Analyzing Client Interactivity in Streaming Media

Cristiano Costa[1], Italo Cunha[1], Alex Borges[1], Claudiney Ramos[1], Marcus Rocha[1]

Jussara Almeida[1], Berthier Ribeiro-Neto[1,2]

[1]Computer Science Department
Federal University of Minas Gerais
http://www.dcc.ufmg.br

[2]Akwan Information Technologies
http://www.akwan.com.br

Belo Horizonte, Brazil

{krusty, cunha, borges, cvramos, mvrocha, jussara, berthier}@dcc.ufmg.br

ABSTRACT

This paper provides an extensive analysis of pre-stored streaming media workloads, focusing on the client interactive behavior. We analyze four workloads that fall into three different domains, namely, education, entertainment video and entertainment audio. Our main goals are: (a) to identify qualitative similarities and differences in the typical client behavior for the three workload classes and (b) to provide data for generating realistic synthetic workloads.

Categories and Subject Descriptors

H.1.0 [**Models and Principles**]: General

General Terms

Measurement, Performance

Keywords

Streaming Media, Workload Characterization

1. INTRODUCTION

The rapid increase in popularity of streaming media traffic in the Internet [7] has led to the development of a number of new applications including online distance education, radio and TV programs. Three key characteristics of streaming media are: the high bandwidth requirements, which motivated the development of new scalable streaming protocols [9, 13], the real time constraints on media delivery, and the possibility of partial or interactive access. By interactive access, we mean that a client may pause, fast forward, rewind, or jump to specific points in the video/audio file. Thus, a streaming media file may not be fully and sequentially requested. To illustrate, clients of a media server delivering educational content may pause the transmission of a lecture to take notes or jump backwards to review a previously watched portion of the lecture.

Despite various previous studies on characterization of streaming media workloads [4, 5, 6, 7, 11, 12, 15, 19], understanding of client behavior is still superficial. This is because previous work has focused on only a few workload aspects, such as arrival process and file access frequencies, and, thus, does not provide a complete

analysis, or because only one type of workload (e.g., educational) has been considered. Thus, it is not clear whether their results hold to other workload domains (e.g., entertainment). Furthermore, a more complete characterization of interactive client behavior is still needed, as interactivity has been shown to significantly impact the scalability of streaming media protocols [5, 14, 17].

This paper provides a more thorough analysis of pre-stored streaming media workloads, focusing on the typical characteristics of client interactive behavior. Our main goal is to provide data for generating more realistic synthetic workloads, which can then be used in the evaluation of alternative media distribution methods. We also confirm previously derived insights into media caching [5], extending them for a richer and more diverse set of workloads.

As in [19], we use a hierarchical approach to deconstruct the workload into two levels: a client session level and an interactive request level. We characterize a long list of workload parameters in each level, including file access frequencies, session inter-arrival times, number, type and duration of interactive requests within a session and amount of media skipped between consecutive client interactions. Moreover, we pay special attention to the temporal variations of the workload and analyze each parameter for a number of selected high load days and also for shorter time periods during which the distributions are expected to remain roughly stable.

Another key point that distinguishes our work from previous studies is that we characterize workloads that fall into three different domains: educational video, entertainment video and entertainment audio. By characterizing this rich set of workloads, we are able to compare our findings and identify those that hold in general and those that are specific to each workload type. The educational workload consists of requests to the eTeach server [3], which delivers high bitrate educational content at one major US University. A one-month log of accesses to eTeach were previously analyzed in [5]. The eTeach logs analyzed in this work cover a longer and more recent time period, allowing us to contrast our results with those reported in [5]. The other three workloads analyzed are for entertainment content, two containing only audio files and a third one containing short video files. These workloads, obtained from two of the largest content and service providers in Latin America, are much heavier than the ones previously studied in the literature, with an average daily load ranging from 34K to 520K user requests.

Key observations from the analysis of our workloads are:

- A large number of client requests is for a small fraction of the media files and the fraction decreases with file size.

- Clients requesting audio content exhibit a very distinct interactive behavior from video clients. Almost 100% of audio sessions start at the beginning of the file and have only one interactive request. Furthermore, clients either listen to the whole audio or stop at an arbitrary position, with roughly equal probability. In contrast, a non-negligible fraction of video sessions starts at arbitrary positions in the file, especially for longer videos. The number of interactive requests within a video session increases with the file size.

- Pause is, by far, the most common client interaction in the video workloads. Jump backwards and jump forwards are roughly equally frequent for longer videos. Furthermore, the probabilities of a client pausing, jumping forwards or jumping backwards seem to strongly depend on the type of his/her previous interaction but *not* on the number of interactions since the beginning of the session. For either pause, jump forwards, or jump backwards, any interaction type is always more frequently followed by an interaction of the same type.

- There seems to be a strong spatial locality in the interactions within a client session. On average, a client skips up to 45 seconds of media between consecutive requests, in our logs.

- Client access patterns to the files in our three entertainment workloads tend to be more evenly distributed in time, whereas accesses to the educational content are usually more skewed towards the middle of the day and weekdays.

- Distribution of file access frequencies is better modeled with the concatenation of two Zipf-like distributions for accesses to audio and educational videos (as in [5]). However, the access frequencies of short entertainment videos can be well approximated with a single Zipf-like distribution.

- Session arrivals for the educational workload are well approximated by either a Weibull or a Lognormal distribution, depending on file size. A heavy-tailed Pareto distribution was found to be a good model for session arrivals at one of the entertainment workloads, whereas exponentially distributed session arrivals were observed in the other two.

- The insights into caching strategies previously drawn in [5] also hold for more recent educational workloads and entertainment workloads. In particular, we found that there is a large fraction of files that are accessed only sporadically, motivating the need to take popularity into account when deciding whether to store new content into a cache. We also found that file segment access frequencies are either roughly uniformly distributed (for the most popular and longer educational files) or skewed towards early segments (for less popular educational files and entertainment content).

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our workloads and the hierarchical model used to characterize them. Overall characteristics of servers' load are analyzed in section 4, and a thorough characterization of file accesses is provided in section 5. Section 6 extends previously derived insights into streaming media caching to our workloads. Section 7 concludes the paper.

2. RELATED WORK

A number of previous streaming media workload characterizations are available in the literature [4, 5, 6, 7, 11, 12, 15, 19]. Studies of interactive client accesses have been done for the MANIC audio content system [15], the low-bitrate Classroom 2000 system [11], the educational eTeach and BIBS media servers [5] and the educational internal server of a large international corporation [12]. These studies have analyzed a number of media workload aspects including session arrival process [5], distributions of ON and OFF times within sessions [5, 11, 12, 15], and frequency of each type of client interaction [5, 11, 12, 15]. The authors in [5] also draw insights into efficient caching strategies and quantify the scalability of a multicast streaming protocol for an interactive workload.

