Evaluation the remote system quality indicators using a mathematical model

V. Zakirov[1][a], E. Abdullaev[1][b]
[1]Tashkent state transport university, Tashkent, Uzbekistan

Abstract: The popularity of web services has increased the demand for quality indicators of remote service systems. One of the indicators of the quality of such a service system is the response time of the system users. This time depends on several indicators, and if it exceeds a certain value, it causes inconvenience to the users of the system. Works based on previously conducted experimental investigations have a limited use. Unlike experimental studies, this research proposes using public service theory models to estimate the response time of user requests. Studies were done to identify system quality indicators using the suggested mathematical model. Based on the model, the waiting time for request responses, the dependence of the waiting time on the number of users, and the dependence of the waiting time on the service system's internal technical indicators were investigated. The studies done showed that the proposed model is completely consistent with the results obtained by the present experimental approach and may be widely employed in research.

Keywords: public service system, mathematical model, response time, request flow, random process, probability model, duration of service, quality indicators.

1. Introduction

After the COVID-19 epidemic, significant changes occurred around the world. At the same time, new ways of organising work have emerged in a variety of disciplines. For example, in the sphere of service supply, there have been significant developments in terms of distant service organisation.

This showed the need to re-analyze and improve the working principles of existing systems in the field. Because during the pandemic, all types of services were organized remotely, which caused a sharp increase in system users and the loads created by them. This led to the fact that this load was not provided with the required quality of service. This situation can happen in all service areas. Therefore, the study of the reasons for the increase in downloads and the correct Organization of methods of servicing them remains one of the important problems of the present day. Reasons for the origin of the problem [1], [2], [3], [23] considered in studies, they can be mainly told that one of the main reasons for the origin of problems is technical problems, which are considered directly related to an increase in the number of users. Because each service device has a capacity limit, consumers who surpass it will experience these problems. As a result, it is critical to examine the system operating methods and the potential for serviceability of the devices indicated in the research [3], [5], [10], and [11]. To do this, it is required to conduct research on service delivery methods based on mathematical models and create models appropriate for the service sector. According to research undertaken in the servicing of user requests, there are various reasons why difficulties emerge. These include limited service device capabilities, a lack of optimisation in user-generated software, and internet connection concerns. These issues were addressed in works [1], [16], [18], [23], [24] and determined through experimental research. These experiments were conducted in many stages and included the following.

According to [1,] during the first phase of the analysis, there were issues with serving requests with the technical aspects of the service device (HDD, RAM), and their ability to serve multiple requests at once was limited, resulting in many incoming requests not being handled. The research used a system server with Intel® Core™i5-2310 CPU @2.90GHz x 4 processor, 8GB Hard disk, 1GB RAM specifications, and Apache JMeter 2.9 software, which is one of the programs for generating concurrent queries. A stream of timed requests was sent and analyzed. In this case, the number of requests was increased and when 51 requests were sent simultaneously, the system was overloaded and lost the ability to service requests.

In this case, the author made several changes to the device's technical parameters (SSD, RAM) to technically reduce the service time, and the service device is Intel® CoreTMi5-2310 CPU @2.90GHz x 4 processors, 8GB of solid state disc, 1GB of RAM, and as a result, the service time has been reduced by 98% compared to the previous one.

In the second stage, web server research was conducted, with the process-based web server (Apache) being replaced by an event-based web server (Nginx). In this scenario, the web servers were installed on a server with an Intel® CoreTMi5-2310 CPU @2.90GHz x 4 processor, 128 GB SSD, and 4GB RAM. When compared to the prior study, the service time was reduced by 34% as a result of this investigation.

The service time was implemented in the following steps of analysis using system software. It also made use of technologies such as caching, Gzip compression, and script optimization. As a result of the optimization technologies, the service time was decreased by 80%, 75%, and 24% compared to the baseline state.

