Investigation of Web Server Access as a Basis for Designing Video-on-Demand Systems[1]

D. Venkatesh and T.D.C. Little

Multimedia Communications Laboratory
Department of Electrical, Computer and Systems Engineering
Boston University, Boston, Massachusetts 02215, USA
(617) 353-9877, (617) 353-6440 fax
{*dinesha,tdcl*}*@bu.edu*

MCL Technical Report 09-01-1995

Abstract–The performance of a video-on-demand server is affected by the dynamics of user accesses behavior. Most existing efforts consider static user request distributions in their design which can lead to poor performance if the accesses are different from that predicted. Even the use of a video store model to characterize user requests fails to account for the interactive nature of access. This suggests that better models for characterizing user behavior are necessary. In the recent past, the World Wide Web has become the most popular means for interactive information delivery. The World Wide Web represents a truly interactive medium with the user having complete control over presentation. Moreover, the performance bottleneck in the World Wide Web is more often the network than the server making it an ideal candidate to understand issues in serving interactive video. In this paper we study access behavior in a World Wide Web server and techniques to apply these observations in the design of a video-on-demand server.

Keywords: video-on-demand, storage servers, world wide web, WWW, access models.

[1] In *Proc. 1st Intl. Symposium on Photonics Technologies and Systems for Voice, Video, and Data Communications*, October 1995, Philadelphia, SPIE Vol. 2617-06. This work is supported in part by EMC Corporation and the National Science Foundation under Grant No. IRI-9502702.

1 Introduction

There has been interest in the recent past in both the academic and commercial communities for building storage servers that can deliver interactive video to customer homes.[8] Existing video-on-demand (VOD) prototypes are limited in their ability to serve a large customer base and there is a significant effort underway to build systems capable of supporting thousands of concurrent customer sessions. Much of the research in this area has focused on overcoming the I/O bandwidth bottlenecks that afflict many of today's magnetic storage media. Because of their relatively superior cost/performance ratios, disks have become the mainstay of storage server design. As a result, significant effort has been invested in developing disk data layout and placement strategies to achieve the maximum available throughput for such systems.[3,8,12] The most commonly used measure of performance for storage systems has thus become the number of concurrent streams that can be supported for a given cost.

However, there are several additional issues that must be considered when evaluating the performance of these systems. These include issues such as sensitivity to loading, response times, cost/session, reliability, and scalability. Because the stored content and user access patterns vary with time, it becomes important to consider the effects of this dynamic behavior on the performance of the system. Most existing studies neglect to consider these dynamics when evaluating system performance. They often assume a homogeneous media set, assume a static user population with a well defined skew in access demands and neglect to consider the short time dynamics of user accesses. For example, the Zipf distribution is most commonly used to characterize the skew in access demands among the set of available videos.[1,5] However, the Zipf distribution more accurately characterizes the long term behavior of access demands.[7] There are bound to be short term variations that affect the performance of the system and must be accounted for by the designer. This is because the demand for a movie can fluctuate with time. Furthermore, these models are not truly representative of the interactive nature of access one would expect in a video-on-demand server. They are often based on access statistics from the video store or a library where the user is often limited by the non-availability of a physical copy and by a limited number of titles.

In this paper, we propose to model user access to a VOD server using the World Wide Web (WWW) as a basis. The WWW has quickly evolved as the most popular means for interactive information dissemination on the Internet. We see the current access paradigm on the WWW in which the user has complete control over the session as a precursor to future interactive VOD systems. As the WWW is currently bandwidth constrained rather than server constrained it becomes an ideal candidate to study access demands at a server.

However, as it stands, much of the traffic flow on the WWW is in the form of short files (gizmo items such as bullets and lines) and as such would not be representative of access to a video-on-demand server. To build a more realistic model, we tracked requests to an on-line conference proceedings accessible from our Web server. Unlike traditional web accesses, access to the the on-line conference proceeding is more similar to requests one might see in a video-on-demand database. This is described in detail in Section 2. It is our belief that these statistics will prove useful to researchers in categorizing access demands to video-on-demand servers when analyzing their performance.

The rest of the paper is organized as follows. In Section 2 we discuss the parameters that can be extracted from the WWW to track the dynamics of user access. We discuss results from our studies on web access and their applicability to the design of a VOD server in Section 3. The paper concludes in Section 4.

