CACHING POLICY OPTIMIZATION FOR RATE ADAPTIVE VIDEO STREAMING

Huiting Su, Shenqian Han, and Chenyang Yang

School of Electronics and Information Engineering, Beihang University, China, 100191
Email: {huitingsu, sqhan, cyyang}@buaa.edu.cn

ABSTRACT

Femtocaching is an effective and low-cost way to address the explosively increased wireless traffic demand driven by video-on-demand (VoD) streaming. Most existing caching policies are optimized for file downloading service, which are not necessarily optimal for the quality of experience of users requesting VoD service. In this paper, the caching policy for VoD service is investigated, where the adaptation of video rate is taken into account. Simulation results demonstrate the performance gain of the proposed caching policies optimized for VoD service over that optimized for file downloading service.

***Index Terms*—** Femtocaching, video-on-demand, QoE

1. INTRODUCTION

Video-on-demand (VoD) is the main driver for the explosion of wireless data traffic [1]. To meet the exponentially increased traffic demands and guarantee the quality of experience (QoE) of users, a concept of femtocaching was proposed in [2]. By caching the popular video files at the small base stations (BSs) without backhaul link (also referred to as "helpers") and streaming the non-cached files from the macro BS, femtocaching is able to dramatically improve the network throughput with low cost and complexity.

Most existing caching policies are designed for file downloading service by optimizing which files should be cached in the helpers, aimed at minimizing the average file downloading delay [2, 3]. Considering the differences between the QoE of file downloading and VoD services, existing optimal caching policies are no longer applicable for VoD service. In this paper, we focus on one of the inherent features of VoD service, i.e., video rate adaptation, and study the content placement problem. Specifically, we consider the scenario with one macro BS and one helper, and investigate how much portion of which video files should be cached at the helper with what video rates, aimed at maximizing the average QoE of users.

Caching placement for rate adaptive videos has received little attention in the literature. In [4] and [5] reactive caching management based on Least Recently Used (LRU) strategy was studied, where the cached video version was optimized subject to the constrains on the available cache size, video processing capability and backhaul capacity at the BS. In [6] the proactive caching policy for a single multi-rate video file was studied given the storage size for the video file, where the number of cached video versions and the corresponding rates were optimized. In [7] the proactive caching policy for multiple video files was investigated by first proportionally allocating the storage space to the video files according to their popularity and then optimizing the cached video rates for every video file.

In this paper we study proactive caching policy for multiple video files, where a user can access the macro BS and the helper to stream a video file, which is different from existing works only considering caching at a single BS [4, 5, 6, 7]. Moreover, we consider the constraint of the total storage size of the helper rather than given the constraint for each video file as in [6, 7]. We first study the optimal caching policy for single user case, and then extend the result to multiuser case. Simulation results demonstrate that the proposed caching policy can improve the QoE of users for VoD service.

2. SYSTEM MODEL

We consider a femtocaching system where a macro BS and a helper serve multiple users. The macro BS is assumed to have ideal backhaul, connected to the core network to gather the requested video files without delay. Each helper has no backhaul but is equipped with a cache with the storage size of C bits. This single helper model is applicable to the system consisting of multiple geographically separated helpers within the coverage of the macro BS, in which scenario each user can only access to one helper and the macro BS. In the sequel, we first consider the single user case and then generalize the model to multiuser case.

Suppose that the user requests the video from a library of F files. Let q_f and T_f denote the popularity (i.e., request probability) and the playback duration of the f-th video file, $f = 1, \dots, F$. For each video, infinite video quality levels are assumed, which is a relaxation of the practical finite quality levels. Thus, the obtained results can be used as a benchmark for the optimization of caching policy with finite quality levels. Furthermore, we assume that the helper has the processing capability to perform video rate down-conversion as considered in [5]. This implies that the helper only needs

This work was supported in part by National High Technology Research and Development Program of China (No. 2014AA01A705) and by National Natural Science Foundation of China (No. 61429101).

to cache a single quality level for each video file. Let $R_{\mathrm{c},f}$ denote the cached video rate of the f-th file.

