SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)
ISSN (Print) 1813-1743, ISSN (online) 2791-0547
Vol. 55:02 (2023).
Doi: https://doi.org/10.26692/surj-ss.v55i02.6491

Analysis of Academic Web Server Traffic and Workload Characterization for Performance Evaluation

SOOMAL JHATIAL[1], AFTAB AHMED CHANDIO[2]*

[1]National University of Modern languages, Hyderabad Campus, Sindh, Pakistan
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Cite this:
Jhatial S, & Chandio AA, (2023) Analysis of Academic Web Server Traffic and Workload Characterization for Performance Evaluation, Sindh Uni. Res.J. (SS), 55(02),

Corresponding author
chandio.aftab@usindh.edu.pk

ABSTRACT

The web 3.0 considers the latest version of world wide web (www) is increasing rapidly nowadays, as the interest of users in online banking, commerce, business, and online learning about the web. Mostly, the requests made by the users have given a high response time due to the large traffic on the web-server. The services provided by the web-server must be available and maintainable, so there is an extremely need to analysis of web-server's performance. The proposed study addressed the problem to analysis of real-world web-server workload characterization. Initially, the dataset was prepared for the analysis technique with the eliminations of anomalies, noises and null values. Next, the workload parameters have identified and defined for exploration. After that the techniques are used to group similar data according to their characteristics and further data analysis has performed to extract meaningful insights. Based on the workload characterization insights, performance evaluation has measured using regression analysis and measure of dispersion.

Keywords: Workload characterization, world wide web, Access log, Web server, Performance evaluation

INTRODUCTION

Workloads can be defined as the set of inputs that a system receives from its environment, during any given period of time (Ferrari D., et al., 1983). Workload analysis has been an active area of research, there has been a numerous study on workload characterization and improving the performance of web server Ferrari D., 1972). It is mandatory to familiar with the workload to determine the performance of any type of system. The efficiency of any given system is achieved via performance evaluation that tells how the system performs well (Williams et al., 2005). Performance evaluation is the process in which resource of system and outputs are determined whether the system is performing at optimal level. The requirement of performance evaluation is the use of workload.

While evaluating a system's performance, a number of matrices are used to determine the results Bhutto, A. et al. (2023) and Chandio, A.A. et al. (2014). Some examples of matrices are response time, throughput and, bandwidth.

The World Wide Web is basically a web; it is a collection of interlinked resources accessed by the internet. The basic part of the web is a web page which is written in HTML (hypertext markup language). Furthermore, the web pages consist of text documents, graphics, picture, sound, video and other software components. The web has a client-server model, the communication between client and web server is in the form of request-response pairs which is initiated by the client. A client requests the document via a web browser, and a web browser sends request to the web server which sends response according to the client's request. The client and web server use Hypertext Transfer Protocol (HTTP) for communication via Transmission Control Protocol (TCP) connection. After sending the response to request, the web server terminates the TCP connection and repeats the same cycle for the next request.

Different higher education institutes of Pakistan have introduced online system to manage student information such as student enrolment, exam registration, grade reporting, online attendance and much more. The traditional educational system converted into online after the outbreak of COVID-19 pandemic. For this purpose, most of the universities also introduced a Learning Management System (LMS) for online education. As a result, the web traffic has been increasing on web servers of universities, so there is need to study the workload characterization in order to analyze the performance of academic web servers.

The motive of this study is to inspect the workload characterization and to analyze the performance of an academic web server. The main objectives of this study are to present workload characterization for a web server, find out the reasons and insights when the web server provides large response time; identify time intervals when workload fluctuates significantly, and lastly provide recommendations for minimizing response time as the web server performance may be improved.

LITERATURE REVIEW

A very good number of the research study is ongoing on the topic of web service workload analysis and characterization for performance optimization Curino, C., et al. (2011), Choudhury, N. (2014), Bhutto, A. et al. (2023), Chandio, A.A. et al. (2014), Chandio, A.A. et al. (2013-March), Chandio, A.A. et al. (2013-April). The study Hossain et al. (2021) attempted to automatic web-based framework for performance analysis of ten e-commerce site. They use webpage test, page speed insights and gtmetrix tool to scan the website URL and record nine parameters including load time, first byte, start render, first content full paint, speed index, largest content full paint, cumulative layout shift, total blocking time and time to interactive parameters. They found that the site7 has lowest score for 'total blocking time' while site10 has 17.78 seconds score for load time which shows that this site takes more loading time as compare to others.

The study Tochukwu et al. (2020) evaluates the performance of the academic web server by the use of bandwidth and response time. The authors analyzed the access log file to design software for checking response time and it can be deployed on any server to gain useful information about a visitor's behavior, and down time of server for making better decisions. Finally, they concluded that decrease in the bandwidth and response time of server depends upon the bandwidth, type of the file and its size and further suggested that to minimizing the bandwidth, the size of images and PDF files may be required.

Another study Saverimoutou et al. (2019) analyzed the influence of new protocol HTTP/2(Hypertext Transfer Protocol version 2) and QUIC (Quick UDP Internet Connections) on web browsing and studied that how to reduce loading time of web pages using these new internet protocols. For the analysis of the influence of new internet protocol and CDN they collected access log files from Top 10,000 Alexa websites 12-month measurement form (from April 2018 to April 2019). In the same study the authors highlighted that the new protocols deployment is slow by public, but in one year its rate increased by 4% and for web browsing the page loading time is reduced by using new protocol for HTTP/2 43.1% and QUIC 38.5%.

