Analysis of Educational Media Server Workloads♣

Jussara M. Almeida[a] Jeffrey Krueger[a] Derek L. Eager[b] Mary K. Vernon[a]

[a]Computer Sciences Department
University of Wisconsin-Madison
{jussara,jkrueger,vernon}@cs.wisc.edu

[b]Department of Computer Science
University of Saskatchewan
eager@cs.usask.ca

ABSTRACT

This paper presents an extensive analysis of the client workloads for educational media servers at two major U.S. universities. The goals of the analysis include providing data for generating synthetic workloads, gaining insight into the design of streaming content distribution networks, and quantifying how much server bandwidth can be saved in interactive educational environments by using recently developed multicast streaming methods for stored content.

1. INTRODUCTION

This paper provides an analysis of the server log data for two media servers in use at major public universities in the United States. Both servers contain higher quality videos than servers or client workloads previously analyzed [1,9,12,14,17]. The *eTeach* system [21] was introduced in September 2000 to deliver the lectures and laboratory demos for a computer science course with an enrollment of 280 students. A key feature of this course is that there are no classroom lectures; the students obtain *all* course content from the server. *BIBS* (*Berkeley Internet Broadcasting System*) [22], in operation since January 1998, provides content for several courses across a number of fields including art, biology, chemistry, computer science, and electrical engineering. BIBS delivers live media multicasts as well as stored videos of lectures that were given on the Berkeley campus. BIBS also has considerably higher load than media servers previously analyzed.

The goals of the workload characterization include providing data for creating synthetic client workloads, gaining insight into streaming media caching strategies, and quantifying how much server bandwidth can be saved in an interactive educational environment by using recently developed multicast streaming methods for stored content (e.g., [11,13]).

Analyses presented in this paper that have not, to our knowledge, previously been reported for media workloads include the distribution of request interarrival times, evaluating stationarity in media access frequencies, measurement of *media segment* access frequencies, analysis of accesses to infrequently requested files and segments, and quantitative measures of the potential for stored media multicast techniques to reduce server load.

Key new observations from the analysis include:

- We find a high degree of similarity in the measures that can be obtained from the logs of both servers.
- During periods of approximately stationary request arrival rate, the BIBS client session arrival process is approximately Poisson, whereas the time between interactive requests within eTeach sessions is more accurately modeled by the heavy-tailed Pareto distribution, as has been found in more traditional Web server workloads [3,16, 19].
- For each server, there are very few periods of stationary relative file access frequency.
- Over periods of reasonably stable relative file access frequency, or over one-day periods, the file access frequencies on either server can be characterized by the concatenation of *two* Zipf-like distributions, rather than the single Zipf-like distribution observed in previous Web and streaming media workloads [2,5,6,9,10].
- On each server, a significant fraction of the new files accessed each hour are not requested again for more than eight hours, which motivates the need to reevaluate the traditional cache-on-first-miss strategy for streaming content.
- For the most frequently accessed files in eTeach, all ten-second segments of the media are accessed nearly equally often, whereas for less popular files, the access frequency is higher for earlier segments of the media.
- The distribution of the media delivered per session or per request within a session depends on media file length. Media delivered per BIBS session typically has a lognormal distribution for short videos and a hybrid gamma/Pareto distribution for longer videos. Media delivered per request within an eTeach session is typically exponential for short videos and a heavier-tailed distribution (i.e., Weibull or Pareto) for videos longer than five minutes.
- Partial file accesses limit the amount of server and network bandwidth that can be conserved by using multicast delivery; however, simulations of a recent multicast method for the requests captured in the eTeach logs show that server bandwidth for the three most frequently accessed files would be reduced by 40-60% during high server load periods.

♣ This work was partially supported by the National Science Foundation under Grant CCR 9975044. Jussara Almeida is partially supported by CNPq/Brazil and by a Lawrence H. Landweber NCR Graduate Fellowship in Distributed Systems.

To appear in *Proc.11th Int'l. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV 2001),* Port Jefferson, NY, June 25-26, 2001.

Section 2 provides background on the server logs and related work. Sections 3-5 provide a characterization of the server load, file access characteristics, and client interactivity, respectively. Section 6 concludes the paper.

2. BACKGROUND

2.1 Server Log Data

All eTeach media files are delivered by a Windows Media Server [15]. Clients view the videos using a version of Windows Media Player embedded in a customized web browser. When a client requests a given lecture or lab demo, the browser starts playing the beginning of the video, and displays the outline and the first slide. Clients can then pause, resume, rewind, fast forward, or skip to a predefined marker in the video that corresponds to a topic in the outline. The server log contains a *separate entry* for each interactive client request that results in delivery of a media stream. The request log thus contains a mixture of requests that start a new client session and requests within a session. Each log entry contains the start position and duration of the request. Thus, interactive client requests can be analyzed in detail.

The BIBS media content is delivered by a RealServer G2 server. To view the videos, clients typically use an ordinary Real Player application, which supports interactive requests; for a small number of courses, a customized browser interface provides a lecture outline with topics linked to markers in the video. The server records one log entry per client session (i.e., per *sequence of client requests* for a given media file) [18]. The start position and duration of each interactive request within a session can be recorded in the log entry, but the Spring 2000 semester logs analyzed in this paper do not contain these optional fields. Thus, in this paper, a BIBS "request" is a session request, and the number of seconds of media sent to the client may not be a contiguous segment of a media file. The BIBS data is analyzed for overall server load and file access characteristics, but not for file segment access characteristics or for measures of the interactivity.

For each server, we define the day to begin at the hour that typically has the lightest load (i.e., 3am for eTeach and 7am for BIBS). Requests for files that are not media files or were not found on the servers were removed from each log.

