Latency Minimization for Content Delivery Networks with Wireless Edge Caching

Thang X. Vu, Lei Lei, Satyanarayana Vuppala, Ashkan Kalantari, Symeon Chatzinotas, and Bjorn Ottersten
The Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg
Email: {thang.vu, lei.lei, satyanarayana.vuppala, ashkan.kalantari, symeon.chatzinotas, bjorn.ottersten}@uni.lu

***Abstract*—Edge-caching has received much attention as an efficient technique to reduce delivery latency and network congestion during peak-traffic times by bringing data closer to end users. In this paper, we investigate the latency performance of content delivery networks with the aid of edge-caching, in which a data centre is serving the users via a shared wireless medium. Firstly, we derive a cache placement design which minimizes the average (buffering) latency during the delivery phase. It is found that the derived placement solution differs from the conventional placement method for throughput minimization. Secondly, for a given cache placement scheme, we optimize the signal transmission in the delivery phase taking into consideration the cached content to minimize the average user latency. Particularly, two optimization problems based on zero-forcing (ZF) and minimum mean square error (MMSE) designs are formulated subject to requesting rate and transmit power constraints. To deal with the non-convexity of the MMSE problem, an iterative algorithm is proposed that approximates the non-convex constraint by its first-order approximation. Finally, numerical results are presented to demonstrate the effectiveness of the proposed designs.**

***Index terms*—** Content delivery networks, latency, edge caching, optimization.

I. Introduction

One of the challenges that future wireless networks will have to tackle with is to support not only the massive increase in network traffics due to the proliferation of mobile devices and data-hungry applications, but also the stringent quality of experience (QoE) requirements. Despite potential high rate of the new network architectures such as heterogeneous networks (HetNets) and cloud radio access networks (CRAN) [1], traffic congestion might occur during peak-traffic times, resulting in poor user QoE. Reducing content delivery latency and hence improving the user QoE is one of the most demanding objectives of the future content delivery networks (CDN). A promising solution to leverage network costs of content delivery is to prefetch the content in distributed storages through out the network, which is known as caching [2]. Since being closer to users, a requested content can be sent directly from the distributed caches without backhaul cost. In this manner, caching allows significant throughput reduction during peak-traffic times and thus reduces network congestion [2], [3].

The research on caching in wireless networks has received much attention recently [4–10]. The basic principle is to consider the caching capability in the signal transmission design in order to optimize the system resources. It is shown in [4] that a cache-aware multicast beamforming design can effectively reduce transmitted power and fronthaul bandwidth in cache-assisted wireless networks. By jointly optimizing routing, channel assignment and caching, the authors in [5] demonstrate the benefit of caching via two sub-problems called restricted master and pricing. The performance of caching wireless device-to-device (D2D) networks are analysed in [6], [7]. In [6], a joint content replacement and delivering scheme is proposed for D2D networks. The throughput-outage tradeoff of the mmWave link underlying D2D networks under a simplified grid topology is derived in [7] for various caching policies. The stochastic performance of cache-assisted wireless networks has recently been studied. Under the Poison point process (PPP) distribution assumption of nodes' location, the authors in [10] derive the ergodic rate and outage performance of the downlink cache-enabled HetNets. In [11], the delivery rate is studied for cluster-centric networks in which small BSs are grouped into disjoint clusters in a hexagon network topology. In this work, the users within a cluster share one common cache which is divided into two parts: one contains the most popular files, and one contains different files which are most locally popular. In [12], a low-complexity greedy algorithm is proposed to minimize the content delivering delay in cooperative caching cloud radio access networks. In [13], a joint design of cloud and edge processing is proposed in the fog radio access network (FRAN) downlink to minimize the delivery times via hybrid hard-soft transfering between the so-called enhanced remote radio head (eRRH) and the data center taking into consideration the cached symbols at the eRRH. In [14], the authors study the tradeoff between the memory at edge nodes and the transmission latency via a novel performance metric named normalized delivery time (NDT). In particular, the authors characterize the information-theoretically optimality on the tradeoff between NDT and caching capacity in the FRAN framework. The authors in [15] derive an achievable upper bound and a theoretical lower-bound for NDT in general interference networks where both transmitter and receiver sides have caching capability.

In this paper, we investigate the benefit of edge caching to improve user QoE in cache-assisted CDN via the signal transmission design taking into account the cached content. Unlike the above works [12–16] which minimize the transmission time, we focus on the *buffering time* - a gap between the moment a user requests a content (e.g., a movie) and when it can start the service (e.g., watching the movie). The motivation behind our work is that in many CDN applications

This work is supported in part by the ERC AGNOSTIC project (R-AGR-3283), the Luxembourg FNR CORE ProCAST project, and the FNR CORE project ROSETTA (11632107).

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

(videos), the users do not have to wait till the complete delivery of the requested content to begin the service. From the user perspective, buffering time latency is one of the most important QoE metrics. Firstly, we propose a cache placement design minimizing the average latency, which appears to be different from the conventional placement scheme that is designed for throughput minimization. Secondly, two optimization problems are formulated corresponding to zero-forcing (ZF) and minimum mean square error (MMSE) based designs to minimize the average (buffering) latency in the delivery phase subjected to some given request rate and limited transmit power constraints. To overcome the non-convexity of the MMSE design, we propose an iterative algorithm that approximates the non-convex constraint by its first-order approximation. Finally, the proposed designs are demonstrated via numerical results.

