Workload Evaluation of an E-learning Server

Diego del Blanco, Carlos E. Palau
Communications Department
Universidad Politécnica de Valencia
ETSI Telecomunicación; Camino de Vera S/N, Valencia 46022
SPAIN

Abstract: - This paper presents an analysis of a three months period of workload data from a large web-based e-learning system. This e-learning server provides on-line materials to the UPV campus at Valencia (Spain) and outside access mainly to Universities in South-America. We characterize user requests and file distribution and determine their impact on system performance. The purpose of our study is to assess scalability and support capacity planning.

Key-Words: workload characterization, content distribution, e-learning, web service

1 Introduction

The phenomenal growth in popularity of the www, the increase in users number and the extension of this tool to different applications, such as e-learning, has made the www traffic the most important one over the Internet. Distance learning over the web has introduced an additional value to this facility, providing a particular kind of access patterns and workload on the servers. This article presents a detailed study of the Technical University of Valencia (Spain) e-learning server. This facility provides content distribution to over 10.000 students of more than 500 courses. The study has been performed during a 105 days period (November 2003 till February 2004).

Workload characterization plays an important role in systems design, mainly allowing us to understand the current state of the system and its possible evolution over time. It is also crucial to the design of new system components. In this paper we will focus on the web server performance, an extended version of the study is available in [1]. We can compare our results with those from previous studies [2-3]. Web server workload is only one of the necessary steps for understanding the changes occurring in Web traffic, although there are different aspects not considered in this paper like web client workloads, network traffic or http analysis that provide better scope of the service. [4]

Much of this recent research activity has been aimed at improving web performance and scalability. The key performance factors to consider are how to reduce the volume of network traffic produced by web clients and servers and how to improve the response time for WWW users. But fundamental to the goal of improving web performance is a good understanding of WWW workloads, while there are several studies; most of them focus on web clients rather than in web servers.[1][5]

A significant difficulty that researchers face in doing a server-based study is the very limited availability of server traces. The few pioneering studies in this area have had to be made from small departmental servers at universities, although there are some studies (none of them related to e-learning) related with large web servers [6-9]

Some of the more significant characteristics we observed in the e-learning server at the Technical University of Valencia are:

- ASP and html files are the basis of the e-learning system.
- The most requested files are image files
- The access patterns follow a Zipf-like distribution
- The user sessions usually are done during labour time.
- There is a small number of files (mainly multimedia) that are responsible of a high bandwidth consumption.

The reminder of the paper is organised as follows. In section 2 we present an overview of the system with its main components and specific parts, analyzing the file distribution and the capture of data. Section 3 presents the data analysis and the paper finishes with the conclusions and future work.

2 Overview of the System

The Web-based e-learning system under study has a multi-tier architecture typical of this specific sites. The tiers of this system are illustrated in Figure 1. All customer requests access the system via the Internet, using the web server. All the tasks to carry out the e-learning activity are installed in a cluster.

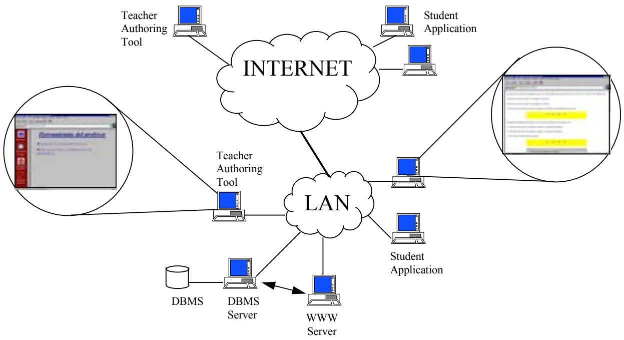

Figure 1 Structure of the system

The application servers prepare responses in HTML format for rendering on users' browsers. These responses are then propagated to the customer via the Web server tier. Networking equipment such as routers are not shown in Figure 1.