Client session duration, object and server popularity and sharing patterns of media objects among clients are analyzed in [7] for the client-side workload of a large university. Locality, dynamics and evolution of the accesses to objects in two enterprise media servers are analyzed in [6, 18]. A thorough characterization of a live streaming media workload is provided in [19].

Although collectively these previous studies cover a large set of different characteristics, most of them focuses on a few specific workload parameters. Key aspects for generating realistic synthetic workloads, such as the dependence between consecutive client interactions within a session, have not been previously analyzed. Furthermore, given the diversity of media workload types, it is not clear whether previous results hold for different workload domains.

To the best of our knowledge, the only previous attempt to compare characteristics of workloads from different domains, such as the one we do here, is the analysis of client accesses to the mMod system, which delivers both educational and entertainment videos at an university in Sweden [4]. However, that work provides a very limited characterization of the workload.

In comparison with previous work, our study: (1) provides a more thorough characterization of client interactive behavior; (2) analyzes the characteristics of a richer and more diverse set of workloads including educational and entertainment content, video and audio content, and (3) analyzes much heavier workloads.

3. METHODOLOGY

The data sources used in this paper are anonymized access logs to the eTeach media server [3], which delivers educational content at a major US University, and to two major Latin America service and content providers. One of these providers is *Universo Online* [1], or simply UOL, one of the largest ISP in Latin America. UOL provides a variety of online services, including streaming media entertainment audio and video services. The other data source, referred to, throughout this paper, as simply ISP, provides an online radio service, delivering, on demand, music files from a large collection. Section 3.1 describes the data sources and provides an overview of the log data used in our analysis. Section 3.2 describes the approach we use to deconstruct the workload into a two-level hierarchy consisting of client sessions, at a higher level, and interactive requests within each session, at a lower level.

3.1 Server Log Data

The eTeach media server [3] delivers educational content at the University of Wisconsin-Madison. Students have no classroom lectures and obtain all course material, including short announcements and variable length lectures directly from the server, accessing it mainly from within campus, using a high bandwidth network.

UOL and ISP deliver, on demand, media content to thousands of users across the Internet. We analyze the workloads of two UOL services: RADIO/UOL, which delivers only audio music files and TV/UOL, which delivers mainly short video files. The music files at RADIO/UOL are organized into pre-defined channels where the users can tune in to retrieve their favorite songs. The video files at TV/UOL consist of a variety of short TV programs, advertisements, interviews and social events. Access to both services is restricted to pre-registered users. ISP, on the other hand, provides a free online radio service offering both pre-compiled and user-defined channels. We refer to this workload as ISP/Audio.

Workload	eTeach	TV/UOL	RADIO/UOL	ISP/Audio
Period	09/02/00 - 10/17/01	01/18/02 - 03/01/02	01/07/02 - 01/30/02	06/01/03 - 06/07/03
Days	411	43	24	7
Total # of unique files	230	42,439	70,479	42,746
Total # of requests	46,958	1,453,117	5,385,822	4,160,889
Total media stored(hours)	80	1,303	4,775	2,765
Total media retrieved(hours)	1,522	25,090	228,018	164,332
Avg. # Requests / Day (CV)	114 (2.3)	33,793 (0.4)	224,409 (0.2)	594,413 (0.2)
Avg. # Sessions / Day (CV)	31 (2.4)	27,718 (0.4)	209,426 (0.2)	572,032 (0.2)
Avg. # Hours Req./ Day (CV)	4 (2.3)	583 (0.4)	9,501 (0.2)	23,476 (0.2)
Avg. # Unique File / Day (CV)	6 (1.3)	5,652 (0.3)	29,286 (0.1)	31,247 (0.03)
Avg. # Clients / Day (CV)	14 (1.9)	5,732 (0.4)	15,194 (0.2)	41,481 (0.1)

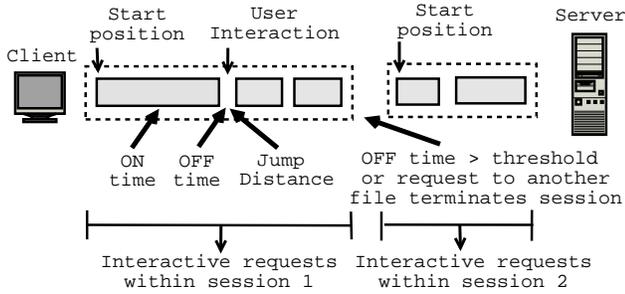**Table 1: Summary of the Workloads (CV = Coefficient of Variation).**

Figure 1: Hierarchical Model of Client Behavior.

Figure 2: Distribution of File Sizes (eTeach).

All four media services deliver content using the Windows Media Server [2]. The eTeach users view the videos using a customized interface which includes synchronized video, outline and slides. Users can pause, rewind, fast forward or jump to a predefined marker in the video corresponding to a topic in the outline.

Our study relies on a set of access logs from each of the aforementioned streaming media services. An overview of each log, as well as overall measures of each workload, are provided in Table 1. Note that, compared with the eTeach workload, the UOL and ISP workloads are approximately two orders of magnitude heavier and show much less daily variation. They are also much heavier than previously analyzed workloads. Also note that the eTeach logs cover a period of over a year. These logs are much longer and more recent than the eTeach logs analyzed in [5].

3.2 Hierarchical Model of Client Behavior

To characterize client interactive behavior, we use a hierarchical model to deconstruct each workload into a collection of client sessions, which, in turn, are broken into sequences of interactive requests. Since client sessions are not explicitly logged in any of our four workloads, we define a client session to be a sequence of interactive requests from the same client to the same media file, provided the time interval between two consecutive requests does not exceed a certain threshold [5, 15].

Figure 1 provides a graphical view of our model. A client session starts with a request to retrieve a file segment, starting at a certain position in the file (*start position*). Each request retrieves a certain amount of media, referred to as the *ON time*. The client think time between two consecutive requests is referred to as the *OFF time*. A new request within the session is triggered by a client interaction (e.g., resume after pause, jump backwards, jump forwards, fast forwarding, rewinding). The amount of media skipped during this in-
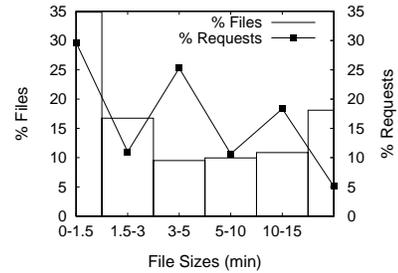
teraction is referred to as the *jump distance*. A pause/resume results in a jump distance equal to 0 whereas a jump forwards (backwards) results in a positive (negative) jump distance. A session ends when either (a) the client issues a request to a different file or (b) the OFF time exceeds a pre-defined threshold. A new session from the same client starts with the next interactive request. In section 5, we characterize each component of this model pointing out similarities and differences among our workloads.

