EVALUATION OF PRACTICAL SCALABILITY OF OVERLAY NETWORKS IN PROVIDING VIDEO-ON-DEMAND SERVICE

Jian-Guang Luo, Yun Tang, Jiang Zhang, Shi-Qiang Yang,

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P.R.China
{luojg03,tangyun98,zhang-jiang03}@mails.tsinghua.edu.cn, yangshq@mail.tsinghua.edu.cn

ABSTRACT

Recently, overlay networks have been proposed to address the problem of scalability in providing video-on-demand (VoD) service. However, from the perspective of service providing, their efficiency has not been carefully studied and still remains far from clear, especially considering the impacts of user interactivities and in the case of multiple files with different and varying popularities on sharing. Towards this end, in this paper, by analyzing more than ***20,000,000*** real workload traces, we first identify two practical factors which we believe have determinant impacts on the scalability: user interactivities and popularity differences among files. Then we further evaluate ***cache-and-relay*** (CR), a representative scheme of overlay networks, with the real workload traces. Simulation results show that CR only save about half of the server bandwidth even when there is no buffer constraint at clients, not so scalable as our original expectation.

1. INTRODUCTION

In past years, many research pioneers have recognized the potential of providing video-on-demand (VoD) service to a large population of users with the development of Internet and information technologies. As known, a major problem of the traditional client/server (C/S) unicast approaches is that they could not scale well due to the high bandwidth consumption at the server side. Therefore, a series of IP multicast based protocols [1] been proposed to overcome the bandwidth bottleneck in C/S, but they are all infeasible for the technical drawbacks and deployment difficulties of IP multicast. Other research efforts advocate proxy/CDN-based solutions [2] to distribute the workload to a number of edge servers, but they are very expensive to deploy. Recently, peer-to-peer (P2P) approaches attract much attention to solve the scalability problem in a distributed and cost-effective manner. By constructing an ***overlay network*** among end nodes, P2P approaches [3, 4, 5, 6] coordinate the resources of peers, exploit the buffer capacity and uplink bandwidth of clients, and thus provide a promising way for scalable VoD service.

This work is supported by the National Natural Science Foundation of China under Grant No. 60273008 and No. 60432030.

Although previous works have shown overlay networks exhibit well scalability in providing VoD service, they suffers from two particular insufficiencies. On one hand, most works either follow the theoretical assumptions or conduct simulation experiments rather than study the practical performance. For example, the universal assumption of sequential access without user interactivities may have a large distance from the reality. As evidenced, a high degree of interactivity has been observed in real workloads of streaming servers [7]. On the other hand, from the perspective of service providing, there are often thousands of or even more video files sharing on streaming server, resulting in that some are hot with high hit rate while others may be less inquired. As known, if the requests of a video clip are few and dispersed over time, the benefit gained from overlay networks certainly diminishes. Therefore, it remains interesting to evaluate the practical scalability of overlay networks in providing VoD service.

In this paper, by collecting more than 20,000,000 practical VoD workload traces over 100 days from CCTV.com, the website of the largest television station in China, we first identify two practical factors which have seldom been considered but determinantly impact the scalability of overlay networks including ***user interactivities*** and ***popularity differences*** among files. Then we evaluate the scalability of cache-and-relay (CR) in providing VoD service in terms of saving server bandwidth. Our main contributions of this paper lie in providing an insightful understanding towards the practical scalability issues and further investigating the scalability of CR in providing VoD service.

The remainder of the paper is organized as follows: In section 2, we analyze the real workload traces and identify the practical factors impacting the scalability of overlay networks. In section 3, we evaluate the scalability of CR through simulations. Finally, we conclude the paper and point out our future work in Section 4.

2. ANALYSIS OF WORKLOAD TRACES

In this section, we analyze the real VoD workload traces collected from CCTV.com, aiming to achieve an insightful understanding of the practical factors which we believe have determinant impacts on the scalability of overlay networks.

On the server, most video clips could be classified into ***news video*** and ***music video*** catalogs while the rests are denoted as ***misc video***. In each catalog, there are thousands of different video files encoded at about 300kbps which last from tens to thousands of seconds. Table 1 lists the numbers of workload traces of each catalog.