The basic structure of the e-learning platform consists in ASP files and supported by a Microsoft SQL Server database. In this platform are installed all different courses, generated in HTML by a tool hosted in another server. The courses generally are made up of text and images, although many of them include multimedia resources like Flash animations, video or sound files or even java applets. The system at the moment does not include any media streaming facility.

The e-learning platform has access restricted by password and, besides to have a complete management system and statistics, it offers several services to the student or tutor:

- User Information and Statistics
- Schedule
- Contents
- Exercises
- Autoevaluations
- Mail
- Forums
- Chat
- News-board

2.1 Server file distribution

The e-learning platform keeps its contents in a very simple directory structure:

- A directory that hosts the ASP programming of the e-learning platform.
- A directory that hosts the Web of the Vicerrectorado de la Universidad Politécnica Abierta.
- A directory that contains the courses in HTML for the e-learning platform, as well as other temporary archives like photos of the users or annexes to the internal mail. Each course is kept in an individual directory keeping all its contents.

The summary of the server structure and contents can be seen in Table 1.

Total Files	85441
Total Directories	5880
Total de Bytes	2652,96 MB
Total de Extensions	85
Average Size of File	31,79 KB
Average Files in directories	14,53
Average Size of directories	462,01 KB

Table1. E-learning contents summary

Figure 2a provides a breakdown of files hosted in the e-learning server, classified using the extension. Almost 60% of the hosted files are HTML files, 24% are both types of more common images (jpg and gif) and the others are different types of downloadable filesprovided by the courses. Only 1% of files are ASP.

Another important information that is needed is server disk occupation. Because there are some types of files that although are not important in percentage, they are in occupied disk space, like compressed, audio or video files. Figure 2b shows distribution in percentage of bytes. As can be seen, the zip-files, that are only 0.35% of total files in server occupy 20% of the space, or Real Media files that being only a 0.14% of the total files occupy 10% of disk space

2.2 Collection of data

The dataset used in this workload characterization study is composed of the access logs collected on a daily basis from the e-learning server of the Technical University of Valencia. For this study we have used access logs of 105 days, from Monday November 17th 2003 to Sunday February 29th 2004. Each access log is in the CL Format [5]. Each entry in the log file provided information regarding the IP issuing the request, the HTTP method (including the resource requested) and the status response, together with the time.

Duration	
Total requests	
Average requests/hour	
Total bytes transferred	
Average bytes transferred/hour	

Table 2. Summary of access log characteristics

A part of the standard logs, we have used some proprietary logs with extra information, extracted from the e-learning application developed at the UPVV. This information has filled some gaps in the information provided by the Common Log Format.