Further research such as Song et al. (2019) compared the workload two different user's mobile and fixed device users, and also presented the statistical model which improves the mobile workload. They used the web log file of an academic web server and find the workload difference between mobile and fixed device. Further they identified the mobile requests had higher success rates as compared to fixed device. The mobile users made less web file requests and more images.

On the other side, the study Xu et al. (2018) suggested a proficient method for web bot traffic detection in a marketplace of large e-commerce and presented detailed analysis on the characteristics of web bot traffic. They applied the method on normal user traffic and web bot traffic. In last, they revealed that the web bot has unique behavioral pattern, generates 10 times more hits than normal user but in a month active for two days.

Authors in Samad et al. (2018) measured the efficiency of web applications which are embedded with caching servers to find an impact on servers with regard to speed and performance of web applications. For the study, the authors find that to speeding up response time of web pages is to reduce the number of embedded objects.

Manoj Kumar (2017) analyzed the web log file from academic web server using a web expert tool to find the user behavior for the purpose of website maintains. They find the different activities of users such as the most used browser is Google Chrome; the download file is PDF and so on.

Summers et. al. (2016) presented the study of workload characterization on online video streaming web server. They developed a prefetch algorithm for analysis of utilization of hard drive and consumption of system memory and characterize a user session using three phases transient, stable and inactive. They observed stable playback sessions with 5% total time spending in transient, 79% in stable, and 16% in inactive phases respectively.

A study Megha P. Jarkad et. al. (2015) proposed a system which predicts future request of users in less time based on users' navigation pattern using Web Usage Mining (WUM) techniques and backtracking algorithm on web log data to reduce the time complexity and improve performance. The steps of system are involved preprocessing of web log file size reduction, classification of users into potential and non-potential, clustering based on similar user behavior using graph partitioning algorithm, and finally using a backtracking algorithm to predict a user's future request.

Furthermore, the study Ahmed Ali-Eldin et. al., (2014) provided the forecasting of the workload evolution over six years of Wikipedia using time series and polynomial splines to study seasonality, trend and page popularity patterns. They indicate the strong seasonality of predictable workload with approximately 2% Mean Absolute Percentage Error (MAPE).

The literature currently focuses on different aspects of web server performance that have been measured for metrics like load time, response times and user behavior. However, when it comes to comprehensive workload evaluation for real-world web servers, there appears to be an important research gap. However, different performance parameters and methodologies examined in earlier studies, it is still needed to investigate factors that contribute to high response times for web servers. In addition, current literature is mainly concerned with the performance of Web sites in specific domains, such as e-commerce sites or academic servers, leaving gaps in understanding the wider impact and possible solutions for optimizing website performance across different domains. This research therefore aims to address these gaps by providing an extensive study of workload characterization for a real-world Web server, as well as insights on why there is a large response time. This study will contribute valuable knowledge on Web server performance evaluations by identifying time intervals in which workload fluctuations in a significant manner, and providing recommendations for reducing response times.

RESEARCH METHODOLOGY

Fig. 1. illustrates the systematic framework of the research methodology, that starts with the pre-processing of access of log file for generating the raw data and workload parameters. The workload has been classified into clustering and data analysis procedures. A CSV file created using raw data which was imported into database, after that the data exploration task was accomplished. On the basis of data analysis, the functional description and visualization techniques were used for workload characterization. Though, the predictive and response variables identified for the regression analysis and measure description. On the basis of box plot the performance evaluations were achieved.

Fig. 2. illustrates the access log file scenario of the system. Firstly, host IP address used for the information requesting purposes with HTTP protocol and also timespan was generated. The request response applied to acquire the knowledge of total number of bytes transferred to the client along with the URL from where information accessed and shows as user panel information. The access log file is used for characterizing, evaluating and usage pattern of the user. It provides the most data which is needed for workload characterization. It is very helpful for detecting server attack; web bots' activities, fixing broken links, measuring bandwidth usage, and user behavior for navigation of any website. It does not store indirect access such as cache memory, proxy server, Ajax and required more space for storage. Access log usually contains the host name, the time and date when a request made, response code, byte transferred from the server and the name of requested documents.

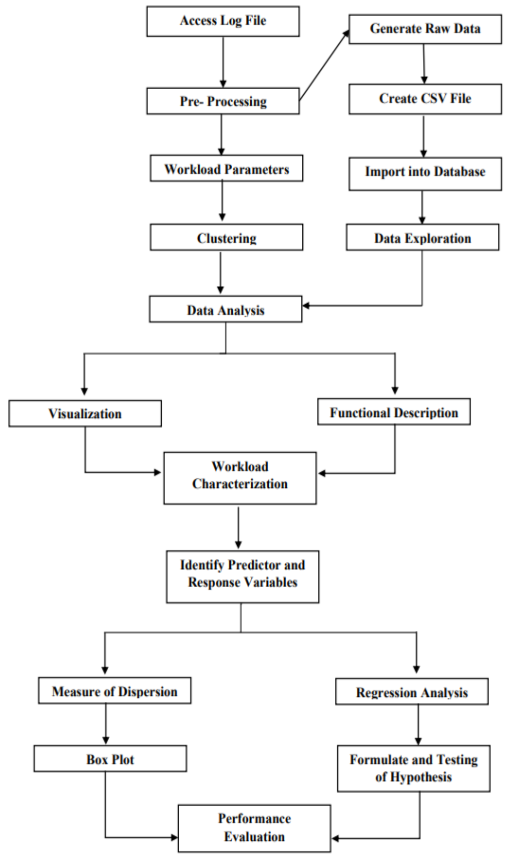

Figure 1. The systematic framework of research methodology

A sample of single entry of access log file is given: **Table 1** displays the action performing for accessing the log file, 182.179.93.148 - - [26/Feb/2014:00:07:33 -0800] "GET/images/logo.jpg HTTP/1.1" 200 1403 "http://www.wikipedia.org" "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64; Trident/5.0)".

