Chapter 1

STREAMING MEDIA CACHING

Jiangchuan Liu
School of Computing Science
Simon Fraser University, British Columbia, Canada
jcliu@cs.sfu.ca

Abstract Streaming media has contributed to a significant amount of today's Internet traffic. Like conventional web objects (e.g., HTML pages and images), media objects can benefit from proxy caching; yet their unique features such as huge size and high bandwidth demand imply that conventional proxy caching strategies have to be substantially revised. This chapter discusses the critical issues and challenges of cache management for proxy-assisted media streaming. We survey, classify, and compare the state-of-the-art solutions. We also investigate advanced issues of combining multicast with caching, cooperating among proxies, and leveraging proxy caching in overlay networks.

Keywords: Streaming Media, Video, Proxy, Caching, Overlay Networks, Multicast

Introduction

With widespread penetration of the broadband Internet, multimedia service is getting increasingly popular among users and has contributed to a significant amount of today's Internet traffic. Media objects can be accessed similar to conventional text and images using a download-and-play mode; but most users prefer to quickly initiate and then continuously play back a media object while it is being downloaded, i.e., to use a *real-time streaming* mode. We have witnessed the initial and incremental deployment of streaming applications like RealNetworks RealPlayer and Microsoft Windows Media Player in recent years. The performance of such applications however is still far from satisfactory, especially during the peak hours.

To reduce client-perceived access latencies as well as server/network loads, an effective means is to cache frequently used data at proxies close to clients. It also enhances the availability of objects and mitigates packet losses, as a

local transmission is generally more reliable than a remote transmission. Proxy caching thus has become one of the vital components in virtually all web systems. Streaming media, particularly those pre-stored, could also benefit significant performance improvement from proxy caching, given their static nature in content and highly localized access interests. However, existing proxies are generally optimized for delivering conventional web objects (e.g., HTML pages or GIF images), which may not meet the requirements of steaming applications. In the following, we list some important and unique features of streaming media and discuss their implications to proxy cache design.

Huge size: A conventional web object is typically on the order of 1K to 100K bytes. Hence, a binary decision works well for proxy caching: either caching an object in its entirety or not caching. In contrast, a media object has a high data rate and a long playback duration, which combined yield a huge data volume. For illustration, a one-hour standard MPEG-1 video has a volume of about 675 MB; caching it entirely at a web proxy is clearly impractical, as several such large streams would exhaust the capacity of the cache. One solution is to cache only portions of an object. In this case, a client's playback needs a joint delivery involving both the proxy and the origin server. To cache which portions of which objects thus has to be carefully managed, such that the benefit of caching outweighs the synchronization overhead of the joint delivery.

Intensive bandwidth use: Streaming nature of delivery requires a significant amount of disk and network I/O bandwidth, sustaining over a long period. Hence, minimizing bandwidth consumption becomes a primary consideration for proxy cache management, even taking precedence over reducing access latencies in many cases. Moreover, the bandwidth bottleneck limits the number of clients that a proxy can simultaneously support; employing multicast delivery and cooperation among proxies thus become particularly attractive for media streaming applications.

High interactivity: The long playback duration of a streaming object also enables various client-server interactions. As an example, recent studies found that nearly 90% media playbacks are terminated prematurely by clients [Chen et al., 2004]. In addition, during a playback, a client often expects VCR-like operations, such as fast-forward and rewind. This implies the access rates might be different for different portions of a stream, which potentially complicates the cache management.

Given these unique features of media objects, novel caching algorithms have been developed in the literature. The objective of this chapter is to review the state-of-the-art caching techniques dedicated to streaming media caching. We begin with discussions on a generic proxy caching architecture and some protocol considerations. The caching strategies for streaming media are classified, examined, and compared in Section 3. Section 4 investigates some advanced issues. Finally, Section 5 concludes the article.

1. Architecture and Protocols for Streaming Caching

Streaming applications generally support diverse client-server interactions and have stringent demands on packet delay and jitter to ensure discontinuity-free playback. To meet these requirements, the Internet Engineering Task Force (IETF) has developed the RTP/RTCP/RTSP protocol suite. A generic system diagram of proxy-assisted media streaming using this suite is depicted in Fig. 1.1.

