An Overview Of Multimedia Proxy Servers

Philip Kwok Chung TSE

Department of Electrical and Electronic Engineering, The University of Hong Kong
Pokfulam Road, Hong Kong SAR, China.

and

Simon Wing Wah SO

Dept. of Information and Applied Technology, Hong Kong Institute of Education
Block D4, 10 Lo Ping Road, Tai Po,
New Territories, Hong Kong SAR, China.

ABSTRACT

This paper gives an overview of the functions of the proxy servers in accessing multimedia objects over the Internet. The proxy servers use the cache replacement policies to select the cold objects to be removed from cache to release space. They use the object partitioning methods to divide each multimedia object into cacheable and non-cacheable parts. When the client bandwidth is insufficient to receive the original object, the proxy servers convert multimedia objects into objects of lower resolution in order to adapt to the client characteristics. When multiple proxy servers are present, they may work together in a cooperative manner to further enhance the cache efficiency or in a distributed manner to evenly spread the load. In summary, proxy servers become the centre of management with respect to the delivery of multimedia objects from web servers to clients over the Internet.

Keywords: Multimedia Proxy Server, System Architecture, Cache Replacement Policy, Object Partitioning, Transcoding, Cooperative Caching, Distributed Proxy Server

1. INTRODUCTION

On the Internet, multimedia objects are stored in the content servers. The clients behind some proxy servers are usually located over a wide area network far from the content server. When clients access multimedia objects from a content server, the content server must have sufficient storage IO bandwidth to retrieve the required objects and allocate enough network resources to deliver the objects to the client [15]. Otherwise, it rejects the client. Thus, popular content server often becomes the bottleneck in delivering multimedia objects.

Proxy servers are usually placed between the clients and the content servers to reduce latency in repeated access and provide firewall protections. They have the disk cache space, network bandwidth, and availability to cache part of the objects for clients, making them good candidates to solve the bottleneck problem. However, large multimedia objects are not cached or only partially cached in current proxy servers, resulting in low cache efficiency.

At present, only a small percentage of Internet traffic delivers multimedia object. When fast optical networks are widely deployed, more clients would be able to access multimedia objects and more multimedia objects will be delivered on the Internet. Thus, the content servers will need to deliver objects to more clients. The contention problem at the content servers would become more severe. Therefore, proxy caches must be enhanced to alleviate the bottleneck problem on popular content servers.

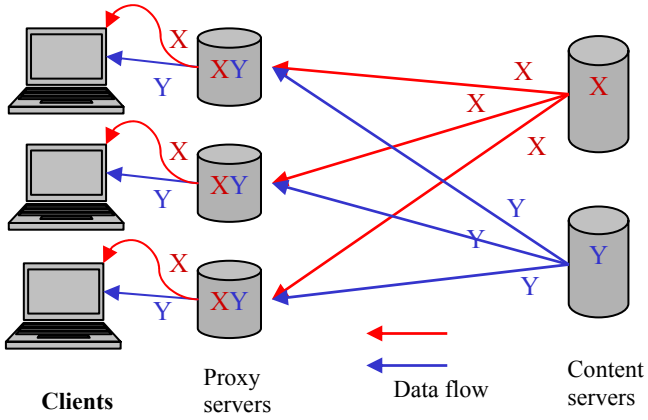

Figure 1. Delivery of multimedia contents over the Internet using proxy servers

This paper gives an overview of the functions of the proxy servers in accessing multimedia objects over the Internet. In the next section, the cache replacement policies being used in proxy servers are described. After that, the object partitioning methods to divide an multimedia object into cacheable and non-cacheable parts are described. Next, the methods that transcode high resolution objects into lower resolution objects are described. Then, we present the cooperative caching method that has been applied to cache objects on proxy servers. Afterwards, we describe a method to distribute proxy server load using a depot. Lastly, we present the summary of this paper.

2. CACHE REPLACEMENT POLICIES

The main contribution or responsibility of proxy servers to the clients is their cache content. The cache content depends on the cost function in the cache replacement policy that determines the cache performance. The cache replacement policy must be optimized to achieve the lowest capacity miss rates on the cache.