The rest of this paper is organised as follows. Section II presents the system model. Section III presents the optimal cache placement design for latency minimization. Section IV optimizes the delivery phase transmission. Section V shows numerical results. Finally, Section VI concludes the paper.

II. System Model

We consider a CDN in which a L-antenna data centre (DC) serving K single-antenna users with $K \leq L$ via the shared wireless medium, as depicted in Figure 1. The DC contains N files of equal size of Q bits (in practice, unequal file size can be divided into trunks of subfiles which have the same size) and is denoted by $\mathcal{F} = \{F_1, \ldots, F_N\}$ the library.

A. Content popularity distribution and caching model

In most practical cases, the content popularity does not a follow uniform distribution. In fact, there are always some files which are more frequently requested than the others. In this paper, we consider practical scenarios in which the content popularity follows a Zipf distribution [17]. The probability of the i-th file being requested from user k is given as

$$q_k(i) = \frac{i^{-\alpha_k}}{\sum_{n=1}^{N} n^{-\alpha_k}}, \tag{1}$$

where α_k is the skewness factor of the Zipf distribution related to the k-th user.

We consider uncoded caching and general cache memories in which the user caches' size can be arbitrary. For convenience, let M_k denote the storage memory (in files) at the k-th user. Parts of the contents are prefetched at the user caches during the placement phase, which occurs during off-peak time [2].

B. Transmission model

During the delivery phase, each user requests a content file from the DC. If (parts of) the requested file is available in the user's cache, it can be served immediately. Otherwise, these file parts will be sent from the DC. Denote $F_{d_1}, \ldots, F_{d_K}$ as the requested files from user $1, \ldots, K$, respectively, and $\bar{F}_{d_1}, \ldots, \bar{F}_{d_K}$ as parts of the requested files which are not at the user cache. First, the DC modulates $\bar{F}_{f_k}$ in to the modulated signal x_k and then sends the precoded signal through the access channels. The received signal at user k is given as

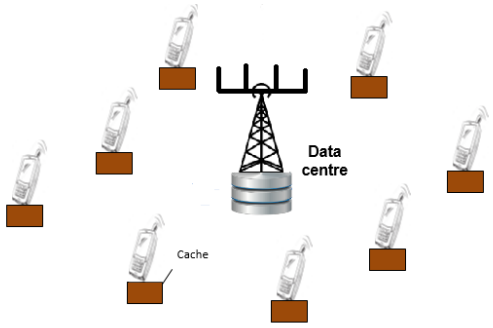

Fig. 1: Content delivery networks with the aid of edge caching.

$$y_k = \mathbf{h}_k^H \mathbf{w}_k x_k + \sum_{l \neq k} \mathbf{h}_k^H \mathbf{w}_l x_l + n_k, \tag{2}$$

where $\mathbf{h}_k \in \mathbb{C}^{L \times 1}$ denote the channel vector from the DC antennas to user k, which follows a circular-symmetric complex Gaussian distribution $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \sigma_{h_k}^2 \mathbf{I}_K)$, $\sigma_{h_k}^2$ is the parameter accounting for the path loss from the DC antennas to user k, $\mathbf{w}_k \in \mathbb{C}^{L \times 1}$ is the precoding vector for user k, n_k is Gaussian noise with zero mean and variance σ^2, and $L \geq K$ is the number of DC's antennas. We consider block fading channels where the channel gains are fixed within a block and independently change across the blocks. The block duration is assumed to be long enough for the users to be served the requested files [4]. Perfect channel state information (CSI) is assumed to be known at the DC. In practice, robust channel estimation can be achieved through the transmission of pilot sequences.

The first term in (2) is the desired signal, and the second term is the inter-user interference. The signal-to-interference-plus-noise ratio at user k is $\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{l \neq k} |\mathbf{h}_k^H \mathbf{w}_l|^2 + \sigma^2}$. The achievable information rate of user k is

$$R_k = B \log_2 \left(1 + \mathrm{SINR}_k\right), 1 \leq k \leq K, \tag{3}$$

where B is the channel bandwidth.

III. Cache placement for latency minimization

The conventional method designs the cache placement to minimize the aggregated throughput on the shared link during the delivery phase, which is the best choice from the content provider's perspectives to minimize operating costs. However, this strategy might not be preferred from the users' perspectives since the users's priority is usually different from that of the content provider. In this section, we design the placement phase in order to minimize buffering time (latency), hence maximizing the user QoE. We first review the placement phase design to minimize the total throughput, and then the proposed design.