To choose the OFF time threshold that triggers the termination of a client session, we analyzed the distribution of OFF times over several time periods. Contrary to the usual observation of distinct distributions for intra-session and inter-session OFF times in simulated sessions, we did not find any clear peak in our real case logs. We experimented with a range of threshold values, varying from 5 minutes to 30 minutes, and obtained qualitatively similar results. The results reported in this paper are for a threshold of 30 minutes.

4. SERVER LOAD CHARACTERISTICS

This section analyzes server load characteristics. The distributions of sizes and bitrates of the files stored at and requested from each server are provided in section 4.1. Daily and hourly load variation patterns are discussed in section 4.2.

4.1 File Characteristics

Figure 2 shows that there is a high variability in the sizes of the eTeach files stored at the server and requested by the clients. These files consist of short announcements (under 5 minutes) and variable length lectures of up to 55 minutes. For the other media services, the distributions of file sizes are more skewed towards smaller objects, as shown in Table 2. The audio workloads contain music files of variable length and short advertisements. The TV/UOL workload contains videos of short advertisements (under 1.5 minutes),

Workload	File Size (min)	% Files	% Requests
TV/UOL	< 1.5	82	61
	1.5 - 5	15	33
	5 - 15	3	6
RADIO/UOL	< 3	27	18
	3 - 5	57	68
	5 - 10	16	14
ISP/Audio	< 3	27	34
	3 - 5	59	55
	5 - 10	14	11

Table 2: Distribution of File Sizes (entertainment workloads).

clips, trailers and commercials (1.5-5 minutes), and news, social events and interviews (5-15 minutes).

In the entertainment workloads, the distributions of file sizes among client sessions are similar to the distributions of file sizes among client requests, shown in Table 2. In eTeach, on the other hand, around 52% of the sessions are for short files (under 5 minutes). The other sessions are roughly uniformly distributed across all size ranges. In comparison with the distribution of requested file sizes (Figure 2), this skewness suggests a larger number of requests within sessions for longer files, as discussed in section 5.

In both audio workloads, all files have low average bitrates (under 50 kbps). The TV/UOL workload has a more variable distribution of file bitrates. Around 87% of the stored files and 49% of the requested files have average bitrates under 50 kbps. Most of the remaining files have intermediate average bitrates of up to 250 kbps. The eTeach files are encoded at higher bitrates: around 91% of the requested files and 54% of the stored files have average bitrates in the range of 300-350 kbps. The average bitrates of the remaining files are lower, mainly in the range of 200-300 kbps. The high bitrates are not a problem for eTeach as its clients access the system mainly from a high bandwidth network within campus.

4.2 Daily and Hourly Load Variations

Figure 3 shows daily load variation, measured in terms of the number of requests and the average amount of media delivered per request, for each workload. The eTeach workload is the lightest one in terms of number of requests. However, the average amount of media delivered per request is around a couple of minutes for eTeach and the audio-only workloads but only one minute for the TV/UOL workload. Note that Figure 3-a shows load variations for eTeach over only 43 days (from Sept. 5th to Oct. 17th, 2001), the longest period of uninterrupted server activity in our 441-day log.

Figure 3-a shows that eTeach presents weekly load patterns with peaks around the middle of the week, when students have to turn in assignments, like observed in [5]. Such load variations are much less pronounced in our entertainment workloads. In fact, accesses to content in our entertainment workloads tend to be more evenly distributed in time, whereas accesses to the educational eTeach files tend to be concentrated around exam or assignment due dates. This result confirms those in [4] for the accesses to the mMod server delivering content at an university in Sweden.

Figures 4-a and 4-b show hourly load variations for typical high load days in eTeach and TV/UOL, respectively. Unlike the eTeach workload, which presents clear daily access patterns, the entertainment workload remains heavy during most of the day, dropping for only a few hours early in the morning. The audio workloads present hourly load variations similar to the one in Figure 4-b. The load peaks reach 18K (59K) requests and 850 (2400) hours of media delivered per hour at RADIO/UOL (ISP/Audio).

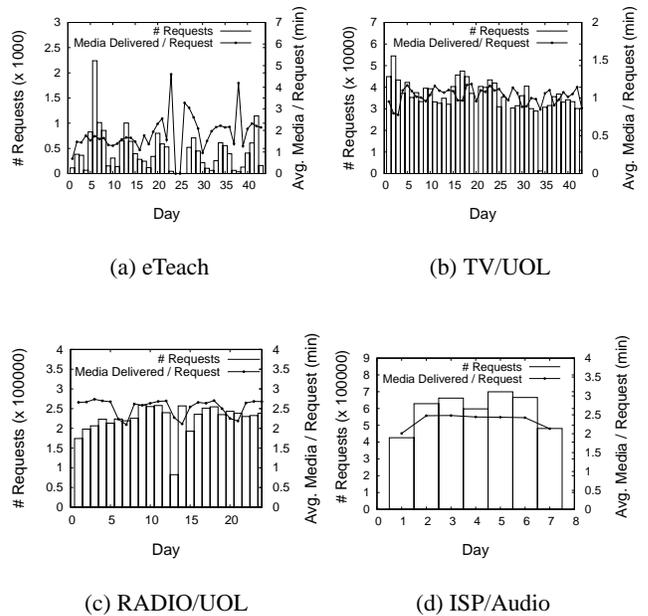Daily and hourly variations in the number of client sessions in

(a) eTeach (b) TV/UOL

(c) RADIO/UOL (d) ISP/Audio

Figure 3: Daily Server Load Variations.
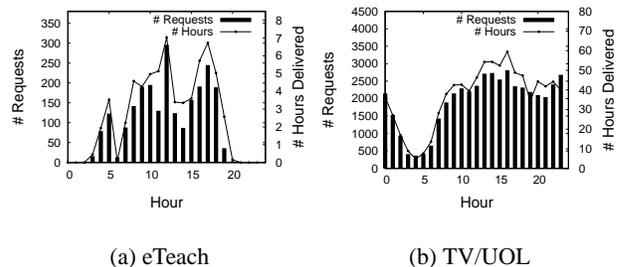

(a) eTeach (b) TV/UOL

Figure 4: Hourly Server Load Variations.

each workload follow similar patterns to those shown in Figures 3 and 4, scaled down by a factor equal to the average number of interactive requests per session. We also measured the number of simultaneous client sessions, observing peaks in the ranges of 10-13, 80-130, 700-1000 and 2500-3000 sessions, for eTeach, TV/UOL, RADIO/UOL and ISP/Audio, respectively. Note that the number of simultaneous client sessions reflects the number of simultaneously open socket connections, which, in turn, has a direct impact on memory allocation and processing overheads at each service.

We selected a number of high load days (such as the ones in Figure 4) for deeper analysis, recognizing that the statistical characteristics of different workload aspects may vary with time. We also analyze the workload for files falling into different size ranges separately, as the distributions may also depend on the file size and content type. Per day and per file size analyses of a large set of workload characteristics are provided in sections 5 and 6.