A video file is composed of a sequence of chunks, each of which is encoded and decoded as a stand-alone unit. The duration of a chunk is generally much longer than the coherent time of the small-scale fading channel, i.e., the channel coding can span across a large number of independent fading channels [8]. Therefore, it is safe to assume that the ergodic capacity is achievable. Let $R_{\mathrm{B},u}$ and $R_{\mathrm{H},u}$ denote the ergodic capacity from the macro BS and the helper to the u-th user, which is referred to as downloading rates. Moreover, consider $R_{\mathrm{B},u} \leq R_{\mathrm{H},u}$ since the helper is usually closer to the user and can allocate a larger bandwidth to users compared to the macro BS by noting that the number of users accessing to the helper is smaller. We assume that $R_{\mathrm{B},u}$ is larger than the video rate required by the minimal video quality level, so that we can ignore the impact of streaming stalling and focus on the impact of video rate adaptation.

Under the constraint on storage size of the helper, partial caching is allowed in the paper, i.e., the helper may only cache a part of a video file. This means that a user may stream the cached part of a video from the helper and the uncached part from the macro BS, respectively. Let x_f denote the cached portion of the f-th video file.

According to [9], it is reasonable to characterize user satisfaction with logarithmic laws for telecommunication services. As such, we adopt the QoE model as a logarithmic function of the user-experienced playback rate [9]

$$\mathrm{QoE}_{*,uf} = \alpha \log(1 + \beta R_{\mathrm{p}_*,uf}), \tag{1}$$

where the subscript "$*$" can be "c" and "uc" corresponding to the cached and uncached parts, respectively, $R_{\mathrm{p}_*,uf}$ is the user-experienced playback rate, depending on the cached video rate $R_{\mathrm{c},f}$ and the downloading rates $R_{\mathrm{B},u}$ and $R_{\mathrm{H},u}$ to be discussed later, and α and β are two scalar constants. Then, the QoE of the u-th user requesting the f-th file can be defined as

$$\mathrm{QoE}_{uf} = x_f \mathrm{QoE}_{\mathrm{c},uf} + (1 - x_f)\mathrm{QoE}_{\mathrm{uc},uf}, \tag{2}$$

where $\mathrm{QoE}_{\mathrm{c},uf}$ and $\mathrm{QoE}_{\mathrm{uc},uf}$ are the QoE corresponding to the cached and uncached parts of the f-th video file, respectively.

3. CACHING POLICY OPTIMIZATION

In this section we design the caching policy by optimizing how large portion of which files should be cached at the helper with what video rates. We first analyze the user-experienced playback rate, based on which the caching policy for single user and multiuser cases are then optimized, respectively.

3.1. User-experienced Playback Rate

Consider that the u-th user requests the f-th video file. To stream the cached part, the user may access to both the helper and the macro BS, because always accessing to the helper to fetch the cached part may be not optimal as will be clear later. When the user accesses to the helper, if the cached video rate $R_{\mathrm{c},f}$ is smaller than the downloading rate $R_{\mathrm{H},u}$, then the playback rate experienced by the user is limited to the cached rate, i.e., $R_{\mathrm{p_c},uf} = R_{\mathrm{c},f}$; otherwise, the helper needs to down-convert the cached video to the rate $R_{\mathrm{H},u}$, and the user-experienced playback rate is $R_{\mathrm{p_c},uf} = R_{\mathrm{H},u}$. Thus, we have $R_{\mathrm{p_c},uf} = \min\{R_{\mathrm{H},u}, R_{\mathrm{c},f}\}$ when the user accesses to the helper. On the other hand, when the user accesses to the macro BS, $R_{\mathrm{p_c},uf}$ is limited by the downloading rate $R_{\mathrm{B},u}$. In summary, for the cached part, the maximum user-experienced playback rate can be obtained as

$$R_{\mathrm{p_c},uf} = \max\{R_{\mathrm{B},u}, \min\{R_{\mathrm{H},u}, R_{\mathrm{c},f}\}\}. \tag{3}$$

For uncached part of the f-th video, the user can only access to the macro BS. Then, we have

$$R_{\mathrm{p_{uc}},uf} = R_{\mathrm{B},u}. \tag{4}$$

3.2. Single user Case

We consider single user case in this subsection, and omit the subscript "u" for notational simplicity.