Management of caches with the help of caching technology is intended to reduce the number of requests to

[a] https://orcid.org/0000-0002-2290-2625
[b] https://orcid.org/0000-0002-8954-9731

the server. As a result, the number of requests to the server decreased, and this, in turn, caused a decrease in service time.

Gzip compression technology reduces service time by employing compression technology to send result files generated in response to user requests. The technology's goal is to lower the size of various forms of data while also reducing the time it takes to transfer them to the user across the network. Because all data is sent over the network as packets, a tiny amount of data results in a small number of packets [4], [6]. This will cut down on the time spent on the request.

The technology of script optimization refers to the software of the service system, and it is meant to alter its sections of the software code from the general code to the necessary code. Because the result of the query served in the database is generated as a result of the software codes, the non-optimality of the codes results in a huge volume of the formed answer [5], [6]. This, in turn, increased the time it took to send them via the network, thus the author observed in the study report that optimizing scripts also influences service time, and making the scripts look the way they should achieve a reduction in service time in his experimental investigations.

Also, researches of this type were carried out in [12], [13], [14], [19], [20], [23], [24], and in them, researches were carried out using experimental methods.

It can also be concluded from the above that the researches related to the service of user requests were carried out only in an experimental - research method.

However, it is now necessary to pre-calculate or evaluate the way the systems perform, as well as the serviceability of the service device. This necessitates the investigation of service approaches based on mathematical models and the development of models appropriate for the service sector.

2. Materials and methods

In this research work, mathematical models were developed for researching the request service system, and based on it, service effectiveness was studied.

As previously stated, in this instance, it can be regarded as a typical flow of requests because user requests enter the system at various times. We believe that the length of their service is governed by a negative exponential law. In this process, requests entering the system are served by V service devices. The system serves requests in a waiting manner, and the number of waiting places is limited to r. Because, according to the service models in [15], serving with an unlimited number of waiting points, the number of requests in the queue part of the system occupies a large volume, the server cannot serve them all, and the system [1], will face the situation of not being able to serve requests like. Therefore, when organizing systems, waiting areas are limited. This, in turn, means the loss of incoming requests to the system when the service device and all queues are busy. In this process, the system serves requests in waiting areas in a FIFO manner. Thus, such a service arrangement corresponds to the $M/M/V/r$ model of user request service based on [Kendal's] specification. The quality indicators of this system are determined as follows. According to it, the probability that all service devices and waiting areas in the system will be occupied corresponds to the case $l = V + r$ and is defined by the following expression (1).

$$p_b = \sum_{i=V}^{l} p_i = \frac{\frac{Y^V}{V!}\frac{V}{V-Y}\left[1-\left(\frac{Y}{V}\right)^{r+1}\right]}{\sum_{i=0}^{V-1}\frac{Y^i}{i!}+\frac{Y^V}{V!}\frac{V}{V-Y}\left[1-\left(\frac{Y}{V}\right)^{r+1}\right]} \quad (1)$$

Here, $Y = \lambda \cdot \bar{t}$ load generated by queries, $\lambda = N \cdot \alpha$ the rate of user-generated requests, N number of system users, α the rate of requests generated by a single user, $\bar{t} = \bar{t_1} + \bar{t_x}$ total time spent serving a single request, $\bar{t_1}$ time spent on user identification, $\bar{t_x}$ service time for one request (according to [3], it is determined based on the technical parameters of the server).

As can be seen from expression (1), if $r = \infty$, this expression becomes Erlang's second or C formula. If so, then it becomes Erlang's first formula. So, when the number of waiting places changes from ∞ to 0, the expression (1) changes in the range of probability of losses according to $M/M/V/\infty$ and $M/M/V/r$ models. Therefore, this expression is a general calculation expression for two models $r = \infty$ and $r = 0$.