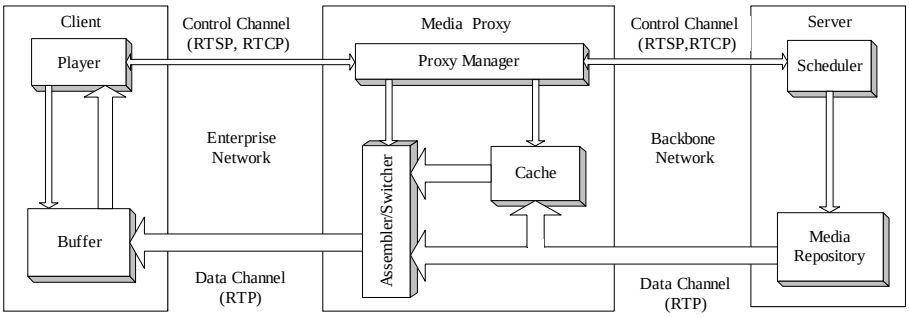

Figure 1.1. A generic system diagram of proxy-assisted media streaming using RTP/RTCP/RTSP.

In this system, the basic functionalities for data transferring are provided by the Real-Time Transport Protocol (RTP), including payload identification, sequence numbering for loss detection, and time stamping for playback control. Running on top of UDP, RTP itself does not guarantee Quality-of-Service (QoS), but relies on its companion, the Real-Time Control Protocol (RTCP), to monitor the network status and provide feedback for application-layer adaptation. The Real-Time Streaming Protocol (RTSP) coordinates the delivery of media objects and enables a rich set of controls for interactive playback. For the proxy-assisted streaming, the proxy has to relay these control messages between the client and the server. The problem is particularly involved if only part of a media object is cached at a proxy. In this case, the proxy must reply to the client PLAY request and initiate transmission of RTP and RTCP messages to the client for the cached portion, while request the uncached portion(s) from the server. Such fetching can be achieved through an RTSP Range request specifying the playback points, as illustrated in Fig. 1.2. The Range request also enables clients to retrieve different segments of a media object from multiple servers or proxies, if needed.

Beside this classical client/server paradigm, peer-to-peer streaming and other overlay streaming paradigms have also attracted much attention recently, which will be discussed in Section 4.C.

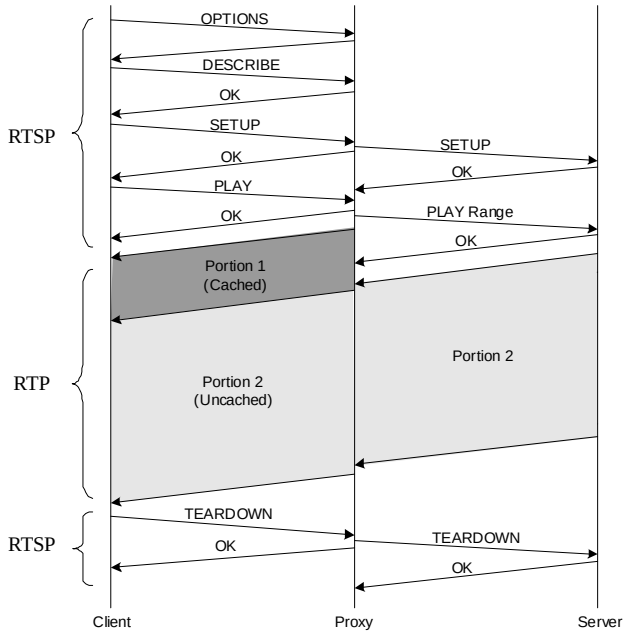

Figure 1.2. Operations for streaming with partial caching.

2. Caching Strategies: Homogeneous Clients and Heterogeneous Clients

Due to the aforementioned features of streaming media objects, media caching has many distinct focuses from conventional web caching. On one hand, since the content of a media object is rarely updated, management issues like cache consistency and coherence are less critical in media caching. On the other hand, given high resource requirements of media objects, effective management of proxy cache resources (i.e., space, disk I/O, and network I/O) becomes more challenging. In this section, we survey the state-of-the-art media caching strategies for both homogenous clients and heterogeneous clients, with an emphasis on how the strategies minimize resource demands.

Stream Caching for Homogeneous Clients

Most existing caching algorithms focus on homogeneous clients, which have identical or similar configurations and capabilities behind a proxy. As such, a single version of an object would match the bandwidth and format demands of all requests to the object. Nevertheless, what to cache (which portions of which objects) and how to manage cache (e.g., cache placement and replacement) at

the proxy remain challenges. According to the selection of the portions to cache, we classify existing algorithms into four categories: *sliding-interval caching*, *prefix caching*, *segment caching*, and *rate-split caching*.