We optimize the caching policy aimed at maximizing the average QoE of the user subject to storage size constraint of the helper, where the average is taken over the random request to the F video files. The optimization problem can be formulated as

$$\begin{aligned} \max_{\{x_f, R_{\mathrm{c},f}\}_1^F} \ & \alpha \textstyle\sum_{f=1}^F q_f \big(x_f \log(1+\beta R_{\mathrm{p_c},f}) + \\ & \qquad (1-x_f) \log(1+\beta R_{\mathrm{B}}) \big) && \text{(5a)} \\ s.t. \ & R_{\mathrm{p_c},f} = \max\{R_{\mathrm{B}}, \min\{R_{\mathrm{H}}, R_{\mathrm{c},f}\}\}, \forall f && \text{(5b)} \\ & \textstyle\sum_{f=1}^F x_f R_{\mathrm{c},f} T_f \leq C && \text{(5c)} \\ & R_{\mathrm{c},f} \geq 0, \forall f && \text{(5d)} \\ & 0 \leq x_f \leq 1, \forall f, && \text{(5e)} \end{aligned}$$

where the objective function is obtained based on (2), (1) and (4), and (5c) is the constraint on the storage size of the helper.

Constraint (5b) is non-convex due to the max-min function. Nevertheless, in single user case we can find that the optimal cached rate $R_{\mathrm{c},f}$ must satisfy

$$R_{\mathrm{B}} \leq R_{\mathrm{c},f} \leq R_{\mathrm{H}} \tag{6}$$

for all f with $x_f > 0$. Otherwise, if $R_{\mathrm{c},f} < R_{\mathrm{B}}$, then the cached video is useless and just a waste of storage resource because fetching the video from the macro BS can achieve a higher QoE; if $R_{\mathrm{c},f} > R_{\mathrm{H}}$, the cached high quality level cannot be streamed to the user due to the limitation of R_{H}, which leads to a waste of storage resource.

With (6), we can convert the non-convex constraint (5b) for all f with $x_f > 0$ as

$$R_{\mathrm{p_c},f} = R_{\mathrm{c},f}. \tag{7}$$

Then, we can rewrite problem (5) as

$$\max_{\{x_f,R_{c,f}\}_1^F} \alpha\sum_{f=1}^F q_f\big(x_f\log(1+\beta R_{c,f})+ (1-x_f)\log(1+\beta R_B)\big) \tag{8a}$$

$$s.t.\ R_B \le R_{c,f} \le R_H, \forall f \tag{8b}$$

$$\sum_{f=1}^F x_f R_{c,f} T_f \le C \tag{8c}$$

$$0 \le x_f \le 1, \forall f. \tag{8d}$$

Problem (8) is still non-convex due to the objective function (8a) and the storage size constraint (8c) are not jointly convex in x_f and $R_{c,f}$. Nevertheless, it is in fact a hidden convex problem by using the change of variables. Specifically, defining $z_f = R_{c,f}x_f, \forall f$, problem (8) can be rewritten as

$$\max_{\{x_f,z_f\}_1^F} \alpha\sum_{f=1}^F q_f\big(x_f\log(1+\beta\tfrac{z_f}{x_f})+ (1-x_f)\log(1+\beta R_B)\big) \tag{9a}$$

$$s.t.\ x_f R_B \le z_f \le x_f R_H, \forall f \tag{9b}$$

$$\sum_{f=1}^F z_f T_f \le C \tag{9c}$$

$$0 \le x_f \le 1, \forall f. \tag{9d}$$

It can be found that problem (9) is convex, whose global optimal solution can be obtained efficiently with convex optimization algorithms. The optimal caching policy is then obtained accordingly for the single user case. Next, we generalize the results to the multiuser case.