A. Content placement design to minimize total throughput

Let $\mathbf{c}_k = \{c_{k,1}, \ldots, c_{k,N}\}, k = 1, \ldots, K$ denote the caching vector at user k, where $c_{k,n}, 0 \leq c_{k,n} \leq 1$ denotes

a (random) fraction of file F_n to be stored at user k' cache. When a user requests file n from the user k, the cost for the backhaul to serve this request is $(1-c_{k,n})Q$ bits. Since the probability of having the file n requested is $q_k(n)$, the average total throughput cost is $\sum_{k=1}^{K}\sum_{n=1}^{N} Q(1-c_{k,n})q_k(n)$. The placement phase optimization problem is formulated as

$$\begin{aligned} \underset{\{c_{k,n}\}_{\forall k,n}}{\texttt{Minimize}} \quad & \sum_{k=1}^{K}\sum_{n=1}^{N} Q(1-c_{k,n})q_k(n) \\ \texttt{s.t.} \quad & \sum_{n=1}^{N} c_{k,n} \le M_k, \forall k,n;\ 0 \le c_{k,n} \le 1, \forall k,n. \end{aligned} \tag{4}$$

Since $\{q_k(n)\}_{n=1}^{N}$ is a decreasingly ordered sequence by Zipf distribution, the linear programming problem (4) has the optimum solution $c_{k,n}^{\star}=1$ if $n \le M_k$ and 0 otherwise. The optimum cache placement design states that the user k should store the whole M_k files which are most popular.

B. Content placement design to minimize latency

One key QoE requirement in CDN is to provide the users a smooth experience of the requested content, e.g., the user does not want to wait while he has already started watching a movie. However, if the request rate is greater than the serving rate, the user has to wait to buffer some content before begin watching the movie to avoid any interruption during his (watching) session.

Denote r_k as the average rate that the DC can support user k, and η_k as the average request rate of user k. A buffering latency occurs if $r_k < \eta_k$, which usually happens in practice when the users request data-hungry services, e.g., high-definition video, and the DC serves a large number of users via a shared medium. Denote $t_n \ge 0$ as the buffering time (latency) when the user request file n. Note that $c_{k,n}Q$ bits of file n are available in the user cache. If $c_{k,n} > 1-\frac{r_k}{\eta_k}$, the user can be served immediately without any interruption during the service. Otherwise, the user waits $Q(\frac{1-c_{k,n}}{r_k}-\frac{1}{\eta_k})$ seconds. Combining the two cases, the buffering time of requesting file n is given as $t_n = \Big(Q(\frac{1-c_{k,n}}{r_k}-\frac{1}{\eta_k})\Big)^{+}$, where $(x)^{+} \triangleq \max(x,0)$.

We want to design the placement phase that minimize the average buffering time. The optimization problem is formulated as follows:

$$\begin{aligned} \underset{\{c_{k,n}\}_{\forall n,k}}{\texttt{Minimize}} \quad & \sum_{k=1}^{K}\sum_{n=1}^{N} \Big(Q\Big(\frac{1}{r_k}-\frac{1}{\eta_k}-\frac{c_{k,n}}{r_k}\Big)\Big)^{+} q_k(n) \\ \texttt{s.t.} \quad & \sum_{n=1}^{N} c_{k,n} \le M_k, \forall k;\ 0 \le c_{k,n} \le 1, \forall k,n, \end{aligned} \tag{5}$$

where the first constraint is due to the memory limit.

Because the caching at the users is independent, the problem (5) can be decoupled into K parallel subproblems:

$$\begin{aligned} \underset{\{c_{k,n}\}_{n=1}^{N}}{\texttt{Minimize}} \quad & \sum_{n=1}^{N} \Big(Q\Big(\frac{1}{r_k}-\frac{1}{\eta_k}-\frac{c_{k,n}}{r_k}\Big)\Big)^{+} q_k(n) \\ \texttt{s.t.} \quad & \sum_{n=1}^{N} c_{k,n} \le M_k;\ 0 \le c_{k,n} \le 1, \forall n. \end{aligned} \tag{6}$$

Lemma 1: The optimum solution of problem (6) satisfies $c_{k,n}^{\star} \le 1-\frac{r_k}{\eta_k}, \forall n$.

Proof: Let $\{c_{k,n}^{\star}\}_{n=1}^{N}$ denote the optimal solution of (6). We will show that under limited cache capacity, i.e., $M_k < N(1-\frac{r_k}{\eta_k})$, $c_{k,n}^{\star} \le 1-\frac{r_k}{\eta_k}, \forall n$. Assuming that there is at least one caching factor in the optimum solution greater than $1-\frac{r_k}{\eta_k}$. Without loss of generality, let $c_1^{\star} > 1-\frac{r_k}{\eta_k}$. Since $M_k < N(1-\frac{r_k}{\eta_k})$, there is at least one caching factor that strictly less than $1-\frac{r_k}{\eta_k}$, denoted by $c_2^{\star}$. There always exist $\tilde{c}_1^{\star}, \tilde{c}_2^{\star}$ such as $c_1^{\star} > \tilde{c}_1^{\star} \le 1-\frac{r_k}{\eta_k}$, $\tilde{c}_2^{\star} > c_2^{\star}$ and $c_1^{\star}+c_2^{\star} = \tilde{c}_1^{\star}+\tilde{c}_2^{\star}$. Consider a new candidate $\tilde{C} \triangleq \{\tilde{c}_1^{\star}, \tilde{c}_2^{\star}, c_3^{\star}, \ldots, c_{k,n}^{\star}\}$ which is a feasible solution of (6) because it satisfies the constraint. However, $\tilde{C}$ results in a objective value smaller than the optimal value of $\{c_{k,n}^{\star}\}_{n=1}^{N}$ since $\tilde{t}_1 = t_1, \tilde{t}_2 < t_2$ and $\tilde{t}_n = t_n, \forall n > 2$, which is is in contrast to the optimum assumption of $\{c_{k,n}^{\star}\}_{n=1}^{N}$. ■