Sliding-Interval Caching [Tewari et al., 1998, Chen et al., 2004]: This algorithm caches a sliding interval of a media object to exploit sequential access of streaming media. For illustration, given two consecutive requests for the same object, the first request may access the object from the server and incrementally store it into the proxy cache; the second request can then access the cached portion and release it after the access. If the two requests arrive close in time, only a small portion of the media object needs to be cached at any time instance, and yet the second request can be completely satisfied from the proxy (see Fig. 1.3). In general, if multiple requests for an object arrive in a short period, a set of adjacent intervals can be grouped to form a *run*, of which the cached portion will be released only after the last request has been satisfied.

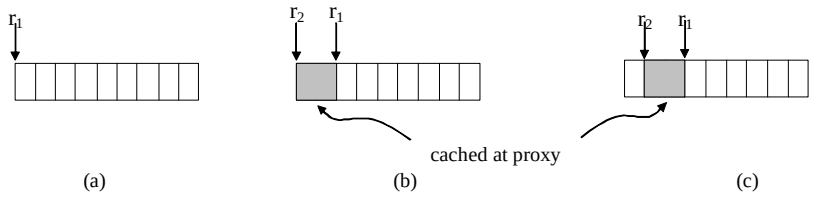

Figure 1.3. An illustration of sliding-interval caching. The object consists of 9 frames, each requiring one unit time to deliver from the proxy to a client. Requests 1 and 2 arrive at times 0 and 2, respectively. To serve request 2, only two frames need to be cached at any time instance. (a) Time 0: request 1 arrives; (b) Time 1-2: frames 1 and 2 accessed by request 1 and cached; request 2 arrives; (c) Time 2-3: frame 3 accessed by request 1 and cached; frame 1 read by request 2 and released.

As the cached portion is dynamically updated with playback, the sliding-interval caching involves high disk bandwidth demands; in the worse case, it would double the disk I/O due to the concurrent read/write operations. To effectively utilize available cache resources, [Tewari et al., 1998] proposed a resource based caching (RBC) policy. The policy characterizes each object by its space and bandwidth requirements, and models the cache as a two-constraint knapsack. A heuristic algorithm was developed to dynamically select the caching granularity of an object with the objective of balancing its bandwidth and space usages. Depending on the object's characteristics and the available resources, the selected granularity could be a sliding interval, a sliding run, or the full object.

Sliding-interval caching can significantly reduce network bandwidth consumption and start-up delay for subsequent accesses. However, as the cached

portion is dynamically updated with playback, the sliding-interval caching involves high disk bandwidth demands; in the worse case, it would double the disk I/O due to the concurrent read/write operations. [Chen et al., 2004] proposed the Shared Running Buffer (SRB) approach, which argues that, with the falling price of memory, it is possible to allocate memory buffers to accommodate media data and thus avoiding the intensive disk read/write.

The effectiveness of sliding-interval caching diminishes with the increase of the access intervals. If the access interval of the same object is longer than the duration of the playback, the algorithm is degenerated to the unaffordable full-object caching. To address this issue, it is preferable to retain the cached content over a relatively long time period. Most of the caching algorithms to be discussed in the rest of this section fall into this category.

Prefix Caching [Sen et al., 1999]: This algorithm caches the initial portion of a media object, called *prefix*, at a proxy. Upon receiving a client request, the proxy immediately delivers the prefix to the client and, meanwhile, fetches the remaining portion, the *suffix*, from the server and relays to the client (see Fig. 1.4). As the proxy is generally closer to the clients than the origin server, the start-up delay for a playback can be remarkably reduced.

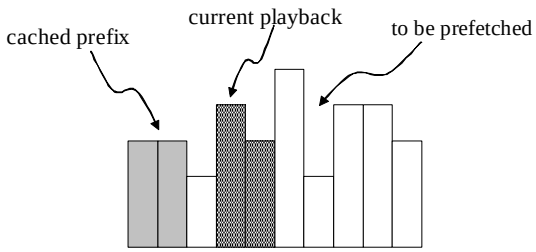

Figure 1.4. A snapshot of prefix caching with workahead smoothing.