2.2 Related Work

There have been several studies of Web workloads [2,3,4,5,6,7,8] but those workloads did not include significant access to media content. Recent studies of client access to MANIC system audio content [17], low-bitrate videos in the Classroom2000 system [14], a mix of education and entertainment videos in the mMod system [1], and education content on an internal server of a large international corporation [12] have provided data on particular aspects of those media server workloads (e.g., example daily variation in server load, distribution of media stream or session durations, relative frequency of various types of interactive requests), and have pointed out that access patterns for educational media may be different than for entertainment media [1].

A parallel study by Chesire *et al.* [9] analyzes the RTSP sessions originating from a large university to numerous media servers on the Internet, and characterizes session duration, object and server popularity, the maximum bandwidth savings due to caching, and the amount of overlap between overlapped consecutive accesses to the same media file.

As the results are presented below, similarities and differences compared with previously reported media workloads are noted.

3. SERVER LOAD CHARACTERISTICS

3.1 Overall Server Load

The overall load for each server is summarized in Table 1. 96% of the eTeach requests are to the stored content for one course with a large enrollment and no classroom lectures, whereas BIBS delivers stored and live content for 11 courses that have classroom lectures. The difference in content is reflected in the relative amount of media delivered per day by each of the servers. The eTeach logs end the day after the first exam for the course. The BIBS logs end during the final exam week for the semester.

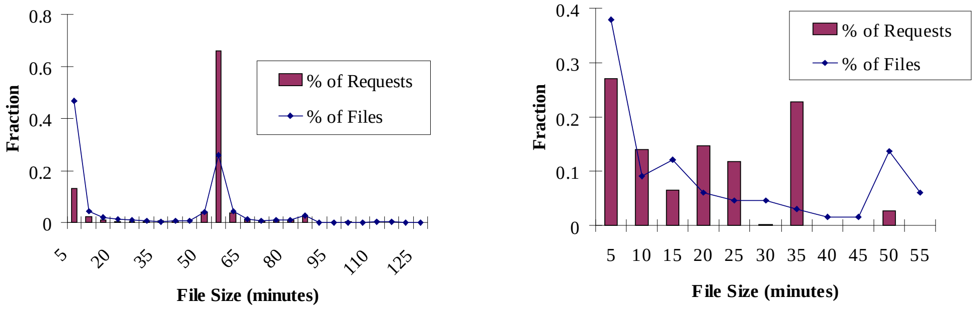
Figure 1(a) shows that eTeach has higher variability in stored and requested file sizes than education media servers previously studied [14, 17], with 25-30% of requests to short videos (i.e., under 5 minutes) and other requests approximately evenly distributed to videos of duration 5-10, 10-15, 15-20, 20-25 and 30-35 minutes. Figure 1(b) shows that a majority (i.e., 70%) of BIBS requests for *stored* content are for 50-55 minute videos, similar to previously studied education server workloads that contain stored content from classroom lectures [14, 17].

Nearly all videos requested on the eTeach server are encoded for 300 kilobits per second. For BIBS, 80% of the requests and more than 95% of the minutes delivered are for videos encoded at 350-500 kb/s. These rates are considerably higher than the rates observed in previous streaming workload studies [1,9,12,14,17].

The load per day for stored content on each server (Figure 2) has a similar profile as in [14]. The BIBS server typically has 500-800 client sessions per weekday, with occasional peaks of up to 2000 sessions, and a typical average of 15–25 min of media per session. The eTeach server typically has 500–1800 interactive requests per

Table 1: Summary Statistics

Statistic	BIBS	eTeach
Log duration (days)	2/1/00 – 5/19/00 (109)	9/12/00 – 10/12/00 (31)
Total number of stored files accessed	1506 files	73 files
Total number of requests for stored (live) content	66,694 sessions (22,411)	17,233 requests
Average (CoV) number of stored requests / day	606 sessions/day (0.27)	538 requests/day (0.86)
Average (CoV) number of stored hours sent / day	177 hours/day (0.61)	15 hours/day (0.81)
Average (CoV) number of stored files requested / day	157 files/day (0.06)	19 files/day (0.13)



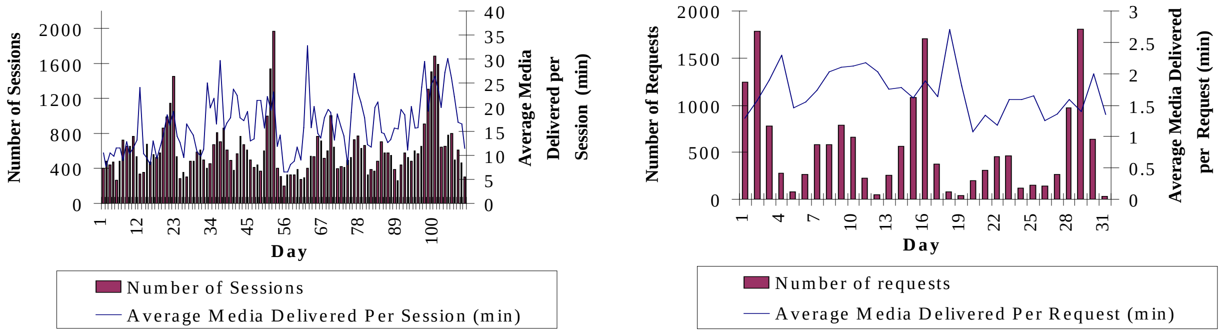
(a) BIBS **(b) eTeach**

Figure 1: Distribution of File Sizes

weekday and average stream duration per request of 1.5–2 minutes.

From the daily usage statistics, we select the days shown in Table 2 for more in-depth characterization of the client requests, recognizing that workloads from various high load days are most useful in evaluating media server design alternatives. Approximately ten other days that have among the highest numbers of client requests are also analyzed for many of the distributions reported in this paper.