2 The WWW Access Model

In this section, we discuss the use of access statistics from a WWW server to model accesses to a video-on-demand database. There have been several studies recently to characterize data delivered across the WWW.[2,4,6] These studies typically study the distribution and nature of documents accessed from a WWW server. However, the characteristics of data traffic from most WWW servers is different from that from a video-on-demand database. There are several reasons for this:

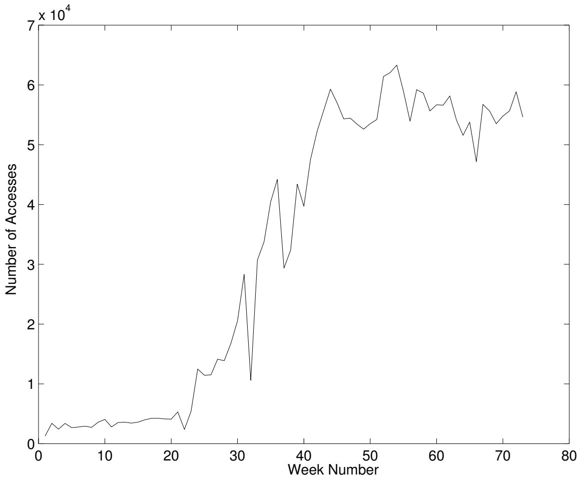

Figure 1: Weekly Accesses to the MCL Server

- The bulk of the traffic flow found in the WWW consists of file transfers. Furthermore, most of these transfers are due to small "gizmo" items such as bullets and lines.[2]

- The growth in the popularity of the WWW has led to an increase in the number of users, thereby increasing the load on the servers. For example, Fig. 1 demonstrates the increase in the number of accessed to the WWW server at the Multimedia Communications Laboratory (MCL) at Boston University. This would be unlike a video server that supports a fixed number of concurrent users.

- Many of the visits to the WWW server are characterized by users "surfing" the Web. As a result, many visits to the WWW server are due to curiosity. While users may surf and preview the titles available on a video-on-demand server, the subsequent request for a movie would not represent random surfing.

- The document sizes are relatively small (usually a few KBytes), and take at most a few seconds to transfer. On the other hand, a customer session in a VOD server can last for the duration of the movie (e.g., 100 mins.).

The WWW access logs can be used to study the relative loading of the server at various times in a day. Fig. 2 demonstrates the access demand on the MCL server for a 24 hour period averaged for a year. This curve was not normalized to take into account the differences in the time-zones of users accessing the system as a majority of the accesses were from North America. It is immediately apparent that the loading of the Web server is not uniform. It is at a minimum during the early morning hours, gradually increases during the day and peaks during the late afternoon after which it falls off. It is foreseeable that accesses to a VOD server will follow a similar pattern, but will be skewed to the right representing the hours when customers are at home.[8]

As discussed, the general nature of accesses to a WWW server makes it inaccurate for modeling accesses to a VOD server. However, it is possible to examine the nature of accesses to a library of relatively long documents and use the statistics to model the loading for a VOD server. To do this, let us examine how we model the nature of accesses to a VOD server.

In a typical VOD scenario, customers access a local database containing a finite number of videos. A reasonable partitioning approach for a video to memory organization is based on the probability of access or video popularity.[9] For example, for a video database containing

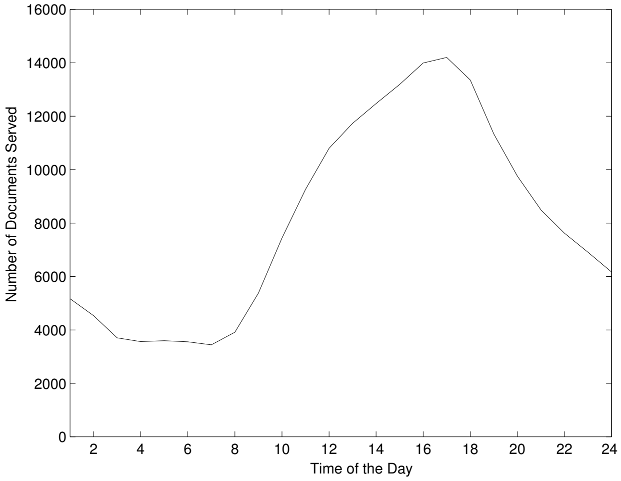

Figure 2: Server Loading vs. Time of Day

K videos, the relative popularities can be defined by a vector

$$P = [p_1, p_2, ..., p_K] \quad \text{where} \quad \sum_{i=1}^{K} p_i = 1$$

The mapping of the videos to popularities changes with time, but remains constant within a reasonable interval (usually a few days/weeks). For example, consider a video database with 10 videos. Let $[p_1, p_2, ..., p_{10}]$ be the popularities of the individual videos in descending order of popularity. A new video that is most popular would map to p_1. As time passes, its popularity drops, until it is discarded and replaced by another video.