5. FILE ACCESS CHARACTERISTICS

This section analyzes each component of the workload model described in section 3.2, namely, file access frequency, session ar-
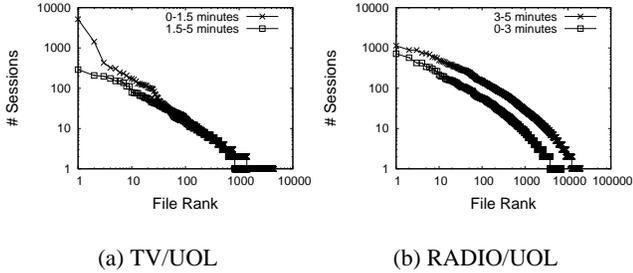
(a) TV/UOL   (b) RADIO/UOL

Figure 5: Daily Distribution of File Access Frequencies.

Figure 6: Distribution of Session Inter-Arrival Times (eTeach, File Size: 5-20 minutes).

rival process, start position, ON and OFF times, number and type of interactive requests within a session, and jump distance. It also shows profiles that illustrate typical interactive patterns of client behavior at each analyzed media service. In this analysis, we identify qualitative similarities and differences among the four workloads. We also contrast our findings with previous results.

5.1 File Access Frequency

We measure file access frequencies in terms of both the number of client interactive requests and the number of client sessions issued to each file, separately for each file size range and on different days. The results for both metrics are qualitatively similar for each workload. Thus we show only the latter in this section.

Figure 5 shows typical log-log plots of the distributions of file access frequencies for the TV/UOL and RADIO/UOL workloads, for different file size ranges, on high load days. These two plots present significantly different patterns of client behavior. Whereas the curves for TV/UOL are roughly linear, the curves for RADIO/UOL present two distinct linear regions. The plots for eTeach and ISP/Audio are analogous to those in Figure 5-b.

A linear curve in a log-log plot of file access frequencies has been modeled by a Zipf-like distribution [20] (Prob (access file i) $= C/i^{\alpha}$, where $\alpha > 0$ and C is a normalizing constant) in a number of previous streaming media workload studies [6, 7, 8, 19]. In contrast, two roughly linear regions (as in Figure 5-b) have been previously observed in the log-log plots of two educational media servers [5]. In that study, the concatenation of two Zipf-like distributions is suggested as a good approximation model.

As in [5], we use the concatenation of two Zipf-like distributions to model file access frequencies in eTeach and in the audio workloads. Table 3 shows the typical range of parameter values for the best fitted combination of two Zipf-like distributions, for these workloads. For each single distribution, it shows the total probability and percentage of files that fall within the corresponding region of the curve and the value of the α parameter. For the TV/UOL workload, Table 3 shows the typical values observed for the α parameter of the single Zipf-like distribution.

Two linear regions in log-log plots of file access frequencies have been addressed in other previous studies [10, 18]. In [10], the authors suggest that they occur in workloads consisting of files whose contents do not change frequently. In this case, the two linear regions could be a consequence of clients requesting the same file at most once. We measured the average number of times each client issues sessions to the same file on a given day, in the four workloads. In eTeach and in the audio workloads, this number was close to 1, in most cases. In the TV/UOL workload, it was slightly higher, varying from 1.28 to 1.42. Thus, although there seems to
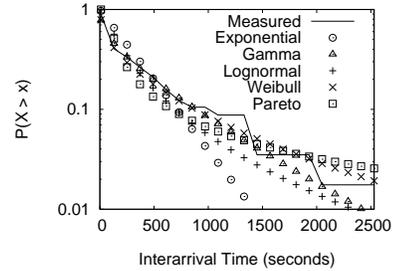
be a correlation between the average number of times a client issues sessions to the same file and the number of linear regions in the plots, the curve pattern can not be completely explained by the immutable nature of the content, as we do not expect file contents to change in one day, in none of our workloads. One possible reason for the single linear region observed in the TV/UOL workload is that some of the files in that workload may be advertisements which are *pushed* to the clients every time they request some content from the server. This is just a conjecture. Further investigation into this topic is left for future work.

Finally, a generalized Zipf-like distribution, based on the application of a Zipf k-transformation to the data, has been proposed to model two linear regions in the log-log plots of long-term file access frequencies in [18]. We found that this method provides a reasonable alternative model for accurately capturing also short-term file access frequency distributions in eTeach and in the audio workloads. However, like for the parameters of the two Zipf-like distributions, the values of the two Zipf k-transformation scale parameters are specific to each data set and to each analyzed day.

5.2 Session Arrival Process

We analyze session arrival process over a large number of periods of roughly stable arrival rate, recognizing that diurnal patterns may impact the aggregated distribution. Since client sessions are not explicitly logged in any of our workloads, we evaluate the arrival process for a session threshold of 30 minutes. Qualitatively similar results are found for other threshold values as well.

To find the distribution that best models session arrivals, we compared the least square differences of the best fitted curves for a set of alternative distributions. We also visually compared the curve fittings at the body and at the tail of the measured data, favoring a better fit for the body, if necessary, as short inter-arrival times have a stronger impact on server capacity planning and content sharing.

We found that the distribution that best fits session inter-arrival times depends on the workload. Session arrivals are exponentially distributed in the TV/UOL and RADIO/UOL workloads, in accordance with previous results [16]. In eTeach, Weibull or Lognormal are the most accurate distributions, depending on file size (consistent with results in [5]). Figure 6 shows an example where a Weibull distribution fits very accurately the distribution of eTeach session inter-arrival times. Session arrivals at ISP/Audio are even more heavy-tailed. A Pareto distribution was found to fit quite well both the *body* and the *tail* of the distributions. Table 4 summarizes these findings providing the observed ranges of values for the mean, standard deviation and distribution parameters for the most popular file size ranges in each workload.

Workload	File Size Range (minutes)	First Zipf			Second Zipf		
		Prob.	%files	α	Prob.	%files	α
eTeach	0-5	0.44-0.81	8-33	0.19-1.6	0.19-0.56	67-92	0.86-2.6
	5-20	0.56-0.88	9-13	0.93-0.97	0.12-0.44	87-91	1.5-2.2
		0.72-0.95	10-25	2.6-3.8	0.05-0.28	75-90	0.96-2.1
	30-40	0.89-0.94	14-25	2.5-4.4	0.06-0.11	75-86	0.85-2.6
TV/UOL	0-1.5	100	100	0.66-0.82			
	1.5-5	100	100	1.0-1.2			
	5-15	100	100	1.0-1.3			
RADIO/UOL	0-3	0.36-0.58	3-5	0.54-0.77	0.42-0.64	95-97	1.1-1.2
	3-5	0.50-0.72	7-11	0.60-0.85	0.28-0.50	89-93	1.3-1.4
	5-10	0.48-0.69	7-13	0.58-0.79	0.31-0.52	87-93	1.3-1.4
ISP/Audio	0-3	0.95-0.96	34-41	0.83-0.85	0.04-0.05	59-66	1.8-2.0
	3-5	0.65-0.72	13-18	0.63-0.70	0.28-0.35	82-87	1.7-2.0
	5-10	0.82-0.88	30-38	0.81-0.86	0.12-0.18	62-70	1.9-2.5

Table 3: Typical Parameters for Daily Distribution of File Access Frequencies (measured in number of sessions).

