Analysis of Enterprise Media Server Workloads: Access Patterns, Locality, Dynamics, and Rate of Change

Ludmila Cherkasova, Minaxi Gupta[1]
Internet Systems and Storage Laboratory
HP Laboratories Palo Alto
HPL-2002-56
March 7th , 2002*

E-mail: cherkasova@hpl.hp.com, minaxi@cc.gatech.edu

workload analysis, enterprise media servers, static locality, temporal locality, sharing patterns, dynamics, clients characterization, CDNs

The main issue we address in this report is the workload analysis of today's enterprise media servers. This analysis aims to establish a set of properties specific for enterprise media server workloads and to compare them with well known related observations about web server workloads. We propose two new metrics to characterize the dynamics and evolution of the accesses, and the rate of change in the site access pattern, and illustrate them with the analysis of two different enterprise media server workloads collected over a significant period of time. Another goal of our workload analysis study is to develop a media server log analysis tool, called **MediaMetrics**, that produces a media server traffic access profile and its system resource usage in a way useful to service providers.

* Internal Accession Date Only Approved for External Publication
[1] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332
A shorter version of this paper is to be published in ACM NOSSDAV 2002, the 12th International Conference on Network and Operating System Support for Digital Audio and Video, 12-14 May 2002, Miami Beach, Florida.

Analysis of Enterprise Media Server Workloads: Access Patterns, Locality, Dynamics, and Rate of Change

Ludmila Cherkasova
Hewlett-Packard Laboratories
1501 Page Mill Road,
Palo Alto, CA 94303, USA
cherkasova@hpl.hp.com

Minaxi Gupta
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
minaxi@cc.gatech.edu

Abstract

The main issue we address in this paper is the workload analysis of today's enterprise media servers. This analysis aims to establish a set of properties specific for enterprise media server workloads and to compare them with well known related observations about web server workloads. We propose two new metrics to characterize the dynamics and evolution of the accesses, and the rate of change in the site access pattern, and illustrate them with the analysis of two different enterprise media server workloads collected over a significant period of time. Another goal of our workload analysis study is to develop a media server log analysis tool, called **MediaMetrics***, that produces a media server traffic access profile and its system resource usage in a way useful to service providers.*

Keywords*: workload analysis, enterprise media servers, static locality, temporal locality, sharing patterns, dynamics, clients characterization, CDNs.*

1 Introduction

Streaming media represents a new wave of rich Internet content. Recent technological advancements in video creation, compression, bandwidths, caching, streaming, and other content delivery technology have brought audio and video together to the Internet as rich media. Products for still (JPEG) and motion (MPEG) pictures are available in consumer markets. This enables potentially anyone to be a producer of rich media content that can be easily distributed and published over the Internet. There are predictions that rich media will significantly add to the user experience, and therefore, will be the Internet's next "killer app."

Video from news, sports, and entertainment sites is more popular than ever. Media servers are being used for educational and training purposes by many universities. Use of the media servers in the enterprise environment is catching momentum too. Enterprises are using more and more rich media to attract prospective customers, improve effectiveness of online advertising, web marketing, customer interaction centers, collaboration, and training.

While there are plenty of reasons to create rich media content, delivering this high bandwidth content over the Internet presents a number of new challenges to system designers. Real-time nature of multimedia content makes it sensitive to congestion conditions in the Internet. Moreover, multimedia streams can consume significant bandwidths and their large sizes require orders of magnitude larger amount of storage at the media servers and proxy caches. Understanding the nature of media server workloads is crucial to properly designing and provisioning current and future services.

Recently, there have been several studies attempting to uncover the multimedia workloads characteristics. However, most of the studies are devoted to the analysis of workloads for educational media servers [1, 2, 3, 12, 13, 16]. One recent study [9] characterizes the workload of a media proxy of a large university. Our paper presents and analyzes the **enterprise media server workloads** based on the access logs from two different media servers in Hewlett-Packard Corporation. Both logs are collected over long period of time (2.5 years and 1 year 9 months). The duration of the logs makes them quite unique and allows us to discover typical and specific client access patterns, media server access trends, dynamics and evolution of the media workload over time.

Web workload studies have identified different types of *locality* in web traffic. *Static locality* or *concentration of references* [5] observes that 10% of the files accessed on the server typically account for 90% of the server requests and 90% of the bytes transferred. *Temporal locality* of references [4] implies that recently accessed documents are more likely to be referenced in the near future. These strongly influence the traffic access patterns seen by the web servers. One goal of our analysis is to characterize **locality** properties in media server workloads and to compare them with

traditional web workloads characterization. Understanding the nature of locality will help in designing more efficient middleware for caching, load balancing, and content distribution systems.

The other questions we address in this paper are tightly related to new trends observed in the evolution of Internet infrastructure such as content distribution networks (CDNs) and overlay networks. CDNs are based on large-scale distributed network of servers located closer to the edges of Internet for efficient delivery of digital content including various forms of streaming media. The main goal of CDN's architecture is to minimize the network impact in the critical path of media delivery.

Access patterns and dynamics of the site have to be taken into account when making a decision about different caching or content distribution systems. For example, if the site is very dynamic, i.e. a large portion of the client requests are accessing new content, (news web sites being a prime example), then CDNs are clearly a good choice to handle the load, because traditional caching solutions will be less efficient in distributing the load due to time involved in propagating the content through the network caches.

Thus, the other question we address in this paper is how to characterize the **dynamics** and **evolution of accesses** at media sites. The first natural step is to observe the introduction of new files in the logs, and to analyze the portion of all requests destinate for those files. We define *new files impact* metric that aims to characterize the site evolution due to new content. It is obtained by computing the ratio of the accesses targeting these new files over time. The definition of "new files" depends on a time scale at which information changes and might be different for different sites. We propose a second *life span* metric to measure the rate of change in the access pattern of the site.

We have developed a tool called **MediaMetrics** that characterizes a media server access profile and its system resource usage in both a quantitative and qualitative way. It extracts and reports information that could be used by service providers to evaluate current solutions and to improve and optimize relevant future components. **MediaMetrics** performs an analysis which is entirely based on media server access logs, which can be from one or multiple servers in a cluster. The tool is written in Perl to process the most common media server log formats: from Windows Media Server and RealNetworks Media Server. In this paper, we highlight most interesting part of the statistics available from our tool.

Key new observations from our analysis include:

- Despite the fact that the two studied workloads had significantly different file size distribution (one set had well represented groups of short, medium, and long videos, while the other set was skewed in long videos range), the clients' viewing behavior was similar for both sets: with 77-79% of media sessions being less than 10 min long, 7-12% of the sessions being 10-30 min, and 6-13% of sessions continued for more than 30 min. This reflects the browsing nature of the most enterprise client accesses.

- Most of the incomplete sessions (i.e. terminated by clients before the video was finished) are accessing the initial segments of media files. The percentage of sessions with interactive requests (such as pause, rewind, or fast forward during the media session) is much higher for medium and long videos.

- Like web workloads, both the media workloads exhibit a high locality of accesses: 14-30% of the files accessed on the server account for 90% of the media sessions and 92-94% of the bytes transferred, and were viewed by 96-97% of the unique clients.

- While there is a significant number of files that are rarely accessed (16% to 19% of the files are accessed only once), these numbers are somewhat lower compared to web server workloads.

- The distribution of clients accesses to media files can be approximated by Zipf-like distribution for both workloads. However, noteworthy is that the time scale plays important role in this approximation. We considered 1-month, 6-month, 1-year and a whole log duration as a time scale for our experiments. For one workload, distribution of clients accesses to media files on a 6-month scale starts to fit Zipf-like distribution. While for the other workload, file popularity on a monthly basis can be approximated by Zipf-like distribution. For longer time scale in the same workloads, the file access frequency distribution does not follow Zipfian distribution.

- Accesses to the new files constitute most of the accesses in any given month. Also, the bytes transferred due to accesses to new files are dominant in both workloads. It makes the access pattern of enterprise media sites resemble the access pattern of the news web sites where the most of the client accesses target new information. We introduce the *new files impact* metric to measure site dynamics due to new files. Moreover, we observed that for enterprise media servers, the tendency of the number of accesses to be increasing or decreasing in nature is strongly correlated with the number of newly added files.

- For both workloads, 51-52% of accesses to media files occur during the first week of their introduction. First five weeks of the files' existence account for 70-80% of all the accesses. We define a *life span* metric to reflect the rate of change in accesses to newly introduced files. Additionally, life span metric reflects the timeliness of the introduced files. Longer life span reflects that media information on a site is less timely and have more consistent percentile of accesses over longer period of time.

The remainder of the paper presents our results in more detail. Section 2 discusses related work, briefly describes the sites we used in our study, and provides a short description of the media server log formats. Section 3 describes the media files length and the distribution of the accesses, client viewing behavior specifics, media files encoding rates, available bandwidth to the sessions, QoS related issues, completed and aborted session characteristics, client clustering, and workload trends. Section 4 provides insight in locality characteristics of studied workloads. Section 5 introduces the new files impact metric to capture the media site dynamics over time, and in particular, the trends in access patterns due to the new files. Section 6 defines the life span metric and measures the rate of the site's access pattern changes. Finally, section 7 presents conclusion and future work.

Acknowledgments: Both the tool and the study would not have been possible without media access logs and help provided by Nic Lyons, Wray Smallwood, Brett Bausk, Magnus Karlsson, Wenting Tang, Yun Fu, John Apostolopoulos, and Susie Wee. Their help is highly appreciated.

2 Background

2.1 Related Work

While web server workloads have been studied extensively [4, 5, 6, 8, 10], there have been relatively fewer papers written about multimedia workload analysis. Acharya et al. [1] characterized non-streaming multimedia content stored on web servers. In their later work [2], authors present the analysis of the six-month trace data from mMOD system (the multicast Media on Demand) which had a mix of educational and entertainment videos. They observed high temporal locality of accesses, the special client browsing pattern showing clients preference to preview the initial portion of the videos, and that rankings of video titles by popularity do not fit a Zipfian distribution.

