IMPROVING HIERARCHICAL VIDEO CACHING THROUGH CROSS-LAYER COOPERATION IN MOBILE NETWORKS

[1]LIXIANG XU, [2]QIANG LING, [3]ZHIWEI ZHANG

[1,2,3]Dept. of Automation, University of Science and Technology of China, Hefei 230027, China
E-mail: [2]qling@ustc.edu.cn

Abstract- With the rapid growth in watching videos from mobile devices, hierarchical caching servers are widely deployed in wireless networks to improve the experience of mobile clients. We propose a novel caching hierarchy with cross-layer cooperations to reduce the video redundancy in traditional inclusive cache hierarchy (ICH). With the help of low-layer caches, the high-layer cache in our proposed cooperated cache hierarchy (CCH) is able to calculate the profit of caching each video in its space and make cache decision that brings the highest performance improvement. Simulations show that our CCH outperforms ICH under all cache sizes.

Index Terms- Video Caching, Cooperation, Wireless Network

I. INTRODUCTION

With the rapid growth in number of smart phones and tablets, access to Internet from mobile devices (MDs) is expected to grow significantly in the next few years. It is reported that mobile data traffic will grow 92% annually from 2010 to 2015, of which up to two thirds is expected to be video [1]. The explosive video data traffic in wireless network drives people to seek for efficient video delivery schemes.

Efforts have been made to improve the user experience of watching videos from MDs. Content delivery networks(CDNs) [2] have been enhanced to place videos close to users. However, high delay remains to be possible when fetching videos from a CDN server because of the congestion in Internet. Furthermore, videos will have to travel through the wireless carrier Core Network (CN) and Radio Access Network(RAN) before reaching the MDs, which brings significant delay and puts high strain on the CN and RAN backhaul.

Caching techniques have been used to improve the video delivery efficiency in CN and RAN. A small cache is introduced in the base station in [3] to reduce the data traffic in CN. The users' preference for different videos is treated as a prior knowledge to cache the most popular videos. This work is further improved in [4] where a three-layer caching structure is employed to reduce data fetched from the CDNs. However, the utility rate of cache space in the hierarchical video caching is highly depressed because of the "inclusive" caching scheme where a higher layer cache includes all of the videos that exist in the associated lower layer caches.

A coordinated edge caching in RAN is proposed in [5]where caches in different base stations can exchange data with the aid of SAE-GW. The cache space of a base station is divided into "local cache" and "shared cached", where the "local cache" only serves the local users while the "shared cached" serves all the users. With the assumption that users' preference for videos in different base stations share a common Zipf-like distribution, the optimal rate of "local cache" depends only on the Zipf-like parameter and the delay off etching videos from different places. Mobile users who share similar preference for videos in [6] are clustered and associated to the same small cell base station to improve the hit ratio of accessing a small cell base station.

Though caching in the base station and hierarchical video caching are both studied in the RAN, cooperation between different layers of hierarchical video caching remains to be a problem. In this paper, we propose a novel hierarchical video caching which improves the utility rate of cache space through cross-layer cooperation in RAN. When videos cached in a lower layer cache changes, a message is sent to the associated higher layer cache. The higher layer cache in the hierarchical video caching tries to eliminate videos that have multiple copies in the lower layer cache to reduce space redundancy. Experiments show that the average delay of watching videos from MDs is highly reduced when cross-layer cooperation is performed.

The rest of this paper is organized as follows. Section 2introduces the system model. The proposed cross-layer cooperation in hierarchical video caching is presented in Section3. In Section 4, the performance of the proposed caching scheme is evaluated, while conclusions are drawn in Section5.