By using Lemma 1, the problem (6) is reformulated as

$$\begin{aligned} \underset{\{c_{k,n}: c_{k,n} \ge 0\}_{n=1}^{N}}{\texttt{Minimize}} \quad & \sum_{n=1}^{N} Q\Big(\frac{1}{r_k}-\frac{1}{\eta_k}-\frac{c_{k,n}}{r_k}\Big) q_k(n) \\ \texttt{s.t.} \quad & \sum_{n=1}^{N} c_{k,n} \le M_k;\ c_{k,n} \le 1-\frac{r_k}{\eta_k}, \forall n. \end{aligned} \tag{7}$$

It is observed that (7) is a linear programming with respect to $\{c_{k,n}\}_{n=1}^{N}$. By taking into account the fact that $\{q_k(n)\}_{n=1}^{N}$ is a decreasing sequence, we can show that the optimum solution of (7) has a form:

$$c_{k,n}^{\star} = \begin{cases} 1-\frac{r_k}{\eta_k}, & \text{if } 1 \le n \le \lfloor \frac{M_k}{1-r_k/\eta_k} \rfloor \\ 0, & \text{if } \lfloor \frac{M_k}{1-r_k/\eta_k} \rfloor < n \le N \end{cases}, \tag{8}$$

where $\lfloor x \rfloor$ denotes the largest integer not exceeding x.

The optimum solution in (8) suggests that it is not necessary to prefetch the whole file for maximizing the average latency. The closer to the request rate the serving rate is, the less parts of a file needs to be cached. This result differs from the solution of (4) since the two methods target different objectives.

IV. Minimization of latency under power constraint

In this section, we design the signal transmission during the delivery phase in order to minimize the total latency. This usually happens when the users' request rates are larger than the channel capacity. It is assumed that the requested content's size is relatively large, such that all users are active during the buffering time. Denote d_k as the file index requested by user k. From Section III.B, we have the latency observed by user k is $Q\Big(\frac{1-c_{d_k}}{R_k}-\frac{1}{\eta_k}\Big)^{+}$, where η_k is the requested rate by user k, c_{d_k} is the fraction of file d_k cached at the user cache, and $R_k = B\log_2\Big(1+\frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum_{l\ne k}|\mathbf{h}_k^H\mathbf{w}_l|^2+\sigma^2}\Big)$.

We aim at minimizing the average latency via beamforming vector design, as follows:

$$\underset{\{\mathbf{w}_k \in \mathbb{C}^L\}_{k=1}^{K}}{\texttt{Minimize}} \quad \frac{Q}{K}\sum_{k=1}^{K}\Big(\frac{1-c_{d_k}}{R_k}-\frac{1}{\eta_k}\Big)^{+} \tag{9}$$

$$\texttt{s.t.} \quad \sum_{k=1}^{K} \| \mathbf{w}_k \|^2 \leq P_{tot},$$

where P_{tot} is the total transmit power. In the following, we will optimize problem (9) based on the two linear beamforming designs: Zero-forcing and MMSE.

A. ZF based design

We first minimize the average latency under the standard ZF beamformer because of its low computational complexity. Let $p_1, p_2, \ldots, p_K$ denote the transmit power dedicated for user k. The precoding vector for user k is given as $\mathbf{w}_k = \sqrt{p_k}\tilde{\mathbf{h}}_k$, where $\tilde{\mathbf{h}}_k$ is the ZF beamforming vector for user k, which is the k-th column of $\mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$, with $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]^T$. By considering the fact that $\mathbf{h}_k^H \tilde{\mathbf{h}}_l = \delta_{kl}$ and denoting $\bar{\eta}_k = \frac{\eta_k}{B}$, we have the resulting optimization problem as

$$\begin{aligned} \underset{\{p_k : p_k \geq 0\}_{k=1}^K}{\texttt{Minimize}} \quad & \frac{Q}{KB} \sum_{k=1}^{K} \Big(\frac{1 - c_{d_k}}{\log_2(1 + \frac{p_k}{\sigma^2})} - \frac{1}{\bar{\eta}_k} \Big)^+ \qquad (10) \\ \texttt{s.t.} \quad & \sum_{k=1}^{K} p_k \| \tilde{\mathbf{h}}_k \|^2 \leq P_{tot}. \end{aligned}$$

Lemma 2: Problem (10) is convex and therefore solvable in polynomial time.

