UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ÍTALLO DE SOUSA SILVA

WORKLOAD CHARACTERIZATION OF A LARGE ECOMMERCE PLATFORM

CAMPINA GRANDE - PB

2023

ÍTALLO DE SOUSA SILVA

WORKLOAD CHARACTERIZATION OF A LARGE ECOMMERCE PLATFORM

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Fabio Jorge Almeida Morais
Co-Orientador: Professor Dr. Thiago Emmanuel Pereira da Cunha Silva

CAMPINA GRANDE - PB

2023

ÍTALLO DE SOUSA SILVA

WORKLOAD CHARACTERIZATION OF A LARGE ECOMMERCE PLATFORM

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Fabio Jorge Almeida Morais

Orientador – UASC/CEEI/UFCG

Professor Dr. Hyggo Oliveira de Almeida

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 14 de Fevereiro de 2023.

CAMPINA GRANDE - PB

RESUMO

Vários trabalhos abordaram a caracterização da carga de trabalho de servidores web. Esses trabalhos resultaram em uma compilação de padrões chamados invariantes, ou seja, observações recorrentes vistas em vários servidores. Embora alguns desses trabalhos tenham se concentrado em sistemas de comércio eletrônico, eles analisaram dados de servidores de pequenas lojas em um curto espaço de tempo no final dos anos 90 e início dos anos 2000. Assim, este trabalho propôs uma caracterização da carga de trabalho de um servidor de uma empresa multinacional de comércio eletrônico e sua comparação com os invariantes anteriores encontrados na literatura. Descobrimos que alguns padrões, como a presença de picos e vales na distribuição da taxa de chegada de requisições ao longo do tempo e sua relação com as horas de trabalho do dia, continuam presentes em servidores de comércio eletrônico modernos. Enquanto isso, outros diminuíram ou desapareceram, como a correlação entre a taxa de chegada de requisições e a latência. Também conduzimos análises não encontradas na literatura, como o impacto da Black Friday na carga de trabalho do servidor e a análise de duas novas métricas: comprimento da fila de pico (*surge queue length*) e contagem de transbordo (*spillover count*). Encontramos uma taxa de chegada mais alta durante a Black Friday do que em dias típicos, uma distribuição assimétrica para o comprimento da fila de pico e uma associação entre a contagem de transbordo e valores elevados de comprimento de fila e latência.

Workload characterization of a large ecommerce platform

Ítallo de Sousa Silva
itallo.silva@ccc.ufcg.edu.br
Federal University of Campina Grande
Campina Grande, Paraíba, Brazil

Fabio Jorge Almeida Morais
fabio@computacao.ufcg.edu.br
Federal University of Campina Grande
Campina Grande, Paraíba, Brazil

Thiago Emmanuel Pereira da Cunha Silva
temmanuel@computacao.ufcg.edu.br
Federal University of Campina Grande
Campina Grande, Paraíba, Brazil

ABSTRACT

Several works covered the characterization of web servers' workload. These works resulted in a compilation of patterns called invariants, i.e., recurrent observations seen in multiple servers. Although some of these works focused on ecommerce systems, they analyzed data from small store servers within a short timespan in the late 90s and early 2000s. Thus, this work proposed a workload characterization of a multinational ecommerce company server and its comparison with the previous invariants found in the literature. We found that some patterns, such as the presence of peaks and valleys in the arrival rate distribution over time and its relation with the working hours of the day, continue to be present in modern ecommerce servers. Meanwhile, others have diminished or disappeared, such as the correlation between the arrival rate and the latency. We also conducted analyses not found in the literature, such as the impact of Black Friday on the server workload and the analysis of two new metrics: surge queue length and spillover count. We found a higher arrival rate during Black Friday than on typical days, a skewed distribution for surge queue length, and an association between the spillover count and high queue length values and latency.

KEYWORDS

Workload characterization. Ecommerce. Modern server infrastructure

1 INTRODUCTION

Online sales in the retail industry have been increasing over the past years [2]. This growth has required companies in this industry to maintain a high level of Quality of Service (QoS) to retain customers. A workload characterization can help improve a web server's QoS by providing information that can guide crucial decisions, such as infrastructure purchasing, service level objectives (SLOs) establishment, and server caching utilization.