II. SYSTEM MODEL

We develop a three-layer hierarchical video caching model in this paper, as show in Fig. 1. CDNs serve as the first layer of the caching system. The huge space of CDNs makes it possible to cache all the videos. Despite the fact that CDN shave been optimized to place videos close to users, a long delay still exists when a video is fetched from the Internet. A proxy

cache in the Packet Data Gateway (PGW)helps to reduce the access delay when a request is served from the proxy server. We denote the proxy cache in the PGW by "PGW cache" in the rest of the paper. PGW cache serves as the second layer of the caching system. Considering the limited storage space of the PGW cache, the videos cachedin PGW are carefully chosen to achieve the optimal performance. With the development of hardware devices, it is possible to introduce cache into the base station. A small cache in the base station will improve the user experience significantly when the most popular videos in the local area are stored. Multiple base station caches serve as the third layer of the caching system. In the

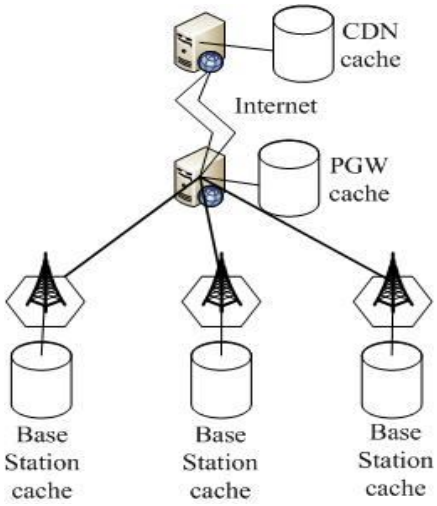


Fig.1.A three-layer hierarchical video caching model in mobile network.

three-layer caching system, multiple base station caches are connected to the same PGW cache and connection between base station caches is performed with the aid of PGW cache.

We denote the delay of serving a request from the base station cache directly by t_0 . t_1 represents the delay of transferring a video between a PGW cache and a base station cache while t_2 represents the delay of fetching a video from the CDN to a PGW cache. When a request is served from the base station cache to which it is connected directly, the delay is t0. If the requested video is not cached in the base station cache, a request is sent to the PGW cache. A hit in the PGW cache means the total delay is $t_0 + t_1$. However, if a miss happens, the PGW cache will check its records about the videos cached in all the base station caches. A request is sent to the base station where the requested video is cached. In this situation, the total delay is $t_0 + 2*t_1$. A request is further sent to the CDN only when no base station cache has the requested video and the total delay will be $t_0 + t_1 + t_2$. Usually t_2 is much larger than t_1 , so the PGW cache always sent its missed requests to the base station caches before fetching videos from the CDN.

III. CACHE HIERARCHY WITH CROSS-LAYER COOPERATION

A. Cache Policy in Base Station Caches

In a hierarchical caching system, each cache of each layerin the hierarchy caches videos according to its received userrequests. The most popular videos in each cache node'sservice area will be cached to achieve high performance. Asalow-layer cache usually serves a sub-region of a high-layercache's coverage, the most popular videos in a low-layercache may be a subset of the most popular videos in ahigh-layer cache. This kind of cache hierarchy isusuallycalled"inclusive cache hierarch" [7]. This certainly will leadtosignificant video redundancy in the caches of differentlayers of the hierarchy, and the cache hit ratio of thecachingsystem will be seriously affected.

In this paper, we propose a cross layercooperation approach in cache hierarchy to depressthe video redundancy.Suppose the cache hierarchy in Fig. 1 consists of one PGW cache and n base station caches. We use P and B_j to represent the PGW cache and base station caches respectively.The request frequency of video v_i in base station cache B_j is denoted by f_{ij} . k_{ij} is a caching flag that showswhether video v_i is cached in base station cache B_j Base station caches make cache decision according tovideos' local requestfrequency. All the f_{ij} in B_j arecompared to each otherand videos with the highest f_{ij} is cached. When a video's request frequencychanges, cache replacement is performed toreplace a less popular video in the cache with a more popular video.