Workload	File Size Range (minutes)	Best Fit	Mean (seconds)	Std. Deviation (seconds)	First Parameter	Second Parameter
eTeach	0-5	Weibull	68 - 320	110 - 520	$\alpha = 0.006 - 0.08$	$\beta = 0.51 - 0.94$
	5-20	Lognormal	459 - 769	726 - 951	$\mu = 5.50 - 6.19$	$\sigma = 0.95 - 1.12$
	30-40	Weibull	349 - 557	510 - 752	$\alpha = 0.004 - 0.009$	$\beta = 0.82 - 0.91$
TV/UOL	0-3	Exponential	1.89 - 2.82	1.98 - 2.89	$\lambda = 0.35 - 0.53$	
	3-5	Exponential	10.9 - 93.6	11.1 - 101.0	$\lambda = 0.01 - 0.09$	
RADIO/UOL	0-3	Exponential	1.14 - 1.94	1.28 - 2.09	$\lambda = 0.51 - 0.87$	
	3-5	Exponential	0.35 - 0.61	0.54 - 0.77	$\lambda = 1.64 - 2.84$	
ISP/Audio	0-3	Pareto	0.25 - 3.80	0.45 - 4.01	$\alpha = 2.49 - 7.23$	$k = 0.83 - 2.56$
	3-5	Pareto	0.11 - 0.24	0.32 - 0.44	$\alpha = 6.25 - 13.40$	$k = 0.79 - 0.85$

(Probability Density Functions: Weibull: $p_X(x) = \alpha \beta x^{\beta-1} e^{-\alpha x^\beta} I_{(0,\infty)}(x)$, Lognormal: $p_X(x) = \frac{1}{x \sigma \sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$,

Exponential: $p_X(x) = \lambda e^{-\lambda x}$ and Pareto: $p_X(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}}$, where $x \geq k$.)

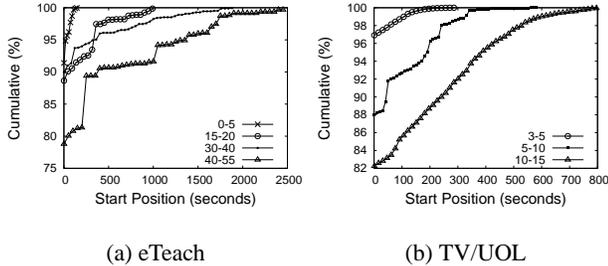**Table 4: Distributions of Session Inter-Arrival Times: Summary.**

(a) eTeach   (b) TV/UOL

Figure 7: Distribution of Session Start Positions.

5.3   Session Start Positions

As described in section 3.2, a session *start position* corresponds to the first file segment requested by a client in its first interaction within a session. Unlike observed in previous studies [4] and assumed in other related work [18], not all sessions in our workloads, especially the video workloads, start at the beginning of the file. In fact, although the distributions of session start positions are skewed towards the beginning of the files, we found a clear correlation between the skew factor, the workload type and the file size.

In the audio workloads, practically all sessions (over 99% at RADIO/UOL and 95% at ISP/Audio) start at the beginning of the file. In the video workloads, on the other hand, a small but significant fraction of the sessions to large files start at different positions. Figures 7-a and 7-b show the distributions for different file size ranges in eTeach and in TV/UOL, respectively. Note that Figure 7-a shows some clear plateaus, corresponding to markers into the eTeach lectures, where clients jump to, at the beginning of a session by clicking on the corresponding topic in the outline.

5.4   ON and OFF Times

We measured the distributions of ON and OFF times within client sessions for different days and file size ranges. Recall that ON times correspond to periods of client activity within a session or, in other words, to the amount of media delivered at each interactive request. OFF times, on the other hand, correspond to quiet periods between consecutive ON times. In our analysis, we use normalized ON times, measured as a fraction of the size of the requested file.

We found that the distributions that best fit ON and OFF times vary with day, file size range and workload. However, looking more carefully into the curves, a common distribution was found to be a good approximation of the measured data, in many cases. Next, we report our main findings with respect to these approximated (but accurate) distributions for the measured ON and OFF times.

The distributions of normalized ON times are heavy tailed for all

Workload	File Size Range (minutes)	Best Fit	Mean (% of file size)	Std. Deviation (% of file size)	First Parameter	Second Parameter
eTeach	0-5	Pareto	54 - 61	39 - 40	$\alpha = 0.23 - 0.32$	$k = 3.9 - 4.7$
	5-20	Weibull	3 - 32	2 - 36	$\alpha = 0.13 - 0.37$	$\beta = 0.62 - 1.5$
	20-55	Weibull	0.94 - 13	1.4 - 24	$\alpha = 0.21 - 1.2$	$\beta = 0.60 - 1.0$
TV/UOL	0-5	Pareto	46 - 85	28 - 41	$\alpha = 0.09 - 0.40$	$k = 3.0 - 6.4$
	5-15	Weibull	18 - 40	23 - 39	$\alpha = 0.03 - 0.15$	$\beta = 0.71 - 0.91$
RADIO/UOL	0-10	Pareto	58 - 79	32 - 42	$\alpha = 0.10 - 0.20$	$k = 2.2 - 4.6$
ISP/Audio	0-10	Pareto	58 - 90	23 - 43	$\alpha = 0.06 - 0.19$	$k = 1.9 - 6.8$

Table 5: Distributions of Session ON Times: Summary.

Workload	File Size Range (minutes)	Best Fit	Mean (seconds)	Std. Deviation (seconds)	First Parameter	Second Parameter
eTeach	0-55	Weibull	55 - 82	134 - 214	$\alpha = 0.08 - 0.14$	$\beta = 0.54 - 0.66$
TV/UOL	0-15	Weibull	25 - 61	69 - 153	$\alpha = 0.09 - 0.17$	$\beta = 0.57 - 0.79$
RADIO/UOL	0-10	Weibull	75 - 125	171 - 242	$\alpha = 0.06 - 0.07$	$\beta = 0.61 - 0.71$
ISP/Audio	0-1.5	Weibull	555	557	$\alpha = 0.01$	$\beta = 0.75$
	1.5-10	Weibull	176 - 236	307 - 381	$\alpha = 0.09 - 0.10$	$\beta = 0.50 - 0.52$

Table 6: Distributions of Session OFF Times: Summary.

file sizes in all four workloads. For short videos and audio files, a Pareto distribution fits well the measured data. For large videos, the distribution of normalized ON times is better modeled by a Weibull distribution. Table 5 summarizes these results.