Proof: We will show that the objective function of (10) is convex. Consider function $f(x) = 1/\log_2(1 + ax)$ in $\mathbb{R}^+$ with $a > 0$. The second-order derivative of $f(x)$ is given as

$$f'(x) = -\frac{a}{\log_2(1+ax)(1+ax)},$$

$$f''(x) = \frac{a^2}{\log_2^2(1+ax)(1+ax)^2} + \frac{a^2}{\log_2(1+ax)(1+ax)^2}.$$

It is verified that the second-order derivative is always positive, thus the objective function is convex in its support. In addition, the function $(x)^+ = \max(x, 0)$ is also convex. Therefore, the objective function of (10) is convex. ∎

B. MMSE based design

Under MMSE precoding, the beamformer vector is of the form $\mathbf{w}_k = \sqrt{p_k}\breve{\mathbf{h}}_k$, where $\breve{\mathbf{h}}_k$ is the k-th column of $\mathbf{H}^H(\sigma^2\mathbf{I} + \mathbf{H}\mathbf{H}^H)^{-1}$. Denote $\alpha_{k,l} = |\mathbf{h}_k^H \breve{\mathbf{h}}_l|^2, \forall k, l$ as the interference factor caused to user k from user l' beamforming vector . The corresponding optimization problem is formulated as follows:

$$\begin{aligned} \underset{\{p_k : p_k \geq 0\}_{k=1}^K}{\texttt{Minimize}} \quad & \frac{Q}{KB} \sum_{k=1}^{K} \Big(\frac{1 - c_{d_k}}{\log_2\Big(1 + \frac{p_k a_{kk}}{\sum_{l \neq k} p_l a_{kl} + \sigma^2}\Big)} - \frac{1}{\bar{\eta}_k} \Big)^+ \qquad (11) \\ \texttt{s.t.} \quad & \sum_{k=1}^{K} p_k a_{kk} \leq P_{tot}. \end{aligned}$$

By introducing arbitrary positive variables $\{x_k\}_{k=1}^K$, we reformulate the above problem as

$$\begin{aligned} \underset{\{x_k, p_k : \geq 0\}_{k=1}^K}{\texttt{Minimize}} \quad & \frac{Q}{KB} \sum_{k=1}^{K} \Big(\frac{1 - c_{d_k}}{x_k} - \frac{1}{\bar{\eta}_k} \Big)^+ \qquad (12) \\ \texttt{s.t.} \quad & \log_2\Big(\frac{\sum_{l=1}^K p_l a_{kl} + \sigma^2}{\sum_{l \neq k} p_l a_{kl} + \sigma^2}\Big) \geq x_k, \forall x \qquad (12a) \\ & \sum_{k=1}^{K} p_k a_{kk} \leq P_{tot}. \qquad (12b) \end{aligned}$$

For convenience, let's denote parameters $A_k = [\sigma^2, \beta_{k,1}, \ldots, \beta_{k,K}]$, $B_k = [\sigma^2, \beta_{k,1}, \ldots, \beta_{k,k-1}, 0, \beta_{k,k+1}, \ldots, \beta_{k,K}]$ and introduce new variables $\mathbf{p} = [1, p_1, \ldots, p_K]^T$ and $\{y_k\}_{k=1}^K$. By decomposing the left-hand-side of (12a) into the subtraction of the two logarithms, the problem (12) is equivalent to

$$\begin{aligned} \underset{x_k, p_k, y_k}{\texttt{Minimize}} \quad & \frac{Q}{KB} \sum_{k=1}^{K} \Big(\frac{1 - c_{d_k}}{x_k} - \frac{1}{\bar{\eta}_k} \Big)^+ \qquad (13) \\ \texttt{s.t.} \quad & \ln(A_k\mathbf{p}) \geq \ln(2) x_k + y_k, \forall k \qquad (13a) \\ & B_k\mathbf{p} \leq e^{y_k}, \forall k \qquad (13b) \\ & A_k\mathbf{p} \leq P_{tot}, \qquad (13c) \end{aligned}$$

where the constraint (13b) is an alternative form of $\ln(B_k\mathbf{p}) \leq y_k$.

Although both constraints (13a) and (13c) of the above problem are convex, solving problem 13 is still challenging since the second constraint is unbounded. To deal with this, we resort to approximate this constraint by its linear approximation as $e^{y_k} \simeq e^{\bar{y}_k}(y_k - \bar{y}_k + 1)$, where $\bar{y}_k$ is any accessible point. The resorted problem is given as

$$\begin{aligned} \underset{\mathbf{p}, \{x_k, y_k\}_{k=1}^K}{\texttt{Minimize}} \quad & \frac{Q}{KB} \sum_{k=1}^{K} \Big(\frac{1 - c_{d_k}}{x_k} - \frac{1}{\bar{\eta}_k} \Big)^+ \qquad (14) \\ \texttt{s.t.} \quad & (13a) \text{ and } (13c) \\ & B_k\mathbf{p} \leq e^{\bar{y}_k}(y_k - \bar{y}_k + 1), \forall k. \end{aligned}$$

It is observed that problem (14) is a convex since the objective function and the constraints are convex, and hence can be solved in an efficient manner by standard solvers, e.g., CVX. Because $e^{\bar{y}_k}(y_k - \bar{y}_k + 1) \leq e^{y_k}, \forall \bar{y}_k$, the resorted problem (14) gives a suboptimal solution of problem (13).