Recent studies on client access to MANIC system audio content [16] and low-bit rate videos in the Classroom2000 system [13] provide the analysis of accesses to educational media servers in terms of daily variation in server loads, distribution of media session durations, and some client interactivity analysis.

Extensive analysis of educational media server workloads is done in [3]. Their study is based on two media servers in use at major public universities in the United States: eTeach and BIBS. The authors provide a detailed study of client session arrival process: the client sessions arrival in BIBS can be characterized as Poisson, and arrivals in eTeach workload are closer to heavy-tailed Pareto distribution. They also observed that media delivered per session depends on the media file length. They discovered different client interactivity patterns for frequently and infrequently accessed files: any video segment is equally likely to be accessed for frequent files, while access frequency is higher for earlier segments in the infrequent videos. The main goal of [3] was to identify the important parameters for generating synthetic workloads.

While all the above papers used media server logs, the study by Chesire et al [9] analyzed the media proxy workload at a large university. The authors presented a detailed characterization of session duration (most of the media streams are less than 10 min), object popularity (78% of objects are accessed only once), server popularity, and sharing patterns of streaming media among the clients.

As the the number of internet users continues to grow, and as the high-speed access methods become more ubiquitous, streaming media starts to occupy more sizable fraction of the Internet's bandwidth. Few recent papers [15, 14, 18] analyze the impact of streaming media on the Internet traffic and the performance of popular Internet real-time streaming technologies.

Our paper builds upon this previous work in a number of significant ways. To our knowledge, this paper is the first study of enterprise media server workloads. Our data is collected over significant period of time, which makes it unique. The duration of this data allowed us to concentrate on the analysis of media server access trends, access locality, dynamics and evolution of the media workload over time, and to propose two new metrics to measure these properties. This type of analysis is new and has not been reported in previous work.

2.2 Data Collection Sites

We use access logs from two different servers:

- **HP Corporate Media Solutions server (HPC)** hosts diverse information about HP: video coverage of major events, keynote speeches, addresses and presentations, meetings with industry analysts, promotional events, product in-

troduction, information related to software and hardware products, and demos illustrating the products usage. Additionally, it has some training and education information. The logs cover almost 2.5 years of duration: from the middle of November, 1998 to the middle of April, 2001. In fact, it is a cluster of media servers. For our analysis, we combined several access logs collected at this cluster. The HPC content is delivered by Windows Media Server [19].

- **HPLabs Media server (HPLabs)** provides information about HP Laboratories, in particular Coffee Talks (monthly, HPLabs wide, hourly meetings), videos of prominent presentations, seminars, meetings, some of the HP wide business related events, Cooltown [1] promotional materials, and some training and educational information. The logs cover 1 year and 9 months duration: from the middle of July, 1999 to the middle of April, 2001. It is an internal server available only for accesses to HP employees. The HPLabs content is delivered by RealServer G2 [17].

2.3 Media Server Log Formats

The media access logs record the information about all the requests and responses processed by a media server. Each line of the access logs provides a description of a user request for a particular media file. Windows Media Server and RealNetworks Media Server have different log formats which we describe in more detail in Appendix A.

For our logs, the transmission protocols used by Windows Media Server and RealNetworks Media Server were UDP and TCP respectively.

The typical fields contain information about the IP-address of the client machine making the request, the time stamp at which the request was made, the filename of the requested document, the advertised duration of the file (in seconds), the size of the requested file (in bytes), the elapsed time of the requested media file when the play ended (a file play can be ended prematurely if the client hit the stop button), the average bandwidth (Kb/s) available to the user while the file was playing, the number of bytes sent by the server, and the number of bytes received by the client etc.

Clients can pause, rewind, fast forward, or skip to a predefined point using a slide bar during their viewing of the requested media files. A *session* is a sequence of client requests corresponding to the same file access and reflecting different client activities during the corresponding file viewing such as pause, fast forward, or rewind actions. We will explicitly distinguish the usage of the term: a session is the access of a particular file and there can be multiple requests within the same session, due to client's interactivity.

Windows Media Server logs contain a separate entry for each client request. Thus, a single session may be comprised of multiple entries in the server access logs. Each log entry has a *start position*, the place where the client started viewing the file; *duration* the client watched the file for; and *client action*, pause/stop/rewind/fast/forward. This is useful information for the analysis of clients' interactive behavior during the media sessions.

RealServer log format allows for similar fields, but unfortunately, the HPLabs access logs did not have these optional fields because the relevant option was turned off. Thus HPLabs workload has only information about client sessions, the client interactivity data are not available for HPLabs workload. There is one entry for each client session in these logs.

3 Workload Characteristization

3.1 Summary Statistics

The overall workload statistics for HPC and HPLabs media servers is summarized in the following Table 1.

	HPC	HPLabs
Duration	29 months	21 months
Total sessions	666,074	14,489
Total Requests	1,179,814	NA
Unique Files	2,999	412
Unique Clients	131,161	2,482
Storage Requirement	42 GB	48 GB
Bytes Transferred	2,664 GB	172 GB

Table 1: Statistics summary for two sites.

In HPC, 471 files corresponded to live streams, while the others were stored content. We excluded them from further analysis.

A glance at the basic statistics shows that HPC media server witnesses more activities and reaches larger client population than HPLabs server. HPLabs server clearly targets more specific, smaller research community at HP, and as a result has a very different, "modest" profile. HPC represents a reasonably busy media server with 300-800 client sessions per weekday and occasional peaks reaching 12000 sessions. HPLabs server is much lighter loaded. By noticing this very obvious difference, it becomes even more interesting whether we can find common properties typical for enterprise workloads in general.

[1] HP's vision of the future, a world where everyone and everything is connected to the web through wired or wireless links.

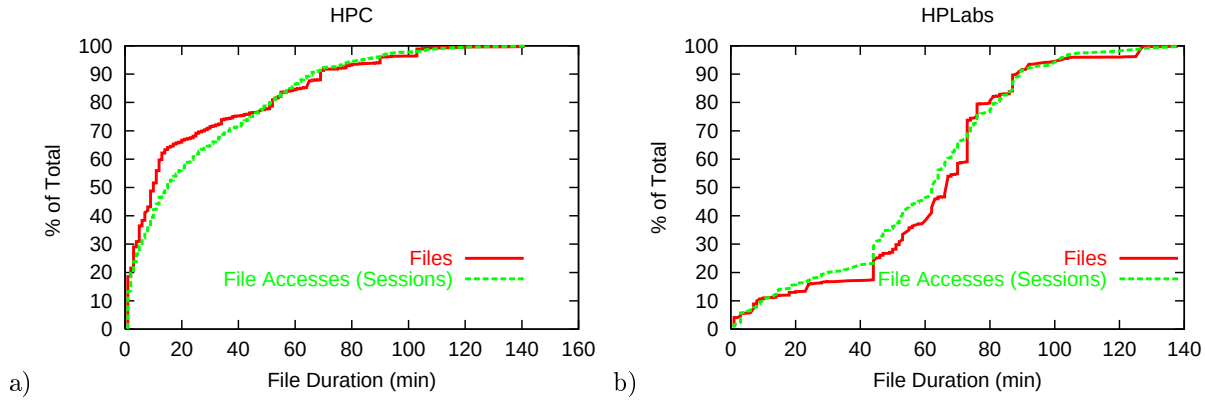


Figure 1: Distribution of file durations and distribution of client sessions to those files: a) HPC and b) HPLabs.

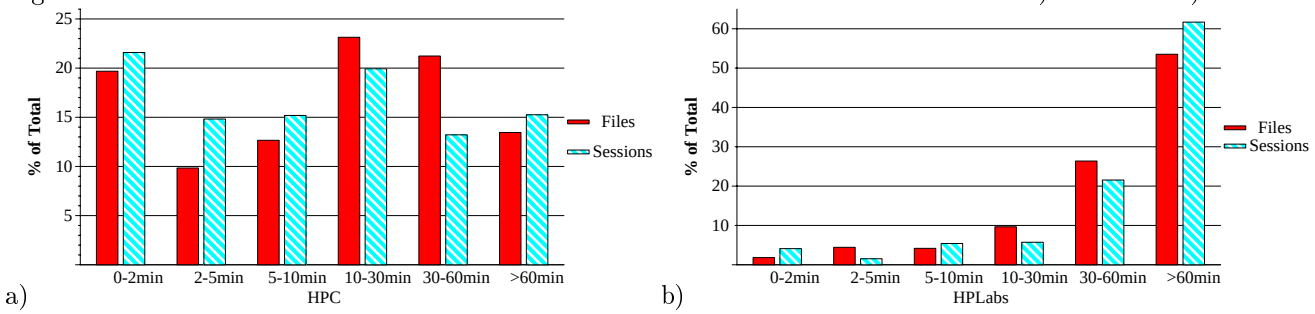


Figure 2: Six classes of file durations and percentage of client sessions to those file classes: a) HPC and b) HPLabs.

3.2 Files and Session Characteristics

In this section, we provide a detailed analysis of advertised duration of media files referenced in the logs and compare that to clients' actual viewing time distribution. The advertised media file duration reflects the total length of the video, while the client can stop viewing or downloading the file by hitting stop button before the video is finished, and it can do so after a sequence of pause, rewind, fast forwards, and using the slide bar to jump to specific sections of the video.

Figure 1 shows the distribution of stored videos for both workloads, and percentage of corresponding accesses to those files.

To simplify the analysis, we created 6 duration classes for considered files: three groups of short videos: 1) less than 2 min, 2) 2-5 min, 3) 5-10 min; one group of medium size videos: 4) 10-30 min, and two groups of long videos: 5) 30-60 min, and 6) longer than 60 min. Figure 2 shows the distribution of stored videos in defined above 6 duration classes, and percentage of corresponding sessions to those files.

Figure 2a) shows that for HPC workload, the content is well represented by videos of different durations: 42% of files belong to a short video group (less than 10 min), 23% of files are in a medium video group, and 34% of files belong to a long video group.

