Cache Aided Wireless Networks: Tradeoffs between Storage and Latency

Avik Sengupta
Hume Center, Department of ECE
Virginia Tech,
Blacksburg, VA 24060, USA
Email: aviksg@vt.edu

Ravi Tandon
Department of ECE
University of Arizona,
Tucson, AZ 85721 USA
Email: tandonr@email.arizona.edu

Osvaldo Simeone
CWCSPR, Department of ECE
New Jersey Institute of Technology,
Newark, NJ 07102 USA
Email: osvaldo.simeone@njit.edu

***Abstract*—We investigate the fundamental information theoretic limits of cache-aided wireless networks, where edge nodes (or transmitters) are endowed with caches that can store popular content such as multimedia files. This architecture aims to localize popular multimedia content by proactively pushing it closer to the edge of the wireless network, thereby alleviating backhaul load. An information theoretic model of such networks is presented, that includes the introduction of a new metric, namely normalized delivery time (NDT), which captures the worst case time to deliver any requested content to the users. We present new results on the trade-off between latency, measured via the NDT, and the cache storage capacity of the edge nodes. In particular, a novel information theoretic lower bound on NDT is presented for cache aided networks. The optimality of this bound is shown for several system parameters.**

***Index Terms*—Caching, 5G, degrees of freedom, latency.**

I. Introduction

Edge processing is one of the emerging trends in the evolution of 5G networks [1]. It refers to the utilization of locally stored content and computing resources at the network edge, i.e., closer to the users. Such localization is particularly appealing for both low-latency or location-based applications as well as multimedia transmissions. A network architecture with edge processing capability is shown in Fig. 1(a). Here, edge nodes (ENs), such as base stations or LTE eNodeBs, are equipped with local caches which can store popular content, most notably multimedia files. The local availability of popular content at the network edge has the potential of reducing the delivery latency as well as the overhead on backhaul connections to content servers. As a result, cache enabled networks have studied extensively in recent literature [2]–[6].

In this paper, we investigate *cache-aided wireless networks*, where ENs are endowed with caching capability to store popular content locally. The design of cache-aided networks involves two key design questions: a) *what to cache*, i.e., how should the storage at ENs be utilized, and which content must be stored; and b) *how to efficiently deliver* the requested content to the users by leveraging the caches at the ENs. The design of *caching policies* is typically done at the long time scale at which users' preferences are invariant and can span many transmission intervals, each corresponding to a set of requests from the users. Hence, the caching policy must be agnostic to the demands of the users as well as to the instantaneous wireless channel conditions. Instead, efficient delivery of requested content to users in each transmission interval, calls for the *design of transmission policies* that

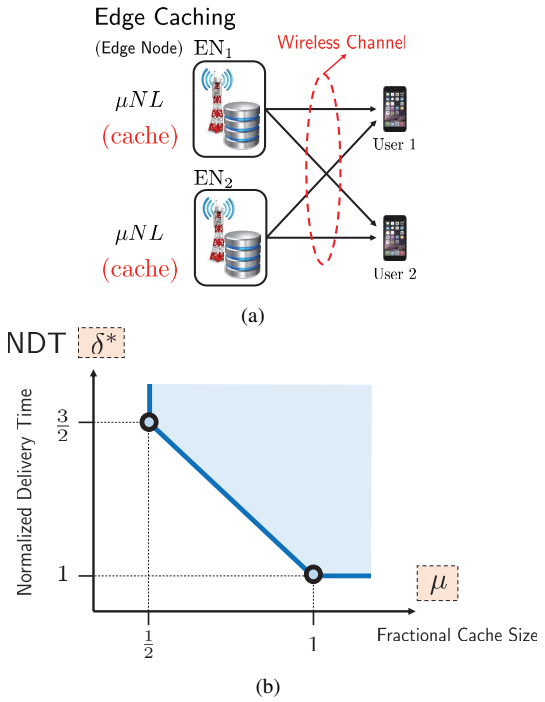

Fig. 1. (a) Information-theoretic model for edge caching for $M = 2$ ENs serving $K = 2$ users; (b) Trade-off between the introduce metric of normalized delivery time (NDT), δ^*, and the fractional cache size μ with full CSI at ENs and users.

utilize the available wireless channel state information (CSI) at the ENs and the instantaneous demands of the users. We first present an information theoretic modeling of cache-aided wireless networks that succinctly captures its new design aspects and constraints. We then develop a new performance measure for such networks termed the *Normalized Delivery Time* (NDT), which measures the worst-case latency incurred by a cache-aided wireless network relative to an ideal system with unlimited caching capability and interference-free links to the users. This helps a latency centric analysis of the high signal-to-noise ratio (SNR) degrees-of-freedom (DoF) performance of the system.