Also, there are many resources on collecting access log file such as server side, client side and proxy server. The dataset used in this study is access logs file collected from web server of University of Sindh, Jamshoro (www.usindh.edu.pk). The access log consists of 21 days data from 31st October to 20th November, 2019 in which total requests were received at web server are 3.58 million. After cleaning data, 250 recodes were removed which is not in correct format so that 3.57 million requests were founded and about 23 GB data is transferred to users.

RESULTS AND DISCUSSIONS

The analysis results of the web server's workload and performance are presented in this section. Data analysis allowed to gain insight on workload patterns, as number of requests and bytes transferred to users. Performance evaluations were conducted using box plots to visualize the distribution of response times. These evaluations provided valuable insights into the web server's performance under varying workloads. By identifying outliers and assessing the central tendency of response times, we gained a better understanding of the server's overall efficiency.

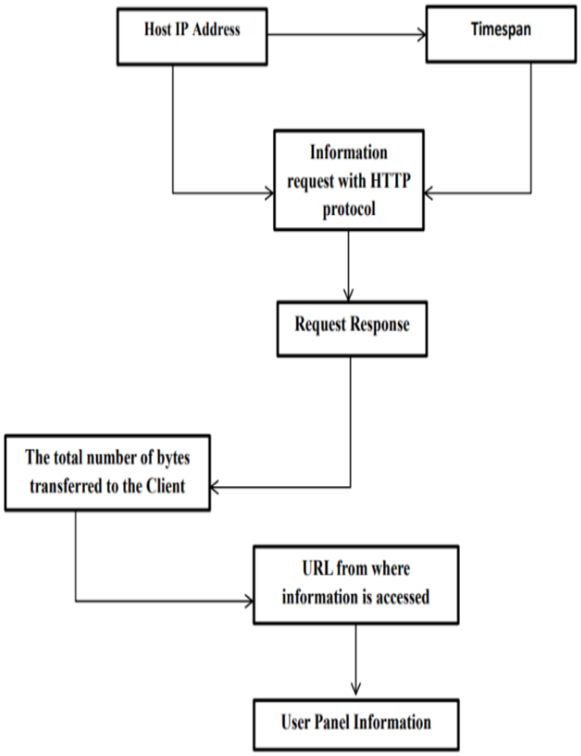
Host IP Address
Timespan
Information request with HTTP protocol
Request Response
The total number of bytes transferred to the Client
URL from where information is accessed
User Panel Information

Figure 2. The www log file scenario

HTTP Method

In the first step of data analysis, the HTTP method is analyzed. **Table 2** shows the request distribution by HTTP method, according to analysis, most incoming requests are made using the POST method, approximately 69.11%. This method, which implies that a significant number of users are engaged in activities linked to handling confidential information, is often accompanied by delicate interactions like file transfers, form submission, or resource creation. The next most common method observed after POST is GET, accounting for approximately 30.82% of the requests and 11.76% of the bytes transferred. It is important to note that the access log also contains a very small number of requests made using HEAD and OPTIONS methods which collectively make up less than 1% of the total dataset. Such fewer common methods can still be a sign of specific user behavior or demands, although they are rarely used.

Table I. Log File Access Entry with description

Example value	**Description**
182.179.93.148	**Who?** The requesting host IP address
26/Feb/2014:00:07:33 -0800	**When?** Timestamp contains Date, Time and Time offset information
GET /images/logo.jpg HTTP/1.1	**What?** What information is requested via HTTP method and HTTP protocol
200	**What is happen?** Response code to conform request is successful or unsuccessful
1403	**How much?** The total number of bytes transferred to the Client.
http://www.wikipedia.org	**From where?** The URL from where information is accessed
Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64; Trident/5.0	**By which means?** Information about client's browser, OS, kernel and user interface

HTTP Protocol Version

In the next phase of analysis, we look at HTTP version adoption by users in resource requests. In particular, the single standard HTTP version is revealed in the findings: HTTP 1.1. This discovery has substantial implications for server performance and security, even though it appears to be straightforward. In terms of protocol diversity within the user base, the disclosure of a single HTTP version, HTTP1.1, implies a significant limitation. The absence of other versions and the potential implications for this are the subject of critical examination. The presence of HTTP 1.1 can indicate the lack of adaptation to more modern and efficient protocols such as HTTP/2 or HTTP/3, while it remains widely supported and functional. This is why further investigation needs to be made into the reasons for this absence and whether it hinders Web servers' ability to optimize data transfers and ensure a responsive user experience. In this study, we acknowledge that it is not possible to investigate the reasons for the lack of other HTTP protocol versions beyond HTTP 1.1 in user resource requests, leaving open questions for future research in this area.

Table 2. Workload by HTTP Method

HTTP Method	% of Requests	% of Bytes Transferred
GET	30.82	11.76
POST	69.11	88.24
HEAD	0.07	0.00
OPTIONS	0.00	0.00
SUBTOTAL	**100.00**	**100.00**

HTTP Response code

Subsequently, HTTP response codes are analyzed in HTTP requests Table III shows that most of the requests (97.32%) resulted in successful (response code 200) transfer of a resource. The web server effectively transmits 99.56% of the total amount of bytes to users. The remaining bytes are associated with other response codes. The second most frequent response code, 404 (Not Found), represents around 1.87% of the requests. This response signals that the requested resource does not exist or is broken, which may happen when users enter an incorrect URL. Other response codes include redirection, and client or server errors, among which 302 (Found) means the requested resource is found but temporarily moved, and 304 (Not Modified) represents that the resource is in the cache, and the previous version is unmodified, so there is no need to retransmit it.