HPLabs workload is strongly skewed in favor of long videos: 7% of videos are in medium group, and 79% of files belong to a long video group.

The interesting characterization is that the percentage of clients accesses is proportional to the percentage of files in each of those file duration categories for both workloads! This implies that each of the file duration groups is equally likely to be accessed by clients. This property is very useful for synthetic workload generation, since it proposes a simple model of defining a media file duration distribution and percentage of corresponding client accesses to those files. For web server workloads, the most of accesses (70%-90%) are for the group of small and medium size documents [5].

However, when we analyzed the actual duration for which clients viewed the videos, the statistics changes dramatically for both workloads as shown in Figure 3. Notice that statistics presented by this graph reflects the overall client viewing time distribution, it is not correlated with the actual media files duration. Most of the viewed media sessions, 50%-60%, were less than 2 min long.

In spite of significant difference in the original file size distribution, the actual duration for which clients viewed the videos, shown in Figure 3, was similar for both sets: with 77-79% of media sessions being less than 10 min long, 7-12% of the sessions being 10-

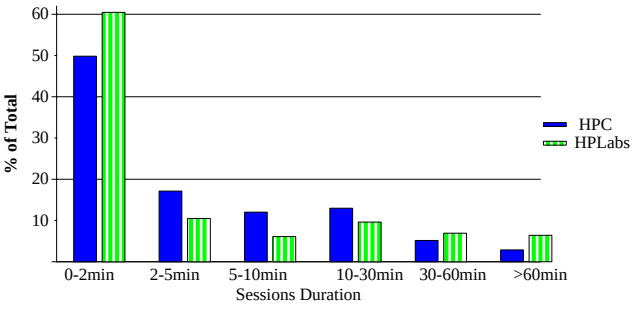


Figure 3: Session duration characterization.

30 min, and 6-13% of sessions continued for more than 30 min. The decision to abort the session is only partially influenced by the available bandwidth for the session (we will show this analysis in more detail in Section 3.4). The observed access pattern rather reflects the browsing nature of the enterprise client accesses, and that often clients are looking for a specific fragment of content in a video, and are not interested in watching it completely. Knowledge of the approximate percent of "browsing" clients helps to estimate and predict the short term load on a server.

3.3 Media Files Encoding Rates and Sessions Available Bandwidth

Both servers, HPC and HPLabs, had videos encoded at different rate.

Table 2 presents the statistics on files encoding rates and their trends over time for both workloads. Videos stored at HPC server had most of the files (59%) encoded at 56 Kb/s rate and lower. However, over the years, the trend is to add more files encoded at higher rate: for example, in 1999 year, only 1.7% of the videos were encoded at a rate between 128-256 Kb/s, while in 2001 this group of videos constitutes already up to 27.8% of total. HPLabs server had most of the files encoded at high bit rate: 67% of all the files are encoded at 256 Kb/s and higher.

Media access logs report the average bandwidth available to the user while the file was playing. Table 3 presents the statistics on session available bandwidth during the different period represented by the logs for both workloads (1999, 2000, 2001 years). HPC media sessions overall had higher available bandwidth to the clients: in 2001, 57.5% of sessions had an average available bandwidth above 56Kb/s (we will call these sessions as *high-bandwidth sessions*). For HPLabs workload, in 2001, high-bandwidth sessions constituted only 21.7% of total.

For HPC workload, most of file encoding rates and average available bandwidth per session show a good allingment as shown in Figure 4. Only the group of videos encoded at rates between 128-256 Kb/s could not meet the requirements. While for HPLabs workload, where the most of the files were encoded at 256 Kb/s and higher, the gap between the demand and available bandwidth is very high: most of the sessions have significant mismatch between the file encoding rates and the available bandwidth.

This information, provided by **MediaMetrics**, could be used by the service providers to analyze the client bandwidth availability for choosing the right encoding rates.

Media access logs also report the number of bytes sent by the server and the number of bytes received by the client. **MediaMetrics** tool uses this information to make an estimate about the percentage of bytes lost during the file transfer, and to implicitly judge the quality of service a client might have experienced. This simple technique can produce useful results when data is transmitted over UDP, because in this case, the difference in sent and received bytes reflects the percentage of the bytes lost on a way to a client. It might be less accurate when data is transferred over TCP because in the presence of congestion, media server will retransmit part of the data to compensate for lost packets, and if those packets were received by the client in time, then the difference in server sent-bytes and the client received-bytes will not always explicitly result in worse QoS. For two workloads under study, HPC data was transmitted over UDP, while for HPLabs workload, the data was transferred over TCP protocol.

The quality of service observed by the low- and high-bandwidth sessions in HPC workload was practically the same: 96.5% of low-bandwidth sessions had 0-5% of bytes loss per session, and there were 97.1% of high-bandwidth sessions with the same quality of service.

For HPLabs workload, the difference in server sent-bytes and the client received-bytes between low- and high-bandwidth sessions was more pronounced: 64.6% of low bandwidth sessions and 88.8% of high bandwidth sessions had 0-5% of bytes loss per session. This numbers reflect the essential role of available bandwidth for viewed media sessions over TCP.

3.4 Completed and Aborted Session Characteristics

We will call a media session as *completed* if during this session the video was watched entirely. For HPC workload, 29% of sessions were completed, while for HPLabs workload, completed sessions accounted for only 12.6% of all sessions. Figure 5 a) shows the overall distribution of completed sessions duration, while Figure 5 b) presents a simplified view of the same dis-

Period	HPC				HPLabs			
Encoding Rate	≤56Kb/s	56-128Kb/s	128-256Kb/s	≥256Kb/s	≤56Kb/s	56-128Kb/s	128-256Kb/s	≥256Kb/s
Files (1999)	73.5%	22.4%	1.7%	2.7%	16%	7%	22%	55%
Files (2000)	56%	27.4%	15.7%	1%	10%	5%	16%	69%
Files (2001)	53%	18.2%	27.8%	1%	13%	2%	17%	68%
All files	59.1%	20.61%	19.4%	0.9%	11%	5%	17%	67%

Table 2: Trends in files encoding rates for both workloads.

Period	HPC			HPLabs		
Bandwidth	≤56Kb/s	56-128Kb/s	≥128Kb/s	≤56Kb/s	56-128Kb/s	≥128Kb/s
Sessions (1999)	57.8%	42%	0.2%	71.4%	14.5%	14.1%
Sessions (2000)	40.3%	52.2%	7.5%	79.1%	16.6%	4.3%
Sessions (2001)	35.8%	57.5%	6.7%	78.3%	18%	3.7%

Table 3: Trends in average available bandwidth per session for both workloads.

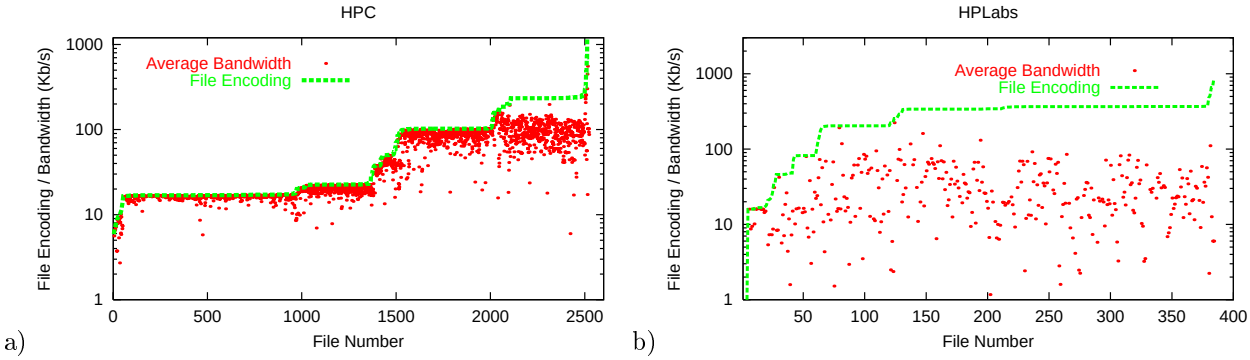


Figure 4: File encoding rates and average available bandwidth of client sessions to those files: a) HPC and b) HPLabs.

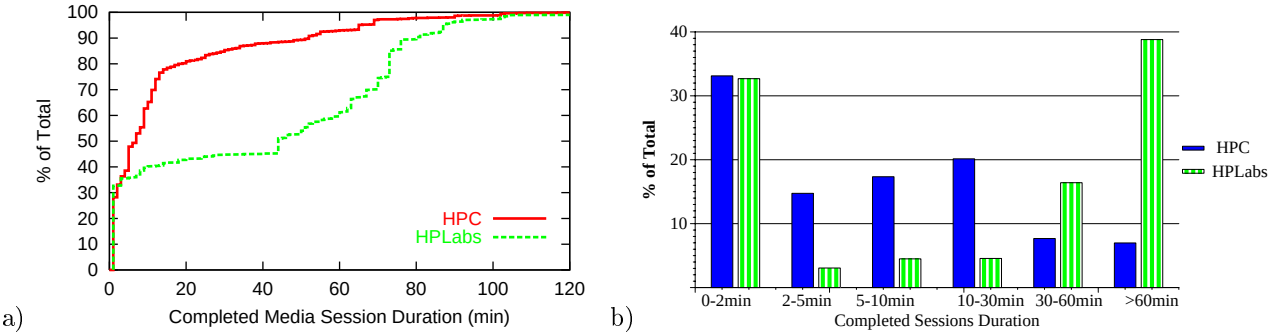


Figure 5: a) Distribution of completed session durations. b) Simplified distribution of completed sessions for six duration classes.

Session Type	HPC			HPLabs		
Bandwidth	≤56Kb/s	56-128Kb/s	≥128Kb/s	≤56Kb/s	56-128Kb/s	≥128Kb/s
All sessions	43.3%	51.1%	5.6%	75%	15.5%	9.5%
Completed sessions	33.9%	60.7%	5.4%	77.3%	11.5%	11.2%
Aborted sessions	47.1%	47.4%	5.5%	74.8%	15.8%	9.4%

Table 4: Distribution of available bandwidth per session for both workloads.

tribution via six duration classes.