A web server workload characterization involves applying statistical techniques, such as descriptive statistics and time series analysis, to describe features and patterns in the server workload. Previous works that focused on characterizing workloads for web servers resulted in a set of patterns that can be observed across many web server workloads, called invariants. Examples of these invariants include a long-tailored response time distribution and a high request success rate [1].

In the early 2000s, some works focused on ecommerce web servers using data usually obtained from small stores and from short time ranges. These works also delivered a set of invariants. However, with the growth and modernization of ecommerce, there is no evidence of whether these invariants are valid for large multinational ecommerce companies' workloads.

Thus, this research aims to do a workload characterization of a modern ecommerce platform, based on data from large ecommerce, and compare the results with achievements of the previous works found in the literature. Some previously registered behaviors, such as the long-tailored response time, are expected to have remained the same. Meanwhile, changes are expected for some behaviors due to evolutions in the technologies used to maintain QoS in servers, such as auto-scaling.

The remainder of this paper is organized as follows. Section 2 discusses the related work and the invariants found by them. Section 3 describes the workload used in the study and the methods applied in the analysis. Section 4 presents the results and their discussion. Finally, Section 5 elucidates the conclusions, limitations, and possible future works.

2 RELATED WORK

Understanding a server workload is essential for evaluating its performance. However, works covering the scenario of an ecommerce server are scarce and restricted to servers with old architectures and low demand, making it difficult to perform tasks such as resource provisioning in these cases.

Table 1 summarizes four works that performed ecommerce workload characterization by presenting the year of publication, the number of servers assessed, and the number of requests and time range for the bigger and longest workload, respectively. The longest time range was of 15 days, which may not be large enough to capture the seasonality of a modern ecommerce and allow a fair generalization of the observations. Also, the number of requests may not be representative for large ecommerce companies in modern days.

Previous works focused on analyzing the workload in three main aspects: request arrival, latency, and sessions. In an ecommmerce context, a session is a sequence of requests made by the same user in a specific time interval.

Regarding the request arrival, Menascé et al. [3], Vallamsetty et al. [5] analyzed its distribution and observed an alternation of peaks and valleys over time. Vallamsetty et al. [5] related this pattern to the period of activity of the ecommerce stores analyzed.

Menascé et al. [3], Suchacka and Dembczak [4], Vallamsetty et al. [5] also analyzed the request time series for self-similarity using different techniques to estimate the Hurst parameter, the main statistic for self-similarity. The methods used were: the variance time plot (VTP) [3], the R/S plot test [5], and the aggregate variance method [4]. All three found a moderate to strong self-similarity.

Table 1: Related works summarization

Authors	Pub. Year	# Servers	# Request	Time range
Menascé et al.	2000	2	3,630,964	15 days
Vallamsetty et al.	2002	2	—[1]	5 days
Wang et al.	2003	1	2,000,000	24 hours
Suchacka and Dembczak	2017	7	89,486	24 hours

[1]The number of request was not provided, but the request rate of the largest server was around 1 request/second

Furthermore, Suchacka and Dembczak [4], Vallamsetty et al. [5] identified highly variable traffic, indicating that the workload had the property of burstiness.

For latency, the works focused on analyzing its distribution for skewness and its correlation with other features, such as the number of requests arriving. Vallamsetty et al. [5] determined that the response time obeys a long-tailored distribution, which can be modeled using a Pareto distribution.

Wang et al. [6] identified the existence of a threshold in the number of requests that lower bounds the latency by analyzing the scatter plot of these features. In addition, they found some correlation between the request rate in one period and the response time in the following periods. They noticed a considerable variation in the latency values regardless of the arrival rate, which could indicate that the type of request can impact response time more than the number of requests.

The analysis regarding the sessions consisted in characterizing their length, request count, and performed operations Menascé et al. [3]. However, our dataset does not allow us to achieve this kind of analysis since it would require a detailed log for each arriving request, and we only had access to an aggregated description. Thus, we will not further detail their methods and results here.

Additionally, in research over general web servers, Arlitt and Williamson [1] established a success rate of around 88% for the requests hitting the server. This success rate was calculated as the proportion of requests with response *Successful* relative to the total.