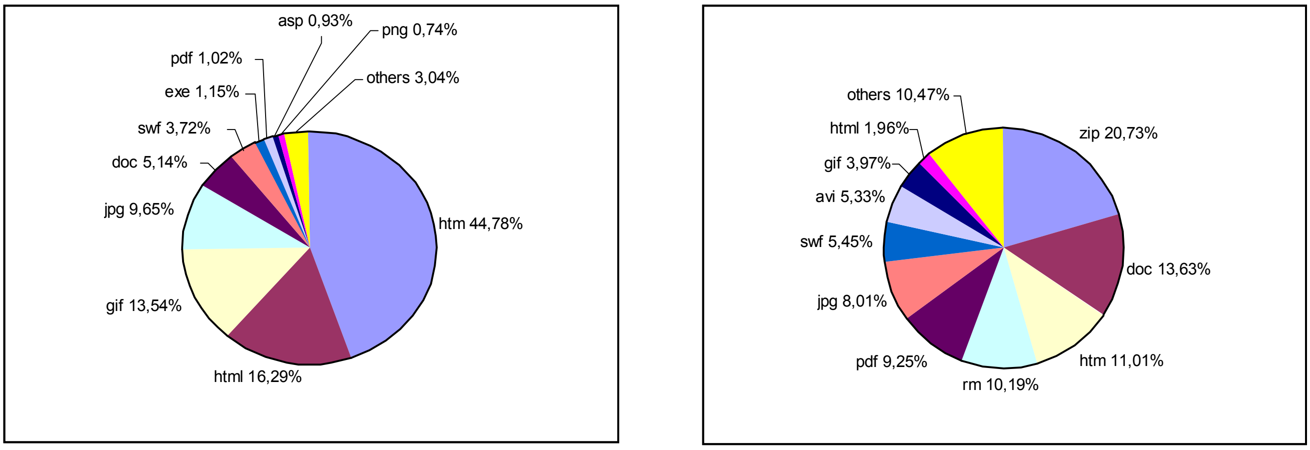

Figure 2. E-learning server file distribution, (a) file type, (b) disk occupation

During this time period some courses where installed in the server, although not all of them were active during the study. 80 out of 200 hundred courses where active at this time. A summary of the data acquired is shown in Table 2.

3 Workload Analysis

This section presents the results obtained from the evaluation and workload characterization (HTTP-level). We have focused on the most meaningful features that serve our aim of identifying methods for improving the scalability and performance of the e-learning system, and also identify patterns in order to model the behaviour of the server and the activity of the clients (teachers and students). The study has been made monitoring the server during 105 days. During this period, some courses have started and others have finished, at the same time between the 70 and 80 simultaneous courses have been active

As we can see in Table 3, the average number of visits is 403 to the day, and each visit generates an average traffic of 750,55 KB; calls an average number of 65 pages and an average number of 264,5 hits per visit, understanding hit as web-object requested.

The total bandwidth used during these 105 days has been of 31082 MB, which makes a daily average of 296,03 MB. This traffic is distributed throughout the diurnal hours with a uniform distribution. We can see that the traffic follows practically the distribution of the labour time schedule. It can be thought that due to the characteristics of e-learning systems, usually used by people who does not have time to study due to their timetable, the logical thing is that the hours of maximum access went from 18 to 24 or even from 6 to 8. Nevertheless we observed that most of the accesses were done during working hours, meaning that people prefer using the resources that they have at work or in the university to download or print the materials and study them at home.

Hits	
Total Hits	11,217,376
Average Hits per Day	106,832
Average Hits per Visit	264.52
Cached Requests	1,951,833
Failed Requests	120,712
Page Views	
Total Page Views	2,757,545
Average Page Views per Day	26,262
Average Page Views per Visit	65.03
Visits	
Total Visits	42,407
Average Visits per Day	403
Total Unique IPs	12,504
Total Visitor Stay Length	17146:42:49
Average Visitor Stay Length	24:15
Bandwidth	
Total Bandwidth	31,082.82 MB
Average Bandwidth per Day	296.03 MB
Average Bandwidth per Hit	2.84 KB
Average Bandwidth per Visit	750.55 KB

Table 3. General information about the workload

The same thing happens with weekends. These days that seem that they would have to be more active for those choosing the e-learning by mutual incompatibility of schedules with their work, are those that less traffic gather. 473 average number of visits per day on weekdays versus 230 average number of visits per day on weekends. It can be observed that according the week advances also decays the traffic, so Mondays are the most active days. Any period of vacation produces a substantial reduction in user accesses and in traffic generated. In our case, Christmas supposed a remarkable workload reduction.