Media sessions with duration under 2 min account for 33% of all the completed sessions for both workloads. While for the rest of completed sessions, their durations reflect the corresponding distribution of media session durations specific to considered workloads as shown in Figures 1 and 2.

HPC media sessions overall had higher available bandwidth to the clients compared to the HPLabs sessions. Additionally, for HPC workload, there was a good allingment between the file encoding and bandwidth availability requirements as was shown in Section 3.3. The HPLabs workload exhibits a significant mismatch between the files encoding and the available bandwidth per session. It explains why, overall, a higher percentage of HPC sessions were completed compared to HPLabs workload (29% of HPC sessions versus 12.6% of HPLabs sessions were completed).

The reasonable question to ask is whether the completed sessions had higher available bandwidth to the clients? Or in other words, whether the aborted sessions were interrupted because of poor available bandwidth?

Table 4 presents the statistics on available bandwidth for completed, aborted, and all the sessions for both workloads. For HPC workload, completed sessions have higher percentage of high-bandwidth sessions. However, the difference in bandwidth is not high enough to assert that sessions were aborted because of the "poor bandwidth" conditions. For HPLabs workload, the bandwidth characteristics of the completed and aborted sessions are similar, which suggest that client will watch the video while he/she is interested in the video content.

Most of the aborted sessions accessed initial segments of media files. The number of sessions which had incomplete accesses to any other segments of the file other than the beginning, depend on the size of the video: less than 1.5% of sessions in short video group accessed any segment of the video other than the beginning, 2.4% of sessions in a medium video group, 4%-7% of sessions in long video group. Clearly, such knowledge about the client viewing patterns may be beneficial when designing media caching strategies.

3.5 Client Interactivity

Windows Media Server log format has a separate entry for each client request. As a result, we are able to get information such as pause, rewind or fast forward activity by the client during the media session. Unfortunately, similar data was not available for HPLabs workload. Analysis of these fields for HPC logs produced very interesting results. First of all, it revealed that 99.9% of the sessions with interactive requests were *high-bandwidth sessions* with available bandwidth greater than 56 Kb/s. Second, that the percentage of sessions that access medium and long videos have much higher interactivity.

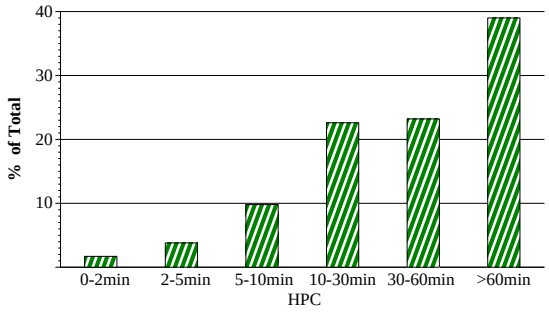


Figure 6: HPC workload: percentage of sessions with interactive requests for different file size classes.

Figure 6 shows that only 15.3% of sessions with interactivity were for a short video group, 22.6% of interactive sessions were for medium size videos, and 62.2% sessions that had client interactivity were for long video group. This statistics helps in better understanding the cliens' viewing behaviors.

3.6 Client Clustering by ASs

MediaMetrics tool provides information about the clients clustering by associating them with various ASs (Autonomous Systems). It also reports the corresponding number of client sessions and percentage of bytes lost for those sessions. Since HPLabs logs only had HP's internal clients, they all belong to the same AS and the results of per AS analysis are not particularly interesting for this case. Here, we present some statistics about the HPC workload.

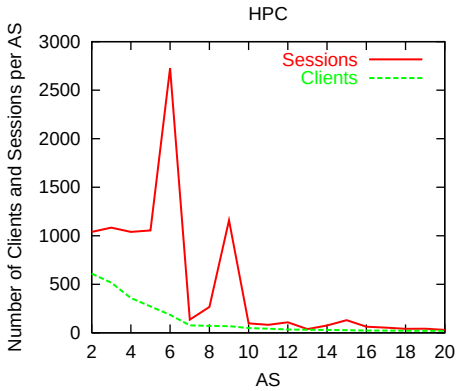


Figure 7: Clients and their sessions clustering by AS.

For HPC logs, client population was spread across 200 different ASs, with 82% of clients being HP in-

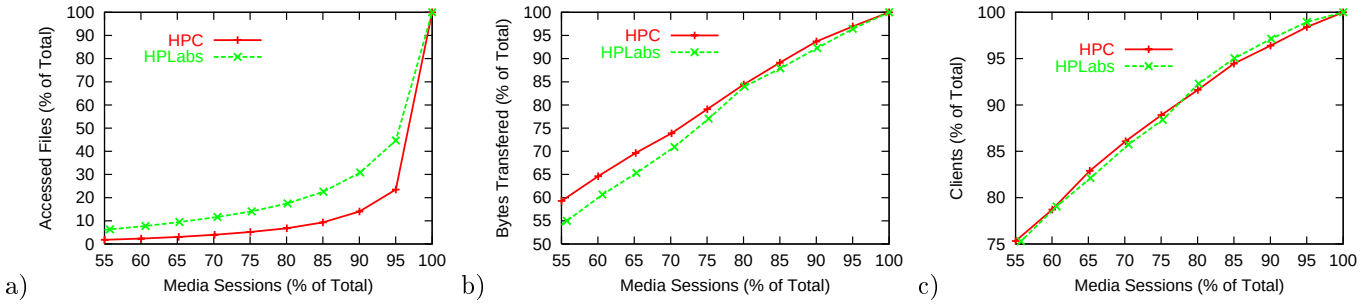


Figure 8: Two workloads compared: a) file set locality, b) bytes-transferred locality c) client set locality.

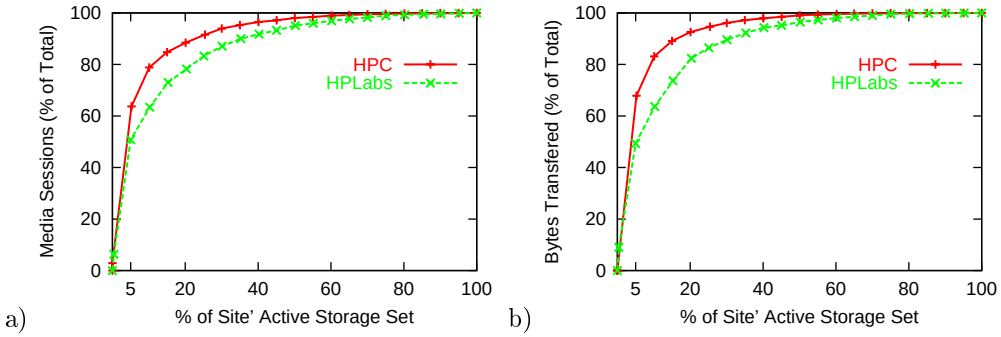


Figure 9: Two workloads compared: storage-set locality and bytes-transferred-storage locality.

ternal. 93% of all the sessions in the logs belonged to internal HP clients. About 45% of all the ASs had just 1 client (with $1-16$ sessions). To show client clustering for ASs that had significant number of clients, we plotted 20 most representative ASs excluding the AS that belonged to HP. Figure 7 shows the number of clients and corresponding number of sessions for the 20 most representative ASs. If we normalize this data then the combined clients from the first 10 ASs account only for 1.6% of clients population and 1.5% of all the sessions. Clearly, the client population profile was dominated by HP internal clients and their activities. For enterprise media servers, this might be a typical usage characterization. Overall, with the spread of CDNs and overlay network technologies, understanding of clients, the content they accessed, and their clustering will play an essential role in deciding efficient placement of edge-servers and the content.

3.7 New Trends Over Time

Analysis of HPC workload over time revealed interesting overall trends in site media content and session characteristics:

- Total number of unique clients accessing media content in each 6 month duration doubled over the duration of our logs.
- Total number of sessions in each 6 month duration also doubled over the duration of our logs.
- Average file size in each 6 month duration increased from less than 7 MB to more than 20 MB in our logs.
- Bytes transferred per session increased from just over 1 MB to over 6 MB in our logs.

4 Locality Characterization

In this section, we will revisit a previously identified invariant for web server workloads. The authors in [5] identified that web traffic exhibits strong concentration of references, "10% of files accessed from the server typically account for 90% of the server requests and 90% of the bytes transferred".

For locality characterization of our logs, we use a table of all files accessed along with their frequency (number of times a file was accessed during the observed period) and the file sizes. This table is ordered in decreasing order of frequency.

Figure 8a) shows the reference locality for the media server access logs used in our study. For both workloads, 90% of the media sessions target 14% of the files for HPC server, and 30% of the files for HPLabs server. This shows high locality of client accesses, though lower than for web workloads. Figure 8b) shows the corresponding bytes transferred due to these media sessions: 94% for HPC site and 92% for HPLabs site. Observed graphs for both workloads

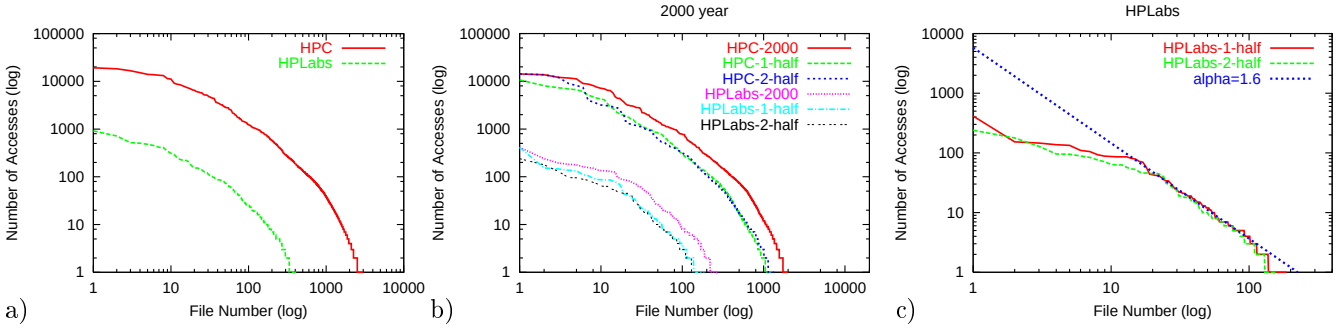


Figure 10: File popularity distribution for both workloads a) over entire period of logs, b) over 2000 year and corresponding first and second 6 months in 2000, c) HPLabs workload 6-month periods with corresponding Zipf-like function fitting.