To ensure discontinuity-free playback with a start-up delay of s, the proxy has to store a prefix of length $max(d_{max} - s, 0)$, where d_{max} is the maximum delay from the server to the proxy. If cache space is abundant, the proxy can also devote some space to assist in performing workahead smoothing for variable-bit-rate (VBR) media [Sen et al., 1999]. With this smoothing cache, the proxy can prefetch large frames in advance of each burst to absorb delay jitter and bandwidth fluctuations of the server-to-proxy path. The delay of prefetching can be hided by the prefix caching. Similar to sliding-interval caching, the content of the smoothing cache is dynamically updated with playback. However, the purposes are different: the former is to improve cache hit for subsequent requests, while the latter is to facilitate workahead smoothing.

Segment Caching [Chen et al., 2004, Wu et al., 2001, Miao and Ortega, 2002, Fahmi et al., 2001]: Segment caching generalizes the prefix caching

paradigm by partitioning a media object into a series of segments, differentiating their respective utilities, and making caching decision accordingly (see Fig. 1.5). Various segment caching algorithms have been proposed in the literature by employing different segmentations and utility calculations. [Wu et al., 2001] suggested grouping the frames of a media object into variable-sized segments, with the length increasing exponentially with the distance from the start of the media, i.e., the size of segment i is 2^{i-1} , which consists of frames $2^{i-1}, 2^{i-1} + 1, ..., 2^i - 1$. The motivation is that a proxy can quickly adapt to the changing access patterns of cached objects by discarding big chunks as needed. The utility of a segment is calculated as the ratio of the segment reference frequency over its distance from the beginning segment, which favors to cache the initial segments as well as those with higher access frequencies. [Chen et al., 2004], however, argued that neither the use of a predefined segment length nor the favorable caching of the initial segments is the best strategy for reducing network traffic. They suggested postponing segmentation as late as possible (called *lazy segmentation*), thus allowing the proxy to collect a sufficient amount of access statistics to improve the effectiveness.

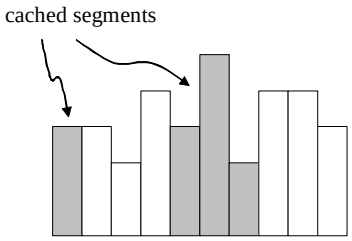

Figure 1.5. An illustration of segment caching.

A salient feature of segment-based caching is its support to VCR-like operations, such as random access, fast-forward, and rewind. As an example, [Fahmi et al., 2001] proposed to cache some key segments of a media object, called *hotspots*, which are identified by content providers. When a client requests the object, the proxy first delivers the hotspots to provide an overview of the stream; the client can then decide whether to play the entire stream or quickly jump to some specific portion introduced by a hotspot. Furthermore, in fast-forwarding and rewinding operations, only the corresponding hotspots are delivered and displayed, while other portions are skipped. As such, the load of the server and backbone network can be greatly reduced, but the client will not miss any important segments in the media object.

Rate-Split Caching [Zhang et al., 2000]: While all the aforementioned caching algorithms partition a media object along the time axis, the rate-split caching partitions a media along the rate axis: the upper part will be cached

at the proxy, whereas the lower part will remain stored at the origin server (see Fig. 1.6). This type of partitioning is particularly attractive for VBR streaming, as only the lower part of a nearly constant rate has to be delivered through the backbone network. For a QoS network with resource reservation, the bandwidth reserved should be equal to the peak rate of a stream; caching the upper part at the proxy clearly reduces the rate variability and improves the backbone bandwidth utilization. A critical issue here is how to select the cut-off rate or, equivalently, the size of the upper part for caching. [Zhang et al., 2000] studied the impact of the cut-off rate for a single stream through empirical evaluation, and found that a significant bandwidth reduction can be achieved with a reasonably small cache space. They also formulated the multiple-stream case as a knapsack problem with two constraints: disk bandwidth and cache space, and developed several heuristics, e.g., caching popular objects only, or caching those with high bandwidth reduction.

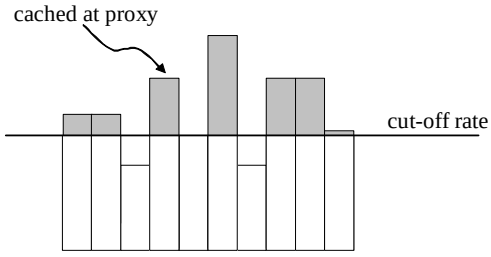

Figure 1.6. An illustration of rate-split caching.