Workload in type of documents

Now, the data is analyzed according to the type of documents of requested resources by users. The categorization of document types has been carried out based on file extensions, resulting in five distinct categories, namely HTML (e.g. .htm, .html), Images (e.g. .jpg, .png, .gif, .jpeg), Documents (e.g. .doc, .txt, .pdf), Embedded (e.g. .css, .js, .php), Query Strings, and Others (e.g. directory). The analysis revealed that most of the requests came from category "other" files, and accounted for 96.84% of total requests. In particular, query strings are represented by 1.48% of

Table 3. Workload by HTTP Response Code

HTTP Response Code	% of Requests	% of Bytes Transferred
200(Successful)	97.32	99.56
206(Partial Content)	0.00	0.00
301(Moved Permanently)	0.00	0.00
302(Found)	0.33	0.14
303(See Other)	0.00	0.00
304(Not Modified)	0.22	0.00
400(Bad Request)	0.02	0.00
403(Forbidden)	0.00	0.00
404(Not Found)	1.87	0.25
500(Internal Server Error)	0.00	0.04
SUBTOTAL	**100.00**	**100.00**

requests as the 2nd most commonly used document type. Query strings, which represent 62.11% of the total number of bytes transferred, are also very prominent. On the contrary, 36.81% of bytes transmitted are allocated to "Other" categories. The remaining byte transfers are allocated as follows: Embedded files (0.68%), which encompass certain PDF files and web bots (ROBOT.txt); Images (0.23%); and Document files (0.17%). Table IV show the statistics of type of documents.

Table 4. Workload by Type of Documents

Type of Documents	% of Requests	% of Bytes Transferred
HTML	0.00	0.00
Embedded	1.16	0.68
Documents	0.05	0.17
Images	0.47	0.23
Query Strings	1.48	62.11
Others	96.84	36.81
SUBTOTAL	**100.00**	**100.00**

Daily traffic volume

Fig. 3(a) shows the percentage of daily traffic volume in aspect to number of requests handled by the website. The y-coordinates of the data show the percentage of traffic volume relative to the number of requests, while the X-coordinates of the data show the number of days. It has been clear that the traffic volume percentage, as regards requests, fluctuates over two distinct periods: from 2 November to 4 November and 9 November to 16 November. From this pattern of traffic, a negative trend line has been analyzed. Three particular peak days were noted in the daily traffic data as compared to cases where users submitted a large number of requests. The following peak days are 03-Nov (22.13%), 02-Nov (17.20%) and 14-Nov (17.01%). The detailed analysis of these specific peak hours shows the main user groups include newcomer students who are frequent visitors to the website for essential admission-related tasks such as checking preliminary entry test results, accessing provisional merit lists, downloading challans, and submitting required documents. These patterns are particularly pronounced during the last months of the year when a university is engaged in major admission activities that include entry tests for bachelor's programs. This test may be directly related to a significant increase in requests on November 2nd, and the results of that test were announced the following day showing an essential role played by these services during this period. It underlines the need to have a robust website infrastructure that can cope with an increase in demand from users related to these important academic events.

Fig. 3(b) shows the percentage of daily traffic volume in aspect to number of bytes transferred (GB) by the web server. In this representation, the y-coordinates indicate the % of traffic volume which is relative to the number of bytes transferred while the x-coordinates represent the number of days. On the 19th and 20th of November, it was found that the largest number of bytes had been transferred.