This dynamic nature of operation suggests to us that the distribution of P is not stationary, but depends on the relative popularities of the videos at any given time. Because the load distribution on the server is dependent on P and affects its performance, it becomes important to consider the requirements imposed by such a distribution when designing a VOD server.

In the following section, we discuss studies conducted on a specific set of accesses to a WWW server that can be useful in characterizing access demands to a VOD server.

2.1 Study of Access Behavior in a Web Server

As described in Section 2, accesses to a WWW server are not typically representative of accesses to a VOD server. However, it is possible to examine a smaller subset of access distributions to understand the issues in designing a VOD server. To build such an access model we tracked accesses to the proceedings of the 5th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'95), available online from the WWW server at the Multimedia Communications Laboratory (MCL) at Boston University.[11] The document hierarchy for the on-line proceedings consists of 95 documents of which approximately 40 are postscript papers. We now describe the access characteristics for this set of on-line proceedings and discuss describe their applicability in a VOD scenario.

To begin with, let us describe the behavior of a typical user accessing the system. A user visiting the conference proceedings site is initially presented with a menu describing the details of the conference. The user has the option of subsequently browsing the abstracts of the papers and selecting a paper to download. After transferring a document, the user has the option of transferring more documents. However, this is treated as an independent request in our study. If the user changes his/her mind during data delivery or the latency of transfer is very large, the transfer is aborted. This model is also similar to that of a user accessing documents from an indexed library.[7] However, the user is not restricted by the number of physical copies.

It is apparent that the series of actions described are similar to the actions of a user accessing a video database.[10] There are other characteristics in serving the conference proceedings that make it similar to a VOD server. It is less likely that a user will download the same paper twice which is similar to a VOD server where the probability that a user may view the same movie more than once is small. We now describe the observed behavior of user accesses and their similarity to accesses from a VOD server.

Fig. 3 illustrates the number of accesses to the server counted on a daily basis. This curve illustrates two distinct regions. The first part represents an initial period of operation when access to the server was restricted to conference participants. The second part represents the period when this restriction was removed and the proceedings were available to the general public. This availability was also publicized by announcing the proceedings to over 1,200 individuals. It is clear from the figure that the there is a general decline in the popularity of the documents with time. It is also clear that this decline is not smooth.

The relative popularities for all the documents in the server over the duration of the test period is illustrated in Fig. 4. Overall, a total of 90 documents (including gizmo files) were served from the server. Fig. 4 illustrates a Zipf distribution fitted onto the popularity curve. We see that the Zipf distribution is a fairly accurate characterization of the cumulative accesses from the server. It was also observed that the most popular documents transferred