It is important to note that the optimal solution of problem (14) heavily depends on parameters $\{\bar{y}_k\}_{k=1}^K$. This raises a question that how to choose the values $\{\bar{y}_k\}_{k=1}^K$ such that the solution of (14) is as close as to the optimal solution of (13). To overcome this problem, we propose an iterative algorithm to improve the performance of problem (14), whose steps are listed in Table I. The principle of the proposed iterative algorithm is to better estimate values $\{\bar{y}_k\}_{k=1}^K$ through iterations.

Proposition 1: The objective function of problem $\boldsymbol{P0}(\mathbf{a} \triangleq \{a_k\}_{k=1}^K)$ in (15) solved by the iterative algorithm in Table I decreases by interations.

Proof: Let $\left(\mathbf{p}_\star^{(t)}, \mathbf{x}_\star^{(t)}, \mathbf{y}_\star^{(t)}\right)$ be the optimal solution of $\mathbf{P0}(\mathbf{a}^{(t)})$ at the t-th iteration. The optimal objective function after iteration t is $L^{(t)} = \frac{Q}{KB}\sum_{k=1}^K \left(\frac{1-c_{d_k}}{x_{\star k}^{(t)}} - \frac{1}{\bar{\eta}_k}\right)^+$. We will show that if $y_{\star k}^{(t)} < a_k^{(t)}, \forall k$, then by using $a_k^{(t+1)} = y_{\star k}^{(t)}$ in the $(t+1)$-th iteration, we will have $L^{(t+1)} < L^{(t)}$. Indeed, by choosing a relatively large initial value $a_k^{(1)}$, we always have $y_{\star k}^{(1)} < a_k^{(1)}, \forall k$.

At the $(t+1)$-th iteration, $f(y; y_{\star k}^{(t)})$ is used in the right-hand side of constraint (15b) instead of $f(y; a_k^{(t)})$, where $f(x;a) = e^a(x+1-a)$ is the first-order approximation of function e^x at a. Consider a candidate $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_K\}$, with $\tilde{y}_k = y_{\star k}^{(t)} - 1 + e^{a_k^{(t)} - y_{\star k}^{(t)}}(y_{\star k}^{(t)} - a_k^{(t)} + 1)$. It is straightforward to verify that $\tilde{y}_k < y_{\star k}^{(t)}$ and $f(\tilde{y}_k; y_{\star k}^{(t)}) = f(y_{\star k}^{(t)}; a_k^{(t)}), \forall k$.

Since $\tilde{y}_k < y_{\star k}^{(t)}, \forall k$, the strictly inequality holds in constraint (15a). Thus, there exits $\tilde{x}_k < x_{\star k}^{(t)}$ which satisfies $\ln(A_k\mathbf{p}) \geq \ln(2)\tilde{x}_k + \tilde{y}_k$. Now consider a new candidate set $\left(\mathbf{p}_\star^{(t)}, \tilde{\mathbf{x}}, \mathbf{a}^{(t+1)} = \mathbf{y}_\star^{(t)}\right)$. This set satisfies all the constraints of problem $\mathbf{P0}(\mathbf{y}_\star^{(t)})$, and therefore is a feasible solution of the optimization problem. Therefore, the objective function at the $t+1$-th iteration is $L^{(t+1)} \leq \frac{Q}{KB}\sum_{k=1}^K \left(\frac{1-c_{d_k}}{\tilde{x}_k} - \frac{1}{\bar{\eta}_k}\right)^+ < \frac{Q}{KB}\sum_{k=1}^K \left(\frac{1-c_{d_k}}{x_{\star k}^{(t)}} - \frac{1}{\bar{\eta}_k}\right)^+ = L^{(t)}$, which completes the proof of Proposition 1. ■

Although Proposition 1 does not prove the optimality of the suboptimal problem (14), it provides a guideline for performing the iterative algorithm.

$$\boldsymbol{P0}(\mathbf{a}): \quad \underset{x_k, p_k, y_k}{\texttt{Minimize}} \quad \frac{Q}{KB}\sum_{k=1}^K \left(\frac{1-c_{d_k}}{x_k} - \frac{1}{\bar{\eta}_k}\right)^+ \tag{15}$$

$$\texttt{s.t.} \quad \ln(A_k\mathbf{p}) \geq \ln(2)x_k + y_k, \forall k \tag{15a}$$

$$B_k\mathbf{p} \leq e^{a_k}(y_k - a_k + 1), \forall k. \tag{15b}$$

$$A_k\mathbf{p} \leq P_{tot}. \tag{15c}$$

C. QoE fairness among users

This subsection studies the fairness QoE among the users by guaranteeing the minimum difference among users' buffering time. We in fact investigate the condition which guarantees the same latency among the users under the ZF design.

Theorem 1: Under the ZF design, all the users can achieve zero-latency if $P_{tot} \geq \sigma^2 \sum_{k=1}^K \| \tilde{\mathbf{h}}_k \|^2 (2^{(1-c_{d_k})\bar{\eta}_k} - 1)$, where $\bar{\eta}_k = \eta_k / B$.