To accelerate the speed of cache replacement, all the accessed videos in a base station are organizedby a *minimumheap* which stores the least frequently requested video at the root node. Upon receiving a request, the base station will check whether the requested video is cached in itslocal cache. If so, the request will be served directly by thebase station; otherwise, the base station will send a requestto the PGW server. This video will be sent to the user oncethe base station receives the response of PGW server. Meanwhile, the request frequency of the requested video will beupdated. If this video is cached in the base station, its position in the *minimum heap*; otherwise,it will compare the request frequency of the requested videoand least frequently requested video in the cache (i.e., thevideo at the root node of the *minimum heap*) to check ifcache replacement is needed. If the requested video has ahigher requested frequency, the least frequently requestedvideo in the cache will be deleted and the requested videowill be cached.

B. Cache Policy in the PGW Cache

Cooperation between thePGW cache and all the base station caches is performed tohelp the PGW cache make the right decision. A base stationwill sent all its f_{ij} and k_{ij} to the PGW cache. Then the PGWcache calculates the profit of caching video v_i in its cachespace. We denote the profit of caching video v_i by F_i. Asa server of the base station, the PGW cache will try to optimizeits cache to serve all the base stations. So F_i must indicates how much benefit is created for all the base stationsby caching v_i in the PGW server. In this paper, we use therequest latency saved by caching v_i in the PGW cache todescribe F_i.

Two different scenarios may be encountered. If $\sum_{j=1}^{n} k_{ij} = 0$, which means none of the base station cacheshas video v_i, all the requests must fetch the video from theCDN cache if it is not cached in the PGW cache. Aftercaching v_i in the PGW cache, all the requests can be servedby the PGW cache. So F_i can be represented by the totaldelay saved by caching v_i in the PGW cache, i.e.

$$F_i = \sum_{j=1}^{n} f_{ij} * (t_0 + t_1 + t_2) - \sum_{j=1}^{n} f_{ij} * (t_0 + t_1) = \sum_{j=1}^{n} f_{ij} * t_2 \quad (1)$$

If $\sum_{j=1}^{n} k_{ij} > 0$, the delay of watching v_i from the basestation caches who have stored v_i is t_0 whereas the delay of watching v_i from other base station caches is $t_0 + 2 * t_1$. If v_i is cachedin the PGW cache, the delay of watching v_i from the basestation caches who has stored v_i remains to be t_0 whereas thedelay ofwatching v_i from other base station caches decreases to $t_0 + t_1$ asthe PGW cache can serve all the requests. So we get

$$F_i = \sum_{j=1}^{n} (1 - k_{ij}) * f_{ij} * (t_0 + 2 * t_1) - \sum_{j=1}^{n} (1 - k_{ij}) * f_{ij} * (t_0 + t_1) = \sum_{j=1}^{n} (1 - k_{ij}) * f_{ij} * t_1 \quad (2)$$

With all the information from all the base station caches,the PGW cache will store the videos with the highest F_i in itscache space to achieve the optimal caching profit. When therequest frequency orcaching flag of any video in any basestation cache changes, this information will be sent to thePGW cache and replacement will be performed to make surethat the PGW cache always make the best choice about itscached videos. Videoredundancy is highly impressed in our proposed cooperated cache hierarchy compared to the inclusive cache hierarchy.

Note that replacement in the PGW cache can be performedquickly when videos are organized with a heap according to F_i. Videos cached in the PGW cache are organized with a*minimum heap* while other videos are organized with a *maximum heap*. When the F_i of a video changes, its positionin the *minimum heap* or the *maximum heap* is updated. Thenthe root node in the *minimum heap* is compared to the rootin the *maximum heap*. If the root node in the maximum heaphas a higher F_i than the root node in the *minimum heap*, areplacement is needed. Both the root nodes in the *maximumheap* and the *minimum heap* are deleted from the originalheap and inserted to the other one.

IV. RESULTS AND DISCUSSION

Some simulations are presented in this section to compare the performance of inclusive cache hierarchy (ICH)and our proposed cooperated cache hierarchy (CCH). Herecaching performance is mainly measured by the Average Request Delay (ARD) of mobile clients over all the base stationcaches.