Summary and Comparison: Tab. 1.1 summarizes the caching algorithms reviewed for homogeneous clients. While these features and metrics provide a general guideline for algorithm selection, the choice for a specific streaming system also largely depends on a number of practical issues, in particular, the complexity of the implementation. In fact, only a few simple algorithms have been employed in commercial systems, though recently-built prototypes have practically demonstrated the viability and superiority of the intelligent algorithms, such as lazy segmentation [Chen et al., 2004].

In addition, we emphasize that these algorithms are not necessarily exclusive with each other, and a combination of them may yield a better performance. For example, segment caching combined with prefix caching of each segment can reduce start-up latency for VCR-like random playback from any key-segment. Combination with conventional data caching algorithms has also been examined.

		Sliding-interval caching	Prefix caching	Segment caching	Rate-split caching
Cached portion		Sliding intervals	Prefix	Segments	Portion of higher rate
VCR-like support		No	No	Yes	No
Resource demand	Disk I/O	High	Moderate	Moderate	Moderate
	Disk space	Low	Moderate	High	High
	Sync overhead	Low	Moderate	High	High
Performance improvement	Bandwidth reduction	High*	Moderate	Moderate	Moderate
	Start-up latency reduction	High*	High	High**	Moderate

* There is no reduction for the first request in a run.
** Assume the initial segment is cached.

Table 1.1. Comparison of the caching algorithms for homogeneous clients

Stream Caching for Heterogeneous Clients

Owing to diverse network models and device configurations, clients behind the same proxy often have quite different requirements on the same media object, in terms of streaming rates or encoding formats. To accommodate such heterogeneity, a straightforward solution is to produce replicated streams of different rates or formats, each targeting on a subset of clients. Though being widely used in commercial streaming system, the storage and bandwidth demands of this approach can be prohibitively high [Liu et al., 2004]. An alternative is to transcode a media from one form to another of a lower rate or a different encoding format in an on-demand fashion [Tang et al., 2002]. The intensive computation overhead of transcoding however prevents a proxy from supporting a large, diverse client population.

Yet a more efficient approach to this problem is the use of layered encoding and transmission. A layered coder compresses a raw media object into several layers: the most significant layer, called *base layer*, contains the data representing the most important features of the object, while additional layers, called *enhancement layers*, contain the data that can progressively refine the quality. A client thus can subscribe to a subset of cumulative layers to reconstruct a stream commensurate with its capability. For layered caching, [Kangasharju et al., 2002] assumed that the cached portions are semi-static and only completed layers are cached. To maximize the total revenue, they developed effective heuristics based on an analytical stochastic knapsack model to determine the cache content. In their model, the client population and the distribution of their capacities are known *a priori*. For layered streaming under

dynamic conditions, [Rajaie et al., 2000] studied segment-based cache replacement and prefecting policies to achieve efficient utilization of cache space and available bandwidth (see Fig. 1.7a). The main objective is to deal with the congestion problem for individual clients. To this end, the proxy keeps track of popularities of each object on a per layer basis. When the quality of the cached layers is lower than the maximum deliverable quality to an interested client, the proxy sends requests to the server for missing segments within a sliding prefetching window. On cache replacement, a victim layer is identified based on popularities, and its cached segments are flushed from the tail until sufficient space is obtained.

A critical drawback of the existing layered streaming systems is that the number of layers is pretty small, typically 2 or 3 only; hence, their adaptation granularity remains coarse. Fortunately, recent development in the coding area has demonstrated the possibility of fine-grained post-encoding rate control. An example is the MPEG-4 Fine-Grained Scalable (FGS) coder with bitplane coding, which generates embedded streams containing several bitplanes and each can be partitioned at any specific rate. As such, for narrowband clients, the proxy can reduce the streaming rate using a bitplane filter; for wideband clients, the proxy can fetch some uncached portion (i.e., higher-order bitplanes) from the server and assemble it with the cached portion to generate a high-rate stream. As illustrated in Fig. 1.7b, the available bandwidth of a client can be almost fully utilized, and, more importantly, both the filtering and the assembling operations in FGS can be done with fast response. Hence, we envision the FGS-based streaming and caching as a very promising solution to media steaming over the Internet comprising highly heterogeneous end-systems. Several caching algorithms have been proposed to minimize the bandwidth consumption and/or improve the client utility [Liu et al., 2004].