(1) the expression can be expressed as the load acting on one device $\eta = \frac{Y}{V}$

$$p_b = \sum_{i=V}^{l} p_i = \frac{1-\eta^{r+1}}{\frac{1-\eta}{E_V(Y)}+\eta-\eta^{r+1}} \quad (2)$$

In this case, the loss of requests in the system occurs when the number of pending requests exceeds r. Therefore, the following expression (3) can be reduced to (4) by the load falling on a single service device.

$$P_{ch} = p_l = \left(\frac{Y}{V}\right)^{l-V}\frac{Y^V}{V!}p_0 = \frac{\left(\frac{Y}{V}\right)^r\frac{Y^V}{V!}}{\sum_{i=0}^{V-1}\frac{Y^i}{i!}+\left(\frac{Y^V}{V!}\right)\frac{1-\left(\frac{Y}{V}\right)^{r+1}}{1-\frac{Y}{V}}} = \frac{\eta^r}{\frac{1}{E_V(Y)}+\frac{\eta-\eta^{r+1}}{1-\eta}} \quad (3)$$

In this case, based on the above formulas, the probability of waiting for requests in the system is determined by the following expression (5).

$$P_{w>0} = p_b - p_{ch} = \frac{Y^V}{V!}\frac{V}{V-Y}\left[1-\left(\frac{Y}{V}\right)^r\right]p_0 = \frac{1-\eta^r}{\frac{1-\eta}{E_V(Y)}+\eta-\eta^{r+1}} \quad (4)$$

Also, the average waiting time for requests to be served is determined by the following expression (5):

$$W = \frac{P_{W>0}}{V-Y}. \quad (5)$$

So, based on the mathematical model described above, it is possible to calculate the quality indicators of the request service system and select the technical indicators of the system based on the analysis of the results

3. Results and discussion

Using the above expressions, we will now research to determine the system's quality indicators.

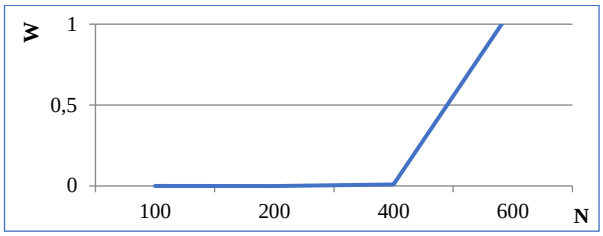

Figure 1. Graph of the dependence of the average waiting time on the number of users

Figure 1 depicts the graph of the average waiting time versus the number of users. The graph shows that as the number of customers increases, so does the average waiting time. When the number of users approaches N, the waiting time skyrockets and the system is unable to reply to queries. This situation supports the conclusion provided in [1]. As a result, the system with the aforementioned indications can provide N users with the desired quality indicators on average. Increasing the number of users from N results in a significant worsening of the system's quality indicators.

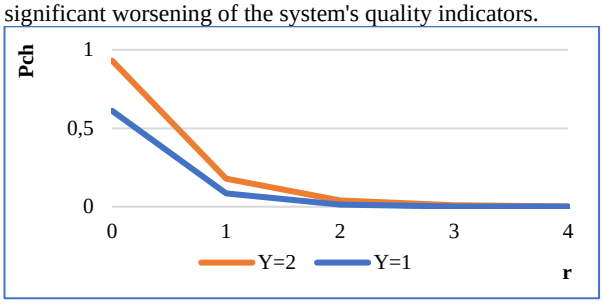

Figure 2. The graph of the dependence of the probability of losses on the waiting areas

Figure 2 depicts the graph of the chance of loss versus the number of waiting areas, which indicates that the probability of loss falls rapidly as the number of waiting spaces grows. This corresponds to the experiment results shown in [1]. Furthermore, the likelihood of losses in this circumstance is determined by the incoming load. The speed of user requests and the service time for one request determines the system load. The service time for one request depends on the following:

- processor power in the service system. It is known that the processor is one of the most basic devices of computer work and performs all its tasks [1], [3], [9], [17], [21]. It is its performance that determines the speed of the system. Because all requests are connected to the memory through the processor and its results are formed. Therefore, service times include the processor's request service time, which is a major part of the total service time;