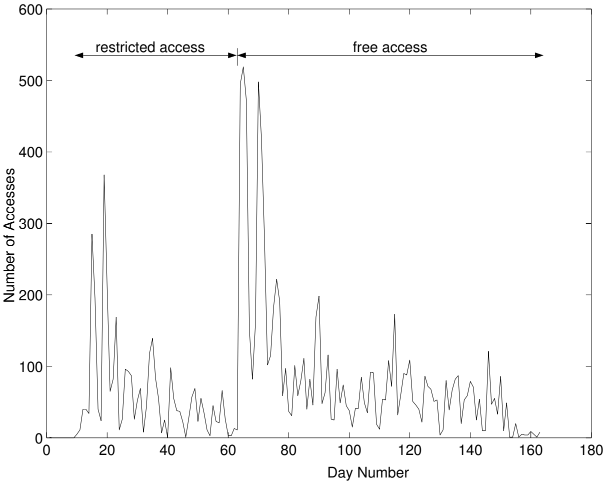

Figure 3: Daily Server Access Statistics

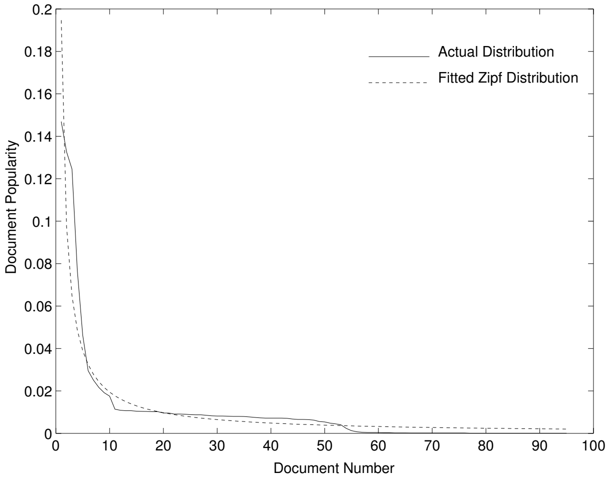

Figure 4: Popularity Distribution of all Documents

from the server were the index files that are smaller in size when compared to the actual papers.

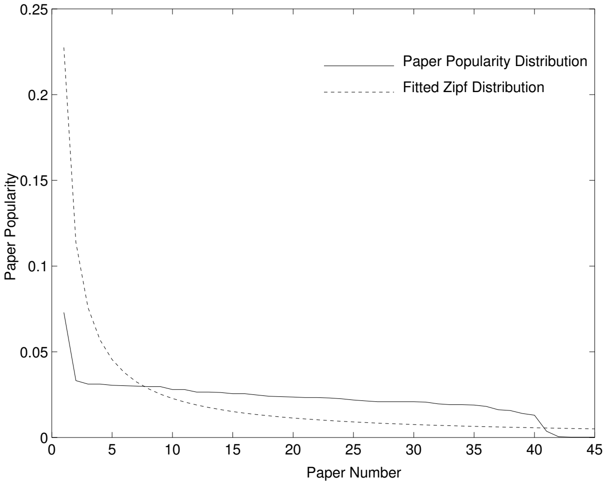

Figure 5: Popularity Distribution for Papers

The bulk of the data transfered from the server is due to the postscript papers (complete documents). Fig. 5 illustrates the relative popularities of these documents over the entire test period. The access distribution for the papers is more uniformly distributed and does not match the Zipf distribution very closely. It is possible that eventually the access skew will map to the Zipf distribution. However, this characteristic is not immediately apparent by the short term distributions.

The graphs described so far characterize the long term access behavior. However, as described earlier, the access demand is non-uniform and changes over time. Fig. 6 is a weekly sampling of the number of requests for the postscript papers. It is clear from this figure that the access demands for the papers are dynamic in nature. While a general skew in the relative popularity of the documents is observed, this skew varies with time.

Until now, we have studied user request behavior and the relative popularity of the various documents. However, they say little about the interactive nature of access. To study this feature, we examined the access logs to determine the number of incomplete file transfers. These transfers represent requests for files by the user that were subsequently terminated. Fig. 7 represents the reneging probability for the different papers. It is clear from this figure that users are more patient with some documents than they are with others.

In order to characterize the reneging behavior, we plotted the number of complete trans-