Figure 3 shows the daily and typical hourly request arrival counts

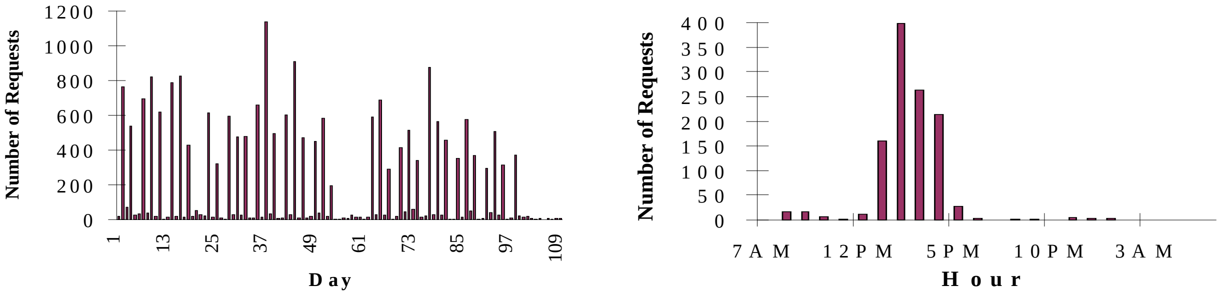


(a) BIBS **(b) eTeach**

Figure 2: Server Load per Day for Stored Content

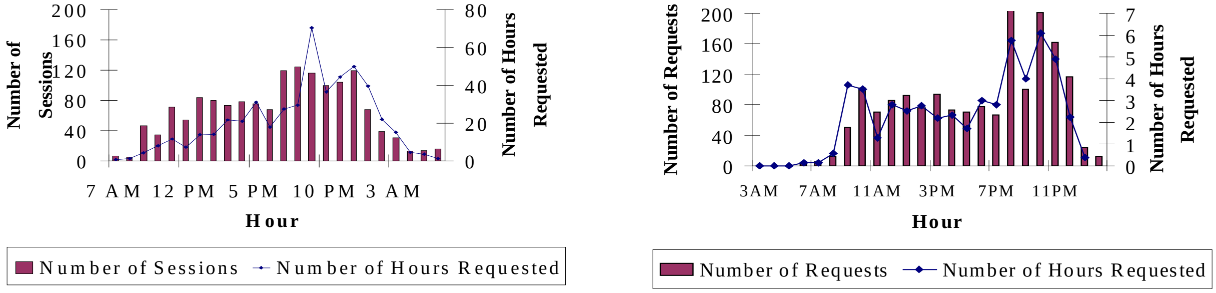
Table 2: Selected High-Load Days

Server	Maximum # requests	Max imum # hours sent	Maximum # mins sent / # requests	Minimum # mins sent / # requests
eTeach	Oct. 10 (1812)	Sep. 27, 54 hrs	Sep. 15, 2.3 min / request	Oct. 3, 1.2 min. / request
BIBS	Mar. 22 (1540)	Mar. 23, 755 hrs	May 14, 30 min / session	Feb. 2, 7.6 min / session



(a) Live Requests per Day **(b) Example Live Requests per Hour**

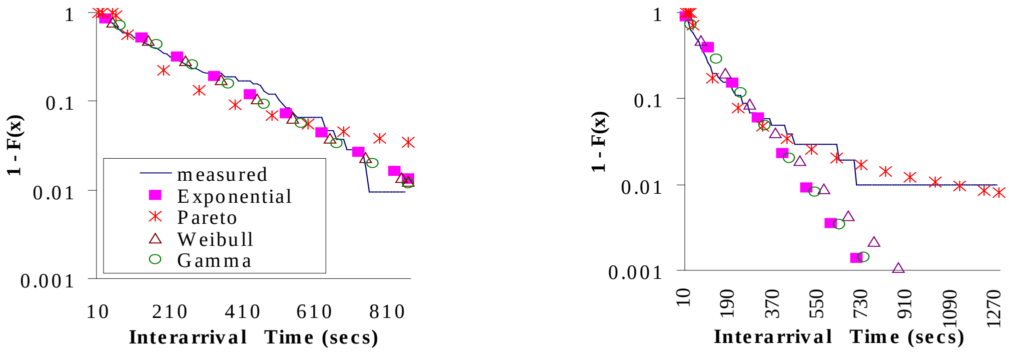
Figure 3: Server Load for Live Content (BIBS)



(a) BIBS 3/22

(b) eTeach 9/27

Figure 4: Server Load per Hour for Stored Content



(a) BIBS 3/23 6pm-11pm

(b) eTeach 10/10 10am-noon

Figure 5: Distribution of Interarrival Times

for live streams delivered by BIBS. The average number of concurrent viewers for live streams typically varies from 15-35, depending on the hour and multicast. Requests for live content are not analyzed further.

3.2 Request Interarrival Times

Figure 4 shows a typical profile of request arrivals per hour for stored content on each server. These profiles are used to find periods of approximately *stationary request arrival rate*, such as 7pm–1am in Figure 4a (March 22, BIBS) or noon – 8pm in Figure 4b (Sept 27, eTeach). Such periods are needed for analyzing the distribution of time between request arrivals. To our knowledge, the distribution of request interarrival times has not previously been analyzed for media servers.

For *all* periods of stationary request arrival rate identified on each server, as well as for *each file* that had a sufficient number of requests during the period, we used curve fitting for the exponential, gamma, Weibull, and Pareto distributions to determine the distribution of time between requests. Examples of the fitted curves are shown in Figure 5. Like other types of Web workloads [3,16,19], client session arrivals (overall and per file) in BIBS were found to be approximately *Poisson*, whereas the time between requests in eTeach has a heavy-tailed *Pareto* distribution in every case except a stationary period on Sept. 27.