One interesting observed trend is that clients usually request a larger fraction of audio and shorter video files. In fact, these files are frequently fully requested, which generates a heavy tail in the distribution of measured ON times. A Pareto distribution captures this tail much more accurately than the other distributions we tested (Exponential, LogNormal, Weibull and Gamma). Pareto was also found to be a good model for the body of the distribution. Note that, in absolute terms, the amount of media retrieved by each interactive request increases with file size, as observed in [5].

A Weibull distribution was also found to be a good fit for the distribution of OFF times, in all four workloads, for all file sizes, as shown in Table 6. Note that heavy tailed distributions were also observed for the ON and OFF times in the workloads analyzed in [5, 15]. However, an exponential distribution was found to be a good fit for the ON times of short files in [5].

5.5 Session Interactive Requests

To create a realistic synthetic workload, one needs a model for the interactive requests a client issues within a session. In particular, one needs the distribution of the number of interactive requests, the relative frequency of each interaction type (e.g., pause, jump forwards, jump backwards) and the distribution of jump distances. This section analyzes these three workload aspects.

Our results show that, like for the session start position, client interactive behavior is strongly correlated to content type and file size. In particular, clients of the video files, especially the longer educational videos, have a highly interactive behavior, issuing a number of requests within the same session. On the other hand, our audio sessions have, on average, only one client request.

5.5.1 Number of Interactive Requests

Figure 8 shows the cumulative distributions of the number of interactive requests within a session for eTeach files of different sizes. The distributions are more skewed towards fewer requests
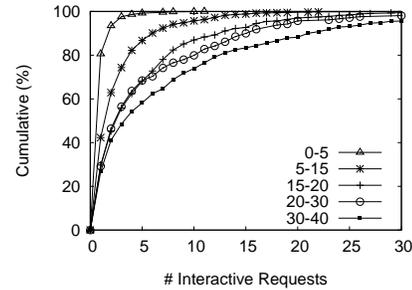

Figure 8: Distribution of Number of Requests per Session (eTeach).

for shorter files. However, note that the probability of long sessions (more than 10 requests) is non-negligible for files longer than 15 minutes. In particular, approximately 22% (28%) of the sessions to 20-30 (30-40) minute files have 10 or more interactive requests. These results are consistent with those in [5, 12], which report a larger number of interactive requests for sessions to longer educational videos. Thus, typical clients of eTeach lectures have a very interactive behavior, issuing many requests within the same session, especially for long videos.

The same overall trend was observed for the sessions to the TV/ UOL entertainment files. However, due to the shorter file sizes, two interactive requests were observed in only 11% of the sessions and around 4% of the sessions have 3 or more interactive requests.

Typical clients of entertainment audio files issue only one request per session. Two or more interactive requests were observed in less than 5% and 2% of all sessions in the RADIO/UOL and ISP/Audio workloads, respectively.

5.5.2 Frequency of Interactions

This section analyzes the relative frequency of each type of client interaction. Because fast forwarding and rewinding are very rare in our logs, accounting for less than 1% of all client interactions,

Workload	File Size	# Reqs / Session	% Pause	% Jump Back	% Jump Forward
eTeach	0-5	1.66	72	20	7
	5-15	2.87	64	19	17
	15-20	4.74	57	25	17
	20-30	5.18	56	26	18
	30-40	7.01	56	22	22
	40-55	4.19	48	17	35
TV/UOL	0-1.5	1.12	92	7	1
	1.5-3	1.24	87	8	5
	3-5	1.29	83	13	4
	5-10	1.28	85	8	7

Table 7: Relative Frequency of Each Type of Client Interaction.
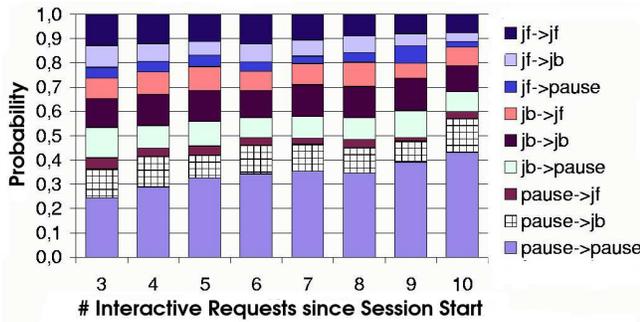
Figure 9: Variations in the Probabilities of Two Consecutive Client Interactions (eTeach).

Interaction i	Interaction $i+1$		
	Pause	Jump Fwd	Jump Back
Pause	0.24 - 0.43	0.01 - 0.05	0.08 - 0.17
Jump Fwd	0.02 - 0.07	0.08 - 0.13	0.02 - 0.09
Jump Back	0.08 - 0.13	0.03 - 0.10	0.11 - 0.16

Table 8: Typical Range of Probabilities for each Pair of Consecutive Client Interactions (eTeach).

(a) Video Files < 5 min   (b) Video Files > 5 min

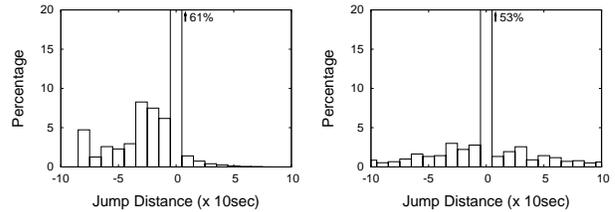**Figure 10: Typical Histograms of Jump Distances.**

we measure only the probabilities of a client pausing, jumping forwards and jumping backwards within a session. Since the vast majority of the audio sessions have only one interactive request, we focus only on the video workloads (eTeach and TV/UOL).

Table 7 shows the average number of interactive requests and the average frequencies of pause, jump forwards and jump backwards observed in the eTeach and TV/UOL sessions. Note that the last request within a session is accounted for as a pause in our analysis. As in [5, 12], the average number of interactive requests somewhat increases with the file size, especially in eTeach sessions. Pause is the most common interaction in both workloads, especially for short files. As file size increases, the percentage of pauses decreases, and the percentage of jump forwards increases. In other words, clients tend to skip more uninteresting file segments as they watch longer videos. The frequency of jump backwards remains roughly stable across all file size ranges in both workloads. For large files, the probabilities of jump forwards and jump backwards are roughly the same (except for eTeach files in the range of 40-55 minutes), as in [5, 11]. In contrast, the educational workload analyzed in [15] had a much stronger predominance of jump forwards.