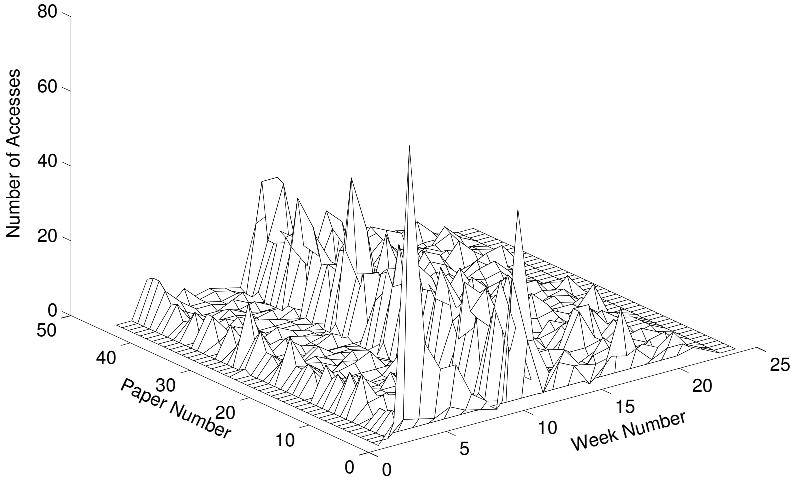

Figure 6: Weekly Access Demand for Papers

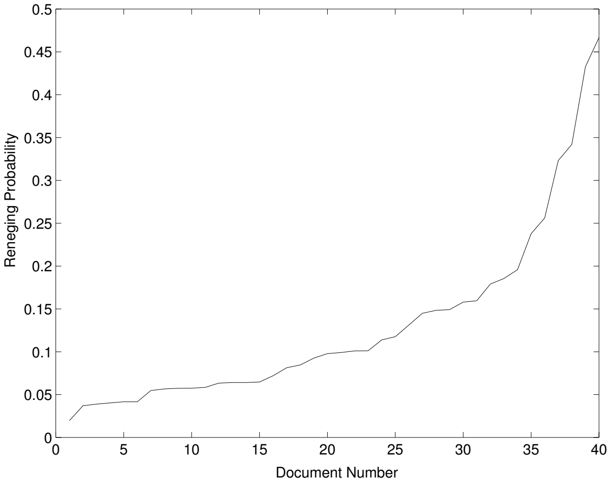

Figure 7: Reneging Probability of Documents

fers with the number of reneged transfers for the entire paper set. The reneging percentage was also compared against the relative document size to examine if the transfer time had any effect on the reneging probability. Fig. 8 illustrates the transfer characteristics for the complete and reneged transfer. There is no direct correlation between the two curves. On an average we found the reneging percentage to be slightly less for the more popular documents.

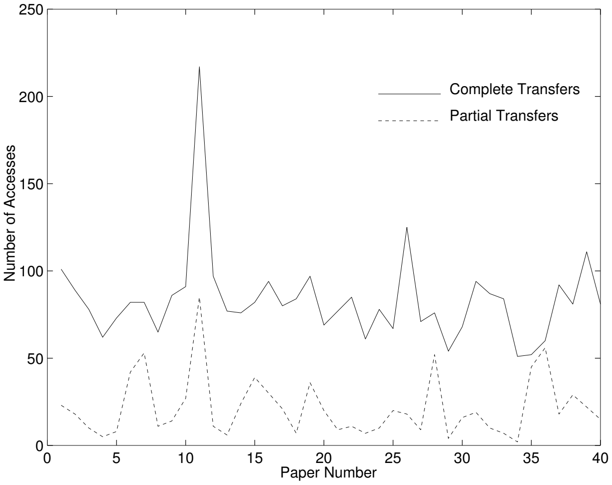

Figure 8: Reneging Behavior on Access Demands

Fig. 9 illustrates the reneging probability as a function of the document size. On the average, larger documents are more likely to be reneged than are smaller documents. However, if a document is popular, it is less likely to be reneged, even if it takes a long time to transfer.

In the next section, we discuss the implications of the observed access statistics, and their applicability in the design of a video-on-demand server.

3 Discussion

We now discuss the implications of the observed access behavior on the design of a VOD server. The design of a VOD server is affected by several factors. Primary among these are the storage and I/O bandwidth capacities necessary to support a given number of simultaneous users and a finite number of videos. The main measure of performance is the ability to support the projected loads at a least cost to the user. We now briefly describe some of the issues that must be addressed by the the designer of a VOD server.

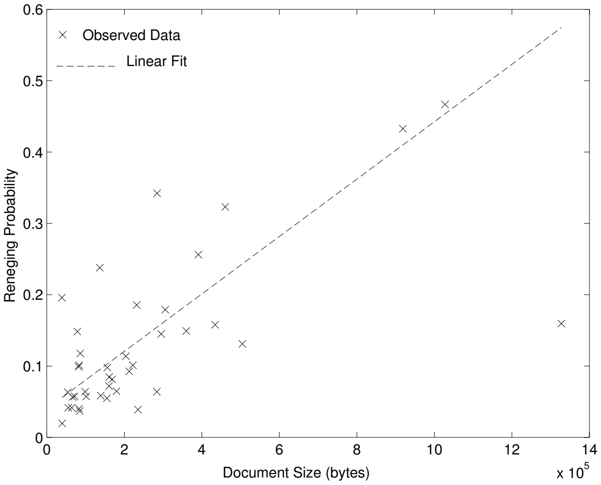

Figure 9: Reneging Probability vs. Document Size

- *Capacity Estimation:* Due to the limited I/O bandwidth capacity of a storage device, supporting the projected access demand requires many storage devices. The estimates for the type of storage devices and the storage hierarchy must also consider the projected loads on the system, media characteristics and the cost of the storage devices. Furthermore, the contents of the database can change over time and the estimate must account for this change.