The period on Sept. 27 had Poisson request arrivals, overall and per file. A possible explanation is that a higher fraction of the request arrivals might be session arrivals rather than interactive requests within a session. Table 2 shows that Sept. 27 was the day with maximum total minutes of media delivered. However, the average number of interactive requests per session (discussed in Section 5) was not notably different than on the other days that were analyzed. Thus, it's unclear why the distribution of request interarrival times is different on that particular day. Session arrivals in eTeach are examined further in Section 5.

3.3 Media Delivered per Session

For both servers, we observed a significant fraction of requests of short duration (less than 3 minutes of video), as in other recent workloads [9,12]. In this section, we further characterize the media delivered per session on the BIBS server, which to our knowledge has not been done in previous media workload studies.

As shown in Figure 4, the average number of minutes delivered per BIBS session fluctuates throughout the day. To determine the mean and the distribution of media delivered per session, which may depend on the length, we have identified several periods of stationary average number of minutes per session for each of the most popular file lengths (i.e., media files shorter than 5 minutes, and files that have length 50-55 minutes).

We performed curve fitting for the measured distribution for each period of stationary average media delivered per session, as illustrated in Figure 6. Table 3 summarizes the distributions that provided the best fit for the file length ranges of the most frequently accessed files. The results show, not surprisingly, that the distribution does depend on the media file length. For videos under five minutes, a single distribution (most commonly lognormal) with mean in the range of 0.75 – 1.5 minutes

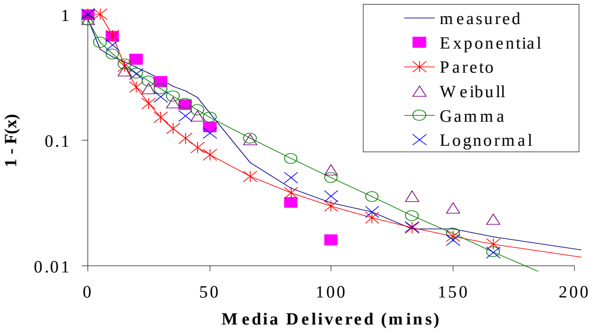


Figure 6: Example Distribution of Media Delivered per Session (BIBS 5/14 11am-10pm, File Size: 50-55mins)

represents the measured data. For 50-55 minute videos, most commonly a hybrid gamma + Pareto distribution (i.e., gamma for the body of the curve and Pareto for the tail) has a better fit. Table 3 provides other distributions that were observed for the given file length range, as well as a summary of the observed (minimum and maximum) means and corresponding standard deviation of the distributions that occurred during the periods of stationary average number of minutes delivered per session for the given file length. During other periods of time, the average number of minutes of media delivered per session was typically in the range of 0.5 – 2.5 for videos under five minutes and 10-30 minutes for the 50-55 minute videos.

4. FILE ACCESS CHARACTERISTICS

To our knowledge, the only previously reported characterizations of media server file access frequencies are the log-log plots of number of accesses versus file rank for an unstated period of time on the mMOD education and entertainment server [1] and for a university one-week trace of RTSP media requests to multiple servers from a given large population [9]. The latter access frequencies are accurately modeled by a single Zipf-like distribution [20], whereas the reported mMOD access frequencies have an unknown distribution that is not Zipf-like. Below we characterize the distribution of file access frequencies on the BIBS and eTeach servers, which depends on the period considered in the analysis. To facilitate synthetic workload generation, we provide separate file access frequency distributions for different file sizes accessed within any given period. We also analyze the distribution of accesses to each ten second segment within a media file on eTeach, as well as the temporal distribution of accesses to infrequently requested files and segments, to provide insight into the design of media caching strategies.

Table 3: Media Delivered per Session (BIBS)

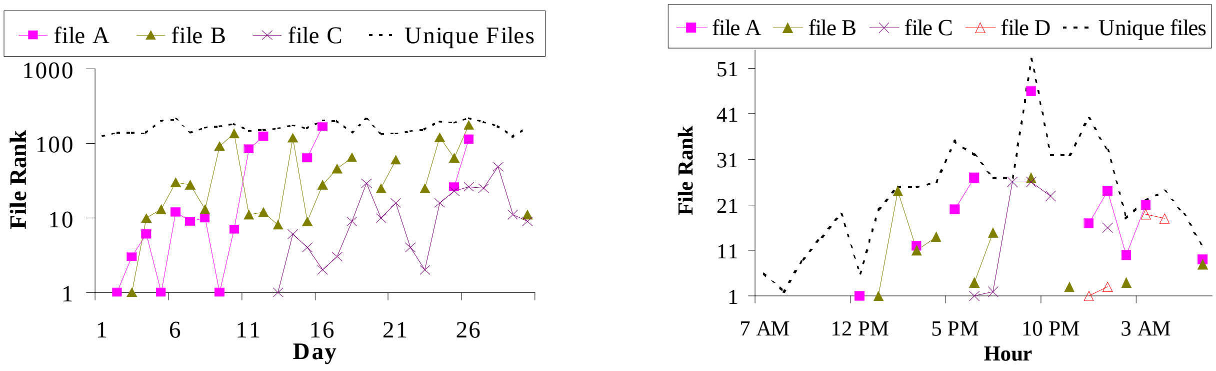
File Size (mins)	Distribution		Mean, σ (minutes)	
	Most Observed	Other	Min Mean, σ	Max Mean, σ
0-5	Lognormal	Exponential	0.74, 0.81	1.5, 2.5
50-55	Gamma + Pareto	Gamma + Lognormal	20, 30	26, 46

4.1 File Access Frequency Distribution

To characterize media file access frequencies on the BIBS and eTeach servers, we first plotted the daily and hourly access frequency rank of various files over many different periods of time for each server. As illustrated in Figure 7 for the BIBS server, there are typically significant changes in the relative popularity of each file per day and even per hour. An interesting topic for future research suggested by the data in Figure 7(a) is whether analysis of historical access frequency per day for a given file (e.g., using machine learning techniques) might be able to predict access frequency on the next day with sufficient accuracy to provide decision support for proxy prefetching strategies.