A. Simulation Setup

The simulation model consists of a cache hierarchy simulator and a user request generator. The cache hierarchy simulator simulates a caching hierarchy to implement differentcaching schemes. The user request generator creates userrequests.

We assume that there are M distinct videos with the samesize. Here the default value is $M = 2,000$. The requestedprobability of videos follows a Zipf-likedistribution[8] [9], where the probability of the $i-th$ most popular videois

$$f_i = \frac{1}{\sum_{j=1}^{M} \frac{1}{j^{\alpha}} * i^{\alpha}}, \quad (3)$$

where $0 \le \alpha \le 1$ is the Zipf parameter which characterizes the degree of skewness. $\alpha = 1$ corresponds to a pureZipf distribution, which is highly skewed, whereas $\alpha = 0$ corresponds to a uniform distribution. The default value of α inour simulation is 0.8. The user requests are assumed to follow a Poisson distribution with the mean inter-arrival timeof $\lambda = 60$ seconds and randomly distributed on each basestation. In our simulations, the CDN cache stores all the videoswhile the PGW cache and base station cache store only asmall part of the videos. We use k_1 to represent the ratio ofthe cache space in PGW cache relative to the size of the totalvideos. The PGW cache has n homogeneous base stationcaches which share the same cache size and user preferencefor videos. The default value of n is 100. We use k_2 torepresent the ratio of the cache space in a base station relativeto the

size of the total videos. The default values of k_1 and k_2 is 0.1 and 0.01 respectively. The definitions of concerned parameters andtheir default values in our simulations are listed in Table Ⅰ.

Table Ⅰ. System Parameters and Default Values

Notation	Definition (Default Values)
M	number of distinct videos (2,000)
α	Zipf-like parameter (0.8)
λ	mean of the inter-arrival time of user requests (60 s)
k_1	the ratio of the cache space in PGW cache relative to the size of the total videos (0.1)
k_2	the ratio of the cache space in a base station cache relative to the size of the total videos (0.01)
n	number of homogeneous base stations (100)
t_0	delay of requests served from a base station cache (20 ms)
t_1	delay between the PGW cache and the base station cache (100 ms)
t_2	delay between the CDN cache and the PGW cache (2000 ms)

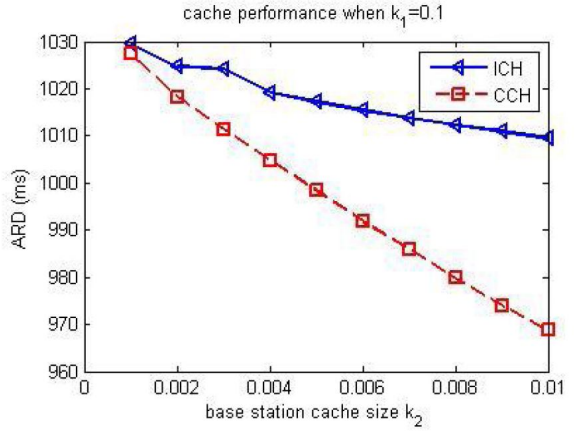


Fig.2. Cache performance under different base station cachesize

B. Simulation Results

The ARDs of ICH and CCH with different base stationcaches are shown in Fig. 2. As the cache space is limited bythe base station's hardware performance, we only simulatethe system's cache performance when $k_2 \le 0.01$.

Results show that our CCH outperforms ICH under all cache sizes, and the advantage of CCH increases with thecache size of the base station. For example, the ARD ofCCH is only 2.1 ms less than that of ICH when $k_2 = 0.001$,but this advantage increases to 40.8 ms when k_2 increase to0.01. These results prove that CCH performs well in reducing the video redundancy of ICH. The video redundancy inICH is slight when the cache space of the base station is small, so the cache performance of CCH is similar to that ofICH. When the cache space of the base station increases, thevideo redundancy in ICH becomes heavier and CCH showsmore significant advantage than ICH.