Next, we analyze the dependency between consecutive interactions of the same client. In other words, we ask: *is the probability of a certain type of interaction higher if the previous interaction, within the same client session, was a pause/jump forwards/jump backwards?* Furthermore, *does this probability change as the client sends more requests within the same session?* To answer these questions, we analyzed the probabilities of each type of interaction, conditioned to the type of the interaction that was issued immedi-

ately before, within the same session. We also analyzed how this probability changes with the number of requests issued since the beginning of the session. In this analysis, we consider only eTeach and TV/UOL sessions with 3 or more interactive requests.

Figure 9 shows, for eTeach, the probabilities of each interaction type, conditioned to the type of the previous interaction, as a function of the number of interactive requests issued by the same client since the beginning of the session. The conditioned probabilities are typical across different file size ranges in eTeach. The legend shows the nine possible combinations of two consecutive interactions. The terms *jf* and *jb* are used to refer to jump forwards and jump backwards, respectively. Note that the marginal probabilities of each interaction type as a function of the number of interactive requests can also be assessed from the results shown in Figure 9.

We draw two key conclusions from these results. First, the probability of a client pausing, jumping forward or backwards does not seem to depend on the number of interactive requests issued by the client since the beginning of the session. The same is also true for the conditioned probabilities. The only exceptions are, perhaps, the probability of a client pausing after a pause and the marginal probability of a client pausing, which increase slightly as the client becomes more interactive. This new result greatly facilitates the generation of realistic synthetic workloads. Second, for either pause, jump forwards or jump backwards, any interaction type is always more frequently followed by an interaction of the same type. In other words, a client usually interacts with the video in the same way repeatedly. Table 8 summarizes these new results providing the range of probabilities found for each sequence of two interactions. Qualitatively similar results were found for the few TV/UOL sessions that have 3 or more interactive requests.

5.5.3 Jump Distances

In this section, we analyze the jump distances of all eTeach and TV/UOL sessions with at least two interactive requests. Recall that the jump distance is the amount of media skipped between two consecutive interactive requests within the same client session.

We found that the average jump distances in either direction increase with file sizes, as one might expect. For short video files (un-

der 5 minutes), the distances of jump backwards are usually longer, with an average around 20 seconds; whereas the average distance of jump forwards is only 7 seconds. For longer files (above 5 minutes), the average jump distances are roughly equal in either direction (around 40 seconds). Figures 10-a) and 10-b) show typical jump distance histograms to illustrate both scenarios. The frequencies for jump distance equal to 0 (i.e., pause) are annotated close to the center bar (x-value = 0) in both histograms.

Our results show that interactive requests in the two video workloads analyzed present strong spatial locality, with an average jump distance in either direction usually under 45 seconds. This result is in sharp contrast with the very long (over 2000 seconds) average jump distance observed in the MANIC educational system [15]. One key implication of short jump distances is that reserving a small amount of space for client buffering and prefetching may reduce server load and client delay significantly for interactive workloads. Furthermore, the scalability of some multicast-based streaming protocols has been shown to degrade in case of high interactivity [5]. Exploring prefetching to improve the scalability of these protocols for interactive workloads is left for future work.

5.6 Profiles of Client Interactive Behavior

This section shows typical profiles of client interactive behavior, focusing now on the file segments (delimited by a start position and an end position) retrieved by a client at each interactive request. We found that, in our workloads, typical clients behave following one of four distinct profiles, illustrated in Figure 11, depending mainly on the type and size of the requested file. The overall conclusions are: (1) the longer the video, the shorter the portions of it requested at each client interaction and (2) audio clients request the files typically from the beginning and either listen to it completely or stop at an arbitrary position, with approximately equal probability.

Each profile in Figure 11 shows the start and end positions of each request issued to a given file on a typical high load day. Requests, identified on the x-axis in the graphs, are sorted first by the start position, and then by the end position, in case of a tie. Figure 11-a shows a typical profile of client accesses to long videos (above 10 minutes). Clients request small portions of the file starting at different points, but there is also a number of requests for the full file. In the particular case of the profile in Figure 11-a, which represents accesses to an eTeach lecture, the start position curve shows some clear plateaus, corresponding to markers in the video.

The client access profile for short videos (under 5 minutes) in Figure 11-b shows a larger number of requests starting from the beginning as well as a larger number of requests for the full file. Clients requesting typical music files (3-5 minutes) either request the full audio or a prefix of arbitrary length, as illustrated in Figure 11-c. Finally, as one might expect, very short files (under 1.5 minutes) are usually fully requested, as shown in Figure 11-d. These results are consistent with those presented in sections 5.3 and 5.4.

6. IMPLICATIONS FOR CACHING

In this section, we analyze the distribution of file segment access frequencies and the accesses to unpopular content. We show that interesting insights into efficient streaming media caching strategies, previously drawn for two educational workloads [5], can be generalized to more recent educational workloads as well as to entertainment audio and video workloads.

6.1 File Segment Access Frequency

We found that the distribution of the access frequencies to 10-second segments of the eTeach educational files depends on the relative popularity of the file, as in [5]. It is roughly uniform for
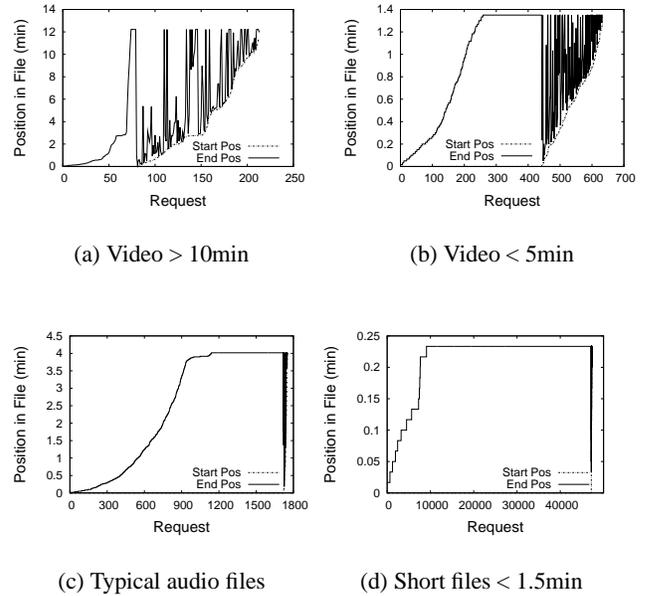

(a) Video > 10min

(b) Video < 5min

(c) Typical audio files

(d) Short files < 1.5min

Figure 11: Typical Profiles of Client Interactive Behavior.