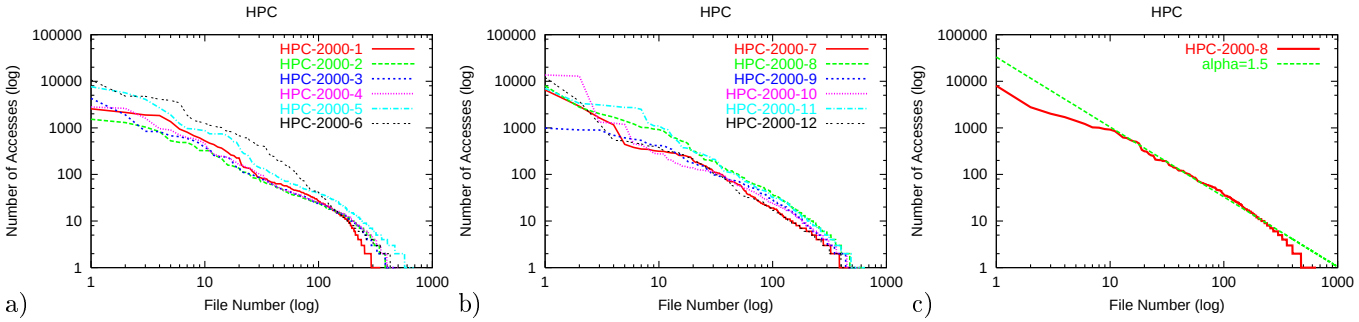


Figure 11: File popularity distribution for HPC workload a) monthly periods, 1st to 6th months b) monthly periods, 7th to 12th months c) 8th month with corresponding Zipf-like function fitting.

are remarkably similar. Figure 8c) shows clients locality for both workloads. It can be interpreted in the following way: at HPC server, 14% of the most popular files (responsible for 90% of accesses) are accessed by 96% of clients. For HPLabs site, 30% of the most popular files are viewed by 97% of the clients.

We also analyzed workload locality from a different angle: what percentage of active storage did the most popular files account for. Here, the *active storage* set is defined by the combined size of all the media files accessed in the logs. For both workloads, we observe a high active storage set locality: 80% to 88% of all sessions are to files that constitute only 20% of the total active storage set as can be seen in Figure 9a). Similarly, 82% to 92% of all transferred, most popular bytes are due to files that constitute only 20% of the total active storage set as can be seen in Figure 9b).

This type of analysis helps in estimating the storage requirements and potential bandwidth savings when using optimizations for the popular portion of the media content. Since these metrics are normalized with respect to the site's active storage set, it allows us to compare different workloads and to identify the similarity inherent to those workloads, independent of the absolute numbers for storage in each workload.

Answering the question: how does the locality characterization in workload vary with a time duration of the logs collection, we found that independent on duration (1-month, 6-month or 12-month durations) both workloads exhibit high locality of client accesses: 90% of the media sessions target 10%-30% of the files for HPC server during corresponding duration interval, and 20%-35% of the files for HPLabs server. This shows a high locality of client accesses in enterprise media server workloads.

Previous studies on web servers and web proxies [7] led to almost universal consensus that web page popularity follows Zipf-like distribution, where the popularity of the i-th most popular file is proportional to $1/i^{\alpha}$. For web proxies, the value of α is typically less than 1, ranging from 0.64 to 0.83, for web servers the reported typical value of α is varying between 1.4-1.6. Paper [9], which analyzes the media proxy workload, reports a Zipf-like distribution for the file access frequencies in their study with $\alpha = 0.47$. Paper [3] approximated educational media server daily workloads using concatenation of two Zipf-like distributions.

Since our workloads under study cover a long period of time, we decided to investigate whether the file access frequencies exhibit the same behaviour on a dif-

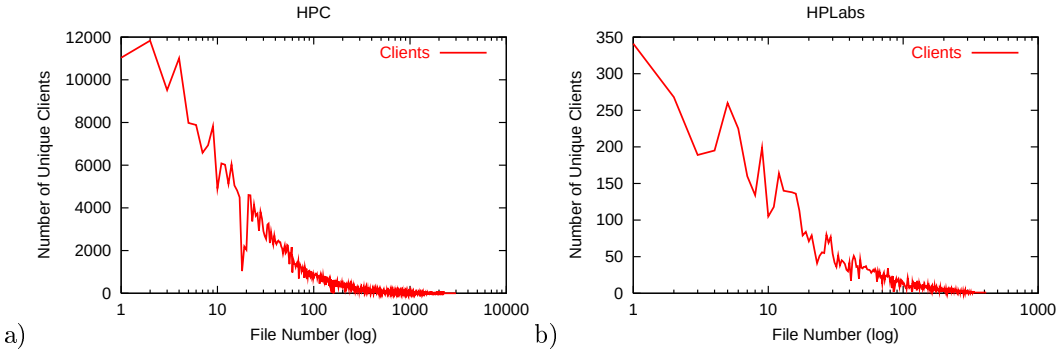


Figure 12: Files sharing statistics: a) HPC and b) HPLabs.

ferent time scale. We considered 1-month, 6-month, 1-year and the entire duration of the logs as a time scale for our experiments.

In order to characterize the distribution of the file access frequencies for workloads under study, we ranked the files by popularity (i.e. the number of accesses to each file), and plotted the results on the log-log scale. Figure 10a) shows the file popularity over entire duration the logs. Both workloads exhibit very similar distribution: the HPLabs curve "follows" the HPC curve, but on a lower scale. This can be explained by almost two orders smaller number of accesses and files in the HPLabs workload. However, both of these curves are far from fitting a straight line of Zipf-like distribution. Figure 10 b) shows files popularity for HPC and HPLabs workloads for the yearly period (of 2000 year), as well as 6-months intervals (the corresponding first half-year and second half-year periods of 2000 year). HPC curves (both 1-year and 6-month) are still far from fitting a straight line of Zipf-like distribution.

However, 6-month curves for HPLabs fit reasonably well with the straight line of Zipf-like distribution when ignoring the first 15-20 files (in [7], authors make similar assumptions about ignoring the top 100 documents and a flat tail at the end of the curve). The straight line on the log-log scale implies that the file access frequency is proportional to $1/i^{\alpha}$. We obtained the values of α using least square fitting: for both 6-month curves $\alpha = 1.6$ works very well. Figure 11 c) shows file popularity distribution for the HPLabs workload corresponding to the 6-month periods of 2000, approximated by Zipf-like function $1/i^{\alpha}$, with $\alpha = 1.6$.

Finally, Figure 11 a) and b) shows files popularity for the HPC workload on a monthly basis. Most of the monthly curves fit straight line reasonably well when ignoring the first 10-15 files and few last files. For different months, value of α is ranging from 1.4 to 1.6. Figure 11 c) shows file popularity distribution for HPC workload during August of 2000, approximated by Zipf-like function $1/i^{\alpha}$, with $\alpha = 1.5$.

The observation that the file access frequencies for the media workloads under study can be approximated by Zipf-like distribution is very useful for synthetic workload generation. It is interesting that the time scale plays important role for this approximation.

The high locality of accesses to specific subset of files and the high concentration of the clients accessing these popular files shown in Figure 8 a) and b) imply that the popular files are widely accessed by many different clients. In HPC workload, first 70 files are accessed by more than 1000 unique clients, with some frequent files accessed by $10,000-12,000$ unique clients. (Note, that for better viewability we used a log scale for file number/rank).

For HPLabs server, degree of sharing is lower (it is expected, because of the smaller clients population), but for the most frequent files it is still very significant: the first 17 files are accessed by $113-341$ unique clients.

The sharing exhibited by the clients' access patterns is essential for designing an efficient caching infrastructure.

Complementary to the characterization of the most frequently accessed files, it is useful to have statistics about the "opposites": the percentage of the files that were requested only a few times, and the percentage of active storage these files account for:

	Files Requested up to 1/5/10 times	Storage Requirements for Corresponding Files
HPC	16% / 38% / 47%	10% / 26% / 34%
HPLabs	19% / 45% / 59 %	17% / 39% / 52%

Table 5: Rarely accessed files statistics.

As the Table 5 shows, 16% to 19% of the files are accessed only once, and 47% to 59% of the files are ac-

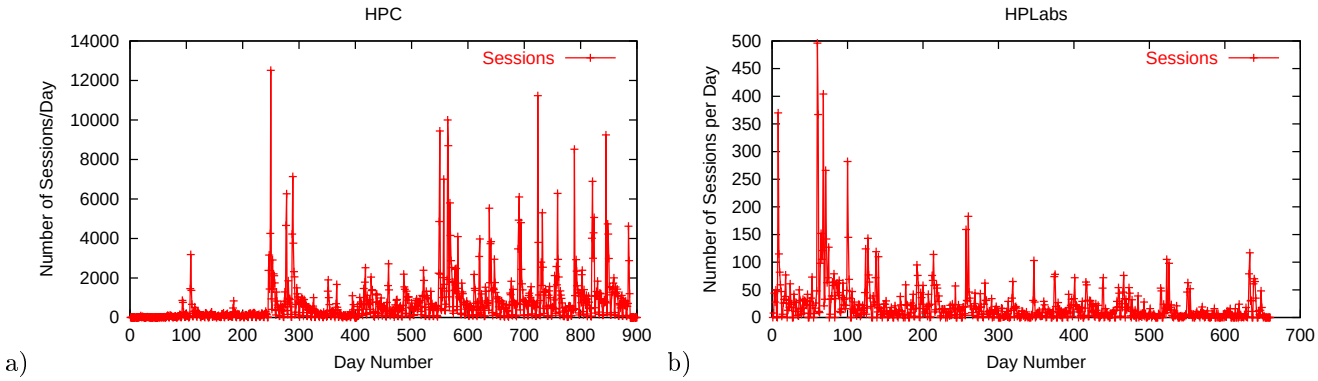


Figure 13: Number of sessions per day: a) HPC and b) HPLabs.

cessed less than 10 times. These rarely accessed files account for quite significant amount of storage: 34% to 52% of total active storage set. These numbers are somewhat lower compared to the web server workloads. For web server workloads, "onetimers" (files accessed only once) may account for 20%-40% of the files and the active storage.

