the term $a_{n,j}$ is computed $\binom{N-1}{s-1}|\mathcal{D}_s|/\binom{N}{s}$ times in the summation $\sum_{\mathbf{d}\in\mathcal{D}_s} \sum_{j\in\mathcal{R}} \frac{a_{n,d_j}}{F}$. Thus, (20) is equal to

$$1 + \frac{s-\sigma}{\sigma} \sum_{j=1}^{N} \sum_{n=1}^{K_T} \frac{a_{n,j}}{NF} \frac{\binom{\sigma}{n}}{\binom{K_T}{n}} \leq \overline{\tau}\left(\mu, s\left(\mathbf{d}\right), \mathcal{Z}\right). \tag{21}$$

Let $\alpha_n = \sum_{j=1}^{N} a_{n,j}/NF$, and $K_T\mu = t$. By minimizing both sides of (21) over all possible uncoded placement schemes, we get

$$\begin{aligned} 1+&\frac{s-\sigma}{\sigma} \min_{\mathcal{Z}} \sum_{n=1}^{K_T} \alpha_n \frac{\binom{\sigma}{n}}{\binom{K_T}{n}} \leq \min_{\mathcal{Z}} \overline{\tau}\left(\mu, s\left(\mathbf{d}\right), \mathcal{Z}\right) \\ s.t. \quad & \sum_{n=1}^{K_T} \alpha_n = 1 \\ & \sum_{n=1}^{K_T} n\alpha_n = t \end{aligned} \tag{22}$$

where the first constraint comes from the total number of bits in library, while the second constraint is to maintain the total size of transmitter caches. Notice that $f_n = \frac{\binom{\sigma}{n}}{\binom{K_T}{n}}$ is a decreasing function of n. Moreover, we can verify that f_n is a discrete convex function of n, since $f_{n+1} + f_{n-1} \geq 2f_n$ in region $1 \leq n \leq \sigma$ [12, Theorem 1]. The objective function is a linear combination of points $\{f_n\}$. Hence, the optimal solution is $\alpha_t = 1$ when t is integer, i.e., $t \in [1 : K_T]$. While for non-integer point of t, we can write $t = \alpha t_1 + (1-\alpha) t_2$, where $t_1 \leq t \leq t_2$. Thus, the optimal solution is $\alpha_{t_1} = \alpha$ and $\alpha_{t_2} = (1-\alpha)$. Therefore, we can proceed the proof to bound the expected NDT for the corner points $t \in [1 : K_T]$, where the expected NDT for non-integer t can be bounded by the linear combination of the nearest two integer points. Thus, we get

$$1 + \frac{s-\sigma}{\sigma} \frac{\binom{\sigma}{t}}{\binom{K_T}{t}} \leq \min_{\mathcal{Z}} \overline{\tau}\left(\mu, s\left(\mathbf{d}\right), \mathcal{Z}\right) \tag{23}$$

To get the best tight bound on the NDT, we maximize the LHS of (23) over all possible values of $\sigma \in \mathcal{F} \triangleq \{1 \leq \sigma \leq \min\{K_T, s\left(d\right)\}\}$.

$$\max_{\sigma\in\mathcal{F}} \frac{t\binom{K_T}{t} + (s-\sigma)\binom{\sigma-1}{t-1}}{t\binom{K_T}{t}} \leq \min_{\mathcal{Z}} \overline{\tau}\left(\mu, s\left(\mathbf{d}\right), \mathcal{Z}\right). \tag{24}$$

Finally, by taking the expectation with respect to $s\left(\mathbf{d}\right)$, we have

$$\mathbb{E}\left[\max_{\sigma\in\mathcal{F}} \frac{t\binom{K_T}{t} + (s-\sigma)\binom{\sigma-1}{t-1}}{t\binom{K_T}{t}}\right] \leq \overline{\tau}\left(\mu\right). \tag{25}$$

This completes the proof of Theorem 1.

V. Conclusion

We have derived a lower bound on the *expected* normalized delivery time for cache-aided interference networks under uniform popularity distribution. Our bound is mainly based on cut-set and genie-aided arguments. For peak NDT, the results have shown that our lower bound is tighter than the bound in [3] for small cache sizes, while both bounds coincide with each other for large cache sizes. Furthermore, We have shown that the feasible region of the expected NDT is bigger than the feasible region of the peak NDT. Hence, the achievable schemes on the peak NDT might no longer becomes order optimal with respect to the new derived bound on the expected NDT.

References

[1] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *International Symposium on Information Theory (ISIT)*, pp. 809–813, IEEE, 2015.

[2] J. Kakar, S. Gherekhloo, Z. H. Awan, and A. Sezgin, "Fundamental limits on latency in cloud-and cache-aided hetnets," in *International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2017.

[3] 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)*, pp. 320–325, IEEE, 2016.

[4] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *International Symposium on Information Theory (ISIT)*, pp. 2029–2033, IEEE, 2016.

[5] J. Zhang and O. Simeone, "Fundamental limits of cloud and cache-aided interference management with multi-antenna base stations," in *International Symposium on Information Theory (ISIT)*, pp. 1425–1429, IEEE, 2018.

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

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

[8] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Transactions on Information Theory*, 2017.

[9] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, "Decentralized coded caching in wireless networks: Trade-off between storage and latency," in *International Symposium on Information Theory (ISIT)*, pp. 2443–2447, IEEE, 2017.

[10] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, "Fundamental limits of memory-latency tradeoff in fog radio access networks under aritrary demands," in *preparation*, 2018.

[11] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," in *International Symposium on Information Theory (ISIT)*, pp. 1613–1617, IEEE, 2017.

[12] K. Murota, "Discrete convex analysis," *Mathematical Programming*, vol. 83, no. 1-3, pp. 313–371, 1998.