3 THE WORKLOAD

3.1 Data collection

The data was obtained from the server of a multinational ecommerce company. Given the service offered by the company, its server can be characterized as both *Business to Consumer* (B2C) and *Business to Business* (B2B).

The data obtained is a time series that summarizes information about the server workload in one-minute granularity. Each feature will be described in the following subsection.

Considering the importance of Black Friday (BF) to the retail industry and the months we had access to, we selected August and November 2021 to analyze. Thus comparing the two months, we could assess the impact of BF on the invariants.

3.2 Data Features

Table 2 describes the features available in the dataset. The first column contains the name of the collected metric, the second briefly describes the features, and the third one holds the aggregation statistic. The aggregation statistic is used to summarize the data in the time frame of one minute. For example, the aggregation statistic for request count in a Load Balancer is the sum, so the value associated to a timestamp will be the sum of request received in the last minute.

Table 2: Data features description

Metric	Description	Statistic
Request count	The number of requests arriving in the server.	Sum
HTTP Code 2xx	The number of 2xx HTTP response codes generated by registered requests.	Sum
HTTP Code 3xx	The number of 3xx HTTP response codes generated by registered requests.	Sum
HTTP Code 4xx	The number of 4xx HTTP response codes generated by registered requests.	Sum
HTTP Code 5xx	The number of 5xx HTTP response codes generated by registered requests.	Sum
Latency	The total time elapsed, in seconds, from the time the load balancer sent the request to a registered instance until the instance started to send the response headers.	Average
Surge Queue Length	The total number of requests that are pending routing.	Maximum
Spillover Count	The number of rejected requests due to a full surge queue.	Average

3.3 Characterization approach

We performed the workload characterization by analyzing the distribution and correlation between the features using descriptive statistic techniques (e.g., histograms, summarization functions, and time series analysis) similar to how it was made in the related work.

Due to a non-disclosure agreement, the precise value of the features will be omitted, and a normalized scale (from 0 to 1) will

be used for the results, complemented with information about the order of magnitude.

4 RESULTS & DISCUSSION

4.1 Success Rate

The success rate (SR) is the ratio between the number of requests successfully attended by a server and the total number of requests received.

We can measure the SR from two perspectives that differ on which requests are considered successful: the server-side and the client-side. For the server-side success rate (SSSR), any request with a response code from an HTTP class different from 5xx is considered successful. Meanwhile, for the client-side success rate (CSSR), only the requests with an HTTP 2xx class response code are considered successful.

Both months analyzed presented an SSSR above 99% and a CSSR above 94%. The overall SSSR and CSSR were also above 99% and 94%, respectively. Arlitt and Williamson [1] observed a CSSR of over 88% in the servers they analyzed. In an over-time analysis, we saw that the SSSR and the CSSR for both months stayed above 99% and 90%, respectively, more than 99% of the time. This difference may be due to the HTTP status code considered, Arlitt and Williamson [1] only took into account the 200 status code, however since we don't have these counts stratified, we considered anything from class 2xx as successful.

We also observed a pattern in the distribution of response codes that repeats over the days in both months: during the early hours of the day (1 am - 9 am), the CSSR drops and usually reaches the smaller values for the day. Figure 1 highlights this behavior by showing the hourly distribution of each response code class: the percentage of response for the 2xx class falls while the percentage for the 3xx and 4xx classes rise. The moment of the day during which this happened coincided with the time with the lowest request arrival rate, as will be seen in the next section.

4.2 Request arrival

The observed arrival rate was in order of magnitude six times bigger than all the related works.

Analyzing the distribution of the number of requests arriving over time, we observed alternating peaks and valleys in both months, as presented in Figure 2. Menascé et al. [3] and Vallamsetty et al. [5] also observed these trends in their works. At November 4th, we can see a high peak in the number of requests, this peak was consequence of a load test, thus this day will not be considered for further analysis in this section.

The valleys coincided with the dawn and early morning (0 - 9 am, UTC-3) in the largest country where the analyzed enterprise operates. During the rest of the day, the distribution of the number of requests arriving is similar. Figure 3 shows the distribution of the number of requests per minute grouped by the hour of the day and highlights this observation.