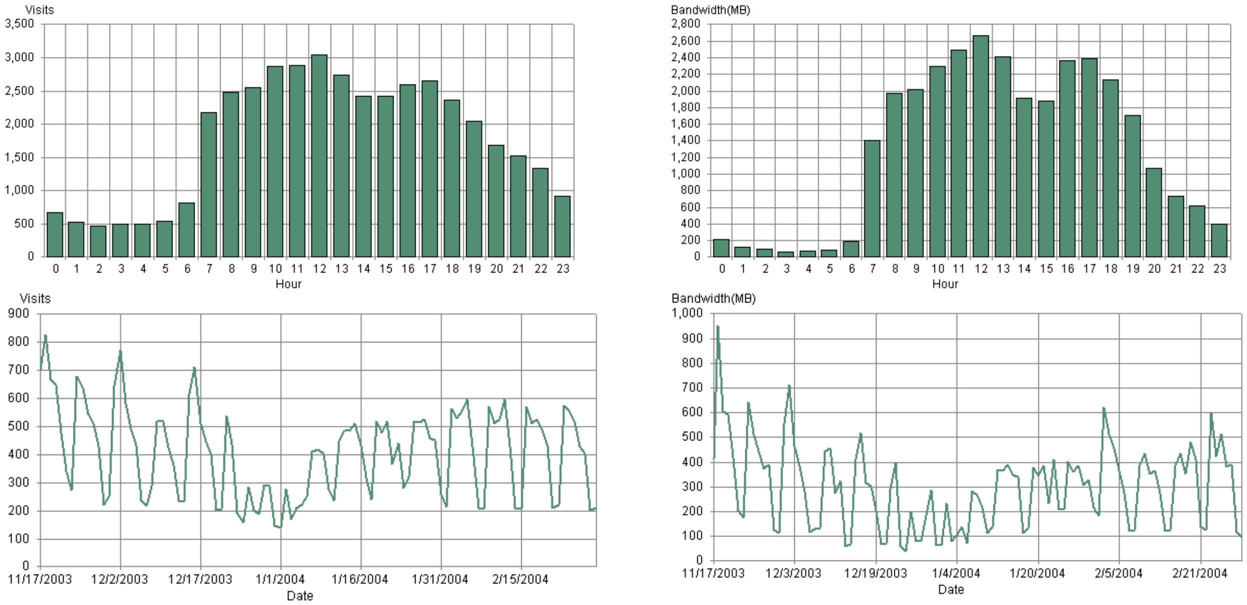

Figure 3. Visits and bandwidth results

The e-learning system at UPV is under continuous development and its usage increases on a daily basis, more students, more teachers and more courses are being deployed, leading to an increase in accesses, web objects and volume of traffic.

Average # of Visits per Day on Weekdays	473
Average # of Hits per Day on Weekdays	134,957
Average # of Visits per Weekend	230
Average # of Hits per Weekend	36,518
Most Active Day of the Week	Tuesday
Least Active Day of the Week	Sunday
Most Active Date	Mon., Feb-02-04
# of Hits on Most Active Date	236,717
# of Visits on Most Active Date	560
BandWidth on Most Active Date	618.60 MB
Least Active Date	Thu., Dec-25-03
# of Hits on Least Active Date	11,088
# of Visits on Least Active Date	155
BandWidth on Least Active Date	36.48 MB
Most Active Hour of the Day	12:00 - 12:59
Least Active Hour of the Day	04:00 - 04:59

Table 4. Summary of activity

Regarding web objects kept at the e-learning server, the most accessed object (identified by the corresponding URL) is the system start page. This main page of the platform includes the html file, icons and logos. It is coherent due to every student must access the system through an authentication process carried out in this page. A part of this element other high accessed pages are those that provide access to all the platform tools (mail, schedule ,board, contents).

The site has a clearly defined entrance page (the home page of the e-learning web server), that accumulates the 77,41% of the entrances, whereas the second page only has a 2% of entrances. Obvious the exits are a little more distributed, although there are many visits that leave from the own page of entrance (34,21%), this does not always means that the visits do not accede to more objects. The explanation is very simple, is that when a visitor logs out from the platform the system sends him to the home page unless he closes the navigator directly. This happens quite often because many of the other habitual pages of exit are ASP pages. This proofs that the users do not log out voluntarily, they just close the browser window, with the security risk that this involves.

Another interesting finding is related to the study of the file types and there hits volume, tables 5 and 6 show the detailed information. The most accessed archive types are all kind images. This is because usually each html file embebdes an average of 3 image files. It can be observed the difference between '% of hits' and '% of bandwidth', the consumed bandwidth is not correlated with the volume of hits. This is due most of the downloaded images are small icons of very few KB, nevertheless, the videos for example, that are only 0.01% of hits, suppose more than 1.5% of the bandwidth. We must consider that the current courses do not include a great number of video files, reason why if audio-visual content increased it would be necessary to consider the weight of these archives.