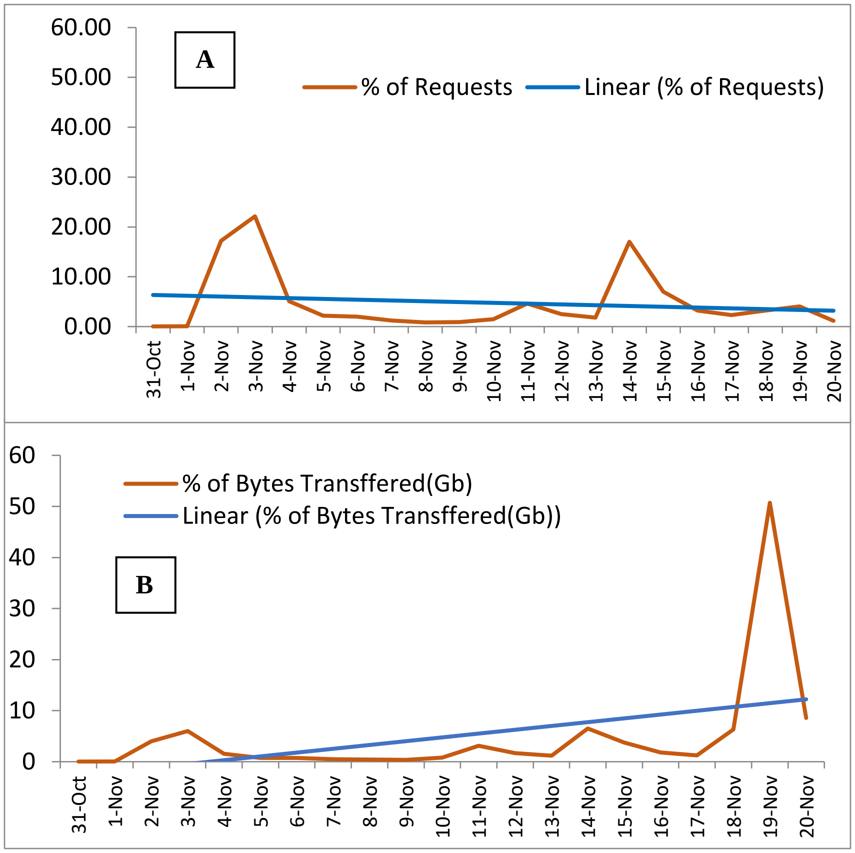

Figure 3. Daily Traffic (A) % of Request (B) % of Bytes Transferred

Hourly traffic volume

For better understanding, the results of hourly traffic have been analyzed, which focus on the number of requests and the number of bytes transferred, as in Fig 4(a) and Fig 4(b), revealing distinct patterns in the hourly traffic volume. In particular, it is clear from these figures that, as observed on 2 November 2019, the busiest hours are between 6:00 and 9:00 p.m.

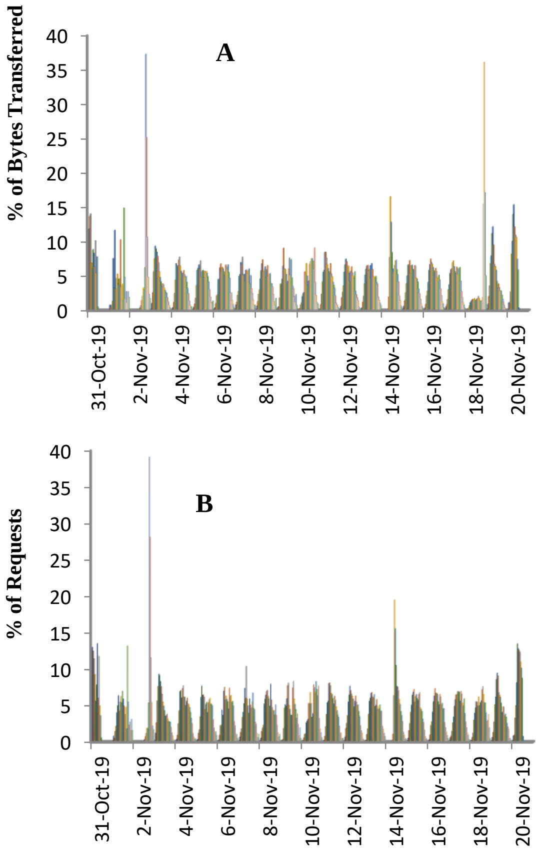

Figure 4. Hourly Traffic (a) % of Request (b) % of Bytes Transferred

Furthermore, on 18 November 2019, the next period of increased activity will be shown in Fig 4(b), from 8:00 to 11:00. As many students had downloaded their Challans from the web portal, which accounts for the significant increase in bytes transferred during those hours.

Furthermore, we distributed the hourly traffic volume into four-time intervals such as Midnight (12:00 AM to 6:00 AM), Morning (6:00 AM to 12:00 AM), Afternoon (12:00 PM to 06:00 PM), and Evening (6:00 PM to 12:00 PM). Their results are shown in Fig. 5 in both aspects (a) number of requests and (b) number of bytes transferred by the server. In Fig 5(a), the highest number of requests was observed during the evening hours of November 2nd, which coincided with the morning's pre-entry test. This is a clear indication of students' interest in checking their test scores online during the evening. At the same time, Fig 5(b) shows a parallel trend in the percentage of bytes transferred, with the most significant data transfer also recorded on the same night, reinforcing the previous observation on the volume of requests.

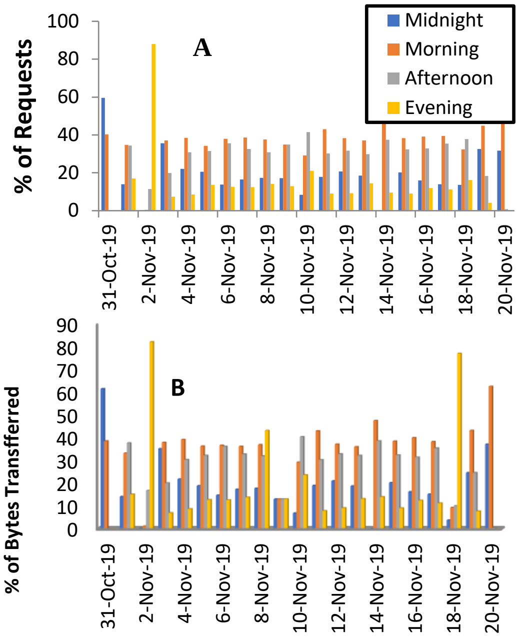

Figure 5. Time Intervals (a) % of Requests (b) % of Bytes Transferred

Box plot Analysis

For analysis performance evaluation, present the spread of the dataset using box blot graphical representation, which has a total of five statistical summary parameters, including maximum, minimum, first quartile, third quartile, and median values. In Fig 6, a box plot is presented to depict the distribution of daily requests (hits). The data show a positive skew,