In Vallamsetty et al. [5], the period with the highest arrival rate for the B2C ecommerce server coincided with the after-office period (night) and with the office hours (morning and afternoon) for the B2B server. Since our ecommerce server can be characterized as both B2C and B2B, the period with the highest activity observed from late morning until midnight is in accordance with the previous findings.

We observed no significant differences in the distribution of the number of requests arriving between the weekdays. However, we noticed that the number is smaller during the weekend for both months. For November, we expanded this analysis by creating a third group containing the days around Black Friday (November 25th - 29th) since this is a big event for ecommerce servers.

Figure 4 shows boxplots for the number of requests arriving per minute for each group in November: weekdays, weekends, and Black Friday (BF). The median request count during the BF is larger than the third quartile of the other two groups, i.e., for half the time during the BF, the number of requests was larger than during 75% of the time of typical days.

4.3 Latency

The average latency had a right-tailored distribution, similar to what Vallamsetty et al. [5] observed in their work. The average latency is less than 100 ms for approximately 35% and 65% of the time for August and November, respectively, and less than 200 ms for more than 94% of the time in both months. The median latency stayed below 8 ms over the time in both months.

As for the 99th percentile, both months presented similar distributions (median close to 2.5 s and the third quartile around 3 s). However, November had a denser tail (more points with high latency).

Differently from the request arrival, the latency did not show any correlation with the time of the day. Its distribution is similar throughout the hours of the day.

We found large variations in latency for different intervals of arriving requests, these variations can indicate that the type of the arriving requests is more important than the quantity, similar to the conclusions in Wang et al. [6]. However, different from Wang et al. [6], we did not find a correlation between the number of requests arriving at a moment and the latency in the following moments. One possible reason for this is the utilization of autoscaling in the server infrastructure that dynamically adjusts the number of replicas and mitigates the overhead of many instances.

4.4 Surge queue length and Spillover count

The surge queue length tracks the number of requests waiting to be attended. Meanwhile, the spillover count is the number of requests rejected due to a full surge queue length.

These metrics could tell us about the availability of a server since a full queue may indicate an unavailable server or a server incapable of handling the arriving workload. We did not find analyses regarding these metrics in the previous works.

Both months presented a skewed distribution with a long tail to the right for the surge queue length. Only August had moments with no queue. The anonymized distribution can be seen in Figure 5.

There were 55 and 203 minutes in August and November, with a spillover count bigger than 0. These moments coincided with values for the surge queue lengths above the third quartile and with thousands of requests. And regarding latency, more than half of