Proof: Under the ZF design, the achievable information rate for user k is $R_k = B\log_2(1 + \frac{p_k}{\sigma^2})$, where p_k is the power factor allocated to user k. We will show that for a given total power P_{tot}, there always exists a power allocation scheme which imposes a latency t to all the users. Indeed, the user k will experience a latency t when $p_k = \sigma^2(2^{\frac{(1-c_{d_k})\bar{\eta}_k}{\bar{\eta}_k t+1}} - 1)$, or equivalently the transmit power of user k is $\| \tilde{\mathbf{h}}_k \|^2 \, p_k = \sigma^2 \| \tilde{\mathbf{h}}_k \|^2 (2^{\frac{(1-c_{d_k})\bar{\eta}_k}{\bar{\eta}_k t+1}} - 1)$.

TABLE I: ITERATIVE ALGORITHM TO SOLVE (15)

1. Initialize a_k, ϵ, $t = 1$, L_{old} and `error`.
2. While `error` $> \epsilon$ do
   2.1. Solve $\boldsymbol{P0}(\{a_k\}_{k=1}^K)$ in (15) to obtain the optimal values $x_k^\star, y_k^\star, \mathbf{p}^\star$,
   and $L^{(t)} = \frac{Q}{KB}\sum_{k=1}^K \left(\frac{1-c_{d_k}}{x_k^\star} - \frac{1}{\bar{\eta}_k}\right)^+$
   2.3. Compute `error` $= |L^{(t)} - L_{old}|$
   2.4. Update $L_{old} = L^{(t)}, a_k = y_k^\star, t := t+1$

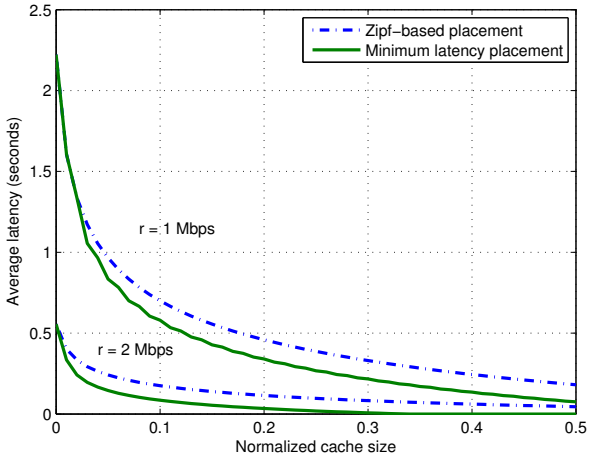

Fig. 2: Average user's latency of different cache placement designs v.s. the normalized cache size (cache size divided by the library size). $M_k = M, r_k = r, \eta_k = 3$ Mbps, $\forall k$.

Now consider a function

$$f(t) = \sigma^2 \sum_{k=1}^K \| \tilde{\mathbf{h}}_k \|^2 \left(2^{\frac{(1-c_{d_k})\bar{\eta}_k}{\bar{\eta}_k t+1}} - 1\right)$$

in $\mathbb{R}^+$. The derivative of $f(t)$ is given as

$$f'(t) = -\sigma^2 \sum_{k=1}^K \frac{(1-c_{d_k})\eta_k^2 \| \tilde{\mathbf{h}}_k \|^2}{(\bar{\eta}_k t+1)^2}\left(2^{\frac{(1-c_{d_k})\bar{\eta}_k}{\bar{\eta}_k t+1}} - 1\right),$$

which is negative in $\mathbb{R}^+$. This indicates that $f(t)$ is a monotonically decreasing function in $[0, +\infty)$. In addition, $f(0) = \sigma^2 \sum_{k=1}^K \| \tilde{\mathbf{h}}_k \|^2 \, (2^{(1-c_{d_k})\bar{\eta}_k} - 1)$ and $f(+\infty) = 0$. Therefore, there always exist one and only one t_0 such as $f(t_0) = P_{tot}$. As a result, all the users achieve zero-latency when $P_{tot} \geq f(0) = \sigma^2 \sum_{k=1}^K \| \tilde{\mathbf{h}}_k \|^2 \left(2^{(1-c_{d_k})\bar{\eta}_k} - 1\right)$ ■

V. NUMERICAL RESULTS

This section presents numerical results to demonstrate the effectiveness of the studied transmission designs. The results are averaged over 100 channel realizations, each accounting for ten independent requests. The library consists of $N = 100$ content files, each is of a length $Q = 10$ Mbits, $K = 8, L = K + 1$. The cache size is $M_k = M, \forall k$. In addition, $\sigma_{h_k}^2 = 1, \forall k$ and $B = 1$ MHz.

Fig. 2 plots the average latency as a function of the normalized cache size (the cache size divided by the library size, i.e., M/N) of the two cache placement methods in Section III. As expected, the proposed cache placement design achieves smaller latency than the Zipf-based scheme. Especially, with the average serving rate of 2 Mbps, the proposed placement design achieve zero latency when the normalized cache size is greater than 0.3.