FGS video also facilitates QoS or quality-based cache replacement. As an example, [Yu et al., 2003] showed that differentiating the utilities of different video blocks in cache replacement can noticeably reduce the end-to-end quality distortion. The use of other advanced scalability tools, in particular, MPEG-4 object scalability, has been investigated in [Schojer et al., 2004, Schojer et al., 2003, Podlipnig and Boeszoermenyi, 2002] as well. They have implemented QBIX-G, a Quality Based Intelligent proXy Gateway, which performs both caching and filtering functionalities. It acts as a general broker accommodating heterogeneous user requirements and video variations. In addition, the MPEG-7 and MPEG-21 standards are employed to ensure inter-operability. The whole system enables a desired What You Need is What You Get (WYNIWYG) video services, which delivers videos to users exactly with the quality they expected.

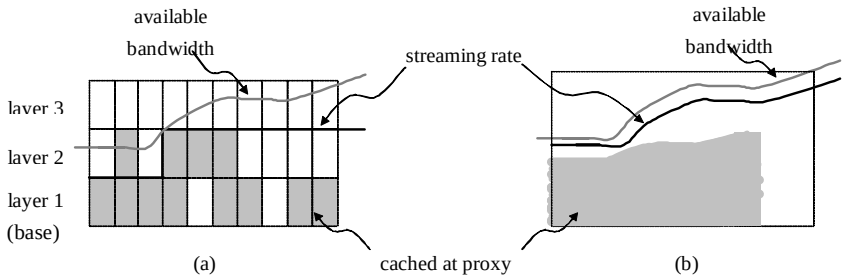

Figure 1.7. Caching for layered streaming. (a) Coarse-grained layering; (b) Fine-grained layering.

3. Advanced Issues

So far we consider a standalone proxy with only unicast delivery. While it can significantly reduce the access latencies and backbone bandwidth demands, the scalability and robustness of this simple architecture are still restricted. We now discuss two effective enhancements: multicast and proxy cooperation; we also address the role of proxy in the recently popularized overlay communication paradigms.

Combining Proxy Caching with Multicasting

Like caching, multicasting also explores the temporal locality of client requests. Specifically, it allows a media server to accommodate concurrent client requests with shared channels through batching, patching, or periodic broadcast. However, multicast delivery suffers from two important deficiencies. First, to save more bandwidth, it is better to accommodate more requests in one multicast channel by using a large batching/patching window, which however leads to long start-up latencies. Second, while IP multicast is enabled in virtually all local-area networks, its deployment over the global Internet remains limited in scope and reach. Hence, it is unlikely that a multicast streaming protocol can be used for geographically dispersed servers and clients.

Interestingly, both deficiencies can be alleviated through the use of proxies. Specifically, a request can be instantaneously served by a cached prefix while waiting for the data from a multicast channel [Ramesh et al., 2001, Wang et al., 2002], and proxies can bridge unicast networks with multicast networks, i.e., employing unicast for server to proxy delivery while batching and/or patching local accesses. [Wang et al., 2002] have derived the optimal length of the prefix to be cached for most typical multicast protocols, and showed that a careful

coupling of caching and multicasting can produce significant cost savings over using the unicast service, even if IP multicast is supported only at local networks.

Cooperative Proxy Caching

In general, proxies grouped together can achieve better performance than independent standalone proxies. Specifically, the group of proxies can cooperate with each other to increase the aggregate cache space, balance loads, and improve system scalability [Acharya and Smith, 2000, Chae et al., 2002, Hofmann et al., 1999]. A typical cooperative media caching architecture is MiddleMan [Acharya and Smith, 2000], which operates a collection of proxies as a scalable cache cluster. Media objects are segmented into equal-sized segments and stored across multiple proxies, where they can be replaced at a granularity of a segment. There are also several local proxies responsible to answer client requests by locating and relaying the segments. Note that, in cooperative web caching, a critical issue is how to efficiently locate web pages with minimum communication costs among the proxies. This is, however, not a major concern for cooperative media caching, as the bandwidth consumption for streaming objects is of orders of magnitude higher than that for object indexing and discovering. Consequently, in MiddleMan, a centralized coordinator works well in keeping track of cache states. On the other hand, while segment-based caching across different proxies facilitates the distribution and balance of proxy loads, it incurs a significant amount overhead for switching among proxies to reconstruct a media object. To reduce such effects as well as to achieve better load balance and fault tolerance, [Chae et al., 2002] suggested a Silo data layout, which partitions a media object into segments of increasing sizes, stores more copies for popular segments, but still guarantees at least one copy stored for each segment.