From Fig. 2, we also conclude that the achieved increasein cache performance of CCH by increasing cache spaceis much more significant than that of ICH. Results showthat increasing the cache space of the base station achievesvery limited improvement incache performance in ICH. Forexample, the ARD of ICH decreases from 1029.44 ms to1009.5 ms when k_2 increases from 0.001 to 0.01, namely tentimes increase in cache space achieves only 20 ms decreasein ARD. In contrast, the decrease in ARD of our CCH is58.7 ms under the same space growth, nearly 3 times of thatof ICH. The reason is that space increase in the base stationdoes not affect the PGW cache in ICH due to its "inclusive"caching scheme. The space increase in the base station justmakes some requests be served from the base station insteadof the PGW cache. However, situations are different in CCH.On one hand, space increase in the base station cache makesthe server of some requests changes from the PGW cacheto the base station. On the other hand, space increase inthe base station moves some videos from the PGW cache tothe base station cache and makes room for other videos. Theefforts paid to decrease video redundancy in CCH make itresponse more positively to the space growth than ICH.

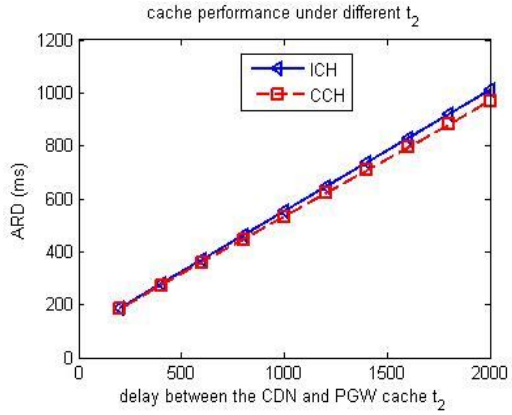


Fig.3. Cache performance under different delay between CDN and PGW cache

To show the affection of dynamic internet condition on different cache policies, we adjust the delay between PGWand CDN and compare the performance of the two cache hierarchies. Note that other system parameters are set as theirdefault values in Table 1. Simulation results are shown inFig. 3. It can be concluded from the results that bad internetcondition will increase the ARD of both hierarchies.It can be concluded from the results that bad internetcondition will increase the ARD of both hierarchies. However, the ARC of CCH stays below ICH under all t_2, and thisdifference of ARD increases with t_2. This demonstrates theadvantage of CCH over ICH in adapting to dynamic internetcondition.

REFERENCES

[1] White Paper, "Cisco Visual Networking Index: Global MobileData,",2010–2015.

[2] G. Pallis and A. Vakali, "Insight and perspectives for contentdelivery networks," Communications of the ACM, 49(1): 101C-106, 2006.

[3] H. AhleHagh, and S. Dey, "Video caching in Radio AccessNetwork: Impact on delay and capacity," Proceedings of IEEEWireless Communications and Networking Conference, 2012:2276–2281.

[4] H. AhleHagh, and S. Dey, "Hierarchical video caching in wireless cloud: approaches and algorithms," Proceedings of IEEEInternational Conference on Communications, 2012: 7082–7087.

[5] Y. Li, Y. Xu, T. Lin, X. Wang, and S. Ci, "A novel coordinatededge caching with request filtration in radio access network,"The Scientific World Journal, 66(11): 1653–1656, 2013.

[6] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho,"Content-aware user clustering and caching in wireless smallcell networks,"Eprint Arxiv,1409: 945–949, 2014.

[7] L. Li, D. Tong, Z. Xie, and J. Lu, "Improving inclusive cacheperformance with two-level eviction priority," IEEE 30th International Conference on Computer Design (ICCD), 2012: 387–392.

[8] J. Choi,A. Reaz, and B. Mukherjee, "A survey of user behavior in VoD service and bandwidth-saving multicast streamingschemes," IEEE Communications Surveys and Tutorials, 14(1):156–169,2012.

[9] A. Abhari, M. Soraya, "Workload generation for youtube,"Multimedia Tools and Applications, 46(1): 91–118,2010.

★★★