Figure 1: Hourly distribution for response code

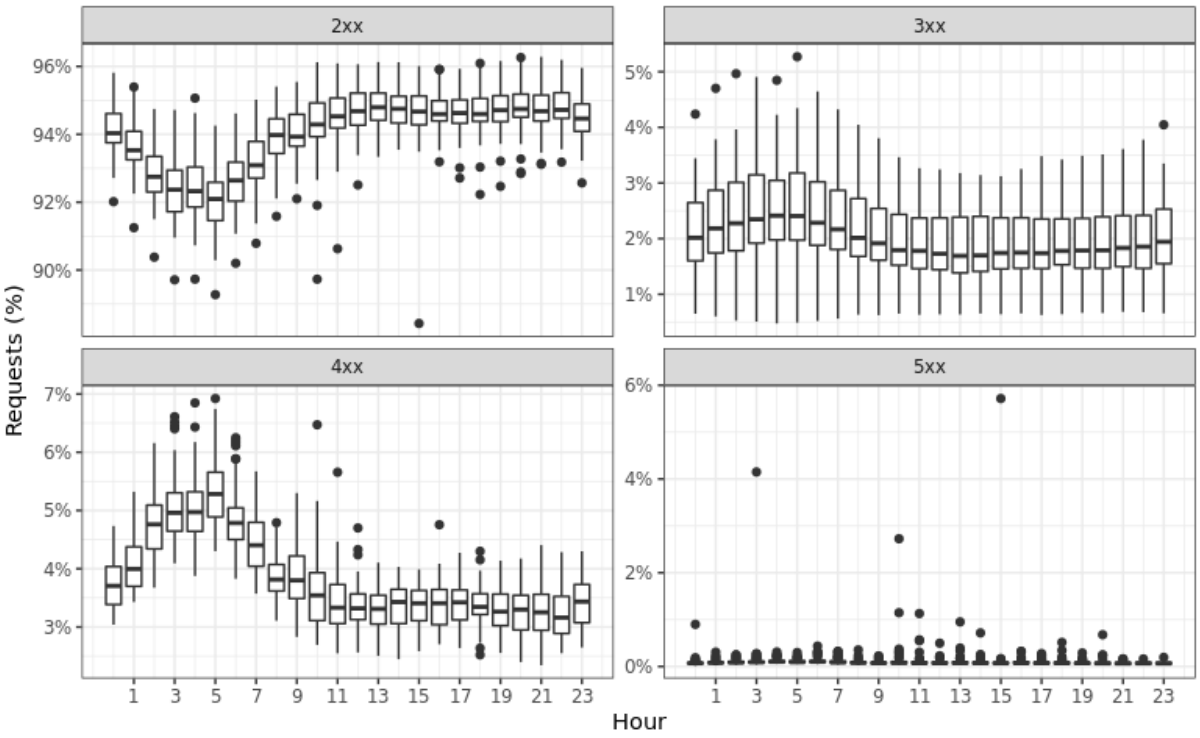


Figure 2: Request arrival distribution over time

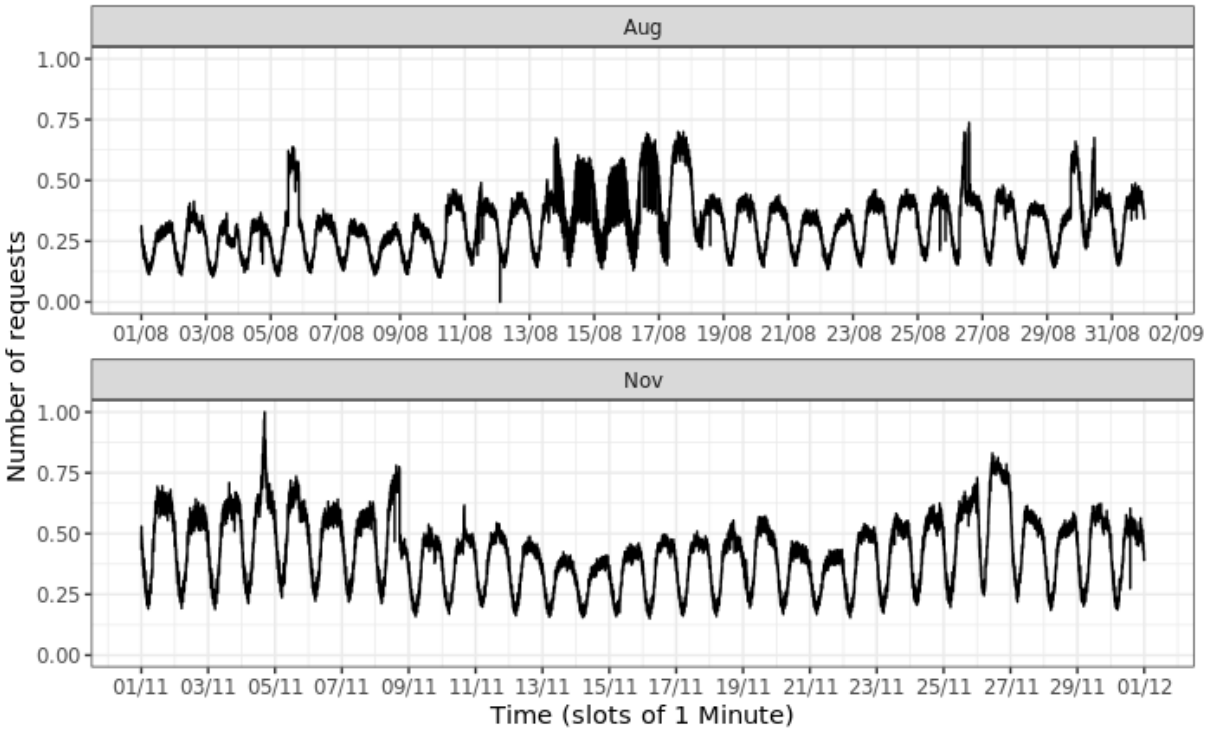