As illustrated in Figure 7(b), we found very few periods of stable file popularity on either server over all of the days that we analyzed. In this section, we characterize the distribution of file access frequencies during each such period. In addition, this paper analyzes the distribution of file access frequencies for many different individual days on each server, recognizing that further research is needed to characterize how the file accesses are distributed throughout the day.

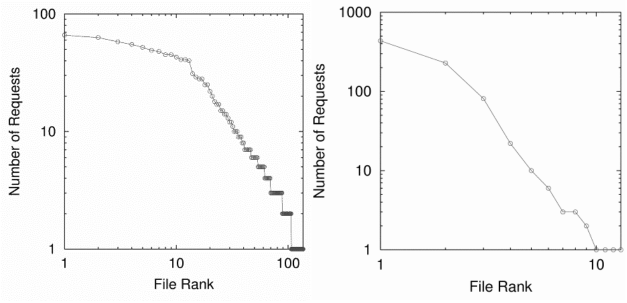
The log-log plot of request count versus file rank is illustrated in Figure 8 for the files accessed during a given day on the BIBS server and on the eTeach server. In Figure 8a (8b), the curve is



(a) Daily, April

(b) Hourly, 5/14

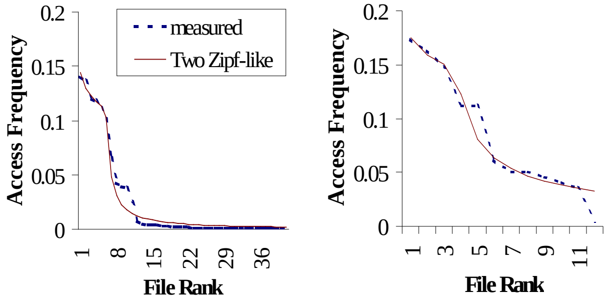
Figure 7: Example Evolution of File Popularity (BIBS)



(a) BIBS 5/9, File Size: 50-55 mins

(b) eTeach 9/27, File Size: 0-5mins

Figure 8: Observed File Access Frequencies (log-log plots)



(a) BIBS 3/23 6pm-1am, File Size: 50-55min

(b) eTeach 10/11 0am-11pm, File Size: 0-5min

Figure 9: Example Distributions of File Access Frequencies

Table 4: Typical Parameter Values for the Distribution of File Access Frequencies (BIBS and eTeach)

File Size (minutes)	Server	Day	First Zipf Distribution			Second Zipf Distribution		
			Total Prob.	# files	α	Total Prob.	# files	α
0–5	eTeach	9/13, 9/14, 9/19, 9/20, 10/3	0.7–0.85	2	2–2.5	0.15–0.3	3–12	1–1.5
		9/21, 10/4	0.6	2	1.7	0.4–0.45	9-17	0.8
		9/26, 9/27	0.84	2	0.9–1.25	0.16	7–11	2
		10/10, 10/11	0.8	7	0.2–0.25	0.2	7	1.2
	BIBS	2/20, 3/23	0.1–0.2	2	0	0.8–0.9	56–85	0.45–0.55
		5/8, 5/9, 5/14	0.2	2-4	0.3-0.33	0.8–0.85	30–60	0.35-0.38
5–10	eTeach	9/27	0.8	2	0.3625	0.2	2	1.6
		10/09	0.85	2	0.5208	0.15	2	10.0471
10-15	eTeach	9/19	0.89	2	3.6	0.11	2	0
		9/20	0.97	2	5.4	0.03	2	2.4
50–55	BIBS	2/21, 2/22	0.5–0.6	6–9	0.4–0.45	0.4–0.5	58–72	0.6–0.65
		5/8, 5/9, 5/10, 5/11	0.5–0.6	12–18	0.2–0.35	0.4–0.5	110–130	0.45–0.55

approximately linear for the first 13 (2) files, and is also approximately linear for the rest of the files. Similarly, for each day that we analyzed on each server, as well as for each period of stable relative file access frequencies, we found that the log-log plot has two distinct approximately linear regions. The two linear regions are less well-defined for the eTeach server, perhaps due to the smaller total number of files accessed per day in eTeach.

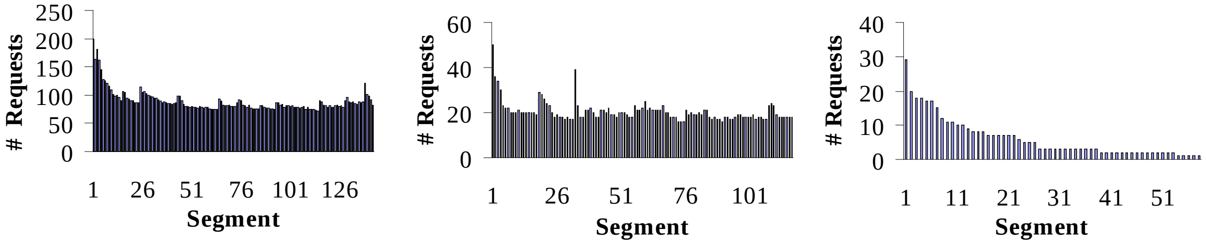
Thus, on *both* servers the distribution of file access frequencies to all media files, or to all media files of a given duration (e.g., under 5 minutes, or 50-55 minutes), can be approximated very closely by a *concatenation* of *two* Zipf-like distributions,[1] as illustrated in Figure 9. This contrasts with previous observations of access frequencies for Web HTML and image files [2,5,6], videos at a video rental store [10], and the one-week multi-server streaming workload from a given client population [9], which are accurately modeled by a single Zipf-like distribution.