Table 1. Number of workload traces

Catalog	News video	Music video	Misc video
Traces	1924423	20055090	1313230

2.1. User Interactivities

Previous studies have shown that user interactivities, such as pause, jump forward/backward, etc., will deteriorate the performance of IP multicast based solutions in providing VoD service [7]. Intuitively, user interactivities will also impact the scalability of overlay networks negatively. For example, when user makes a jump forward operation, it may not be able to get service from its current supplying peer any more, but has to find a substitute parent or simply go to streaming server for help, and thus more workload will possibly be imposed on the server. Therefore, we take the analysis to the degree of user interactivities as the first step.[1]

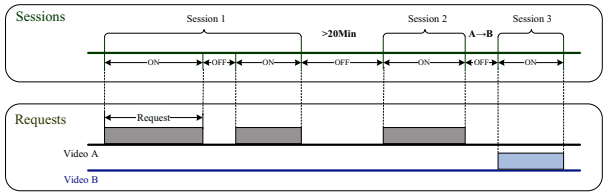


Fig. 1. Relationship between requests and sessions

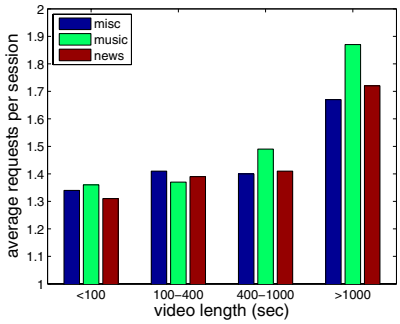


Fig. 2. Average requests per session

In order to catch the degree of user interactivities, we use the same session model as in [8], in which a session is defined as a sequence of requests to the same file from the same IP address while each OFF time is no greater than 20 minutes.

[1]Fast forward/backward operations are not taken into account in this paper since they appear too few in our traces.

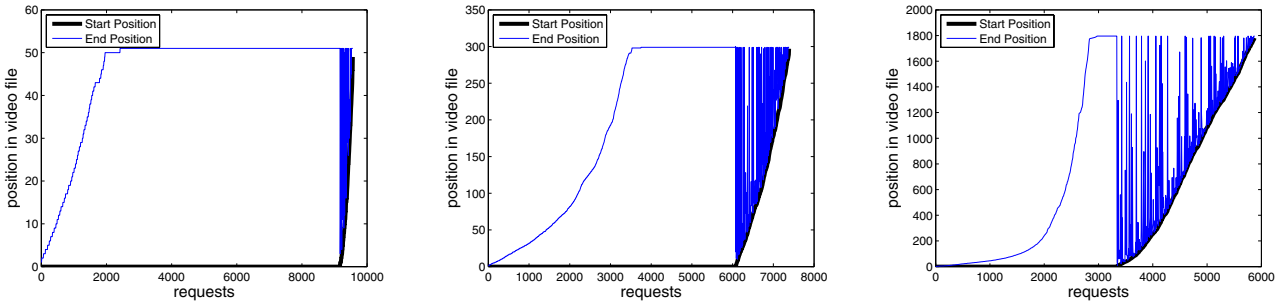
The relationship between requests and sessions is depicted in Fig.1. We intentionally depict the average requests per session over different video file lengths in Fig.2. Note that at the first glance, the longer the video file is, the more the average requests are in one session, indicating that more user interactivities are involved. It is not surprising if we recognize users may not be patient enough to go through the whole video clips. Instead, when the video is longer, they prefer skips or jumps to where it is more interesting. In order to verify this speculation, we further analyze the start and end position of requests for three randomly selected video files with different lengths in Fig.3. Fig.3(a) denotes that when the video length is only 52 seconds, most requests starts from the very beginning of the file, and most of these requests stay till the end of the file. As the comparison, in Fig.3(c), it could be easily observed that nearly half of the requests do not start from the beginning while only a few requests enjoy with whole file when the video is with the length of 1796 seconds. This "the longer the video, the more interactivities" characteristics of end users reminds us an appropriate length of video clip when it comes to the practical design of VoD system.

2.2. Popularity differences

To provide a VoD service, there are usually a number of video files sharing on one streaming server. Users can choose any video at any time, fully on their own decisions. Since the fundamental design philosophy of overlay networks expects the aggregation of requests from users so that the benefits of mutual cooperation accrue, the popularity differences among video clips will surely affect the scalability of overlay networks. In this subsection, we analyze the popularity differences among video files in following two aspects.