In the literature, many cost functions have been proposed and studied for multimedia objects in a single proxy cache environment [1-3, 11, 17, 20, 22, 24, 26]. Traditional cache replacement policies consider recency and frequency in the cost function of the cached objects to replace the coldest object from the cache. Xiang et. al. add the delays into the cost function [26]. Acharya et. al. add resolution size [1]. Paknikar et. al. add the object size and layer number [17]. Wu et. al. [24] and Aggarwal [2] increase the segment length when the position of a segment is far from the beginning of a video. This is advantageous when many users may stop playing the media after only a few initial blocks [2, 24]. Using these cost functions, either the cache hit rate or byte hit rate is optimized for each individual proxy server. In general, the cost function can be described in [3] as

$$\text{CacheValue} = \frac{\left(d_i^{r1} * n_i^{r2}\right)}{\left(t_i^{r3} * s_i^{r4}\right)}, \qquad (1)$$

where d_i is the latency to fetch an object i, n_i is the number of references made to i since it has been brought into the cache, t_i is the last reference time, s_i is the size of object i. r1, r2, r3, r4 are constants with default value r1=0.1 and r2=r3=r4=1. Objects with the smallest cache value will be deleted in order to release storage space for the new objects.

Web prefetching obtains the web data that a client is expected to need on the basis of data about that client's past surfing activity. It reduces access latency by actively preloading data for its clients. The prediction by partial match (PPM) model makes prefetching decisions by reviewing the URLs that similar clients have visited. The model structures these URLs in a Markov predictor tree.

The standard PPM model builds a tree for every visited URL. A fixed threshold is used to limit the length of each prediction branch. In order to achieve accurate prediction on future requests, the tree being built needs to be very large and it takes up too much space.

The longest repeating sequence (LRS) PPM reduces the size of the prediction tree by storing only long branches with frequently accessed URL predictors. The tree size is reduced at the expense of lower prediction accuracy.

The popularity-based (PB) PPM ranks URL's relative popularity (RP) into four grades:

- Grade 3, $10\% < RP \leq 100\%$
- Grade 2, $1\% < RP \leq 10\%$
- Grade 1, $0.1\% < RP \leq 1\%$
- Grade 0, $RP \leq 0.1\%$

It assigns long branches to popular URLs and shorter branches to less popular URLs [7]. Only frequently accessed paths are thus kept to reduce the storage requirement of the predictor tree.

Although web prefetching can reduce the latency in accessing predictable objects, they may incur an overhead of extra traffic due to two reasons: First, the inaccurate prediction of web accesses. Second, the dynamic content may easily become stale and outdated after prefetching.

Caching and web prefecthing at the proxy server reduce the number and latency of accesses for static contents from the web server. Further investigations will be needed to improve prediction accuracy and cache performance on dynamic web contents. The cache value of prefetched objects should also be defined differently from the cache value of the accessed objects to achieve optimal cache efficiency.

3. OBJECT PARTITIONING

If the local proxy server caches a part of the object, the local proxy will directly deliver the cached part and access the missing part from the content server. Although this is done in a transparent way to the client, the position and size of the cached part relative to the requested object's origin and size affects the response time of new streams and the minimum network bandwidth needed to deliver the object.

In the literature, three object partitioning methods, namely Leader, Staging, and Hotspot, have been proposed (Figure 2) and studied for individual proxy servers [9, 21, 27].

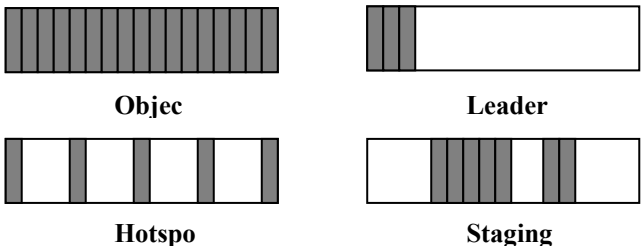

Figure 2. Object partitioning methods

Some research has been done on delivering QoS multimedia streaming by caching the leader (or front part) of the objects [18-19, 21]. In this method, each object is divided into a front part and a rear part. The front part is placed in the proxy cache that is the nearest to the client and the rear part is placed in other LAN proxy servers. This method allows more front parts to be cached close to clients and reduce start up latency. The rear part in other LAN proxy caches could reduce network traffic for repeated accesses from the same LAN.

A hotspot of an object can provide a preview of the object. The size of a hotspot is only a fraction of the object size. It would be useful to provide previews of many objects from local proxy cache. Hotspot caching thus improves storage retrieval efficiency and scalability of a single proxy server [9]. The proxy server's physical proximity to clients reduces the response time to a random seek for previewing a video.

In video staging, the bursty portion of video above certain bandwidth to proxy server is cached in advance to reduce the necessary WAN bandwidth for usage [27]. Heuristic algorithms are proposed to choose the videos to be staged according to the access probability of the video or the highest reduction in WAN bandwidth.