- *Resource Allocation:* The resource allocation problem deals with mapping the stored media objects to storage devices in a manner that maximizes the availability of information to the user. Replication and striping can be used to satisfy access demands when the I/O bandwidth of a single device is insufficient to support the projected access demands. As updates occur dynamically, they must be transparent to the user. In other words, existing sessions must continue receiving services when database updates occur.

- *Resource Replacement:* When new objects manifest themselves in the MMIS server, storage space must be freed by replacing old objects. The policies used for object updating must ensure that the replaced information has a small chance of being accessed. However, this factor must also consider the object size and bandwidth requirements.

Most existing VOD designs employ the mechanical disk as a medium to serve video. Some of the approaches that are currently being employed for building disk-based VOD servers are described below.

- *Replication:* A storage device has a limited capacity for supporting concurrent CM streams. To support several users, a media object may have to be replicated across several devices.[9]

- *Caching:* The use of store and forward mechanisms requires buffering to store data. Depending on predicted loads, the system may have to store the data locally (cache them) to support future requests. Caching techniques take advantage of the observations of locality of reference and temporal locality to improve I/O performance.

- *Striping:* Striping and RAID technology attempts to take advantage of the combined bandwidths of the multiple units to improve performance as well as reliability. Data are retrieved in parallel from multiple disks to increase the ability to support concurrent sessions.[13]

Other factors that must be considered in the design of a VOD server include load balancing issues which are concerned with making sure that accesses to storage devices are evenly distributed and reliability. We now discuss the implications of the results gathered via the WWW model from Section 2 on the design of a VOD server.

We begin by considering the characteristics of user requests observed during the test period. When the papers were on-line, but had not been widely publicized, the number of accesses was moderate. However, after the availability of the proceedings was announced, there was a spurt of activity and the load on the server increased. This gradually decreased and eventually reached a steady state. However, it was observed that the decline was not smooth, but jagged. It was also observed that while all papers were not uniformly popular, the distribution of popularities was markedly different from the Zipf distribution. It was also observed that different papers were popular on different weeks. The reneging behavior demonstrated that there was a correlation between the document size (and hence the transmission latency) and reneging probability. However, users were willing to wait a little longer for the more popular documents.

There are several inferences one can draw from these observations that can be useful to the designer of video-on-demand servers. It is clear that a movie's popularity demonstrates a gradual decline with time. However, this behavior is not smooth and can change drastically, as was illustrated by the sudden increase in the number of requests after the public announcement. Such a scenario can occur in a VOD database as a result of a favorable review or a publicity blitz. A true-VOD server must be capable of absorbing such load fluctuations with little visible effect to the user. For a server employing a caching technique the implications of such sudden load surges imply a phase of operation during which the probability that a user's request is blocked increases. In a striped system, the admission control algorithm must modify its behavior to ensure that the disks are evenly loaded. A load fluctuation also necessitates a redistribution of video copies in a replicated system.

There are some features in the on-line proceedings that would not occur in a VOD server. All documents became accessible to the users at the same time in our study. This is similar to a VOD system in which all titles are released at the same time. While the chances of 50 movies being released simultaneously are remote, it is often the case that 4-5 movies are released simultaneously. Even though all new movies are not equally popular, it is possible that there is sufficient demand for all of them. However, when we observe the request distributions for the more popular papers, it is apparent that different papers are popular at different times. This variation in popularity must be absorbed by the server. These results are also useful when estimating the capacity of the VOD server. We can use the statistics to estimate the number of movie copies that are necessary to satisfy a given access demand.

The daily access behavior can be used to schedule operational periods and develop a pricing policy for the video server. The number of requests to the video server will be least during the early morning hours. This time can be scheduled to perform estimates on future loads and reorganize the video server to satisfy future demands. A pricing policy can be developed to attract customers during the slack periods by offering them lower costs. Our study on reneging shows that users are willing to wait for a longer time for a popular object. This characteristic can be used in developing batching approaches to aggregating sessions to conserve server I/O bandwidth during peak loads. These issues are relevant to the design of most interactive domains with small segments and frequent updates including interactive news databases, movies, distance learning, catalog and shopping.