Figure 3: Request arrival distribution grouped by hour for both months

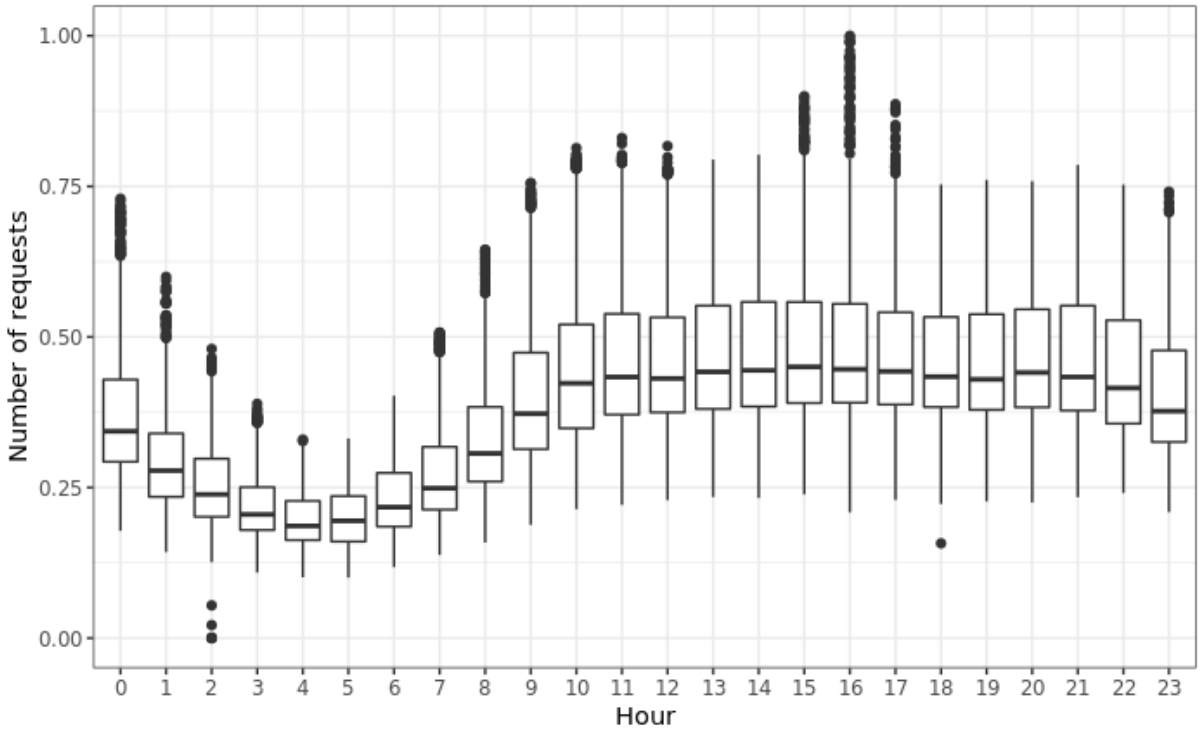


Figure 4: Request arrival distribution grouped by type of the day for November, 2021