The locality properties of the client references as well as the complementary knowledge about the rarely accessed files are very important in designing the media proxy caching strategies and efficient content placement on distributed media servers and media proxies. As part of our future work in this direction, we intend to explore the temporal locality of client accesses as well as the degree of file sharing among the clients. We expect that this information will serve as a basis for using media delivery optimizations such as multicast.

5 Dynamics and Evolution of Media Sites

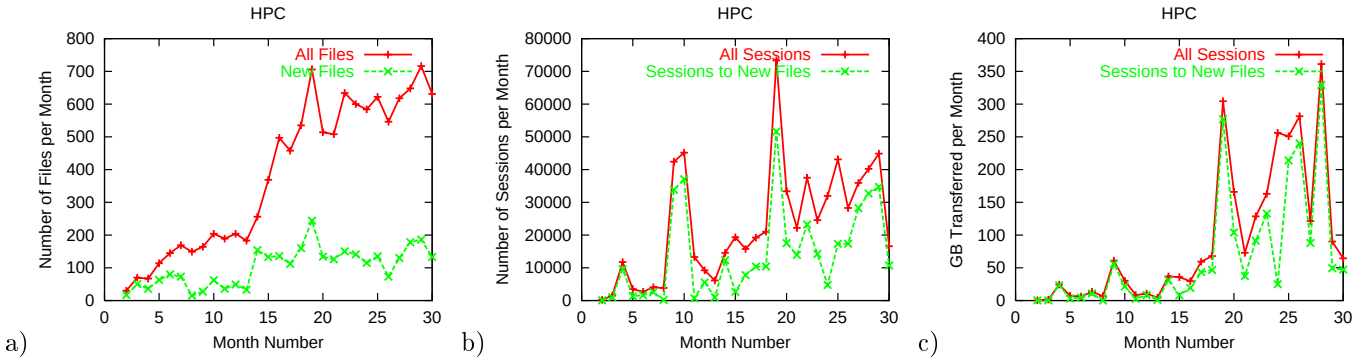
In this section, we investigate specific file access patterns discovered through the analysis of the two workloads under study.

First of all, we observed that the traffic to both the HPC and the HPLabs sites is very bursty. Figure 13 shows total number of media sessions per day for the entire duration of the logs. Some days exhibit two orders of magnitude higher number of sessions for both workloads. We will relate this burstiness to sessions accessing new files later in the section.

For enterprise web servers studied before [8], daily traffic amount was much more stable and predictable. Other studies of different media workloads [9, 3] contained similar observations about media traffic burstiness, but the degree of burstiness observed was smaller, and more correlated with the day of the week, especially for educational media workloads. Both of the mentioned studies analyzed workloads over significantly shorter interval of time: workload in [9] was a week long, and two workloads in [3] were 1 to 3 months long.

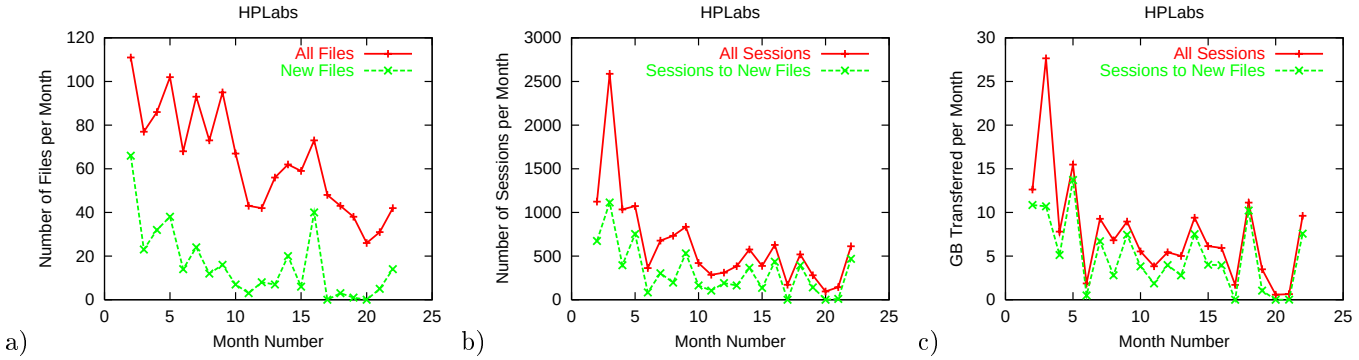
Since our logs provide information about the client accesses over a long period of time, one of the main goals of this study is to characterize the dynamics and evolution of media sites over time. The first natural step is to observe the introduction of new files in the logs, and to analyze the portion of all requests destinate for those files. We define a metric called **new files impact**, to characterize the site evolution due to new content, by computing the ratio of the accesses targeting these new files over time. Figures 14 a) and 15 a) show two curves for HPC and HPLabs workload respectively. The curves show all the files which were accessed in a particular month, and all the new files which were accessed in the same month. We define a file being *new* if it was not ever accessed before, based on the information in the access logs. The HPC site has an explicit growth trend with respect of total number of files accessed per month, and consistently steady amount of new files added to the site during each month.

The growth of total number of files accessed each month for HPLabs site is "negative". Since this was unexpected, we asked the team supporting this site whether there were specific reasons for the trend we observed. Specifically, we wanted to know if there is a significant number of new video files that "nobody watches" and hence the logs don't contain any information about them or if the actual new media content on that site decreased over time. The team explained that lately they had been adding only a limited number of new files because they are working on a transition plan to upgrade the entire site design and equipment. So, the "negative" trend in the addition of new files to the site was observed correctly.



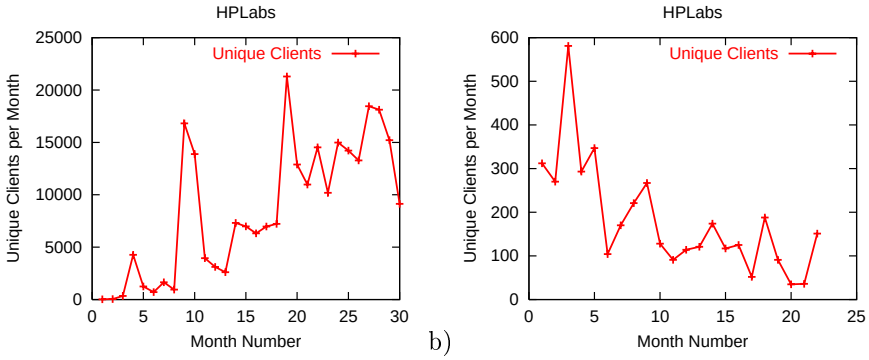
a) b) c)

Figure 14: HPC workload: a) all and new files per month, b) all sessions and sessions to new files per month, c) all bytes transferred and bytes transferred due to new file accesses per month



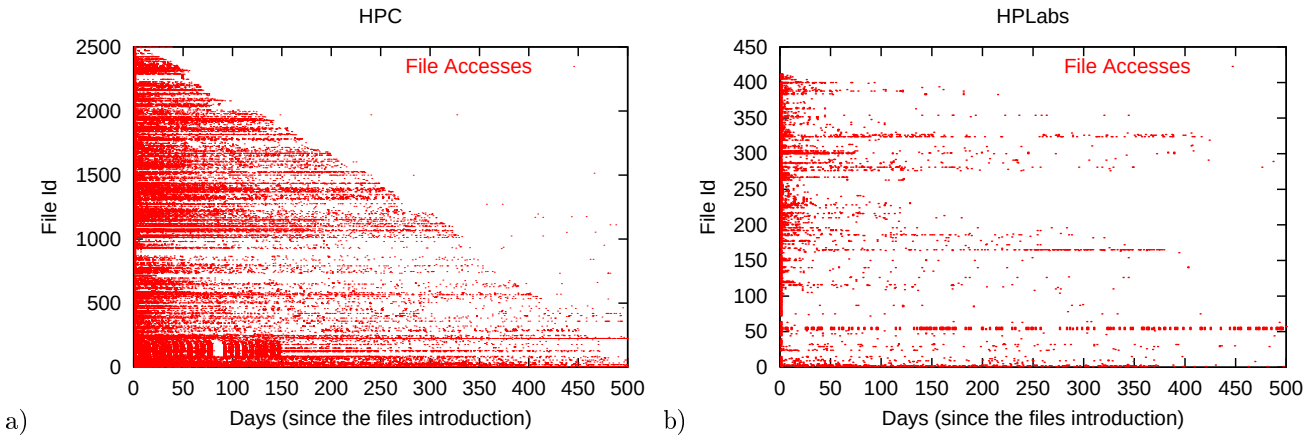
a) b) c)

Figure 15: HPLabs workload: a) all and new files per month, b) all sessions and sessions to new files per month, c) all bytes transferred and bytes transferred due to new file accesses per month



a) b)

Figure 16: Unique clients per month a) HPC and b) HPLabs.



a) b)

Figure 17: Files accessed on different days after their introduction.