Example 1. To illustrate the NDT performance metric, consider the set-up of Fig. 1(a), in which two ENs, labeled as EN_1 and EN_2 are deployed to serve two users over a shared wireless channel. We assume that there is a library of N popular files, each of size L bits, and each EN can cache at most μNL bits. In other words, $\mu \in [0, 1]$ denotes the *fractional cache size*, i.e., the ratio between the available per-

978-1-4673-9457-4/16/$31.00 ©2016 IEEE

file storage at an EN and the total size of all the files. For the example shown in Fig. 1(a), the information-theoretically optimal trade-off between NDT $\delta^*(\mu)$ and the fractional cache size μ is shown in Fig. 1(b). To explain the operating points on this curve, first consider $\mu = 1$, i.e., the case when both ENs can store all files, and full cooperative transmission is possible from the ENs, i.e., via zero-forcing beamforming for any set of users' requests. This yields an NDT of $\delta^*(1) = 1$, implying that the latency performance is the same as that of the ideal interference free system. On the other hand, at $\mu = 1/2$, which is the smallest cache for enabling the delivery of any set of requests, the NDT increases to $\delta^*(1/2) = 3/2$, and is achieved via interference alignment, thus revealing the performance loss due to decrease in the fractional cache size.

Related Work: Cache-aided interference channels were first investigated in a recent work by Maddah Ali and Niesen [7], [8], who introduced the problem and investigated it for $M = 3$ ENs and $K = 3$ users, and presented an upper bound on the NDT for this specific setting of $M = K = 3$. However, no attempt was made in [7] to develop lower bounds on NDT or to show the optimality of the scheme.

Main Contributions: To the best of our knowledge, this paper is the first to develop information theoretic lower bounds (converse) on latency in cache-aided wireless networks. The main questions we investigate in this work are the following: *What is the optimal caching-transmission policy as a function of the fractional cache size μ? What is the optimal trade-off between the system performance (measured in terms of NDT), and the fractional cache size μ?* The main contributions of this paper are as follows:

- We first present an information-theoretic modeling of cache enabled wireless networks and develop the NDT to measure the latency performance of such networks. For a class of practically relevant caching policies, namely uncoded caching, with full CSI at the ENs, we develop information theoretic lower bounds on the NDT.
- We show that the presented lower bounds on the NDT are optimal for the setting of $M = 2$ ENs and $K = 2$ users. Together with the upper bound in [7], we partially characterize the NDT trade-off for $M = 3$ ENs and the $K = 3$ users. In addition, we show that our lower bounds are optimal for extremal values of μ for general problem parameters.
- Finally, we investigate the impact of CSI availability at the ENs on the NDT. For the case of $M = 2$ ENs and $K = 2$ users, we illustrate the impact of delayed or no CSI at the ENs on the resulting NDT.

II. System Model

We consider a cache-aided wireless network where M edge-nodes (EN) are connected to a total of K users. The ENs can cache content from a library of N files, $F_1, \ldots, F_N$, where each file is of size L bits, for some $L \in \mathbb{N}^+$. Formally, the files F_n are independent identically distributed (i.i.d.) as:

$$F_n \sim \text{Unif}\{1, 2, \ldots, 2^L\}, \quad \forall n = 1, \ldots, N. \tag{1}$$

Each EN is equipped with a cache in which it can store μNL bits, where the fraction μ, with $0 \leq \mu \leq 1$, is referred to as the *fractional cache size*. It is required that the collective cache size of the M ENs be large enough to completely store the entire library of N files. In this way, all user requests can be completely serviced by the ENs. Thus, we impose the condition that $M \times \mu NL \geq NL$, i.e., $\mu \geq 1/M$. Therefore, it suffices to focus on the range $\mu \in [1/M, 1]$. In each transmission interval, a user can request any file from the library and these requests are served by the ENs. The channel between EN_m and user k, in a given transmission interval is denoted by $h_{km} \in \mathbb{C}$, where $k = 1, \ldots, K$ and $m = 1, \ldots, M$. The coefficients are assumed to be drawn i.i.d. from a continuous distribution and to be time-invariant with each transmission interval.

Definition 1 (Policy)**.** A caching, edge transmission, decoding policy $\pi = (\pi_c, \pi_e, \pi_d)$ is characterized by the following three mapping functions.

a) Caching Policy π_c: The caching policy is defined by a function, $\pi_c^m(\cdot)$, at each edge node EN_m, $m = 1, 2, \ldots, M$, which maps each file to its cache storage

$$S_{m,n} \triangleq \pi_c^m(F_n) \quad \forall n = 1, 2, \ldots, N. \tag{2}$$

The mapping is such that $H(S_{m,n}) \leq \mu L$ in order to satisfy the cache capacity constraints. The total cache content at EN_m is given by $S_m = (S_{m,1}, S_{m,2}, \ldots, S_{m,N})$. Note that the caching policy π_c allows for arbitrary coding within each file. However, it does not allow for inter-file coding and is hence a special case of a more general caching policy which might allow for arbitrary inter-file coding. Furthermore, the caching policy is kept fixed over multiple transmission intervals and is thus agnostic to user requests and to channel coefficients h_{km}.