4 Conclusion and Directions for Future Research

In this paper, we described the use of the WWW access logs as a basis to predict loading behavior on a video-on-demand server. Our study considered several issues that affect the performance of such a server including daily loading pattern, document popularity distribution, reneging probabilities and the dynamics of user accesses. Our study demonstrated that the nature of access to a database server is highly dynamic. It was shown that while some behaviors (such as accesses during a 24 hour period) are predictable, the same cannot be said for the relative popularity of the different documents on a given day.

The performance of a VOD server will be significantly affected by the dynamics of user accesses. Building a server based on a static user access distribution may yield a system that performs poorly when user preferences change. We are currently involved in an effort to scale up the observed results and apply them in the design of a VOD server. This includes finding alternate and larger information sources that will let us more accurately model the behavior of user accesses to video servers. We are examining the performance of various storage architectures in this context. These results will be evaluated on a video server currently under development at the Multimedia Communications Laboratory.

5 Acknowledgements

We thank the authors of *Perl* and *WWWSTAT* who have provided the tools that simplified our task of analyzing WWW server logs.

References

[1] K.C. Almeroth and M.H. Ammar, "The role of Multicast Communication in the Provision of Scalable and Interactive Video-on-Demand Service," *Proc. 5th Intl. Workshop on Network and Operating System Support for Digital Audio and Video, (NOSSDAV'95)*, Durham, NH, April 1995, pp. 267-270.

[2] A. Bestavros, R. Carter, M. Crovella, C. Cunha, A. Heddaya, and S. Mirdad, "Application-Level Document Caching in the Internet," *Proc. 2nd Intl. Workshop on Services in Distributed and Networked Environments (SDNE'95)*, Whistler, Canada, 1995.

[3] S. Berson, S. Ghandeharizadeh, R. Muntz, and X. Ju, "Staggered Striping in Multimedia Information Systems," *Proc. ACM SIGMOD*, 1994, pp. 79-89.

[4] L.D. Catledge and J.E. Pitkow, "Characterizing Browsing Strategies in the World-Wide Web," *Electronic Proc. 3rd Intl. World-Wide Web Conference*, http://www.igd.fhg.de/www/www95/proceedings/proceedings.html, 1995.

[5] A. Dan, D. Sitaram, and P. Shahabuddin, "Scheduling Policies for an On-Demand Video Server with Batching," *Proc. 2nd ACM Intl. Conf. on Multimedia (ACM Multimedia'95)*, San Francisco, October 1995, pp. 15-23.

[6] E. Katz, M. Butler, and R. McGrath, "A Scalable HTTP Server: The NCSA Prototype," *Electronic Proc. 1st. Intl. World-Wide Web Conference*, http://www.cern.ch/WWW94/Welcome.html, 1994.

[7] F.W. Lancaster, "The Measurement and Evaluation of Library Services," *Information Resources Press*, Washington D.C., 1977.

[8] T.D.C. Little and D. Venkatesh, "Prospects for Interactive Video-on-Demand," *IEEE Multimedia*, Vol. 1, No. 3, Fall 1994, pp. 14-24.

[9] T.D.C. Little and D. Venkatesh, "Probabilistic Assignment of Movies to Storage Devices in a Video-on-Demand System," *ACM/Springer Multimedia Systems*, Vol. 2, No. 6, January 1995, pp. 280-287.

[10] T.D.C. Little, G. Ahanger, H.-J. Chen, R.J. Folz, J.F. Gibbon, A. Krishnamurthy, P. Lumba, M. Ramanathan, and D. Venkatesh, "Selection and Dissemination of Digital Video via the Virtual Video Browser," *Journal of Multimedia Tools and Applications*, Kluwer Publications, Vol. 1, No. 2, June 1995, pp. 149-172.

[11] NOSSDAV'95, "Electronic Proceedings of the 5th Intl. Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'95), *http://spiderman.bu.edu/nossdav95/NOSSDAV95.html*, April 1995.

[12] P.V. Rangan, H.M. Vin, and S. Ramanathan, "Designing an On-Demand Multimedia Service," *IEEE Communications Magazine*, Vol. 30, No. 7, July 1992, pp. 56-64.

[13] H.M. Vin, A. Goyal, A. Goyal, and P. Goyal, "An Observation-Base Admission Control Algorithm for Multimedia Servers," *Proc. IEEE Intl. Conf. on Multimedia Computing and Systems, (ICMCS'94),* Boston, MA, May 1994, pp. 234-243.