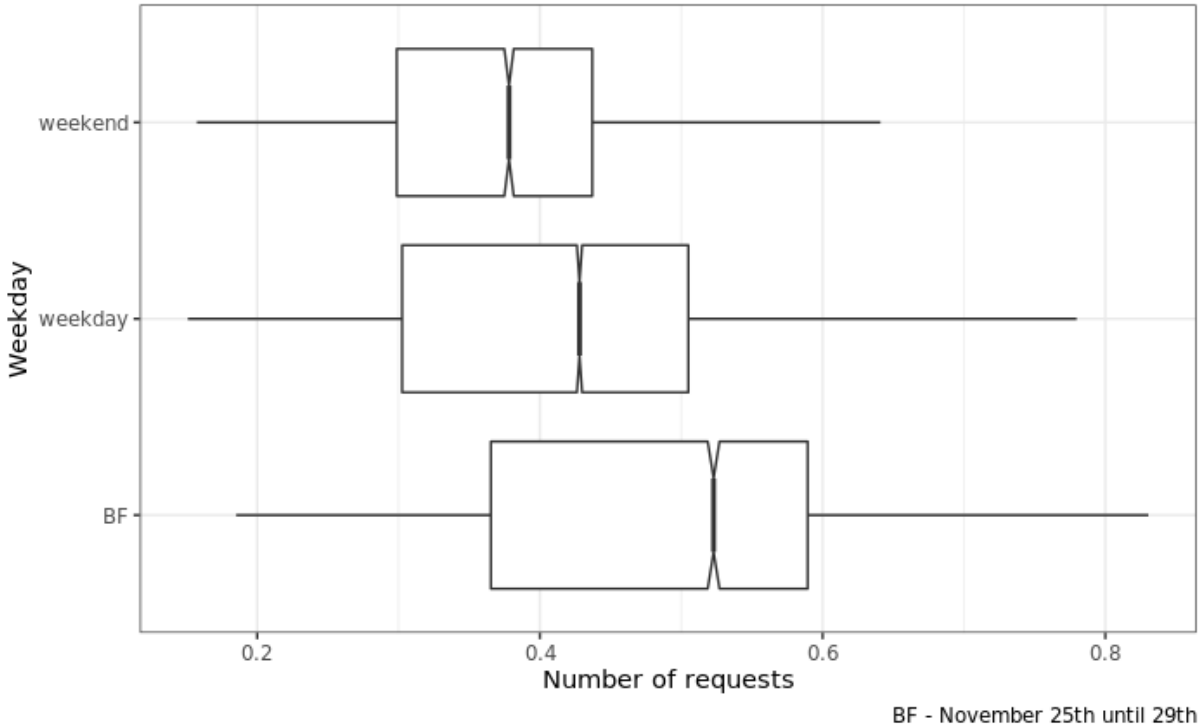
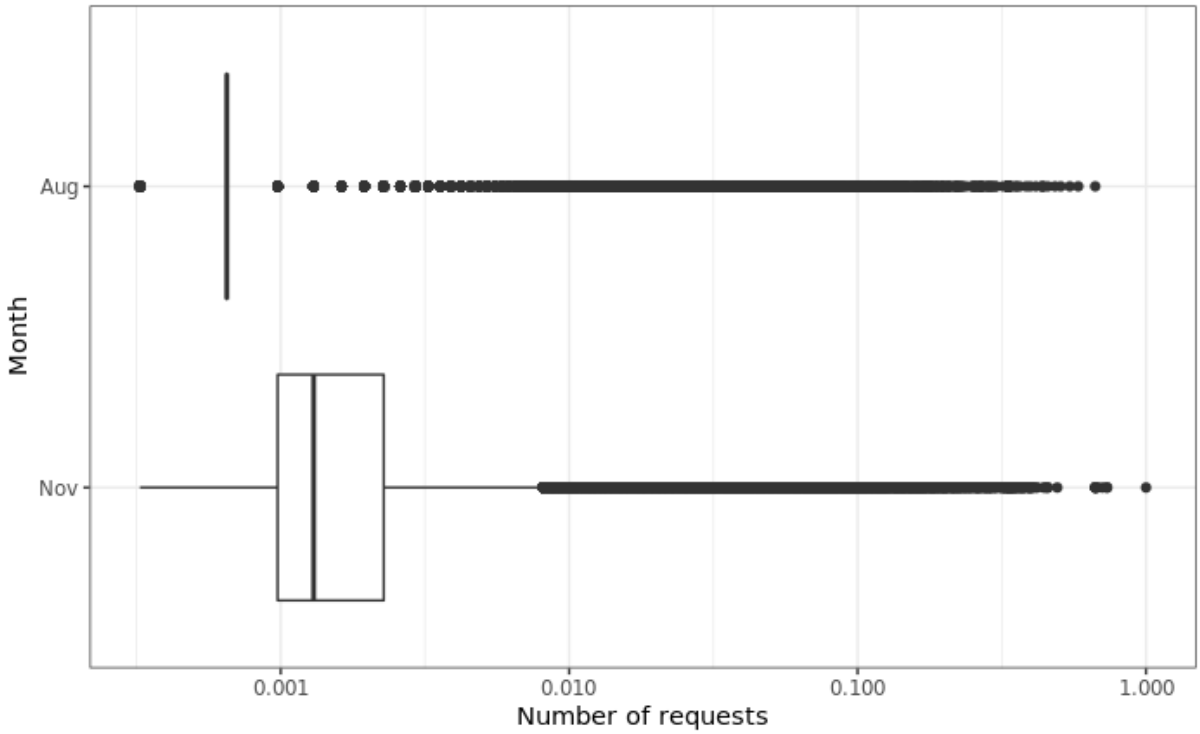


Figure 5: Distribution of surge queue length (log scale)



these values coincided with latency values in the respective month's third quartile.