b) *Edge Transmission Policy π_e*: During the delivery phase of each transmission interval, each receiver $k \in \{1, 2, \ldots, K\}$ requests one of the N files. The demand vector is denoted by $\mathbf{D} \triangleq (d_1, \ldots, d_K) \in \{1, \ldots, N\}^K$. Knowing the demand vector $\mathbf{D}$, the global CSI $\mathbf{H} = \left\{\{h_{km}\} : {}^{k=1,\ldots,K}_{m=1,\ldots,M}\right\}$, denoting the channel coefficient between every user and EN, and having access only to its local cache content, S_m, the edge-node EN_m uses an edge transmission policy, $\pi_e^m(\cdot)$, which encodes the cache content, S_m, to output a codeword

$$\mathbf{X}_m^{T^{(\mathbf{D},\mathbf{H})}} = (X_m[t])_{t=1}^{T^{(\mathbf{D},\mathbf{H})}} = \pi_e^m(S_m, \mathbf{D}, \mathbf{H}), \tag{3}$$

which is transmitted to the users. Here, $T^{(\mathbf{D},\mathbf{H})}$ is the duration or block-length, of the channel coding policy based on a demand vector $\mathbf{D}$ and the channel realization $\mathbf{H}$. An average power constraint of P is imposed on each codeword, i.e.

$$E\left[\left(X_m[t] - E[X_m[t]]\right)^2\right] \leq P, \quad \forall t. \tag{4}$$

We assume that full CSI is available at all ENs and users. The issue of performance losses incurred due to degraded CSI is briefly addressed in Section III-B.

c) *Decoding Policy π_d*: Each user $k \in \{1, 2, \ldots, K\}$, receives a channel output given by:

$$\mathbf{Y}_k^{T^{(\mathbf{D},\mathbf{H})}} = (Y_k[t])_{t=1}^{T^{(\mathbf{D},\mathbf{H})}} = \sum_{m=1}^{M} h_{km} \mathbf{X}_m^{T^{(\mathbf{D},\mathbf{H})}} + \mathbf{n}_k^{T^{(\mathbf{D},\mathbf{H})}}, \tag{5}$$

where the noise $\mathbf{n}_k^{T^{(\mathbf{D},\mathbf{H})}} = (n_k[t])_{t=1}^{T^{(\mathbf{D},\mathbf{H})}}$. Each noise term $n_k[t] \sim \mathcal{N}(0,1)$ is a zero mean, unit variance Gaussian random variable which is i.i.d. across time and users. Each user $k \in \{1,2,\ldots,K\}$, has a decoding policy $\pi_d(\cdot)$, which maps the channel outputs, the receiver demands and the channel realization to the estimate

$$\widehat{F}_{d_k} \triangleq \pi_d^k\left(\mathbf{Y}_k^{T^{(\mathbf{D},\mathbf{H})}}, \mathbf{D}, \mathbf{H}\right) \tag{6}$$

of the requested file F_{d_k}. The caching, edge transmission and decoding policies together form a policy $\pi = (\pi_c^m, \pi_e^m, \pi_d^k)$ for the cache-aided wireless network. The probability of error of the policy π is defined as

$$P_e = \max_{\mathbf{D}} \max_{k\in\{1,\ldots,K\}} \mathbb{P}\left(\widehat{F}_{d_k} \neq F_{d_k}\right). \tag{7}$$

A policy is said to be feasible if, for almost all channel realizations $\mathbf{H}$ of the channel, i.e., with probability 1, we have $P_e \to 0$ when $L \to \infty$.

Definition 2. (Delivery time per bit) The *average achievable delivery time per bit* for a given feasible policy is defined as

$$\Delta(\mu, P) = \max_{\mathbf{D}} \limsup_{L\to\infty} \frac{\mathbb{E}_{\mathbf{H}}\left[T^{(\mathbf{D},\mathbf{H})}\right]}{L}, \tag{8}$$

where the expectation is over the channel realizations $\mathbf{H}$.

While $\Delta(\mu, P)$ generally depends on the power level P, as well as on μ, we next define a more tractable metric that reflects the latency performance in the high SNR regime.

Definition 3. (**NDT**) For any achievable $\Delta(\mu, P)$, the *normalized delivery time* (NDT), is defined as

$$\delta(\mu) = \lim_{P\to\infty} \frac{\Delta(\mu, P)}{1/\log P}. \tag{9}$$

Moreover, for a given μ, the minimum NDT is defined as

$$\delta^*(\mu) = \inf\left\{\delta(\mu) : \delta(\mu) \text{ is achievable}\right\}. \tag{10}$$

Remark 1. The delivery time per bit $\Delta(\mu, P)$ is normalized by the term $1/\log P$. This is the delivery time per bit in the high SNR regime for an ideal baseline system with no interference and unlimited caching, in which each user can be served by a dedicated EN which has locally stored all the files. An NDT of δ^* indicates that the worst-case time required to serve any possible request $\mathbf{D}$, is δ^* times larger than the time needed by this ideal baseline system.