Each of the above methods has its merits. The Leader methods reduce the start up delay in viewing an object. The staged part minimizes the amount of WAN bandwidth. The hotspot method quickly provides a logical preview prior to viewing the object.

Current object partitioning methods only cater for the delivery of cached content to the local clients. When multiple proxy servers are present within a region, it would be helpful if each proxy server caches a different part of the object. The union of the cached contents in multiple proxy servers can thus form a bigger portion of the object and reduce the size of the missing parts that must be retrieved from the server. Further investigation is needed to optimize the object partitioning method for best cache efficiency.

4. TRANSCODING FUNCTIONS

There is much research on enhancing individual proxy servers for multimedia streaming [6, 10, 16, 26]. The use of proxy servers, as a virtual server between the server and the client, was proposed to manage server client connections, data delivery, and transcoding [1, 5, 16]. The proxy server becomes the center of control to handle the client bandwidth limitation, transcoding, and cache management.

One of the main research issues on multimedia object streaming is the provision of QoS delivery. A multimedia stream is a group of periodic requests that are separated from each other with a fixed time interval. In order to support QoS delivery to the client, the network bandwidth at the client needs to be able to support the variable data rate of the object stream. Proxy servers are continuously connected to the Internet at high bandwidth links, but the clients may be connected only with a modem or mobile device. Thus, the proxy server transcodes objects into its lower resolution in order to reduce the object size and data rate requirements at the client.

Transcoding is an irreversible process of adapting the spatial and temporal resolutions of objects (Figure 3) to that of the user specification, client limiations, and wireless device characteristics [4, 10, 12]. A transcoded objects cannot be converted into another object with a higher one in any resolutions. Thus, the transcoding needs to be chosen carefully to avoid repeatedly accessing the original object from the server.

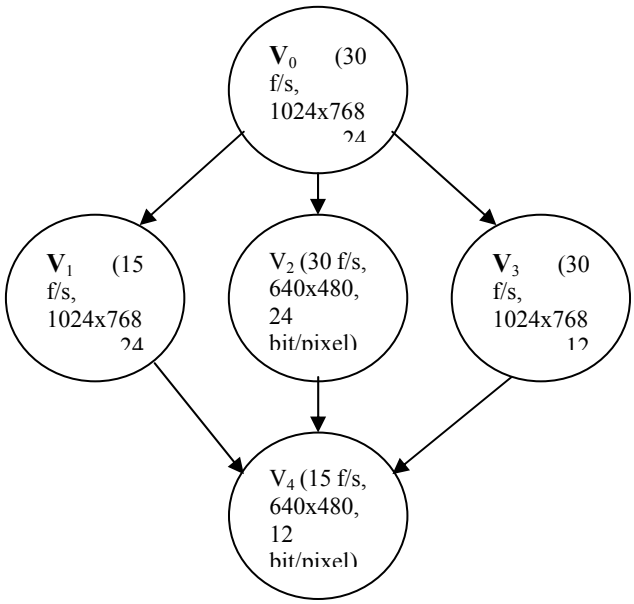

Figure 3. Transcoding is an irreversible process to reduce the object resolution. An object is filtered in one or more of the resolutions.

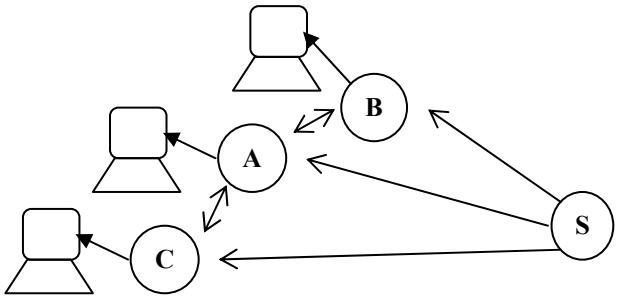

Figure 4. Cooperative proxy servers B and C help content server S in delivering documents to proxy A.

After transcoding, the proxy server should cache the transcoded objects only. Otherwise, repeated transcoding can also erode the processing power of the proxy server. Since different clients at different network link capacity may access from the same server, the proxy server may need to cache different transcoded objects for different clients. Further investigation will be needed to selectively cache different transcoded objects in order to improve the overall cooperative cache performance.

5. COOPERATIVE CACHING MECHANISMS

The only found research work in cooperative caching of multimedia objects is an investigation of the Leader method in the hierarchical proxy server [18]. As there are no other works on caching multimedia objects in multiple proxy servers, we shall describe the caching web documents on cooperative proxy servers.