Figures 14 b) and 15 b) show graphs for HPC and HPLabs workload respectively: the number of all sessions per month and the number of sessions to the new files in this month. These graphs reflect that the accesses to the new files constitute the most or a very significant portion of all accesses, excluding a few months that were exceptions. Figure 14 c) and Figure 15 c) show very similar trends for the bytes transferred per month and the bytes transferred due to the accesses to new files. Since the number of new files added per month plays a crucial role in defining the site dynamics, evolution, and growth trends, evaluating the **new files impact** metric becomes important.

Figures 14 c) and 15 c) show the number of unique clients per month accessing each of the HPC and HPLabs site correspondingly. Again, the graphs are correlated with the trends of the sessions to each site's new files. Thus, the client population of enterprise media site strongly depends on the amount of new information regularly added to the site.

Dynamics of the enterprise *web sites* exhibits much more stability in terms of the accesses to the "old" documents. Only about 2% of the monthly requests are to the new files added that month as shown in [8]. Differently, the access pattern of enterprise *media sites* resembles with the access pattern of the *news web sites* where most of the client accesses target newly added information.

6 Life Span of File Accesses

In this section, we attempt to answer the following question: how much does the popularity of the file and frequency of file accesses changes over time? The answer to this question is critical for designing prefetching or server-push algorithms, as well as for design of efficient content distribution strategies in CDN network for media content.

First, we plotted the histograms of accesses for the most frequent files. Most of histograms had lognormal-like curve with majority of accesses occurring during the first 1-4 weeks and with a sharp decline in the number of sessions after that. File access pattern with stable, non-declining amount of sessions over long period of time was much less typical.

Figure 17 shows the distribution of accesses to files since they were introduced at the site. On the x-axis are shown days after the files introduction; y-axis represent the file ids, e.g. the dot (100,20) means the the file with id=100 was accessed on 20th day after its first introduction. The graphs in Figure 17 a) and b) reflect two interesting observation: in spite of existence of the "long-lived" files (i.e. the files which are still accessed 1.5 years later after their introduction, the most of the accesses happen in the first two months.

Enterprise media server workloads exhibit high locality of references. As has been shown in Section 4, it was observed that 90% of the media server sessions target only 14%-30% of the files. Thus, this small set of files has the strong impact on the media site performance and its access patterns. We define the *core-90%* as the set of most frequently accessed files that makes up for 90% of all the media sessions. From the performance point of view it is these *core* files one should concentrate on to obtain good performance as most of the accesses are to them. Along with understanding the dynamics of all the files at the site, we would like to see whether the core files exhibit some specific properties.

We define a *life duration* for a particular file to be the time between the first and the last accesses to this file in the given workload. Figure 18 shows the distribution of file *life duration* for both workloads.

There are two curves on the graphs representing a

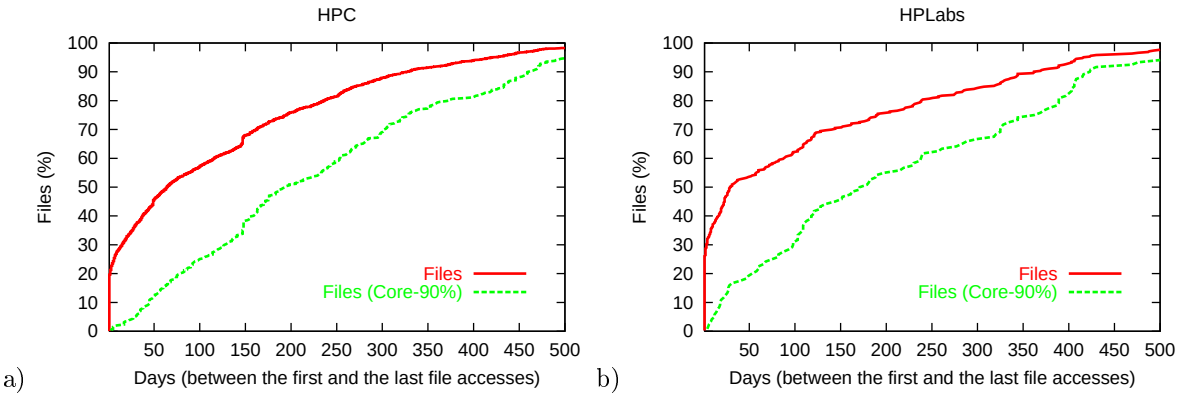


Figure 18: Days between the first and last file accesses.

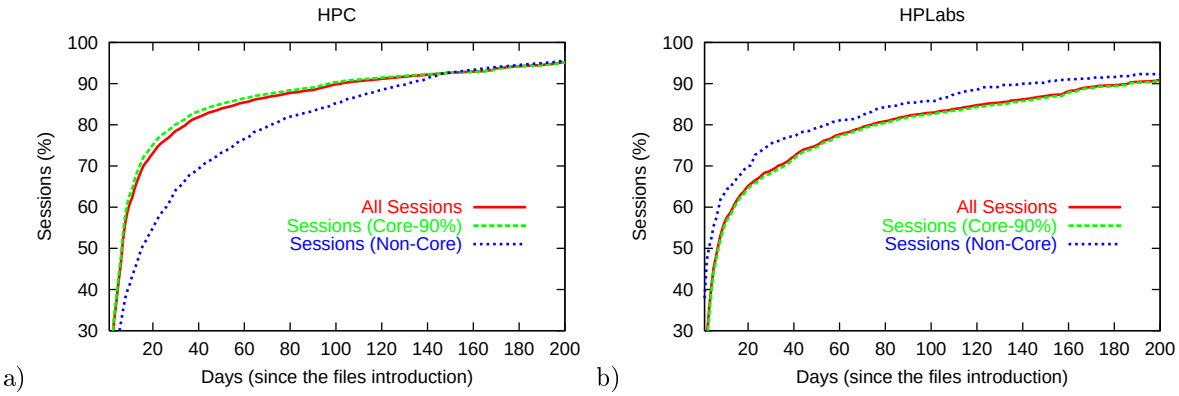


Figure 19: Percent of sessions on days between the first and last file accesses.

life duration distribution for all the files and for the core files. Our analysis shows that high percentage of all files have a short life duration: files that "live" less than a month constitute 37% of all the files in the HPC workload and 50% of all the files in the HPLabs workload (this number is partially so high, because 16-19% of all the files are accessed only once as reported in Section locality). 73% of all the files for both workloads have a life duration less than 6 months. Only 10% of files for the HPL site and 8% for the HPC site live longer than a year. As for the frequently accessed files, much higher percentage of them live longer compared to the life duration of all the files. And additionally, the "short-lived" frequent files in the graphs are mostly represented by the recently introduced files.

For the files of different life duration, we introduce a new metric, called a **life span** metric, which is defined as the cumulative distribution of accesses to the files since their introduction at a site.

Figure 19 shows the life span of the file's accesses for both workloads. The x-axis reflects the days since the files introduction; the y-axis represents the cumulative percentage of all the file accesses up to this day (relative to the total number of all the sessions over the entire duration of the logs).

For HPC (HPLabs) workload, 52% (51%) of all the sessions occur during the first week of files existence, 68% (61%) of all the sessions occur during two weeks of the files existence, 74% (66%) - during three weeks of files existence, 77% (69%) - during four weeks of files existence, 80% (70%) - during five weeks of the files existence. Thus, HPLabs site has longer life span for their files than HPC site.

Above statistics can be interpreted in a different way, reflecting the rate of changes of the accesses in a given workload: 52% (51%) of all the sessions occur during the first week of file existence, followed by only 16% (10%) of accesses during the second week, decreasing 6% (5%) of accesses during the third week, and only 3% (1%) - of accesses for 4th and 5th weeks since the file introduction.

Life span of core-90% files is almost identical with life span of all the files. It is not surprising, because by definition the core-90% files represent 90% of all the accesses to the site. Their properties have major impact on characteristics of life span for the whole site.

As for the rest of the files (non-core files), their properties are different for the HPC and the HPLabs workloads. For example, for the HPC workload, 70% of the sessions to non-core files occur during first 42 days after the files introduction, while for the HPL workload, 70% of corresponding sessions occur during the first 21 days after the introduction of the files.

The life span metric is a normalized metric. The files could have been individually introduced at different times. The metric reflects the rate of change of the file access pattern during the files' existence at the site. Moreover, the life span metric reflects the timeliness of the introduced files. Longer life span means that media information on a site is less timely and has more consistent percentile of accesses over a longer period of time. Life span metric allows one to interpolate the intensity of the client accesses to the new and the existing files over a future period of time.

We believe that locality properties, access patterns of newly introduced files, and their life span are critical metrics in defining the efficient caching infrastructure and future content delivery systems.

7 Conclusion and Future Work

Media server access logs are invaluable source of information not only to extract business related information, but also for understanding traffic access patterns and system resource requirements of different media sites. Our tool **MediaMetrics** is specially designed for system administrators and service providers to understand the nature of traffic to their media sites. Issues of workload analysis are crucial to properly designing the site, and its support infrastructure, especially for large, busy media sites.

Our analysis aimed to establish a set of properties specific for the enterprise media server workloads and compare them with the well known related observations about the web server workloads. In particular, we observed high locality of references in media file accesses for both workloads. Similar to previous web workloads studies, our analysis of the media file popularity distribution revealed that it can be approximated by Zipf-like distribution with α parameter in a range 1.4-1.6. The interesting new observation is that the time scale plays an important role in this approximation. We considered 1-month, 6-month, 1-year and the entire duration of the logs as a time scale for our experiments. For the HPLabs workload, the distribution of the clients accesses to the media files on a 6-month scale starts to fit Zipf-like distribution. While for the HPC workload, the file popularity on a monthly basis can be approximated by Zipf-like distribution. For longer time scale in the same workloads – the file access frequency distribution does not follow a Zipfian distribution.

We introduced the *new files impact* metric for enterprise media workloads, which reflects that accesses to the new files constitute most of the monthly accesses, and the bytes transferred due to the accesses to the new files account for most of the transferred bytes. Also, we observed that the growth trend in the site accesses directly depend on the amount of the newly added files.