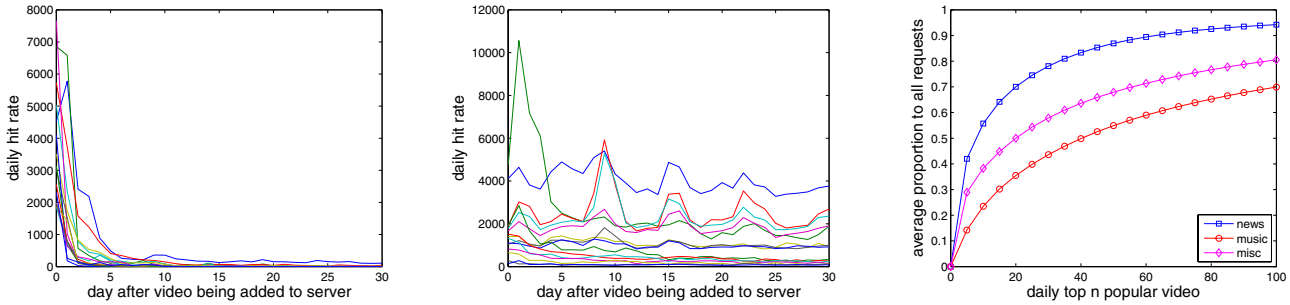
Popularity evolution: After a video file is added to the streaming server, its popularity, reflected as the hit rate, will change over time. This so called popularity evolution denotes the popularity variance of single files. For the reason of increasing the cooperation chance among peers, if the requests to a given video file aggregate in a short period of time, the streaming server will be able to save more bandwidth on this file from the deployment of overlay networks. Fig.4(a) and 4(b) depict the 30-day popularity evolution over a random set of 15 news and music video clip samples respectively. Observe that most requests for news clips arrive in only several days after their initial launch on the server. This "time-efficacy" characteristic of news catalog is potentially helpful for the design of overlay approaches. However, as evidenced in Fig.4(b), music clips does not exhibit the same characteristic as the news ones.

Popularity skewness: As lots of existing works, e.g. [8], we observe the popularity skewness among video objects on the VoD server. We depict the CDF curve of average requests proportion of daily top 100 hot video clips above three catalogs in Fig.4(c). As shown, popularities of news video clips



(a) video length = 52 sec (b) video length = 300 sec (c) video length = 1796 sec

Fig. 3. Distribution of start and end position of requests to video clips



(a) popularity evolution (news videos) (b) popularity evolution (music videos) (c) popularity skewness

Fig. 4. Popularity differences among video objects

are more skew than music videos, while misc catalog presents a considerably moderate degree. Besides, more than 40% requests stick with the top five most popular clips in news, while the proportion declines to only about 15% in music catalog. Since overlay networks expect an aggregation of requests, we could guess that streaming server of news video will benefit more from the deployment of overlay networks.

3. SCALABILITY OF CACHE-AND-RELAY

In this section, we evaluate the scalability of cache-and-relay (CR) proposed in [3, 4] in terms of saving server bandwidth through simulations with the real workload traces.

3.1. Methodologies

Several different schemes of overlay networks have been proposed to provide on demand streaming service, among which cache-and-relay (CR) is a representative one. In CR, after a client joins a multicast session, it caches the stream content just played out, and if needed, relays that stream to neighboring clients. Since the buffer capacity is limited, the client overwrites its buffer in a circular manner, i.e. it replaces the oldest data in its buffer with newly received content.

In our simulations, the uplink capacities of end hosts are not limited, and the client is only allowed to stream from VoD server if no other peers hold the content segments it needs. If there are multiple requests exist in a single session, we assume that the client keeps staying in the system during the OFF time between sequential requests, and thus is still able to relay the video content already cached in its buffer. Obviously, our simulation results in terms of saving server bandwidth provide a strict upper bound that CR could achieve. To make the question clear, we reasonably simplify the heterogeneity of end hosts by assuming they are of identical buffer capacity.

Note our workload traces are collected from a large and busy VoD server, on which there are more than 10,000 video files on sharing. During the 100 days we observed, the average bandwidth cost of the server is about 150Mbps, i.e. the daily throughput of the server is 1600G bytes. Since the efficiency of overlay networks increase with the busyness of VoD server in general, we believe our simulations make out the practical scalability of overlay networks to some extent.