When multiple proxy servers cooperate to serve client requests, they must have capabilities to exchange their cached contents. This cooperation mechanism may induce overhead that could erode its benefits. Thus, the cooperation mechanisms must be carefully designed to facilitate quality of service (QoS) retrievals from other proxy servers at minimal overhead.

In the literature, three proxy cooperation mechanisms, namely hierarchical, directory-based, and hash-based, have been proposed (Figure 4) and investigated for Web documents [4, 8, 13-14, 18-19, 23, 25]. The cooperation mechanism involves some overhead that affects the cache performance significantly.

The hierarchical approach builds a proxy tree with one parent proxy server above a number of child proxy servers [4, 18-19]. Each level of parent proxy keeps some contents that are commonly accessed by its child proxies. A proxy server forwards its cache misses through its parents and finally to the content server. A significant amount of extra cache space is needed in the parent proxies to store the commonly accessed cache contents. When the proxy tree is deep, the miss penalty (extra time in searching through the parent proxies) erodes the reduction in miss rate.

The proxy servers in the directory-based approach exchange their directory contents with other cooperative proxy servers [4, 18-19]. The proxy servers thus need extra space to store the directory of other proxy caches and extra bandwidth to exchange the directory contents.

In the hash-based approach, each document is assigned to one of the cooperative proxy servers [14, 23]. If a document is fetched, it is either cached in the assigned proxy server or not cached at all. Thus, this approach minimizes the amount of cache space and the message overheads, but it involves many false positive hits.

Cooperative proxy caching has been shown to be positive for Web document retrieval. The investigations on Web documents is however limited to a regional level. It is expected that cooperative proxy caching becomes more appealing for streaming multimedia objects [23].

Individual proxy cache may not be large enough to cache several multimedia objects, but the aggregate size of multiple proxy servers can be large enough to cache many popular objects. Thus, cooperative caching can be used to enhance the cache performance on multimedia objects.

Cooperative caching of multimedia objects will affect the local hit ratio and byte hit ratio of proxy caches. Each proxy server may decide the percentage of cache for its local clients and for the cooperative proxy servers. This percentage can be adjusted by tuning the cache replacement function.

6. DISTRIBUTED PROXY SERVERS

Another approach to use multiple proxy servers at a client is to add a depot between the client and the proxy servers [13]. This depot distributes and balances the load by scheduling the TCP sessions via proxy servers in a round-robin manner as shown in Figure 5. Two proxy servers with different access speeds were investigated in [13]. The performance of the proxy system falls in-between of two stand-alone proxy servers and is not optimal.

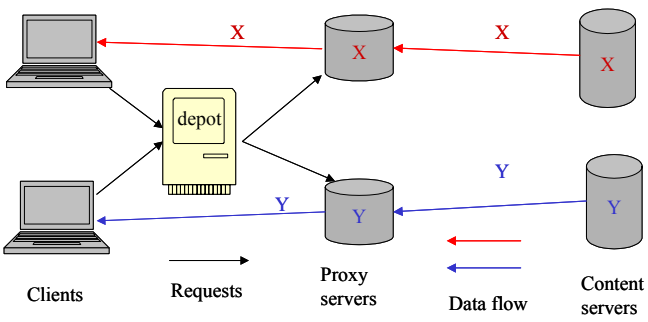

Figure 5. A depot between clients and proxy servers in distributed proxy servers.

7. SUMMARY

We have given a brief overview of the functions of multimedia proxy servers in this paper. Proxy servers are becoming the centre to manage the delivery of multimedia contents from servers to clients. Their involvement can increase or decrease dynamically according to the number of active clients. They cache parts of object that in a cooperative manner to minimize the number of streams at the servers. They adapt the object resolutions to the client and network characteristics. They cooperate with each other to reduce the load at the content servers and on the network.

8. ACKNOWLEDGEMENT

This research is supported by the Areas of Excellence IT Scheme established under the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E/01/99).

9. REFERENCES

[1] S. Acharya et.al., "Systematic Multiresolution and its Application to the World Wide Web," **Proc. of 15th International Conference on Data Engineering**, 1999, pp. 40-49.

[2] C. Aggarwal, "On Disk Caching of Web Objects in Proxy Servers," **ACM CIKM**, 1997, pp. 238-245.

[3] H. Bahn *et. al.*, "Efficient Replacement of Nonuniform Objects in Web Caches," **IEEE Computer**, Vol. 35, No. 6, 2002, pp. 65-73.

[4] V. Cardellini *et. al.*, "Collaborative Proxy System for Distributed Web Content Transcoding," **Proc. ACM CIKM**, 2000, pp.520-527.