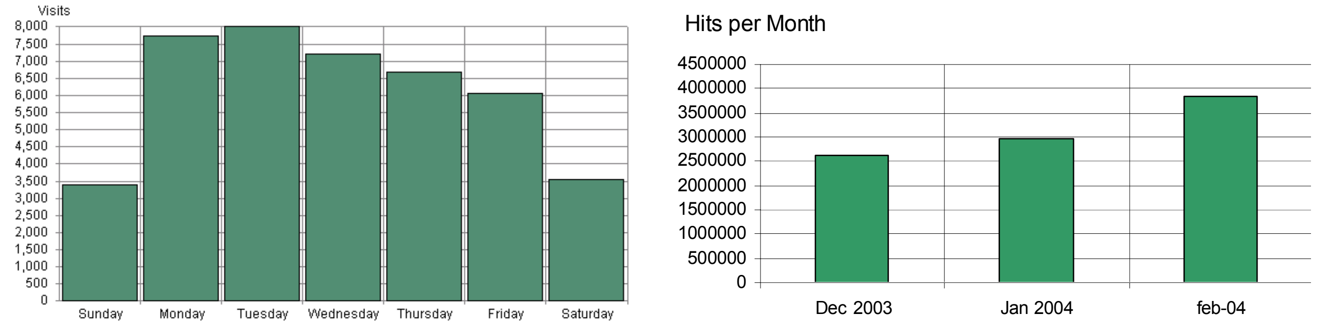

Figure 4. Mean visit number during the week and hits per month

The entire e-learning server contains 85441 web objects and throughout 105 days a total amount of 32323 archives has been accessed, ie.: 37,8% .Also, considering only the 81 web objects with higher % of bandwidth they suppose more then 50% of total used bandwidth. This means that 0.095 % of files of the server consume 50% of bandwidth, providing a typical Zipf-like distribution.

File type	% hits	% bandwidth
Page objects	24,58	21,90
Image files	69,52	46,26
Video files	0,01	1,68
Audio files	0,02	0,01
Download files	0,64	12,40
Other files	7,12	17,75

Table 5. Summary of accesses per file type

File extension	% hits	% bandwidth
GIF	64,04	33,34
ASP	19,18	4,69
JPG	5,47	12,83
HTML/HTM	2,97	7,94
PHP	0,49	2,17
EXE	0,32	0,65
PDF	0,28	11,24
WAV	0,02	0,01
TXT	0,02	0,01
OTHERS	0,04	0,01

Table 6. Summary of accesses per file extension

Not every request performed by the clients results in a successful hit, the E-learning platform is based on HTTP and it returns error message codes like any other web server. 1.07% of total requests result in an error message generation. Most of them are 404 error Object Not Found (81,45%) or 500 Internal Server Error (17,01%), (table 7). Analyzing these errors, we see that most of them are produced by insolvent attacks of hackers and virus. Requests of directions like "/scripts/..%5c../winnt/system32/cmd.exe" or "admin/$" produce most of these errors. In any case, an approximated percentage of 5% of the errors takes place by requests of nonexistent objects, generally annexed with the erroneous URL.

Error types	% total
404 Object not found	81,45
500 Internal server error	17,01
400 Bad request	0,49
403 Forbidden access	0,42
423	0,25
Other errors	0,37

Table 7. Error type distribution

The user utilization of the system is another important feature that has to be studied in order to evaluate the workload of the e-learning web server. The users access to the system and are allowed to access the content of the courses they have signed in, until they log-out. The average duration of the visits is 24 minutes and 15 seconds, taking as limit from session closing, 30 minutes. Nevertheless, this average does not agree with most of the visits (50,43%), that are visits of less than 2 minutes. Figure 5 shows the summary of client session duration. Generally, clients that access the system to print a unit; to see if they have messages; verify if there is something new in the course or for making an examination. However, those short visits are compensated with those very long visits from people who really study in front of the computer (almost a 3% of the visits are longer than 3 hours).