We defined a *life span metric* to reflect the rate of change in the accesses to the newly introduced files. For the studied workloads, 51%-52% of the accesses to the media files occur during the first week of their introduction. This stresses a high temporal locality of the accesses in media server workloads which is consistent with observations in other media workload studies.

Additionally, we also discovered some interesting facts about the clients' viewing behavior. Despite the fact that the two studied workloads had significantly different file size distribution, the clients' viewing behavior was very similar for the both sets: 77%-79% of media sessions were less than 10 min long, 7%-12% of the sessions between 10-30 min, and only 6%-13% of sessions continued for more than 30 min. This reflects the browsing nature of most of the enterprise client accesses. We also found that the percentage of sessions with interactive requests are much higher for medium and long videos.

In our future work, we are planning to exploit the locality properties of the client references and the specifics of client viewing behavior for designing efficient media proxy caching strategies, appropriate content placement at distributed media servers and media proxies, as well as in using multicast for better bandwidth utilization.

References

[1] S. Acharya, B. Smith. An experiment to characterize videos stored on the web. In Proc. of ACM/SPIE Multimedia Computing and Networking 1998, January 1998.

[2] S. Acharya, B. Smith, P.Parnes. Characterizing User Access to Videos on the World Wide Web. In Proc. of ACM/SPIE Multimedia Computing and Networking. San Jose, CA, January 2000.

[3] Almeida, J. M., J. Krueger, D. L. Eager, and M. K. Vernon. Analysis of Educational Media Server Workloads, Proc. 11th Int'l. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV 2001), June 2001.

[4] V. Almeida, A. Bestavros, M. Crovella, A. Oliviera. Characterizing reference locality in the WWW. In *Proc. 4th Int. Conf. Parallel and Distributed Information Systems (PFIS)*, pp.92–106. IEEE Comp. Soc. Press,1996.

[5] M. Arlitt and C. Williamson. Web server workload characterization: the search for invariants. In Proceedings of the ACM SIGMETRICS '96 Conference, Philadelphia, PA, May 1996.

[6] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in web client access patterns: characteristics and caching implications. Technical Report, Boston University, TR-1998-023, 1998.

[7] L.Breslau, P.Cao, L.Fan, G.Phillips, S.Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. In Proceedings of IEEE INFOCOM 1999, March 1999.

[8] L. Cherkasova, M. Karlsson: Dynamics and Evolution of Web Sites: Analysis, Metrics and Design Issues. In Proceedings of the Sixth International Symposium on Computers and Communications (ISCC'01), Hammamet, Tunisia, July 3-5, 2001.

[9] M.Chesire, A.Wolman, G.M.Voelker, H.M.Levy. Measurement and Analysis of a Streaming Media Workload. Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems, San Francisco, CA, March 26-28, 2001.

[10] F. Douglis, A. Feldmann, B. Krishnamurthy, and J.Mogul, "Rate of change and other metrics: A live study of the World Wide Web". In Proceedings of the USENIX Symposium on Internet Technologies and Systems, December 1997.

[11] Content Networking. Inktomi Corp. http://www.inktomi.com.

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

[13] N.Harel, V. Vellanki, A. Chervenak, G. Abowd, U. Ramachandran. Workload of a Media-Enchanced Classroom Server.
In Proceedings of IEEE on Workload Characterization, October, 1999.

[14] D. Loguinov, H. Radha. Measurement Study of Low-bitrate Internet Video Streaming. In Proceedings of the ACM SIGCOMM Internet Measurement Workshop San Francisco, California, USA, November 2001.

[15] A. Mena, J. Heidemann. An Empirical Study of Real Audio Traffic. In Proceedings of the IEEE Infocom, p. 101-110. Tel-Aviv, Israel, March, 2000.

[16] J.Padhye, J.Kurose. An Empirical Study of Client Interactions with Continuous-Media Couseware Server. Proc. 8th Int'l. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV 1998), July 1998.

[17] RealServer administration Guide – RealSyztem G2. RealNetworks, Inc., Nov.1998.
http://docs.real.com/docs/serveradminguideg2.pdf

[18] Wang, M. Claypool, Z.Zuo. An Empirical Study of RealVideo Performance Across the Internet. In Proceedings of the ACM SIGCOMM Internet Measurement Workshop San Francisco, California, USA, November 2001.

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

A APPENDIX

A.1 Windows Media Server Log Format

An entry in the Windows Media Server logs looks like:

```
15.128.147.57  1999-12-30  08:00:01
jujug57.grenoble.hp.com
mms://desktvplus.cup.hp.com/carly/dec23_99.asf
0  231 1 200 {ef93d2da-77d2-11d3-99be-0060b0c2ea9a}
6.4.5.809  fr-FR  -  -  mplayer2.exe  6.4.5.809
Windows_NT4.0.0.1381  Pentium  231  489154  16821
mms  UDP  Voxware_MetaSound  -  -  495492  495492
903  903  0  0  0  0  0  0  2  10  100
15.14.200.234  desktvplus.cup.hp.com  3  0
```

In our analysis, we used only a subset of entries, which we describe below.

Entry	Value
1st	15.128.147.57
2nd	1999-12-30
3rd	08:00:01
5th	mms://desktvplus.cup.hp.com/carly/dec23_99.asf
6th	0
7th	231
8th	1
20th	489154
21st	16821
27th	495492
28th	495492
42rd	3
43th	0

Names used for describing the entries below are the same as used by Microsoft.

- c-ip: 1st entry in the logs, the client IP address. A client connected via proxy provides a client proxy server IP address.
- date: 2nd entry in the logs, the date (in Greenwich mean time) when an entry is generated in the log file.
- time: 3rd entry in the logs, the time (in Greenwich mean time) when an entry is generated in the log file.
- cs-uri-stem: 5th entry in the logs, the name of the file that is playing, an .asf file for a unicast and an .asx file for a multicast. Notice that this field is actually a URL, so the file name has to be extracted.
- c-starttime: 6h entry in the logs, the timestamp (in seconds) of the stream when an entry is generated in the log file.
- x-duration: 7th entry in the logs, the length of time a client played content prior to a client event (fast-forward, rewind, pause, stop, or jump to marker). A log entry is generated whenever one of these client events occurs.
- c-rate: 8th entry in the logs, the mode of Windows Media Player when the last command event was sent.
    - 1 = Windows Media Player was paused or stopped during a play, fast-forward, rewind, or marker jump operation.
    - -5 = Windows Media Player was rewound from a play, stop, or pause operation.
    - 5 = Windows Media Player was fast-forwarded from a play, stop, or pause operation.
- filelength: 20th entry in the logs, the length of the file (in seconds). This value is 0 for a live stream.
- filesize: 21st entry in the logs, the size of the file (in bytes). This value is 0 for a live stream.
- avgbandwidth: 22nd entry in the logs, the average bandwidth (in bits per second) at which the client was connected to the server.
- sc-bytes: 28th entry in the logs, the bytes sent by the server to the client.
- c-bytes: 29th entry in the logs, the number of bytes received by the client from the server. For a unicast, c-bytes and sc-bytes must be identical. If not, packet loss occurred.
- s-totalclients: 43rd entry in the logs, the number of clients connected to the server (but not necessarily receiving streams).
- s-cpu-util: 44th entry in the logs, the average load on the server processor (0%-100%). If multiple processors exist, this value is the average for all processors.

A.2 Real Media Server Log Format

An entry in the Real Media Server logs looks like:

```
15.0.143.56 - - [06/Jul/1999:15:12:58 -0700]
"GET coffee/HPL-990621.rm PNA/10"
200 893325 [WinNT_4.0_6.0.6.33_play32_LF60_en-US_686]
[8a3ca9b0-0717-11d3-9aa6-00108300ef70]
[Stat1: 228 0 0 0 0 16_Kbps_Music_-_High_Response]
[Stat2: 15936 10192 0 0 0 0 0 0 0 0 9422
16_Kbps_Music_-_High_Response]
74612852  2932  54  0  0  1
```

In our analysis, we used only a subset of entries, which we describe below.

Entry	Value
1st entry	15.0.143.56
3rd entry	[06/Jul/1999:15:12:58 -0700]
4th entry	GET coffee/HPL-990621.rm
12th entry	74612852
13th entry	2932
2nd field in 11th entry	10192
7th entry	893325

Names used for entries mostly the same as used by REAL. Only in few cases special names are used, when some of the values were derived (to make the data analysis uniform across different server types: Windows Media Server and Real Media Server). The entries are described below in the same as for Microsoft Media Server logs, for consistency.

- IP_address: 1st entry in the logs, the IP address of client.
- timestamp: 3th entry in the logs (2nd entry is -, for compatibility with the web server logs), the time that client accessed the file in the format

  `[<dd>/<Mmm>/<yyyy>:<hh>:<mm>:<ss><TZ>]`

  where TZ is the time zone expressed as the number of hours relative to the Coordinated Universal Time (Greenwich, England). Notice that this time stamp contains both date and time fields.
- GET filename: 4th entry in the logs, the name of the file requested by client. Notice that this is the file name, not the URL.
- file_size: Depending on how many options are present in the logs, this would be the 11th entry (if none or one option is present), 12th entry (if two options are present – as is the case for our logs), or 13th entry (if all three options are present in the logs). It provides the file size in bytes (the file requested by the client).
- file_time: The entry next to the file size entry, contains the duration of the file being played by the client.
- Client BPS: 2nd field in the 11th log entry (notice that 11th log entry is optional in the logs, may be absent), the bandwidth available to the client when it was viewing the file. Notice that this figure varies depending on the network conditions.
- bytes_sent: 7th entry in the logs, the number of bytes transferred to client during play. This field may be lower than the total size of the media file, indicating partial playback of the file. If this field is consistently low for some or all media files, it may mean that RealPlayers are able to connect to your server, but are unable to play files.
- Bytes Received: derived value. The 1st field in the 10th log entry (notice that 10th log entry is optional, may be absent) contains the *total packets* sent by the server. The 3rd field in the 10th log entry contains the number of *missing packets*. From this information and some additional information (such as bytes_sent by the server and total packets sent by server), we can derive the bytes received by the client).