Streaming Caching in Overlay Networks

So far, we have focused on the client/server paradigm for media streaming, and proxies act as intermediaries between them. Generalizing the proxy functionalities into every end-host will shift the system to the recently popularized overlay communication paradigms, such as peer-to-peer communication or application-layer multicast. There have been many pioneering efforts on overlay streaming, which have demonstrated the superior scalability and deployability of these overlay systems; as an example, DONet/CoolStreaming [Zhang et al., 2004], an experimental overlay streaming software, has attracted over 30000 distinct user (in terms of unique IP addresses) with over 4000 being simultaneously online. The enormous buffer capacities distributed in end-hosts also enable efficient client-side caching and sharing to improve content avail-

ability as well as to support asynchronous streaming [Xu et al., 2002, Jin and Bestavros, Padmanabhan et al., 2002, Cui et al., 2004].

Nevertheless, we are aware that, in contrast to the reliable and dedicated servers or proxies, the loosely-coupled autonomous end-hosts can easily crash, leave without notice, or even refuse to share its own data. Given that a media playback lasts a long time and consumes huge resources, we believe dedicated proxies will still play an important role in building high-quality media streaming systems; in particular, strategically placed proxies may effectively assist the construction and maintenance of large-scale overlays. On the other hand, we may also leverage the overlay paradigm in proxy design. As an example, [Guo et al., 2004] suggested a proxy and its clients be structured into a peer-to-peer system to collaboratively serve local streaming requests. Their work focused on local area collaboration. [Ip et al., 2004] extended it to a two-level streaming overlay, called Cooperative Proxy-Client Caching System, or COPACC. In COPACC, a cluster of proxies forms an overlay in wide-area networks while each proxy collaborates with local client caches to form a local overlay. These two overlays well complement each other: The proxy-level overlay provides a dedicated storage and reliable service, while the local overlay provides a scalable storage for caching and reduces the load of the proxy. COPACC embeds an efficient indexing and searching algorithm for video contents cached across different proxies or clients, as well as a signature verification mechanism, which can effectively identify and block malicious clients. It also makes effective use of multicast delivery in local regions to further reduce the cost of the system.

4. Summary

Proxy caching is an effective means to reduce access latencies as well as resource consumptions for networked applications. Due to the unique features of media objects like huge size and high bandwidth demand, a number of novel streaming caching solutions have been reported in the literature. This article serves as a pioneer survey to this field, though it by no means covers all aspects. Plenty of research issues have yet to be addressed, e.g., caching over the wireless mobile Internet, for large-scale dynamic overlays, and with advanced video coding/indexing schemes such as multiple description coding [Padmanabhan et al., 2002] and MPEG-7/21 standards, as well as security and privacy for cached media objects, to name but a few. We envision that streaming media caching remains a fertile area, and both theoretical and practical solutions to the listed problems are urged with rising demands on ubiquitous multimedia services throughout the world.

References

Chen, S., Shen, B., Wee, S., and Zhang, X. (2004). Designs of high quality streaming proxy systems. In *Proceedings of of IEEE INFOCOM'04*, Hong Kong.

Tewari, R., Vin, H. M., Dan, A., and Sitaram, D. (1998). Resource-based caching for Web servers. In *Proceedings of SPIE/ACM Conf. on Multimedia Computing and Networking (MMCN'98)*, San Jose, CA.

Sen, S., Rexford, J., and Towsley, D. (1999). Proxy prefix caching for multimedia streams. In *Proceedings of IEEE INFOCOM'99*, New York, NY.

Wu, K. L., Yu, P. S., and Wolf, J. L. (2001). Segment-based proxy caching of multimedia streams. In *Proceedings of World Wide Web Conference (WWW10)*, Hong Kong.

Miao, Z., and Ortega, A. (2002). Scalable proxy caching of video under storage constraints. *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1315-1327, Sep. 2002.

Fahmi, H., Latif,M., Sedigh-Ali, S., Ghafoor, A., Liu, P., and Hsu, L. (2001). Proxy servers for scalable interactive video support. *IEEE Computer*, 43(9): 54-60.

Zhang, Z.-L., Wang, Y., Du, D., and Su, D. (2000). Video staging: A proxy-server-based approach to end-to-end video delivery over wide-area networks. *IEEE/ACM Transactions on Networking*, 8(4): 429-442.

Liu, J., Chu, X., and Xu, J. (2004). Proxy Cache Management for Fine-Grained Scalable Video Streaming *Proceedings of IEEE INFOCOM'04*, Hong Kong.