3.3. Multiuser Case

In multiuser case, we consider that the macro BS and the helper serve U users. The caching policy is optimized to maximize the weighted sum average QoE of multiple users, which can be formulated based on (5) as

$$\max_{\{x_f,R_{c,f}\}_1^F} \alpha\sum_{u=1}^U w_u\sum_{f=1}^F q_f\big(x_f\log(1+\beta R_{p_c,uf})+ (1-x_f)\log(1+\beta R_{B,u})\big) \tag{10a}$$

$$s.t.\ R_{p_c,uf} = \max\{R_{B,u}, \min\{R_{H,u}, R_{c,f}\}\}, \forall u, f \tag{10b}$$

$$\sum_{f=1}^F x_f R_{c,f} T_f \le C \tag{10c}$$

$$R_{c,f} \ge 0, \forall f \tag{10d}$$

$$0 \le x_f \le 1, \forall f, \tag{10e}$$

where w_u is the priority weight associated with the u-th user.

The difficulty of solving problem lies in the non-convex constraint (10b). In multiuser case, we cannot simplify this constraint as (6) in single user case, because it may happen that the cached video rate $R_{c,f}$ is higher than $R_{H,u}$ for some users and lower than $R_{B,u}$ for other users. In the following, we strive to find a suboptimal solution to the problem by introducing a user access strategy. Specifically, we restrict that the u-th user for arbitrary u, who is requesting the f-th video file, must access to the helper to stream the cached part, even though accessing to the macro BS may achieve a higher video quality which may happen when $R_{B,u} > R_{c,f}$. Although this accessing constraint is suboptimal, it is very likely to be applied in practical systems owing to offloading the traffic load of the macro BS.

Then, constraint (10b) can be simplified as

$$R_{p_c,uf} = \min\{R_{H,u}, R_{c,f}\}, \forall u, f \tag{11}$$

and problem (10) can be reformulated as

$$\max_{\{x_f,R_{c,f}\}_1^F} \alpha\sum_{u=1}^U w_u\sum_{f=1}^F q_f\big(x_f\log(1+\beta R_{p_c,uf})+ (1-x_f)\log(1+\beta R_{B,u})\big) \tag{12a}$$

$$s.t.\ R_{p_c,uf} \le R_{H,u}, \forall u, f \tag{12b}$$

$$R_{p_c,uf} \le R_{c,f}, \forall u, f \tag{12c}$$

$$\sum_{f=1}^F x_f R_{c,f} T_f \le C \tag{12d}$$

$$R_{c,f} \ge 0, \forall f \tag{12e}$$

$$0 \le x_f \le 1, \forall f. \tag{12f}$$

By using the change of variables, i.e., defining $z_f = x_f R_{c,f}$ and $y_{uf} = x_f R_{p_c,uf}, \forall u, f$, we can rewrite problem (12) as

$$\max_{\{x_f,z_f,y_{uf}\}} \alpha\sum_{u=1}^U w_u\sum_{f=1}^F q_f\big(x_f\log(1+\beta\tfrac{y_{uf}}{x_f})+ (1-x_f)\log(1+\beta R_{B,u})\big) \tag{13a}$$

$$s.t.\ y_{uf} \le x_f R_{H,u}, \forall u, f \tag{13b}$$

$$y_{uf} \le z_f, \forall u, f \tag{13c}$$

$$\sum_{f=1}^F z_f T_f \le C \tag{13d}$$

$$z_f \ge 0, \forall f \tag{13e}$$

$$y_{uf} \ge 0, \forall u, f \tag{13f}$$

$$0 \le x_f \le 1, \forall f. \tag{13g}$$

It can be found that problem (13) is convex, whose optimal solution can be obtained readily, which provides the optimal caching policy for the considered user access strategy. Given the obtained caching results, $\{x_f, R_{c,f}\}$, one can also further update the access strategy to improve the QoE, i.e., accessing to the macro BS if $R_{B,u} > R_{c,f}$.