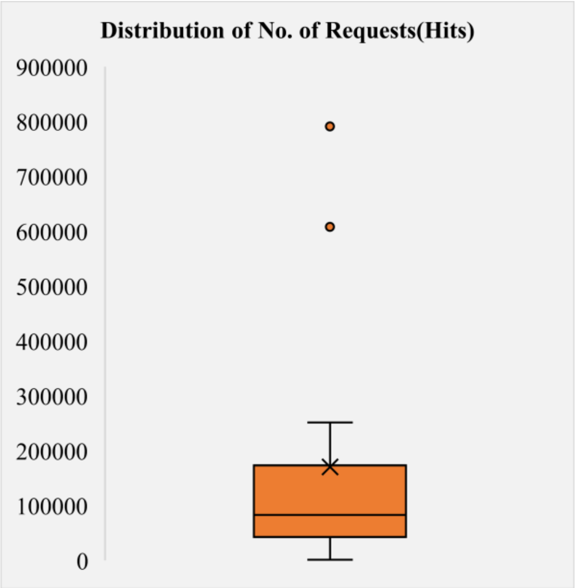

Figure 6. Distribution of dataset according to number of Requests

indicating that the majority of daily requests are relatively low, with only occasional cases of very high requests. However, it is important to pay attention to the presence of strong upper outliers, occurring in three out of 21 days with daily requests ranging from 6 million to 8 million. These outliers suggest a large increase in requests on certain days. This increase can be attributed to the days when universities conduct admission tests. During these times, there may be significant increases in web traffic and this can lead to more requests from visitors due to access to the university's website for registering, checking exam test scores, or obtaining admission information.

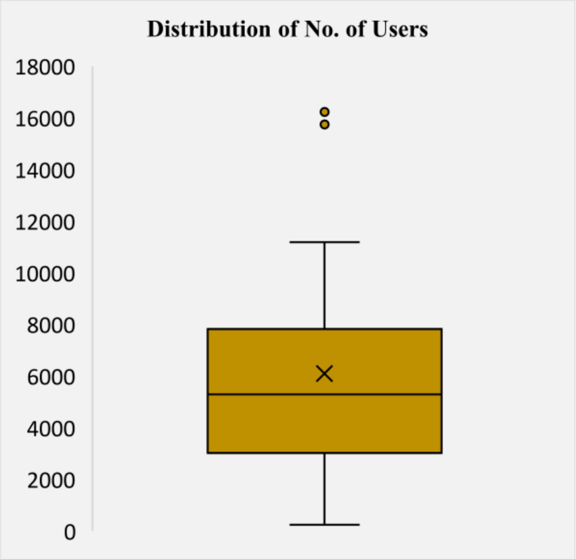

Figure 7. Distribution of dataset according to number of users

In Fig 7, the dataset for daily users shows a positive skew, implying that the majority of users have low daily requests, while some have high daily requests. Approximately 75% of the users register below 8000 daily, and there is a slight difference between the mean and median, indicating a slight rightward skew in the distribution.

The distribution of the transferred MBs is positively skewed in Fig 8, with the majority of MBs being small in size. The mean is closer to the upper whisker, indicating the influence of larger values. A strong upper outlier is caused by users accessing large files during the admission process.

Regression Analysis

Regression is a statistical technique to determine the relationship between independent and dependent variables, as prediction would be achieved. In this study, linear regression and multiple regression were used.

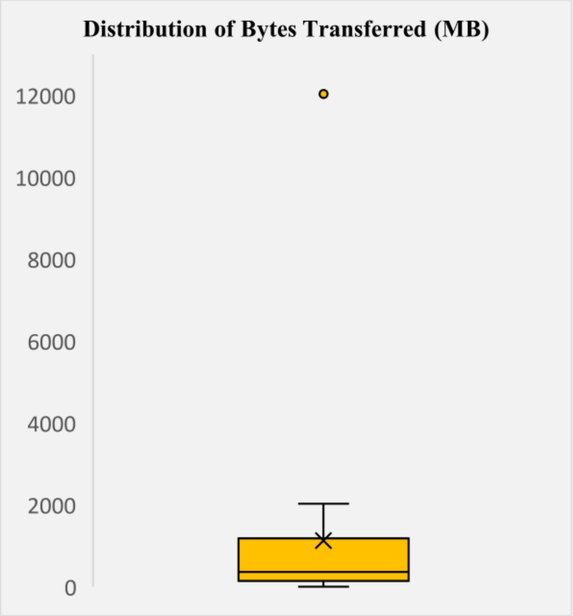

Figure 8. Distribution of dataset according to number of bytes transferred

The first observation states that the number of daily requests submitted by users is increased by the number of daily users who accessing the website. To test the above observation, linear regression is used, in which two variables, independent (i.e., predictor) variable number of daily users and dependent (i.e., response) variable numbers of daily requests were set. In Fig. 9, there was strong positive relationship between these variables (i.e., correlation coefficient) was found 92%. The fitted linear regression model is:

Let R be the number of requests, U be the number of users, and B be the bytes transferred (in MB).

The regression equation can be written as:

$R= -115485 + 46.92 \times U$

From the above model, -115485 represent the value of dependent variable, however independent variable value is zero. Basically, 46.92 indicates that how change is occurred when independent variable changes. Further, we examine the goodness-of-fit measure (R Square) for linear regression models and found that 85% of the data was fitted to the regression model. The significance level was set to be 0.05(5%) and p-value is 2.27E-09. It indicates that the regression model as a whole is statistically significant ($R2 = .85$, $F(1, 19) = 110.90$, $p < .2.27E\text{-}09$). Hence, it was found that number of users significantly predicted number of requests ($\beta = 46.92$, $p < 2.27E\text{-}09$).