Fig. 3 compares the latency performance of the ZF and MMSE -based designs developed in Section IV. The most popular files are cached based on Zipf distribution. The total transmit power is limited to 10 dB. The results are shown

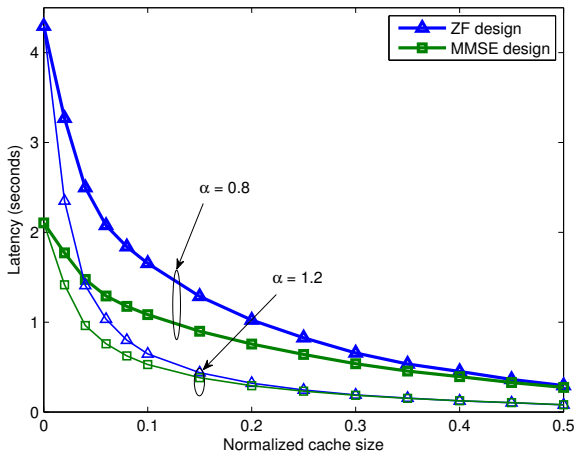

Fig. 3: Latency performance of the ZF and MMSE designs v.s. normalized cache size. Total transmit power $P_{tot} = 10$ dB, $\eta = 5$ Mbps.

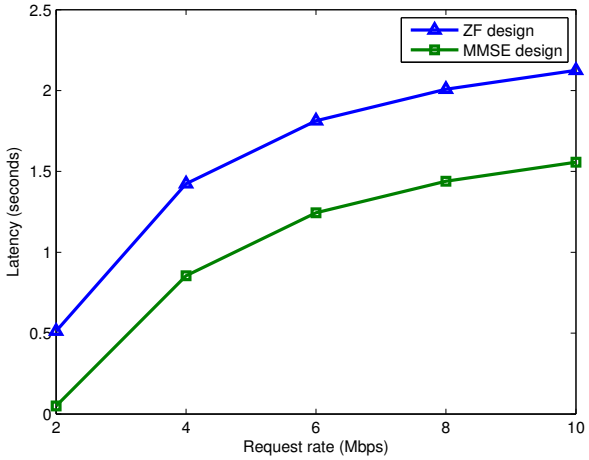

Fig. 4: Latency performance of the ZF and MMSE designs v.s. requesting rate η. Normalized cache size equals to 10%, $P_{tot} = 10$ dB.

for the normalized cache size up to 0.5 since the library size is usually much larger than the cache size in practice. It is shown that the latency is dramatically reduced as the cache capacity increases. This is because with a larger cache size, there is higher chance of requesting a file which is already available in the cache. It is also shown that the MMSE design is more efficient than the ZF, especially in the small cache size regime. Another observation is that higher Zipf exponent results in smaller latency. In this case, the content distribution is more focused on a small number of popular files. As the result, it is higher probable that the requested file has been cached.

Fig. 4 presents the latency as a function of the request rate at the normalized cache size equals to 10%. Similar conclusion is observed that the MMSE design achieves smaller latency that the ZF. Furthermore, the users have to wait longer when the request rate increases, which is reasonable since the total transmit power is limited.

VI. Conclusions

We have analysed the latency performance of content delivery networks in the presence of edge caching capability. First, we proposed a cache placement design that minimizes the average buffering time over Zipf distribution of request. We then formulated two optimizations corresponding to ZF and MMSE techniques which minimize the average buffering time in the delivery phase under transmit power constraint. We then proposed an iterative algorithm to overcome the non-convexity of the MMSE problem. Numerical results suggested to employ the MMSE design in such limited transmit power conditions.

References

[1] T. X. Vu, H. D. Nguyen, T. Q. S. Quek, and S. Sun, "Adaptive cloud radio access networks: compression and optimization," *IEEE Trans. Signal Process*, vol. 65, no. 1, pp. 228–241, Jan. 2017.

[2] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

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

[4] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118 – 6131, Sept. 2016.

[5] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug 2016.

[6] M. Gregori, J. Gmez-Vilardeb, J. Matamoros, and D. Gndz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[7] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[8] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Energy-efficient design for edge-caching wireless networks: When is coded-caching beneficial?" in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2017, pp. 1–5.

[9] T. X. Vu, S. Chatzinotas, and B. Ottersten "Energy Minimization for Cache-assisted Content Delivery Networks with Wireless Backhaul," *IEEE Wireless Commun. Lett.*, vol. pp, no. pp, pp. 1–1, 2018.

[10] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[11] Z. Chen, J. Lee, T. Q. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[12] T. X. Tran and D. Pompili, "Octopus: A cooperative hierarchical caching strategy for cloud radio access networks," in *Proc. IEEE Int. Conf. Mobile Ad Hoc Sensor Systems*, Oct. 2016, pp. 154–162.

[13] S. H. Park, O. Simeone, and S. Shamai, "Joint cloud and edge processing for latency minimization in fog radio access networks," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2016, pp. 1–5.

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

[15] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Info. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[16] T. X. Vu, S. Chatzinotas, and B. Ottersten "Edge-Caching Wireless Networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, to appear.

[17] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *IEEE INFOCOM*, Mar. 1999, vol. 1, pp. 126–134.