Remark 2. We observe that the NDT in (10) is proportional to the inverse of the more conventional degrees of freedom (DoF) metric DoF(μ) defined in [7], [8], namely $\delta^*(\mu) = K/\text{DoF}(\mu)$. In this paper, we opted for definition (10), rather than resorting to the DoF metric, as we believe that it more clearly reflects the operational meaning in terms of delivery latency. We also recall that [7], [8] adopted the metric $1/\text{DoF}(\mu)$ based on the observation that the latter is a convex function of μ, unlike the function DoF(μ). Finally, we note that the NDT can be extended to more general scenarios for which a direct functional dependence with the DoF cannot be established [9].

Remark 3. Following the same arguments in [7], [8], it can be seen that the minimum NDT, $\delta^*(\mu)$, is a convex function of μ. In fact, consider any two caching policies π_1, requiring storage μ_1, and π_2, requiring storage μ_2. Given a system with storage $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$, for any $\alpha \in [0,1]$, the system can then operate according to policy π_1 using an α-fraction of the cache and of time on the channel to the users, and with policy π_2 for the remaining part of the cache and of time, achieving an NDT of $\delta^*(\mu) \leq \alpha\delta^*(\mu_1) + (1-\alpha)\delta^*(\mu_2)$. Thus, the convexity of $\delta^*(\mu)$ follows from the possibility of implementing the outlined cache-sharing and time-sharing scheme.

III. Main Results and Discussion

In this work, we aim to provide fundamental limits for the NDT of an $M \times K$ cache-aided wireless system. To this end, an information-theoretic lower-bound on the NDT of the system is presented in the following section. Section III-B instead briefly discusses the impact of imperfect CSI at the ENs.

A. Lower Bounds on NDT with Perfect CSI at the ENs

In this section, we consider cache-aided wireless systems where perfect CSI is present at the ENs and users.

Theorem 1. *For a cache-aided wireless system with M ENs, each with a fractional cache size $\mu \in [1/M, 1]$, K users and a library of $N \geq K$ files and with perfect CSI at both ENs and users, the NDT is lower bounded as*

$$\delta^*(\mu) \geq \max_{\ell\in 1,\ldots,\min\{M,K\}} \frac{K - (M-\ell)^+(K-\ell)^+\mu}{\ell}, \tag{11}$$

where the function $(x)^+$ is defined as $(x)^+ = \max\{0, x\}$.

To the best of the authors' knowledge, Theorem 1 provides the first converse for the $M \times K$ cache-aided wireless system. The proof of Theorem 1 is presented in Appendix A. To provide further insight into the lower bound in Theorem 1, we present here, a short proof sketch. As shown in Appendix A, the channel outputs of ℓ users, along with the cache contents of $(M-\ell)^+$ ENs is sufficient in the high-SNR regime to decode any K requested files. By bounding the joint entropy of these random variables and utilizing the cache storage, caching policy and decodability constraints, one obtains the lower bound on the optimal NDT $\delta^*(\mu)$. Varying the parameter ℓ leads to the family of lower bounds in Theorem 1. Based on this lower bound, we next expound on the optimal characterization of $\delta^*(\mu)$ for some cache-aided wireless systems.

Corollary 1. *For a cache-aided wireless system with M ENs, each with a fractional cache size $\mu \in [1/M, 1]$, K users and a library of $N \geq K$ files, we have*

$$\delta^*(\mu) = \frac{M+K-1}{M} \quad \textit{for} \quad \mu = 1/M, \tag{12}$$

which can be achieved by leveraging interference alignment techniques for a $M \times K$ X-channel. Further, we have

$$\delta^*(\mu) = \frac{K}{\min\{M,K\}} \quad \textit{for} \quad \mu = 1, \tag{13}$$

which can be achieved by using zero-forcing beamforming for a $M \times K$ broadcast channel.

Proof. To prove the corollary, we show that a policy with a NDT matching the lower bound in Theorem 1 can be identified for both $\mu = 1/M$ and $\mu = 1$.

NDT at $\mu = 1/M$: For $\mu = 1/M$, we substitute $\ell = 1$ in (11) to get

$$\delta^*(1/M) \geq K - \frac{(M-1)(K-1)}{M} = \frac{M+K-1}{M}. \tag{14}$$

To obtain an upper bound on NDT, we consider the following policy. For $\mu = 1/M$, each file can be split into M non-overlapping fragments $F_n = \{F_{n,1}, F_{n,2}, \ldots, F_{n,M}\}$ each of size L/M bits. The fragments $F_{n,m}$ are stored in the cache of EN_m for $n = 1, \ldots, N$ [7]. Thus, the cache storage for each EN is NL/M bits and the total amount of data stored in the caches of all ENs is NL bits. Next, when a file is requested by any user k, each of the ENs have a fragment $F_{d_k,m}$ to transmit to the user. The $M \times K$ system then becomes an X-channel for which, a reliable sum-rate of $\frac{MK}{M+K-1}\log(P)$, neglecting $o(\log(P))$ terms, is achievable by interference alignment [10], [11]. Thus the achievable delivery time per bit, in Definition 2, is approximately given by

$$\Delta(\mu, P) = \lim_{L\to\infty} \frac{1}{L} \cdot \frac{KL}{\frac{MK}{M+K-1}\log(P)} = \frac{M+K-1}{M\log(P)}. \quad (15)$$

And hence, we have the achievable NDT

$$\delta(\mu) = \lim_{P\to\infty} \frac{\Delta(\mu, P)}{1/\log(P)} = \frac{M+K-1}{M}. \quad (16)$$

Thus, we have the upper bound

$$\delta^*(1/M) \leq \frac{M+K-1}{M}. \quad (17)$$

Combining (14) and (17) shows that the lower bound in Theorem 1 is tight for $\mu = 1/M$.