The number of files, the total access probability and the Zipf parameter α in each region depend on the period analyzed and the file size. Table 4 provides typical values for these parameters.

4.2 Segment Access Frequency Distribution

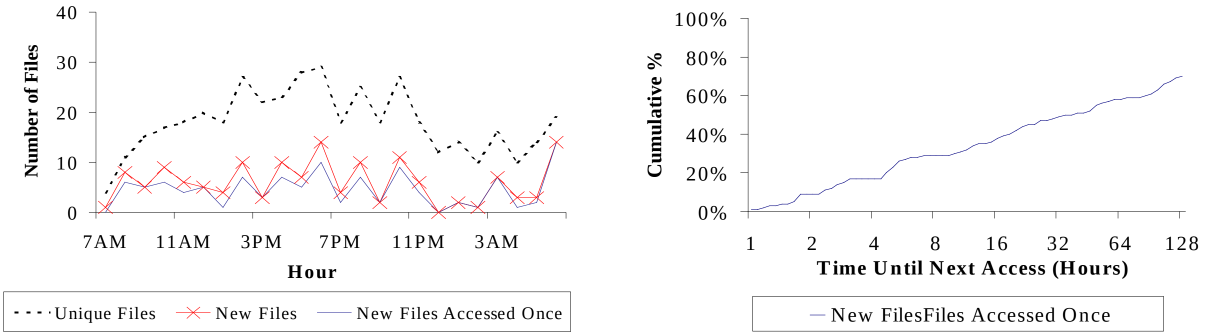
Recent papers have suggested prefix caching and other partial file caching strategies for media files. To provide insight into the design of partial file caching strategies, this section analyzes the distribution of the number of accesses to each ten-second segment of the eTeach media files. (As noted in Section 2, the BIBS logs analyzed in this paper do not provide data on which segments within a media file are accessed during a client session.)

[1] In these Zipf-like distributions, access frequency for file of rank i is equal to C/i^{α}, $\alpha > 0$, where C is a constant such that the total access frequency for the files in the distribution is equal to the measured total access frequency to the given files.



(a) Most Accessed File (b) 4th Most Accessed File (c) 6th Most Accessed File

Figure 10: Example Distributions of Accesses to 10-second File Segments (eTeach, 9/27)



(a) Number of File Accesses per Hour (b) Distribution of Time Until Next Access

Figure 11: Accesses to New Files Each Hour (BIBS 3/23, n=4)

Figure 10 illustrates the observed number of accesses to each ten second segment, for each of three files that have different file access frequency rank on a given day. An interesting result illustrated in Figures 10(a) and (b) is that, although typically 50% of eTeach requests have stream duration less than one minute, each ten-second segment of the most frequently accessed media files is accessed approximately equally often. This was true on *every day that we analyz*ed. On the other hand, media files that have lower access frequency rank (e.g., below the median for each day analyzed) have skewed access frequency with earlier segments of the media typically being accessed more frequently than later segments, as illustrated in Figure 10(c). An important question is whether similar observations hold for other media servers, or whether the observations will hold as media servers evolve in sophistication. If so, it may not be necessary to measure access frequency to each file segment in order to determine which segments should be cached, particularly for the most frequently accessed files.

4.3 Infrequently Requested Files

Another important consideration when caching streaming media files at a proxy server is that inserting a new file in the cache requires a significant disk write overhead. If each request for a file that is not cached at the proxy causes the file to be placed in the cache, significant write overhead might be incurred even for files that are not accessed at least one more time before they are evicted from the cache. In this section, we provide some insight into the significance of this issue for the BIBS and eTeach server workloads. Parallel work by Chesire et al. [9] shows that 78% of the media files accessed during a one-week period were accessed by the client population studied only once during the week.

Table 5: Summary of Accesses to New Media Files (BIBS)

Day	n	Median # new	Average # new accessed once / # new	Time Until Next Access			
				% > 4 hr	% > 8 hr	% > 16h	% > 32h
2/2	2	5	0.70	62	54	48	38
	8	5	0.67	71	64	58	51
3/22	2	10	0.78	71	59	50	40
	8	6	0.81	84	72	67	61
3/23	2	8	0.70	73	64	54	44
	8	4	0.79	82	76	68	57
5/14	2	14	0.81	65	54	39	26
	8	10	0.84	74	64	49	34

We define the *new media files* (or new media segments) accessed in a given hour to be the files (or segments) that were not accessed in the previous *n* hours, where *n* = 1, 2, 4, or 8. Figure 11(a) shows that for *n* = 4 and a given day on the BIBS server, a very high fraction (i.e., 70-100%) of new files accessed during each hour were accessed *only once* in the hour. Furthermore, Figure 11(b) shows that a significant fraction of these new files accessed only once (i.e., 70%) are not accessed again within the next 8

Table 6: Summary of Accesses to New Media Segments (eTeach)

Day	n	Median # new	Avg # new accessed once / # new	Time Until Next Access			
				% > 4 hr	% > 8 hr	% > 16h	% > 32h
9/15	2	10	0.82	75	66	66	59
	8	4	0.82	72	71	71	63
9/27	2	10	0.81	69	68	62	60
	8	8	0.78	80	78	67	64
10/3	2	2	0.70	83	83	79	56
	8	2	0.87	84	84	79	57
10/10	2	12	0.62	58	53	53	52
	8	5	0.7	64	55	55	54

hours. Similar results were obtained for *n* = 1, 2, 4, and 8, and for every day that we analyzed on both the BIBS and eTeach servers, as shown for the BIBS server in Table 5. The evidence is similar for accesses to files or to *one-minute segments* within the videos on the eTeach server, as shown in Table 6. These results imply that caching a media file or segment based on a single client request, without further information about the file popularity, may reduce disk bandwidth available to deliver files to clients with little benefit in many cases. This motivates the need to reevaluate the traditional cache-on-first-access strategy for media files.