3.2. Simulation Results

In this subsection, we first investigate the overall bandwidth saving of streaming server with our collected real workload traces. Fig.5 depicts the sever bandwidth saved from overlay networks over increasing buffer size of clients. Obviously, overlay networks become more efficient when the buffer size

increases at end hosts. However, even when there is no buffer constraint at client side, CR save only about half of the server bandwidth compared to traditional C/S approaches. Recall that our simulations only evaluate an upper bound of CR in reducing server bandwidth cost. In practical system, the efficiency of CR probably is worse than the simulation results. Considering the high busyness degree of our VoD server, the efficiency of CR is far from our original expectation.

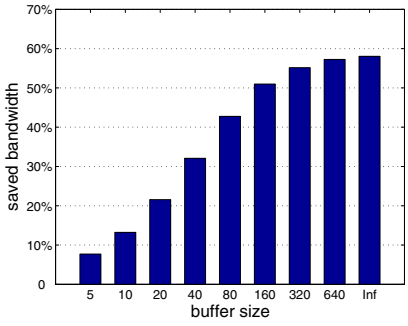


Fig. 5. Saved bandwidth vs. client buffer size

Further, in order to verify our speculation of the impacts of popularity skewness in subsection 2.2, we illustrate the percentage of server bandwidth saved from overlay networks over the proportion of requests to daily top 5 video clips of news catalog in Fig.6. As shown, a higher proportion of requests to top 5 clips, which indicates a more skew popularity, generally results in a higher percentage of saved bandwidth. Similiar results could be observed in music and misc video catalogs, which validate our speculation well.

4. CONCLUSION AND FUTURE WORK

In this paper, we first identify user interactivities and popularity differences among video objects through analysis to a large amount of real workload traces. Our main findings include: (a)user interactivities increases with the video length; (b)requests to news catalog are more aggregative in time and more skewly distributed among files than music video catalog. We believe these results will be helpful to the design of practical VoD systems. Then we evaluate cache-and-relay in terms of saving server bandwidth with the real workload traces. Simulation results show that CR only save about half of the server bandwidth even when there is no buffer constraint at clients, not so scalable as our original expectation.

This paper is only the first step of our work. Our future work is to design a more scalable scheme under our understanding of practical issues in providing VoD service and develop a practical on demand streaming system.

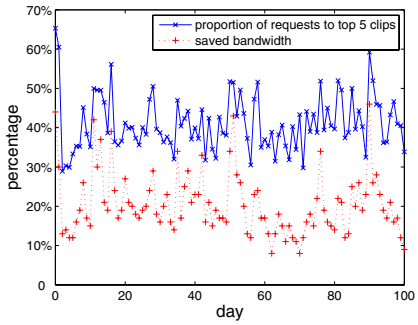


Fig. 6. Saved bandwidth vs. proportion of requests to daily top 5 video clips of news catalog (buffer size = 30 sec)

5. REFERENCES

[1] Derek Eager, Mary Vernon, and John Zahorjan, "Bandwidth skimming: A technique for cost-effective video-on-demand," *Proceedings of SPIE MMCN*, pp. 206–215, January 2000.

[2] Yuewei Wang, Zhi-Li Zhang, David Hung-Chang Du, and Dongli Su, "A network conscious approach to end-to-end video delivery over wide area networks using proxy servers," *Proceedings of IEEE INFOCOM*, pp. 660–667, April 1998.

[3] Yi Cui, Baochun Li, and Klara Nahrstedt, "ostream: Asynchronous streaming multicast in application-layer overlay networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 91–106, January 2004.

[4] Azer Bestavros and Shudong Jin, "Osmosis: Scalable delivery of real-time streaming media in ad-hoc overlay networks," *Proceedings of IEEE ICDCSW*, pp. 214–219, May 2003.

[5] Tai T. Do, Kien A. Hua, and Mounir A. Tantaoui, "P2vod: Providing fault tolerant video-on-demand streaming in peer-to-peer environment," *Proceedings of IEEE ICC*, pp. 1467–1472, June 2004.

[6] Yang Guo, Kyoungwon Suh, Jim Kurose, and Don Towsley, "P2cast: Peer-to-peer patching scheme for vod service," *Proceedings of International World Wide Web Conference*, pp. 301–309, May 2003.

[7] Marcus Rocha, Marcelo Maia, Italo Cunha, Jussara Almeida, and Sergio Campos, "Scalable media streaming to interactive users," *Proceedings of ACM Multimedia*, pp. 966–975, October 2005.

[8] Jussara M. Almeida, Jeffrey Krueger, Derek L. Eager, and Mary K. Vernon, "Analysis of educational media server workloads," *Proceedings of IEEE NOSSDAV*, pp. 21–30, June 2001.