NDT at $\mu = 1$: For $\mu = 1$, substituting, $\ell = \min\{M, K\}$ into (11), we get

$$\delta^*(1) \geq \frac{K}{\min\{M, K\}}. \quad (18)$$

When $\mu = 1$, each EN has a cache storage of NL bits, i.e., each EN can completely store the entire library. Hence the ENs can cooperatively transmit to the users using broadcast techniques such as zero-forcing to achieve a reliable sum-rate of $\min\{M, K\}\log(P)$, neglecting $o(\log(P))$ terms [12]. Thus, the delivery time per bit is approximately given by

$$\Delta(\mu, P) = \lim_{L\to\infty} \frac{KL/L}{\min\{M, K\}\log(P)} = \frac{K/\log(P)}{\min\{M, K\}}. \quad (19)$$

And hence, we have the achievable NDT

$$\delta(\mu) = \lim_{P\to\infty} \frac{\Delta(\mu, P)}{1/\log(P)} = \frac{K}{\min\{M, K\}}. \quad (20)$$

Thus, we have the upper bound

$$\delta^*(1) \leq \frac{K}{\min\{M, K\}}. \quad (21)$$

Combining (18) and (21), shows that the lower bound in Theorem 1 is tight at $\mu = 1$. □

Based on the results of Corollary 1, we establish the optimal NDT for a system with $M = K = 2$ as stated in the following corollary.

Corollary 2. *For a cache-aided wireless system with $M = 2$ ENs, $K = 2$ users and $N \geq 2$ files, the optimal NDT is given by*

$$\delta^*(\mu) = 2 - \mu, \quad \forall \mu \in [1/2, 1]. \quad (22)$$

For this 2-EN, 2-user system, the two corner points $\mu = 1/2$ and $\mu = 1$ are achievable as per Corollary 1. Instead, for $\mu = 1/2$, the system is a $2-$user X-channel which has a sum-DoF of $4/3$, i.e., $\delta(1/2) = 3/2$. Again, at $\mu = 1$, the system becomes a broadcast channel which has a sum-DoF of 2, i.e., $\delta(1) = 1$. All points on the line joining these two achievable points can be achieved through cache and time sharing between the two schemes as stated in Remark 3 in Section II. Next, considering the lower bound from Theorem 1 and using $\ell = 1$, we get (22), which is the line joining the two achievable corner points. Thus, Theorem 1 completely characterizes the optimal NDT $\delta^*(\mu)$ of the cache-aided wireless system with $M = K = 2$ and $N \geq 2$. This is illustrated in Fig. 1(b). We next present an application of Theorem 1 to obtain partial characterization of the NDT for a system with $M = K = 3$.

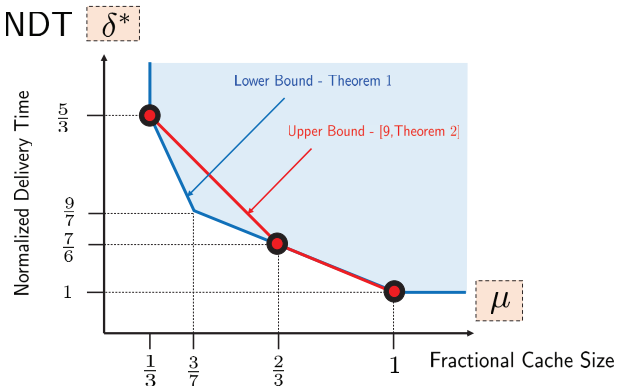

Fig. 2. Lower and upper bounds on the NDT for a cache-aided wireless system with $M = K = 3$.

Corollary 3. *For a cache-aided wireless system with $M = 3$ ENs, $K = 3$ users and $N \geq 3$ files, we have*

$$\delta^*(\mu) = \begin{cases} 5/3 & \text{for } \mu = 1/3, \\ 3/2 - \mu/2 & \text{for } 2/3 \leq \mu \leq 1 \end{cases}$$

$$3 - 4\mu \leq \delta^*(\mu) \leq 13/6 - 3\mu/2 \text{ for } 1/3 \leq \mu \leq 2/3. \quad (23)$$