Tang, X., Zhang, F., and Chanson, S. T. (2002). Streaming media caching algorithms for transcoding proxies. In *Proceedings of 31st International Conference on Parallel Processing (ICPP'02)*.

Kangasharju, J., Hartanto, F., Reisslein, M., and Ross, K. W. (2002). Distributing layered encoded video through caches. *IEEE Transactions on Computers*, 51(6), pp. 622-636.

Rejaie, R., Yu, H., Handley, M., and Estrin, D. (2000). Multimedia proxy caching mechanism for quality adaptive streaming applications in the Internet. In *Proceedings of IEEE INFOCOM'00*, Tel Aviv, Israel.

Ramesh, S., Rhee, I., and Guo, K. (2001). Multicast with cache (Mcache): An adaptive zero-delay video-on-demand service. In *Proceedings of IEEE INFOCOM'01*, Anchorage, AK.

Wang, B., Sen, S., Adler, M., and Towsley, D. (2002). Optimal proxy cache allocation for efficient streaming media distribution. In *Proceedings of IEEE INFOCOM'02*, New York, NY.

Acharya, S. and Smith, B. C. (2000). Middleman: A video caching proxy server. In *Proceedings of 10th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'00)*.

Chae, Y., Guo, K., Buddhikot, M. M., Suri, S., and Zegura, E. W. (2002). Silo, rainbow, and caching token: Schemes for scalable, fault tolerant stream caching. *IEEE Journal on Selected Areas in Communications*, 20(7), pp. 1328-1344.

Xu, D., Hefeeda, M., Hambrusch, S., and Bhargava, B. (2002). On peer-to-peer media streaming. In *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS'02)*, Wien, Austria.

Jin, S. and Bestavros, A. (2002). Cache-and-relay streaming media delivery for asynchronous clients. In *Proceedings of the 4th International Workshop on Networked Group Communication (NGC)*, Boston, MA, USA.

Padmanabhan, V. N., Wang, H. J., Chou, P. A., and Sripanidkulchai, K. (2002). Distributing streaming media content using cooperative networking. In *Proceedings of 12th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'02)*, Miami, FL, USA.

Cui, Y., Li, B., and Nahrstedt, K. (2004). oStream: Asynchronous Streaming Multicast in Application-Layer Overlay Networks. *IEEE Journal on Selected Areas in Communications*, 22(1), pp. 91-106.

Guo, L., Chen, S., Ren, S., Chen, X., and Jiang, S. (2004). PROP: a scalable and reliable P2P assisted proxy streaming system. In *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS'04)*, Tokyo, Japan.

Hofmann, M. , Ng, T. E., Guo, K., Paul, S., and Zhang, H. (1999) Caching Techniques for Streaming Multimedia over the Internet. *Technical Report, Bell Labs*.

Chen, S., Shen, B., Yan, Y., Basu, S., and Zhang, X. (2004). SRB: Shared running buffers in proxy to exploit memory locality of multiple streaming media sessions. In *Proceedings of the 24th IEEE International Conference on Distributed Computing Systems (ICDCS)*.

Zhang, X., Liu, J., Li, B., and Yum, T.-S. P. (2004). CoolStreaming/DONet: A dData-driven overlay network for live media streaming. *Technical Report, Chinese University of Hong Kong*.

Ip, A. T.-S., Liu, J., and Lui, J. C.-S. (2004). COPACC: A cooperative proxy-client caching system for on-demand media streaming. *Technical Report, Chinese University of Hong Kong*.

Yu, F., Zhang, Q., Zhu, W., and Zhang, Y.-Q. (2003). QoS-adaptive proxy caching for multimedia streaming over the Internet. *IEEE Trans. on Circuit and System for Video Technology*.

Schojer, P., Böszörmenyi, L., and Hellwagner, H. (2004). QBIX-G: A Quality Based Intelligent proXy Gateway. em Technical Reports of the Institute of Information Technology, University Klagenfurt,TR/ITEC/04/2.16.

Schojer, P., Böszörmenyi, L., Hellwagner, H., Penz, B., and Podlipnig, S.(2003). Architecture of a quality based intelligent proxy (QBIX) for MPEG-4 videos.

In *Proceedings of the 2003 ACM World Wide Web Conference (WWW'2003)*, Budapest,Hungary.

Podlipnig, S. and Böszörmenyi, L. (2002). Replacement strategies for quality based video caching. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Lausanne, Switzerland.