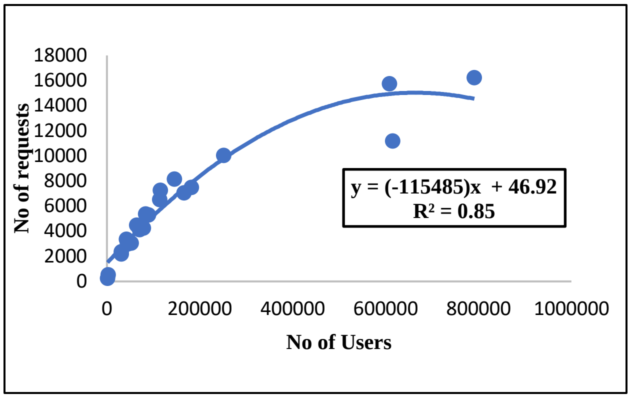

Figure 9. Relationship between users and number of Requests

Next, second observation state that the number of daily requests made by users is influenced by the number of bytes transferred (MB) to users. In Fig. 10, linear regression was used to test the above observation. The independent variable is the number of daily requests and dependent variable is number of bytes transferred (MB) to users. In this analysis, 11% weak positive correlation coefficient was found and the fitted regression equation is:

$$B=1.23\times10^{-6}+0.92\times R$$

The R-squared is 0.01, which indicates that 1% of the variance in the bytes transferred (MB) can be explained by the number of requests. The overall regression was not statistically significant (R2 = .01, F (1, 19) = 0.21, p < .65). It means that the number of daily requests not significantly predicted Bytes Transferred (MB) (β = 1.23E-06, p < .65).

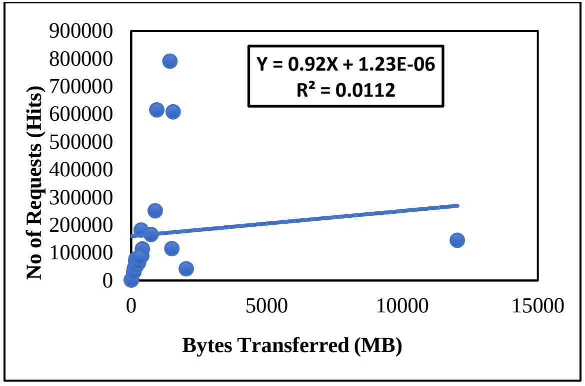

Figure 10. Relationship between Requests and Bytes Transferred (MB)

The last observation state that the type of documents requested by users and the number of bytes transferred (MB) to users have a significant effect on workload of the web server. For the testing of above observation, multiple regression was used in which number of bytes transferred (MB) is a dependent on type of documents such as images, HTML, Embedded, Documents, Query String and others files. The correlation coefficient between these variables was found 100% (0.99) which is perfectly strong positive and regression model is:

Bytes Transferred (MB) = (-294.077) + 0.004×Images + (-104.632) × HTML+4.68×Embedded+ (-0.13) ×Query String+0.39×Documents+0.002×Others

The above model was 100% (0.99) fitted to our dataset. Next, we examine each variable and found that the independent variable Documents, Embedded, Query string, and others are statistically significant to bytes transferred (MB) whereas HTML and Images is not significant to it. Finally, it is state that the above regression model is statistically significant to (R2 = 0.99, F (6, 14) = 1862, p < 1.73E-19) so that the type of documents requested by users can predict the bytes transferred (MB) to users. Fig. 11 shows the results of above analysis.

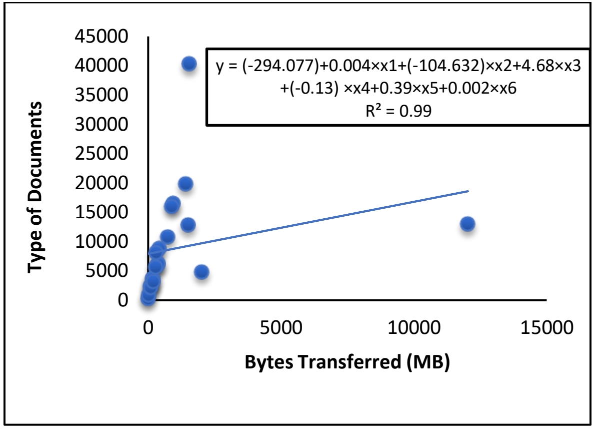

Figure 11. Relationship between type of Documents and Bytes Transferred (MB)

From the above observations, it is highlight that on the highlighted days of the dataset, the webserver was accessed much higher than the other days due to the following reasons, such as, (a) on 2nd Nov., the university-based pre-entry test for the admission process was conducted; (b) on 3rd Nov., the results of the pre-entry test exam were announced online; (c) on 11th Nov., the first provisional merit list of various degree programs was declared online; (d) on 14th November at morning time interval, the university accepted the application from candidates for objection on first provisional merit list. The candidates downloaded the admission fees challan documents on the same day, and it definitely increased the user's traffic volume.

CONCLUSIONS

Write This study presented workload characterization, which evaluates the performance of web servers. The twenty-one-day access log is collected from the University of Sindh, Jamshoro (https://www.usindh.edu.pk) web server which has 3.58 million total requests and 23.73 GB data transferred. Firstly, request data was analyzed in aspect to the HTTP method, the POST method found as greater in against to others, in which 69.131% of requests were submitted and 88.24% of bytes were transferred to users. Finally, this study presented the daily and hourly traffic in which the maximum number of requests were submitted by users on 02-Nov.-19 and 03-Nov.-19, whereas most bytes were transferred on 19-Nov.-19. We also divided the daily traffic into four-time intervals and observed the most requests submitted by users and the number of bytes transferred at evening time intervals of the country.