The bounds in Corollary 3 are illustrated in Fig. 2. The lower bounds on NDT used in Corollary 3 are obtained from Theorem 1, by setting $M = K = 3$ system, yielding

$$\delta^*(\mu) \geq 3 - 4\mu \quad \text{for} \quad \ell = 1, \quad (24)$$

$$\delta^*(\mu) \geq 3/2 - \mu/2 \quad \text{for} \quad \ell = 2, \quad (25)$$

$$\delta^*(\mu) \geq 1 \quad \text{for} \quad \ell = 3. \quad (26)$$

As for upper bounds, we adapt the results in [7, Theorem 2] to obtain the following achievable NDT:

$$\delta^*(\mu) \leq \begin{cases} 13/6 - 3\mu/2 & \text{for } 1/3 \leq \mu \leq 2/3, \\ 3/2 - \mu/2 & \text{for } 2/3 \leq \mu \leq 1. \end{cases} \quad (27)$$

The two corner points for $\mu = 1/3$ and $\mu = 1$ of the achievable NDT in (27) are achieved similar to Corollary 1. The inner point at $\mu = 2/3$, instead uses a novel interference alignment and zero-forcing scheme to achieve a $\delta(\mu) = 7/6$ [7]. It can be seen from Fig. 2 that the lower bound coincides with the upper bound at $\mu = 1/3$ and for the range $2/3 \leq \mu \leq 1$. Hence, the proposed lower bound in conjunction with the recent result from [7], partially characterizes the optimal NDT versus μ trade-off for the $M = K = 3$ system as summarized in Corollary 3. For the regime $1/3 \leq \mu \leq 2/3$, characterizing the optimal NDT remains an open problem.

B. Impact of Imperfect CSI on the NDT Trade-off

In this section, we investigate the impact of CSI availability at the ENs and its impact on the NDT. When CSI is delayed,

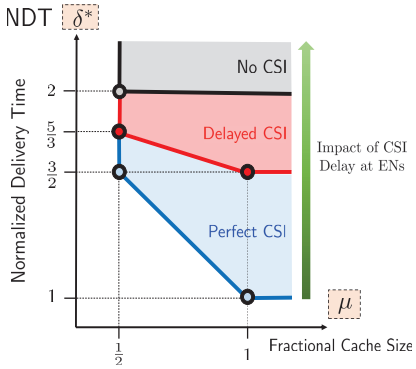

Fig. 3. Effect of delayed or no CSI on the NDT for $M = K = 2$.

at time t, ENs only have access to $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_{t-1}$, i.e., the CSI of the previous $t-1$ slots. For illustration, we consider the system with $M = K = 2$ and $N \geq 2$ with $\mu \in [1/2, 1]$. For the case of perfect CSI the optimal NDT can be characterized as in Fig. 1(b). Next we look at the achievable NDT for the case of delayed and no CSI respectively.

a) Delayed CSI: For the case of delayed CSI, consider the corner point $\mu = 1/2$ where the system behaves like a 2×2 X-Channel. It is known for the 2×2 X-Channel with delayed CSI that a sum-DoF of $6/5$ is achievable [13]. As a result, an NDT of $\delta(\mu) = 5/3$ is achievable. Compared to the perfect CSI case, the NDT thus incurs a loss due to delayed CSI. Next, consider the corner point $\mu = 1$, where the system reduces to a 2×2 broadcast channel with delayed CSI. It is known that for such a system, a sum-DoF of $4/3$ is achievable [14], i.e., a NDT of $\delta(\mu) = 3/2$ is achievable. The optimality of this trade-off is, again, an open problem. However, the achievability illustrates the loss incurred due to delay in CSI availability.

b) No CSI: In case of no CSI, it is known that the optimal scheme is transmit using time-division to each user in a separate slot [15]. Therefore a sum-DoF of 1 can be achieved, i.e., an NDT of 2 can be achieved which is optimal for all values of $\mu \in [1/2, 1]$. The NDT trade-offs for perfect, delayed and no CSI are shown in Fig. 3.

IV. Conclusions

In this paper, we studied the fundamental information-theoretic limits of cache-aided wireless networks where network edge nodes are endowed with cache storage. We first proposed an information-theoretic model for such a network and we introduced the metric of normalized delivery time (NDT), which captures the worst-case latency in delivering file requests to users. We presented the first known information theoretic lower bounds for a general $M \times K$ cache-aided wireless networks with perfect CSI. Based on this result, we showed that the optimal NDT for some system parameters can be characterized by the use of known transmission schemes such as interference alignment and zero-forcing beamforming. Finally, we also demonstrated the effect of imperfect (delayed or no) CSI at the ENs and users on the NDT for cache-aided wireless systems.

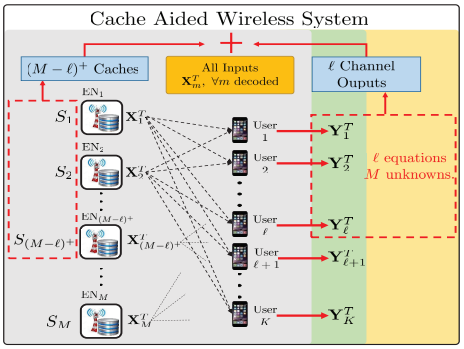

Fig. 4. Edge-caching set-up for the proof of Theorem 1.