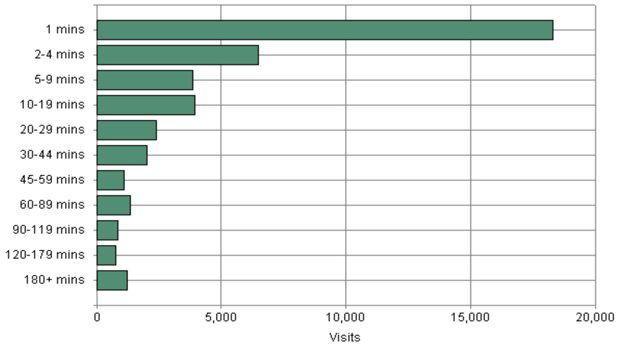

Figure 5. Summary of session duration

4 Conclusion

We have seen that traffic that takes place in a e-learning system, has a distribution that follows the schedule labour unlike which it would be possible to hope.

Due to the functional but relatively austere design of the e-learning platform of the UPA, the bandwidth that users of this system generate by session is not excessive (approximately 700KB) as long as he does not accede to audio-visual elements. Nevertheless, if we want to add to many multimedia files to the server will be necessary to anticipate the important bandwidth increase required.

There are more than 80000 files in the server. Throughout 105 days it was acceded to one third part of them, but nevertheless, 50% of the bandwidth so concentrate in 81 of them and 50% of hits in only 48 archives. We can conclude of it and of which most of these archives are images that do not change habitually, that is very probable that leaving those archives being stored in client's cache is freed quite bandwidth.

Finally, traffic that generates attacks to the system is of the order of 1% of hits. This is worrisome, it does not saturate the server but it supposes an enormous amount of attacks superior to 1000 daily.

References:

[1] V. Almeida, A. Bestavros, M. Crovella and A. de Oliveira, ‘‘Characterizing Reference Locality in the WWW’’, Proceedings of 1996 International Conference on Parallel and Distributed Information Systems (PDIS ‘96), pp. 92-103, December 1996.

[2] M. Arlitt and C. Williamson, ‘‘Internet Web Servers: Workload Characterization and Performance Implications’’, IEEE/ACM Transactions on Networking, Vol. 5, No. 5, pp. 631-645, October 1997.

[3] P. Barford and M. Crovella, ‘‘Generating Representative Web Workloads for Network and Server Performance Evaluation’’, Proceedings of ACM SIGMETRICS ’98, Madison, WI, pp. 151-160, June 1998.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, “Web Caching and Zipf-Like Distributions: Evidence and Implications”, Proceedings of IEEE Infocom ‘99, New York, NY, March 1999.

[5] R. Fielding, J. Gettys, J. Mogul, H. Frystyk-Nielsen, L. Masinter, P. Leach, and T. Berners-Lee, ‘‘RFC 2616 - Hypertext Transfer Protocol - - HTTP/1.1’’, June 1999.

[6] C. Roadknight, I. Marshall and D. Vearer, “File Popularity Characterisation”, Proceedings of the 2nd Workshop on Internet Server Performance (WISP ‘99), Atlanta, GA, May 1999.

[7] M. Arlitt and T. Jin, ‘‘Workload Characterization of the 1998 World Cup Web Site’’, IEEE Network, Vol. 14, No. 3, pp. 30-37, May/June 2000.

[8] M. Arlitt and C. Williamson, ‘‘Internet Web Servers: Workload Characterization and Performance Implications’’, IEEE/ACM Transactions on Networking, Vol. 5, No. 5, pp. 631-645, October 1997.

[9] V. Padmanabhan and L. Qiu, “The Content and Access Dynamics of a Busy Web Server”, Proceedings of ACM SIGCOMM 2000, Stockholm (Sweden), August 2000.