5. CLIENT INTERACTIVITY

5.1 Multicast Delivery

In eTeach, 90% of the requests are for fewer than three minutes of video, indicating a high degree of interactivity. To quantify how much server bandwidth can be saved by multicast delivery in eTeach, we simulated the Closest Target multicast technique described in [11] for client traces during periods of stationary total request rate, varying from 10 to 70 requests per hour.

For the three files with highest request count in each period, Table 7 compares the average number of concurrent multicast streams to the average number of concurrent unicast streams needed to deliver the file during the period. The results show that although many stream durations are short, the bandwidth savings for the most popular files are between 40-60% for each of the periods simulated. Further bandwidth savings might be expected for higher client loads (e.g., courses with higher enrollment), although the precise way in which the bandwidth savings scale with the size of the client population are difficult to predict.

The results in Table 7 indicate that multicast delivery methods may be useful even for media servers that have significant client interactivity. Parallel work [9] reaches a similar conclusion after showing that, for requests from a given client population in which a significant fraction of the streams had short duration, two overlapped requests for the same media file tend to have a high fraction of overlap.

Table 7: Server Load Reduction for Multicast Delivery (eTeach)

Period	File rank	Avg # Concurrent Streams		
		Unicast	Multicast	Savings
9/27, 10pm-11pm	1	3.28	1.34	59%
9/27, 10am-5pm	2	0.33	0.18	45%
9/27, 2pm-5pm	2	0.32	0.14	57%
9/27, 10am-7pm	3	0.11	0.06	46%
10/3, 4pm-9pm	1	0.27	0.14	49%
10/3, 3pm-4pm	2	0.69	0.32	54%
10/10, 10am-11am	1	1.38	0.79	43%
10/10, 10am-12am	1	1.48	0.89	39%
10/10, 5pm-6pm	3	0.42	0.24	44%

5.2 Session Characteristics

To create a synthetic workload model for streaming media servers, one needs a session request arrival distribution (Section 3.2), a file access frequency distribution (Figure 1 and Section 4.1), and a model of the interactive requests that occur during the session. This section characterizes interactive requests that occur during sessions on the eTeach server, since interactive requests within a session are not recorded in the BIBS logs.

We analyze the average number of interactive requests per session, the frequency of each interactive request type (e.g., pause, jump forward, etc.), and the distribution of ON and OFF times in the session. Previous characterizations of interactive media sessions [12, 14, 17] have computed some of these statistics only for the aggregate of all sessions in the client workload that was analyzed, whereas we provide a summary of the variations in the statistics obtained for each of different days and for each file that had more than 20 sessions during the given day. Similarities and differences between the previously reported aggregate statistics and the more detailed eTeach session characteristics are noted as the results are presented below.

Sessions, which consist of a sequence of alternating media ON and OFF times for a given client, are not identified in the eTeach logs. Analysis over many periods did not yield clear peaks in the OFF time distribution as observed for simulated sessions in which the inter-session OFF times have a distinct distribution from the intra-session OFF times. On the other hand, similar to results in [17], OFF times is eTeach are heavily skewed, with typically 90% of OFF times being under 4 minutes and very few off times between 4 and 20 minutes. We thus define a session to be a sequence of requests to the *same file* from the same IP address such that each OFF time is no greater than 20 minutes. The session characteristics (discussed below) are nearly identical for smaller upper bounds on the OFF time such as 4 minutes.

Fast forward and rewind operations are extremely rare in eTeach (i.e., less than 0.5% of all requests in the logs) because the customized browser interface encourages use of markers in the

Table 8: Summary of Interactive Requests Per Session Per File (eTeach)

File Size (minutes)	# Sessions	Avg # Interactive Requests Per Session (min, typical, max)	% Pause (min, typical, max)	% Jump Forward (min, typical, max)	% Jump Back (min, typical, max)
0-5	992	0.35, 0.6-1.0, 1.8	5, 14-30, 55	0-3.5, 35	29, 45-95
5-15	449	1-3, 5.5	30, 40-55,70	12, 20-35, 50	11, 20-35
15-25	517	3-6	35-50, 60	15, 20-35, 45	20-35
30-35	411	2.5,4-5, 9	30-40 & 60-75	10-20 & 40-50	15-30

Table 9: Distribution of ON Times Per File (eTeach)

File Size (min)	Distribution		Mean, σ (minutes)	
	Most Observed	Other	Min Mean, σ	Max Mean, σ
0-5	Exponential	Pareto	0.36,0.17	0.87,0.81
5-35	Pareto Weibull	Lognormal Gamma + Pareto	0.9, 1.4	3.7,3.8

Table 10: Distribution of OFF Times Per File (eTeach)

File Size (min)	Most Observed Distribution	Mean, σ (minutes)	
		Min Mean, σ	Max Mean, σ
0-5	Pareto Exponential	0.46,0.84	1.0,0.89
5-35	Pareto Lognormal Weibull	0.43,0.88	2.1, 5.8

lectures. For each other type of interactive request, Table 8 provides the range of frequency typically observed over all sessions for a given file in the given file size range, as well as the minimum and maximum observed frequency if outside the typical range. As in [12], the average number of interactive requests per session increases somewhat with the file length. Although the average number varies from day to day, the average is always under 10 interactions per session, similar to the workload in [12] where 83% of the sessions have four or fewer interactions. For short files, jump backwards is the most common client interaction. For larger file sizes, pauses become more frequent, and like the workload in [14] jump forward has approximately the same frequency as jump backward. In constrast, the workload analyzed in [17] had a strong predominance of jump forwards.