Appendix A
Proof of Theorem 1

To obtain a lower bound on the NDT, we fix a specific request vector $\mathbf{D}$, namely one for which all requested files $(F_1, ..., F_K) = F_{[1:K]}$ are different and a given channel realization $\mathbf{H}$. Note that this is possible given the assumption $N \geq K$. For any integer a and b with $a \leq b$, we define the notation $[a : b] = (a, a+1, \ldots, b)$. We denote as T the time $T^{(\mathbf{D},\mathbf{H})}$ as per Definition 1 of any given feasible policy $\pi = (\pi_c, \pi_e, \pi_d)$ which guarantees a vanishing probability of error P_e as $L \to \infty$ for the given request $\mathbf{D}$ and channel $\mathbf{H}$. Our goal is to lower bound T in order to obtain a lower bound on the minimum NDT $\delta^*(\mu)$. To this end, consider the channel output in (5), where $\mathbf{Y}_k^T, \mathbf{X}_m^T$ and $\mathbf{n}_k^T$ are $1 \times T$ row vectors.

To obtain the lower bound on NDT, we make the following key observation, which is illustrated in Fig. 4. Given any set of $\ell \leq \min\{M, K\}$ output signals $\mathbf{Y}_k^T$, say $\mathbf{Y}_{[1:\ell]}^T$, and the content of any $(M-\ell)^+$ caches, say $S_{[1:(M-\ell)^+]}$, all transmitted signals $\mathbf{X}_{[1:M]}^T$, and hence also all the files $F_{[1:K]}$, can be resolved in the high-SNR regime. This is because: (*i*) from the cache contents $S_{[1:(M-\ell)^+]}$ one can reconstruct the corresponding inputs $\mathbf{X}_{[1:(M-\ell)^+]}^T$; (*ii*) neglecting the noise in the high-SNR regime, the relationship between the variables $\mathbf{Y}_{[1:\ell]}^T$ and the remaining inputs $\mathbf{X}_{[(M-\ell)^+:M]}^T$ is given almost surely by an invertible linear system as in (5). We use this argument in the following:

$$\begin{aligned} KL &= H\left(F_{[1:K]}\right) \overset{\text{(a)}}{=} H\left(F_{[1:K]}|F_{[K+1:N]}\right) \\ &= I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^T, S_{[1:(M-\ell)^+]}|F_{[K+1:N]}\right) \\ &\quad + H\left(F_{[1:K]}|\mathbf{Y}_{[1:\ell]}^T, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right) \end{aligned} \quad (28)$$

where (a) follows from the fact that all files are independent of each other. The first term in (28) can be upper bounded as follows:

$$\begin{aligned} &I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^T, S_{[1:(M-\ell)^+]}|F_{[K+1:N]}\right) \\ &= I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^T|F_{[K+1:N]}\right) \\ &\quad + I\left(F_{[1:K]}; S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^T, F_{[K+1:N]}\right) \\ &\overset{\text{(a)}}{\leq} I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^T|F_{[K+1:N]}\right) + H\left(F_{[1:\ell]}|\mathbf{Y}_{[1:\ell]}^T\right) \\ &\quad + H\left(S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^T, F_{[1:\ell]\cup[K+1:N]}\right) \\ &\quad - H\left(S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^T, F_{[1:N]}\right) \end{aligned}$$

$$\overset{(b)}{\leq} h\left(\mathbf{Y}^T_{[1:\ell]}\right) + L\epsilon_L + H\left(S_{[1:(M-\ell)^+]}|F_{[1:\ell]\cup[K+1:N]}\right) - H\left(S_{[1:(M-\ell)^+]}|\mathbf{Y}^T_{[1:\ell]}, F_{[1:N]}\right) - h\left(\mathbf{Y}^T_{[1:\ell]}|F_{[1:N]}\right)$$

$$\overset{(c)}{\leq} \ell T \log\Big(2\pi e\big(\Lambda P + 1\big)\Big) - h\left(\mathbf{n}^T_{[1:\ell]}|F_{[1:N]}\right) + \sum_{i=1}^{(M-\ell)^+} H\Big(S_{i,[1:N]}|F_{[1:\ell]}, F_{[K+1:N]}\Big) + L\epsilon_L$$

$$\overset{(d)}{\leq} \ell T \log\big(\Lambda P + 1\big) + (M-\ell)^+(K-\ell)^+\mu L + L\epsilon_L, \quad (29)$$

where, the steps in (29) are explained as follows:

- Step (a) follows from careful expansion of the second term in the previous step and that conditioning reduces entropy.
- Step (b) follows from the fact that $\mathbf{Y}^T_{[1:\ell]}$ are continuous random variables and that dropping the conditioning in the first term increases entropy. We apply Fano's inequality to the second term where ϵ_L is a function, independent of P, which vanishes as $L \to \infty$.
- Step (c) can be explained as follows. The first term is upper bounded by the use of Lemma 1 stated below. The parameter Λ is a constant dependent only on the channel parameters and is defined in Lemma 1. The third term is zero since the cache contents $S_{[1:(M-\ell)^+]}$ are functions of the files $F_{[1:N]}$. Moreover, given all the files, the channel outputs are a function of the channel noise at each receiver.
- Step (d) follows from the fact that the channel noise is i.i.d. across time and distributed as $\mathcal{N}(0,1)$.