After presenting workload characterization, the box plot used to show the distribution of the dataset, in which the number of requests, number of users, and the number of bytes transferred were found as positively skewed and upper outliers. In the last, the regression analysis is also used to test the impact of different dependent variables on independent variables. We found that the number of users can directly influence the number of requests made by the users, however, the number of bytes transferred is reflected in the type of documents requested by the users. We found in our study that PDF and image document files have large sizes, and users access the above types of document files frequently. So, there is a need for compressing to reduce the size of downloading files and remove broken links from web pages. On the other side, we observed that the loading time of a web page is reflected by the number of embedded files on the particular web page. Combine CSS and JS files into one file and minimize embedded files on web pages that would also help to speed up the web page's loading time.

In the last, we suggest that it would be better to increase the bandwidth of the web server during special occasions. Another effective optimization way is to use a website cache and prefetching resources is also useful.

ACKNOWLEDGEMENTS

This work was supported for her MPhil studies at IMCS, University of Sindh, Jamshoro, Pakistan. Authors would like to thanks Information Technology Services Center (ITSC) University of Sindh Jamshoro for helping in collection of weblog Dataset.

REFERENCES

Bhutto, A., Chandio, A. A., Luhano, K. K., & Korejo, I. A. (2023). Analysis of Energy and Network Cost Effectiveness of Scheduling Strategies in Datacentre. Cybernetics and Information Technologies, 23(3), 56-69. Retrieved from https://dl.acm.org/doi/abs/10.2478/cait-2023-0024

Curino, C., Jones, E. P., Madden, S., & Balakrishnan, H. (2011). Workload-aware database monitoring and consolidation. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 313-324.

Chandio, A. A., Zhang, F., & Memon, T. D. (2014). Study on LBS for characterization and analysis of big data benchmarks. Mehran University Research Journal of Engineering and Technology, 33(4), 432-440. Retrieved from https://publications.muet.edu.pk/article_detail_abstract.php?p_id=967

Chandio, A. A., Korejo, I. A., Khuhro, Z. U. A., & Memon, F. N. (2013-March). Clouds based smart video transcoding system. *Sindh University Research Journal-SURJ (Science Series)*, *45*(1). Retrieved from https://sujo.usindh.edu.pk/index.php/SURJ/article/view/5502

Chandio, A. A., Yu, Z., Syed, F. S., & Korejo, I. A. (2013-April). A case study on job scheduling policy for workload characterization and power efficiency. *Sindh University Research Journal-SURJ (Science Series)*, *45*(A-1), 23-28.

Choudhury, N. (2014). World wide web and its journey from web 1.0 to web 4.0. International Journal of Computer Science and Information Technologies, 5(6), 8096-8100.

Eldin, A. A., Rezaie, A., Mehta, A., Razroev, S., de Sjöstedt-de Luna, S. S., Seleznjev, O., ... & Elmroth, E. (2014, March). How will your workload look like in 6 years? analyzing wikimedia's workload. In 2014 IEEE international conference on cloud engineering, 349-354. IEEE.

Ferrari, D. (1972). Workload characterization and selection in computer performance measurement. *Computer*, *5*(4), 18-24.

Ferrari, D., Serazzi, G., & Zeigner, A. (1983). Measurement and tuning of computer systems.

Hossain, M. T., Hassan, R., Amjad, M., & Rahman, M. A. (2021). Web Performance Analysis: An Empirical Analysis of E-Commerce Sites in Bangladesh. International Journal of Information Engineering & Electronic Business, 13(4), 47-54, DOI: 10.5815/ijieeb.2021.04.04.

Jarkad, P. M. P., & Bhonsle, M. (2015). Improved Web Prediction Algorithm Using Web Log Data. International Journal of Innovative Research in Computer and Communication Engineering, 3(5).

Kumar, M. (2017, April). Analysis of visitor's behavior from web log using web log expert tool. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2, 296-301. IEEE.

Saverimoutou, A., Mathieu, B., & Vaton, S. (2019, June). Influence of internet protocols and CDN on web browsing. In 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 1-5. IEEE.

Samad, H., Hanizan, S. H., Din, R., Murad, R., & Tahir, A. (2018, May). Performance evaluation of web application server based on request bit per second and transfer rate parameters. In Journal of Physics: Conference Series,1018(1), 012007. IOP Publishing.

Song, Y. D., & Mahanti, A. (2019). Comparison of mobile and fixed device workloads in an academic web server. In 2019 IEEE International Symposium on Measurements & Networking (M&N), 1-6. IEEE.

Summers, J., Brecht, T., Eager, D., & Gutarin, A. (2016). Characterizing the workload of a Netflix streaming video server. In 2016 IEEE International Symposium on Workload Characterization (IISWC), 1-12 . IEEE.

Tochukwu, N. J., & Mary, O. E. C. (2020). Performance Evaluation of Web Servers using Response Time and Bandwidth. Performance Evaluation, 9(12), 133-138.

Williams, A., Arlitt, M., Williamson, C., & Barker, K. (2005). Web workload characterization: Ten years later. *Web content delivery*, 3-21.

Xu, H., Li, Z., Chu, C., Chen, Y., Yang, Y., Lu, H., ... & Stavrou, A. (2018). Detecting and characterizing web bot traffic in a large e-commerce marketplace. In Computer Security: 23rd European Symposium on Research in Computer Security, ESORICS 2018, Barcelona, Spain, September 3-7, 2018, Proceedings, Part II 23,143-163. Springer International Publishing.