There are too few sessions for each file during any period of stationary client arrival rate, even on high load days, to determine the distribution of session interarrival times for the file. Instead, we analyze the eTeach session interarrival time distribution for all sessions that started in a period of stable session arrival rate. Curve fitting of the measured data indicates that, unlike the Poisson session arrival process in BIBS, session arrivals at eTeach have either a Weibull distribution or a combination of Weibull for the body and Pareto for the tail (except on September 27, which had Poisson session arrivals).

The distribution of session ON times is summarized in Table 9. As with the distribution of media delivered per session in BIBS, the distribution of ON times in an eTeach session depends on the file length. The distribution is exponential (or occasionally lognormal) with mean under one minute for file lengths under 5 minutes, and becomes heavier tailed as the file length increases. OFF times have similar distributions, as shown in Table 10, with exponential for small files, and Pareto, lognormal or Weibull for larger files. Similar distributions of ON and OFF times were observed for the workload in [17]. Note that some of the heavy tailed distributions of OFF times in eTeach have maximum OFF time under 20 minutes; thus, the heavy tail in not due to spurious intersession OFF times in the sessions defined for the analysis

6. CONCLUSIONS

This paper has presented an extensive analysis of the client workloads for educational media servers at two major universities. The workloads were characterized in terms of daily and hourly session arrival rate, session interarrival time distributions, distribution of media delivered per session, distribution of file and file segment access frequencies, number of interactive requests per client session, and media ON and OFF times within a client session. These characteristics can be used to create synthetic media client workloads for evaluating alternative delivery and caching algorithms. Furthermore, we found that on each server, a high fraction of the new files or new file segments accessed during a given hour are accessed only once during the hour and are not accessed again for at least another eight hours. This implies that the traditional cache-on-first-access strategy may incur excess disk write overhead for caching streaming media. Trace simulations of the interactive requests for the most popular files in eTeach showed that although 90% of the media streams have duration under three minutes, multicast delivery techniques could reduce the server bandwidth used to deliver those files by 40-60%.

The logs analyzed in this study represent a snapshot of a particular class of streaming servers (i.e., educational servers at universities). Education server content is rapidly evolving in sophistication, as is client sophistication in viewing the content. The results in this paper provide data about current server workloads and a benchmark against which future server workloads can be compared.

7. ACKNOWLEDGMENTS

The authors thank Mike Litzkow, Anirban Mahanti, Larry Rowe, and David Sundaram-Stukel for helpful comments on the analysis in this paper. Larry Rowe and Mike Litzkow provided the logs and other valuable information about the BIBS and eTeach systems, respectively.

8. REFERENCES

[1] S. Acharya, B. Smith and P. Parnes, Characterizing User Access To Videos On The World Wide Web, in *Proc SPIE/ACM Conf. on Multimedia Computing and Networking*, San Jose, CA, Jan. 2000.

[2] V. Almeida, A. Bestavros, M. Crovella and A. de Oliveira, Characterizing Reference Locality in the WWW, in *Proc. IEEE Int'l. Conf. on Parallel and Distributed Information Systems*, Miami Beach, Dec. 1996.

[3] M. Arlitt and C. Willliamson, Web Server Workload Characterization: The Search for Invariants, in *Proc. ACM Sigmetrics*, Philadelphia, PA, May 1996.

[4] M. Arlitt and T. Jin, Workload Characterization of the 1998 World Cup Web Site, *IEEE Network*, Vol.14, No. 3, May/June 2000.

[5] P. Barford, A. Bestavros, A. Bradley and M. Crovella, Changes in Web Client Access Patterns: Characteristics and Caching Implications, *World Wide Web Journal, Special Issue on Characterization and Performance Evaluation*, Dec. 1998.

[6] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, Web Caching and Zipf-like Distributions: Evidence and Implications, *Proc. IEEE INFOCOM '99*, New York City, NY, March 1999.

[7] M. Crovella and A. Bestavros, Self-similarity in World Wide Web Traffic: Evidence and Possible Causes, *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, Dec. 1997.

[8] C. Cunha, A. Bestavros and M. Crovella, Characteristics of WWW Client-based Traces, Technical Report TR-95-010, Computer Science Dept., Boston Univ., June 1995.

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

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

[11] D. L. Eager, M. K. Vernon and J. Zahorjan, Minimizing Bandwidth Requirements for On-Demand Data Delivery, *Proc. 5th Int'l. Workshop on Multimedia Information Systems*, Indian Wells, CA, Oct. 1999.

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

[13] K. A. Hua, Y. Cai and S. Sheu, Patching: A Multicast Technique for True Video-on-Demand Services, *Proc. 6th ACM Int'l. Multimedia Conf.*, Bristol, U.K., Sept. 1998.

[14] N. Harel, V. Vellanki, A. Chervenak, G. Abowd, U. Ramachandran, Workload of a Media-Enhanced Classroom Server, *Proc. IEEE Workshop on Workload Characterization*, Oct. 1999.

[15] *Windows Media Services SDK, Version 4.1*, Microsoft Corporation. http://msdn.microsoft.com/workshop/imedia/windowsmedia/sdk/wmsdk.asp

[16] V. Paxson and S. Floyd, Wide-Area Traffic: The Failure of Poisson Modeling, *IEEE/ACM Trans. on Networking*, Vol. 3, No. 3, June 1995.

[17] J. Padhye and J. Kurose, An Empirical Study of Client Interactions with a Continuous-Media Courseware Server, *Proc. NOSSDAV '98*, July 1998.

[18] *RealServer Administration Guide – RealSystem G2*, RealNetworks, Inc., Nov. 1998.
http://docs.real.com/docs/serveradminguideg2.pdf

[19] W. Willinger, M. Taqqu, R. Sherman and D. Wilson, Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, *Proc. ACM SIGCOMM '95*, Cambridge, MA, Aug. 1995.

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

[21] http://eteach.cs.wisc.edu/index.html

[22] http://bmrc.berkeley.edu/bibs