Next, considering the set-up in Fig. 4, the second term in (28) can be upper bounded as follows:

$$H\left(F_{[1:K]}|\mathbf{Y}^T_{[1:\ell]}, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right) \leq L\epsilon_L + Tc, \quad (30)$$

where ϵ_L is a function, independent of P and vanishes as $L \to \infty$. Further, c is a constant term independent of signal power P and file size L and is dependent only on the noise variance and the channel coefficients. The proof of (30) is omitted due to lack of space and provided in detail in [16]. Substituting (29) and (30) into (28), we have

$$KL \leq \ell T \log\left(\Lambda P + 1\right) + (M-\ell)^+(K-\ell)^+\mu L + L\epsilon_L + Tc$$
$$= \ell T \log(P)\left[1 + \frac{\log(\Lambda + \frac{1}{P}) + \frac{c}{\ell}}{\log(P)}\right] + (M-\ell)^+(K-\ell)^+\mu L + L\epsilon_L \quad (31)$$

Rearranging (31), we get the following

$$\frac{T\log(P)}{L}\left[1 + \frac{\log(\Lambda + \frac{1}{P}) + \frac{c}{\ell}}{\log(P)}\right] \geq \frac{K - (M-\ell)^+(K-\ell)^+\mu - \epsilon_L}{\ell}. \quad (32)$$

Now, using (32), we first take the limit of $L \to \infty$ such that $\epsilon_L \to 0$ as $P_e \to 0$. Further, taking the limit $P \to \infty$, for the high-SNR regime, we have

$$\delta^*(\mu) \geq \lim_{\substack{P\to\infty \\ L\to\infty}} \frac{T\log P}{L} \geq \frac{K - (M-\ell)^+(K-\ell)^+\mu}{\ell}, \quad (33)$$

where the term $\frac{\log(\Lambda+\frac{1}{P})+\frac{c}{\ell}}{\log(P)}$ vanishes under the limit $P \to \infty$. Optimizing the bound in (33) over all possible choices of $\ell = 1, 2, \ldots, \min\{M, K\}$ completes the proof of Theorem 1.

Lemma 1. *For the cache-aided wireless system under consideration, the differential entropy of any ℓ channel outputs $\mathbf{Y}^T_{[1:\ell]}$ can be upper bounded as*

$$h\left(\mathbf{Y}^T_{[1:\ell]}\right) \leq \ell T \log\Big(2\pi e\big(\Lambda P + 1\big)\Big), \quad (34)$$

where the parameter Λ is a function of the channel coefficients in $\mathbf{H}$ and is defined as

$$\Lambda = \left(\max_{k\in\{1,\ldots,\ell\}}\left[\sum_{m=1}^{M} h_{km}^2 + \sum_{m\neq\tilde{m}} h_{km}h_{k\tilde{m}}\right]\right).$$

The proof of Lemma 1 is omitted due to lack of space and is provided in full detail in [16].

References

[1] F. Boccardi, R. W. Heath Jr., A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

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

[3] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59(12), pp. 8402–8413, Dec. 2013.

[4] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *International Symposium on Wireless Communication Systems*, Aug 2014, pp. 917–921.

[5] A. Sengupta, R. Tandon, and T. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *IEEE International Symposium on Information Theory*, June 2015, pp. 1691–1695.

[6] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 355–370, Feb 2015.

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

[8] ——, "Cache-aided interference channels," *arXiv: 1510.06121*, Oct 2015. [Online]. Available: http://arxiv.org/abs/1510.06121

[9] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in Fog radio access Networks," *To be submitted to IEEE International Symposium on Information Theory*, Dec 2015.

[10] V. R. Cadambe and S. A. Jafar, "Interference alignment and the degrees of freedom of wireless X-networks," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 3893–3908, Sept 2009.

[11] A. S. Motahari, S. Oveis-Gharan, M. A. Maddah-Ali, and A. Khandani, "Real interference alignment: Exploiting the potential of single antenna systems," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4799–4810, Aug 2014.

[12] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept 2006.

[13] A. Ghasemi, A. S. Motahari, and A. K. Khandani, "On the degrees of freedom of X-channel with delayed CSIT," in *IEEE International Symposium on Information Theory Proceedings*, July 2011, pp. 767–770.

[14] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4418–4431, July 2012.

[15] C. S. Vaze and M. K. Varanasi, "The degree-of-freedom regions of MIMO broadcast, interference, and cognitive radio channels with no CSIT," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5354–5374, Aug 2012.

[16] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," *arXiv:1512.07856*, Dec 2015. [Online]. Available: http://arxiv.org/pdf/1512.07856v1.pdf