[5] S. Chandra *et. al.*, "Multimedia Web Services for Mobile Clients Using Quality Aware Transcoding," **ACM WoWMoM**, Seattle, WA, USA, 1999, pp. 99-108.

[6] Y. Chang and N.C. Hock, "Providing Quality of Service Guarantee in Internet by a Proxy Method," **Proceedings of IEEE TENCON**, Vol. 3, 2000, pp. 51-54.

[7] X. Chen and X. Zhang, "A Popularity-based Prediction Model for Web Prefetching," **IEEE Computer**, Vol. 36, No. 3, March 2003, pp. 63-70.

[8] S. Dykes and K.A. Robbins, "A Viability Analysis of Cooperative Proxy Caching," **Proc. IEEE INFOCOM**, Vol. 3, 2001, pp. 1205-1214.

[9] H. Fahmi *et. al.*, "Proxy Servers for Scalable Interactive Video Support," **IEEE Computer**, Vol. 34, No. 9, 2001, pp. 54-60.

[10] K. Ham *et. al.*, "Wireless-adaptation of WWW Content over CDMA," **IEEE International Workshop on Mobile Multimedia Communications**, 1999, pp. 368-372.

[11] S. Hosseini-Khayat, "Replacement Algorithms for Object Caching," **ACM Symposium on Applied Computing**, 1998, pp. 90-97.

[12] A. Kassler *et. al.*, "Filtering Wavelet based Video Streams for Wireless Inter-working," **IEEE International Conference on Multimedia and Expo**, vol. 3, 2000, pp. 1257-1260.

[13] K. Law *et. al.*, "A Scalable and Distributed WWW Proxy System," **IEEE Intl Conf. on Multimedia Computing and Systems**, 1997, pp. 565-571.

[14] K. Lee *et. al.*, "On the Sensitivity of Cooperative Caching Performance to Workload and Network Characteristics," **ACM SIGMETRICS**, Vol. 30, No. 1, 2002, pp.268-269.

[15] H. Ma and K.G. Shin, "Multicast Video-on-demand Services," **ACM SIGCOMM Computer Communication Review**, Vol. 32, No. 1, 2002, pp.31-43.

[16] K.D. Nam and H. Lee, "Design of a Virtual Server for Service Internetworking over Heterogeneous Networks," **IEEE Pacific Rim Conference on Networking**, Vol. 1, 1997, pp. 406-409.

[17] S. Paknikar *et. al.*, "A Caching and Streaming Framework for Multimedia," **ACM Multimedia Conference**, 2000, pp. 13-20.

[18] S. Park *et. al.*, "A Proxy Server Management Scheme for Continuous Media Objects Based on Object Partitioning," **IEEE ICPADS**, 2001, pp. 757-762.

[19] Y.W. Park, K.H. Baek, and K.D. Chung, "Reducing Network Traffic Using Two-layered Cache Servers for Continuous Media Data on the Internet," **IEEE Computer Software and Applications Conference**, 2000, pp. 389-394.

[20] S. Sohoni *et. al*., "A Study of Memory System Performance of Multimedia Applications," **ACM SIGMETRICS**, 2001, pp. 206-215.
[21] R. State and O. Festor, "Active Network based management for QoS assured multicast delivered media," **IEEE ICATM**, 2001, pp.123-127.
[22] Z. Su *et. al*., "A Prediction System for Multimedia Pre-fetching in Internet," **ACM Multimedia Conference**, 2000, pp.p3-12.
[23] A. Wolman *et. al*, "On the Scale and Performance of cooperative Web proxy caching," **ACM Symposium on Operating Systems Principles**, Vol. 34, No. 5, Dec. 1999, pp. 16-31.
[24] Kun-Lung Wu *et. al*., "Segment-Based Proxy Caching of Multimedia Streams," **ACM WWW10 Conference**, 2001, pp. 36-44.
[25] S. Wu and C.C. Liao, "Virtual Proxy Servers for WWW and Intelligent Agents on the Internet," **Proc. of the Thirtieth Hwaii Intl Conf on System Sciences**, Vol. 4, 1997, pp. 200-209.
[26] Z. Xiang *et. al*., "Cost-Based Replacement Policy for Multimedia Proxy across Wireless Internet," **IEEE Global Telecommunications Conference**, Vol. 3, 2001, pp. 2009-2013.
[27] Z. Zhang *et. al*., "Video Staging: A Proxy-Server-Based Approach to End-to-End Video Delivery over Wide-Area Networks," **IEEE/ACM Trans. On Networking**, Vol. 8, No. 4, August 2000, p.429-442.