4. SIMULATION RESULTS

In this section we evaluate the performance of the proposed caching policy for VoD service. For comparison, we also simulate the existing caching policy for file downloading service proposed in [3], where it is shown that caching the most popular files is the optimal policy for the single-helper scenario. Thus, in the simulation the “file downloading” policy is set to cache the most popular files until the storage is fully used,

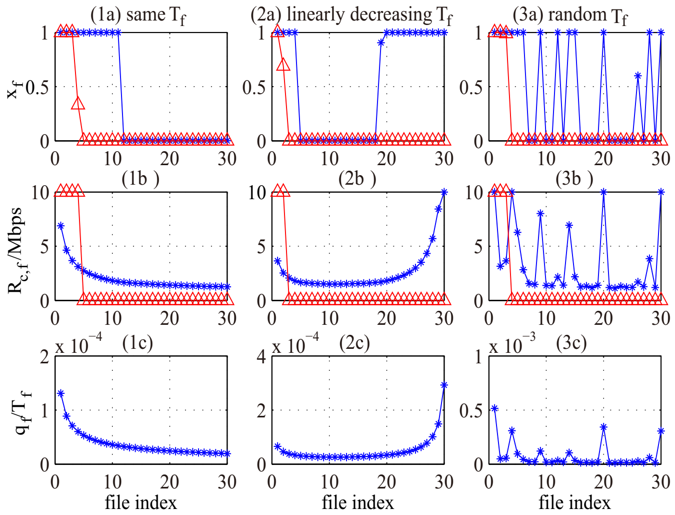

Fig. 1. Comparison of the caching policies designed for VoD (marked by $*$) and file downloading (marked by $\triangle$).

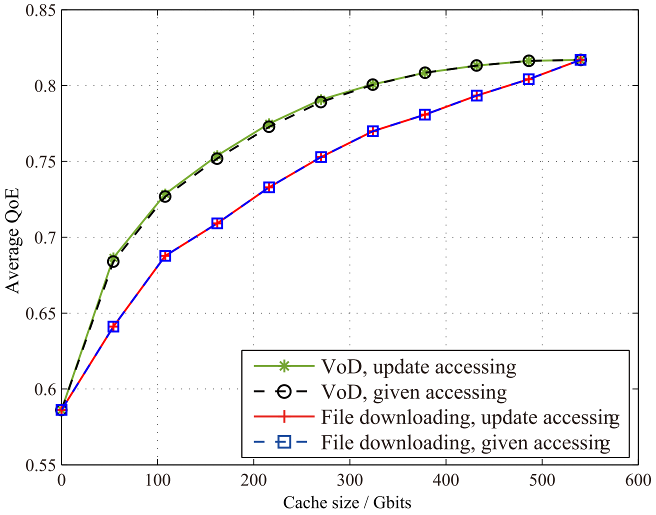

Fig. 2. Average QoE of the two caching policies.

where the cached video rate is set as the maximum downloading rate from the helper to all users. Note that the existing caching policies for VoD are designed for the scenarios with single file or single BS [6, 7], which are not applicable to the considered femtocaching system.