(a) Most Popular Files (Educational)

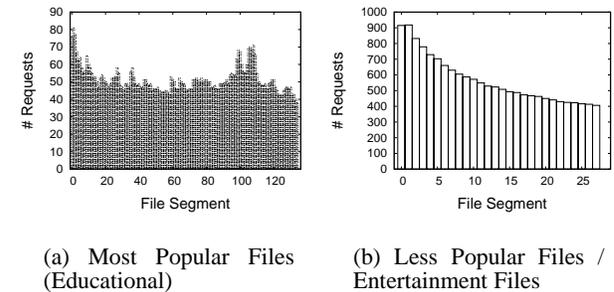(b) Less Popular Files / Entertainment Files

Figure 12: File Segment Access Frequencies.

the most popular files (usually lectures above 15 minute long), as shown in Figure 12-a, and skewed towards early segments for less popular files, as in Figure 12-b. For the entertainment workloads, on the other hand, a skewed distribution was most commonly observed for the accesses to all files, except to very short files, which are usually fully accessed, as discussed in sections 5.4 and 5.6.

The distribution of segment access frequencies has a direct impact on the design of efficient caching strategies. The decision between full file versus partial file caching and the most cost-effective method for estimating segment access frequencies depends on this distribution. If it is roughly uniform (as in Figure 12-a), a single (file) measure is enough to capture the distribution of segment access frequencies, and full file caching is the best strategy for unicast delivery. In the cases where a skewed distribution was found, the curve is usually well behaved (as shown in Figure 12-b) and can be roughly approximated using only two or three measures. Moreover, caching of a prefix may be a better strategy for unicast delivery.

6.2 Accesses to Unpopular Content

In [5], the authors conjecture that traditional caching strategies that insert new content into the cache without evaluating its histori-

Workload	Avg # new files	% new files accessed once	Time Until Next Access			
			% > 4hr	% > 8hr	% > 16hr	% > 32hr
eTeach	3	76	66	47	42	40
TV/UOL	316	85	84	74	63	47
RAD./UOL	1486	87	68	56	43	25
ISP/Audio	1559	87	57	39	26	12

Table 9: Summary of Accesses to *New* Files.

cal popularity information may result in significant disk write overhead. This is because they may frequently insert unpopular media content into the cache. This insight was drawn from the observation that a large amount of content is accessed only very sporadically.

We re-evaluated this conjecture for the more recent eTeach educational workload and for the three entertainment workloads. Table 9 summarizes our results. Column 2 shows the average number of *new* files accessed each hour on a typical day. A *new* file is a file that was not accessed in the previous n hours. We show results for $n = 4$. Similar results were also found for other values of n. On average, from 76% to 87% of the *new* files accessed each hour are acccessed only once on that hour (column 3). Furthermore, a significant fraction of those files are not accessed again in the next 4, 8, 16 or even 32 hours (columns 4-7). Qualitatively similar results were observed for accesses to *new* file segments as well. These results show that the conjecture introduced in [5] also holds for our educational and entertainment workloads.

7. CONCLUSIONS AND FUTURE WORK

This paper provides a thorough characterization of pre-stored streaming media workloads, focusing on client interactive behavior. The workloads analyzed are more diverse and significantly heavier than the ones previously studied. They fall into three categories, namely, educational, entertainment video and entertainment audio.

In our audio workloads, typical sessions consist of only one request for a file prefix. In our video workloads, clients are much more interactive, issuing a number of requests within the same session. The degree of interactivity increases with file size. Furthermore, we found that the probability of a client pausing, jumping forwards or backwards depends strongly on which interaction he/she issued immediately before within the same session, but *not* on the number of requests issued since the beginning of the session. This result greatly facilitates the generation of realistic synthetic workload. We also found that clients tend to interact with a video in the same way (e.g., pausing), repeatedly, within a session.

Possible directions for future work include characterizing other educational and entertainment workloads, further analyzing the correlation between different workload parameters and generating more realistic synthetic workloads.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] http://www.uol.com.br.

[2] http://www.microsoft.com/windows/windowsmedia.

[3] eTeach - Learning on Demand. http://eteach.cs.wisc.edu/index.html.

[4] S. Acharya, B. Smith, and P. Parnes. Characterizing User Access to Videos on the World Wide Web. In *Proc. MMCN*, San Jose, CA, Jan. 2000.

[5] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon. Analysis of Educational Media Server Workloads. In *Proc. NOSSDAV*, Port Jefferson, NY, June 2001.

[6] L. Cherkasova and M. Gupta. Characterizing Locality, Evolution, and Life Span of Accesses in Enterprise Media Server Workloads. In *Proc. NOSSDAV*, Miami Beach, FL, May 2002.

[7] M. Chesire, A. Wolman, G. M. Voelker, and H. M. Levy. Measurement and Analysis of a Streaming Media Workload. In *Proc. 3rd USENIX Symp. on Internet Technologies and Systems*, San Francisco, CA, Mar. 2001.

[8] A. Dan, D. Sitaram, and P. Shahabuddin. Scheduling Policies for an On-Demand Video Server with Batching. In *Proc. ACM Multimedia*, San Francisco, CA, Oct. 1994.

[9] D. L. Eager, M. K. Vernon, and J. Zahorjan. Bandwidth Skimming: A Technique for Cost-Effective Video on Demand. In *Proc. MMCN*, San Jose, CA, Jan. 2000.

[10] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, Modeling and Analysis of a Peer-to-Peer File-Sharing Workload. In *Proc. SOSP*, Bolton Landing, NY, Oct. 2003.

[11] N. Harel, V. Vellanki, A. Chervenak, G. Abowd, and U. Ramachandran. Workload of a Media-Enhanced Classroom Server. In *Proc. 2nd Annual Workshop on Workload Characterization*, Austin, TX, Oct. 1999.

[12] L. He, J. Grudin, and A. Gupta. Designing Presentations for On-Demand Viewing. In *Proc. ACM Conf. on Computer Supported Cooperative Work*, Philadelphia, PA, Dec. 2000.

[13] K. Hua, Y. Cai, and S. Sheu. Patching: A Multicast Technique for True Video-on-Demand Services. In *Proc. ACM Multimedia*, Bristol, U.K., Sept. 1998.

[14] S. Jin and A. Bestavros. Scalability of Multicast Delivery for Non-sequential Streaming Access. In *Proc. ACM SIGMETRICS*, Marina Del Rey, CA, June 2002.

[15] J. Padhye and J. Kurose. An Empirical Study of Client Interactions with a Continuous-Media Courseware Server. In *Proc. NOSSDAV*, Cambridge, UK, July 1998.

[16] V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, 3(3), June 1995.

[17] H. Tan, D. L. Eager, and M. K. Vernon. Delimiting the Range of Effectiveness of Scalable On-Demand Streaming. In *Proc. Int'l Symp. on Computer Performance Modeling and Evaluation*, Rome, Italy, Sept. 2002.

[18] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat. Medisyn: A Synthetic Streaming Media Service Workload Generator. In *Proc. NOSSDAV*, Monterey, CA, June 2003.

[19] E. Veloso, V. Almeida, W. Meira, A. Bestravos, and S. Jin. A Hierarchical Characterization of a Live Streaming Media Workload. *IEEE/ACM Trans. on Networking*, Sept. 2004.

[20] G. K. Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.