5 CONCLUSION

The servers' architecture changed over the years, so it is expected that these changes would impact the behavior observed in an application. This work analyzed the workload of a modern large ecommerce web server to find patterns and compare them with the previous observations found in the literature. The following list compiles the finding of this work:

- The success rate for both months analyzed stayed above 94% in both server-side and client-side perspectives.
- The arrival rate distribution over time followed an alternating pattern of peaks and valleys. The valleys corresponded with the dawn and early morning hours, a period of low commercial activity in the largest country where the enterprise works.
- The weekdays presented a similar distribution for the arrival rate. Meanwhile, the weekend had a lower arrival rate compared to them.
- The Black Friday surrounding days impacted the arrival rate. These days had higher arrival rates them the others days of both months.
- Latency presented a skewed distribution with a long tail on the right. Moreover, it did not correlate with the time of the day or the number of requests arriving at the current or previous moment.
- The surge queue length had a skewed distribution with a long tail on the right. The spillover count, as expected, coincided with high values in the surge queue length (cause) and high values of latency (consequence).

The major limitation faced by this work was that we did not have access to a complete description of each arriving request on the server. We had access to an aggregation of this information that did not allow us to perform analysis based on sessions and type of request, as some previous works did. Therefore, some possible future works derived from this could be:

- Extend the analysis by including the facet of sessions and type of requests;
- Analyze the correlation between the workload and the server utilization (e.g., CPU utilization, memory, I/O);
- Analyze the workload for burstiness and self-similarity;
- Check if the workload is suitable for machine learning predictive algorithms given is seasonality.

ACKNOWLEDGMENTS

I would like to thank my advisors, Fabio and Tiago, for their guidance and advice during this research and my undergraduate studies, and all my co-workers for their help with revisions and moral support.

Additionally, I could not have finished this journey without the support of my best friend, Thiago, to whom I do not have enough words to express my gratitude. I also would like to say thanks to my former roommates Gabriel, João Marcelo, Thiago Yuri and

Ada (the cat), with whom I divided many moments in this journey. Lastly, I would like to thank my parents, their belief in me kept me motivated to pursue my goals.

This work has been funded by MCTIC/CNPq-FAPESQ/PB (EDITAL Nº 010/2021) and by VTEX BRASIL (EMBRAPII PCEE1911.0140).

REFERENCES

[1] Martin F. Arlitt and Carey L. Williamson. 1996. Web Server Workload Characterization: The Search for Invariants. In *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '96)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/233013.233034

[2] Daniela Coppola. 2022. E-commerce as share of total retail sales worldwide 2015-2025. https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/

[3] Daniel Menascé, Virgílio Almeida, Rudolf Riedi, Flávia Ribeiro, Rodrigo Fonseca, and Wagner Meira. 2000. In Search of Invariants for E-Business Workloads. In *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC '00)*. Association for Computing Machinery, New York, NY, USA, 56–65. https://doi.org/10.1145/352871.352878

[4] Grażyna Suchacka and Alicja Dembczak. 2018. Verification of Web Traffic Burstiness and Self-similarity for Multiple Online Stores. In *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology – ISAT 2017*, Leszek Borzemski, Jerzy Świątek, and Zofia Wilimowska (Eds.). Springer International Publishing, Cham, 305–314. https://doi.org/10.1007/978-3-319-67220-5_28

[5] Udaykiran Vallamsetty, Krishna Kant, and Prasant Mohapatra. 2003. Characterization of E-Commerce Traffic. *Electronic Commerce Research* 3, 1 (01 Jan 2003), 167–192. https://doi.org/10.1023/A:1021585529079

[6] Qing Wang, Dwight Makaroff, H. Keith Edwards, and Ryan Thompson. 2003. Workload Characterization for an E-Commerce Web Site. In *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON '03)*. IBM Press, 313–327.