In single user case, the spectral efficiency of the macro BS and the helper is set as 3 and 5 bps/Hz, respectively, as in [3]. Considering that the macro BS covers more users than the helper, we set the bandwidth allocated by the macro BS and the helper to the user as 0.2 and 2 MHz, respectively, which leads to the downloading rates from the macro BS and the helper as $R_{\mathrm{B},u} = 0.6$ and $R_{\mathrm{H},u} = 10$ Mbps. In multiuser case, we set the downloading rate from the helper, $R_{\mathrm{H},u}$, to follow uniform distribution between 2 Mbps and 20 Mbps, and then set the downloading rate from the macro BS, $R_{\mathrm{B},u}$, to follow uniform distribution between 0.2 Mbps and $\min(10 \text{ Mbps}, R_{\mathrm{H},u})$. In this manner, we can ensure $R_{\mathrm{B},u} \leq R_{\mathrm{H},u}$ for all u. We consider $F = 30$ video files, whose popularity follows Zipf distribution with the parameter of 0.56 [3]. We consider different playback durations of the video files to analyze their impact. The parameters in the QoE function are set as $\alpha = 1$ and $\beta = 10$. The priority weights of the users are set as 1.

Figure 1 shows the proposed and existing caching policies designed for VoD and file downloading, respectively in single user case, where the storage size of the helper is set as 30 Gbits. The three columns correspond to three configurations for the playback durations of the video files, and the three rows show the cached portion x_f, cached rate $R_{\mathrm{c},f}$, and the ratio q_f/T_f, respectively. The x-axis is the index of the video files, ranking from one (the most popular) to F (the least popular) based on popularity. In the first column all the video files have the same playback duration of 15 min, in the second column the playback duration of the f-th video file is set as $31 - f$ min, and in the third column random playback durations are drawn uniformly from 0 to 30 min. We can observe from the first and second rows that the proposed caching policy prefers to cache more video files but with lower video rates compared to the policy designed for file downloading. Moreover, by comparing the second and the third rows, we can find that the proposed caching policy is related to the ratio q_f/T_f. A video file with low popularity may still be cached if its playback duration is short.

Figure 2 compares the average QoE achieved by the proposed and existing caching policies as a function of the storage size of the helper, where $F = 30$, $U = 10$, and random playback durations varying from 0 to 30 min are considered. We can see that the proposed policy performs the same as the existing policy when the storage size is small so that few videos can be cached or when the storage size is very large so that all video files can be cached. For medium storage size, the proposed policy can achieve a higher average QoE of the users even when we restrict that the users must access to the helper to fetch the cached videos. The performance can be further improved when removing the accessing restriction, but the gain is marginal because the downloading rate from the helper is statistically much larger than that from the macro BS, making the probability of not accessing to the helper is small.

5. CONCLUSIONS

In this paper we optimized the caching policy for VoD service. We obtained the optimal policy for the single user case and a suboptimal policy for the multiuser case. The results show that caching the most popular video files, which is optimal for file downloading service in the single-helper scenario, is no longer optimal for VoD service. Both the file popularity and the playback duration will affect the caching policy. Simulation results show the performance gain of the proposed caching policies.

6. REFERENCES

[1] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[2] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, 2012.

[3] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[4] H. Ahlehagh and S. Dey, "Adaptive bit rate capable video caching and scheduling," in *Proc. IEEE WCNC*, 2013.

[5] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Transactions on Networking*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[6] W. Zhang, Y. Wen, Z. Chen, and A. Khisti, "QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1431–1445, Oct. 2013.

[7] Yumei Wang, Xiaojiang Zhou, Mengyao Sun, Lin Zhang, and Xiaofei Wu, "A new QoE-driven video cache management scheme with wireless cloud computing in cellular networks," *Mobile Networks and Applications*, p. 1, Feb. 2016.

[8] D. Bethanabhotla, G. Caire, and M. J. Neely, "Adaptive video streaming for wireless networks with multiple users and helpers," *IEEE Transactions on Communications*, vol. 63, no. 1, pp. 268–285, Jan. 2015.

[9] P. Reichl, B. Tuffin, and R. Schatz, "Logarithmic laws in service quality perception: Where microeconomics meets psychophysics and quality of experience," *Telecommun. Syst.*, pp. 1–14, 2011.