- time to load HTML documents into the browser. In remote systems, data exchange between the client and the server is carried out through browsers. And it sends the query results to the client device in an HTML file format that the browser can understand [5], [6], [22]. This, in turn, requires the result of the request to be sent to the user via the Internet. In this process, the size of the file and the speed of the Internet connection mean that the service times will be longer or shorter;

- to the number of HTTP requests. Each request sent to the system server requires a certain amount of time from the server to be served [3]. Therefore, the increase in requests to the system causes the system to gradually fill up service devices and waiting areas [7], [8]. This, in turn, increases the volume of the queue of requests as a result of the increase in the intensity of requests to the system, and this causes an increase in the service time. Sometimes, a sudden increase in the number of requests will cause the system to lose its ability to serve;

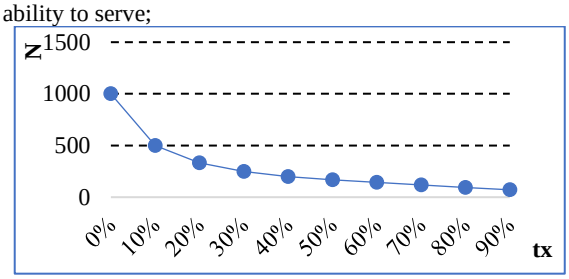

Figure 3. The graph of the dependence of the number of system users on the service time of one request

Figure 3 is a graph of the dependence of the number of system users on the service time of a single request, in which the service time of a single request is shown in the order of increasing percentages. As can be seen from the graph, the number of users of the system decreases as the request service time increases. If we approach this issue from the other side, we can see that it is possible to increase the number of users by reducing the time of serving one request. But the service time for this one request depends on a number of factors, including processor power, technical parameters of other server devices, optimization of the service software, internet speed, etc.

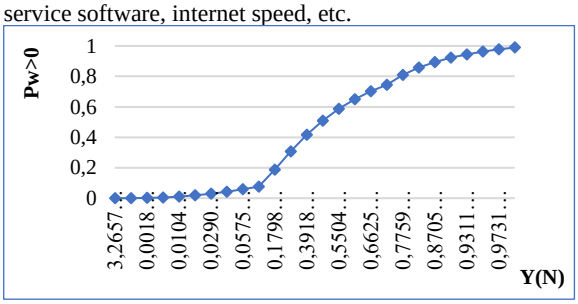

Figure 4. Graph of dependence of the probability of losses on the falling load

It is well known that an unexpected increase in system demands leads the system to behave somewhat differently than typical operating operations. Such conditions were vividly witnessed during the COVID-19 epidemic. In the following section of our investigation, we will look at the status of the service system, which is similar to the situation that occurred during the pandemic, in which the number of downloads or requests into the system has increased drastically. This condition is depicted in figure 4, which depicts the graph of the probability of losses as a function of falling load. The graph shows that as the load grows significantly, the likelihood of losses increases significantly as well, and the system's quality indicators significantly decline. For the reasons that the system's waiting rooms and service devices operate at maximum capacity due to a sudden spike in demand, as previously stated. Request queue times go longer as a result and the system's serviceability drops.

4. Conclusion

In summary, the mathematical model of the proposed user request service system is based on mass service theory and allows for the computation of the following system quality indicators:

The system's efficiency can be determined by implementing strategies to reduce the load and duration of service requests, calculating the likelihood of waiting for requests, calculating the likelihood of lost requests, and calculating the number of users based on quality measures.

Simultaneously, the results produced using the proposed model are completely consistent with the results of the experimental testing. This allows for the verification of the suggested mathematical model's correctness, as well as the identification and optimization of the quality indicators of the user request service system.

References

[1] Manchanda P. Analysis of optimisation techniques to improve user response time of web applications and their implementation for MOODLE // Advances in Information Technology: 6th International Conference, IAIT 2013, Bangkok, Thailand, December 12–13, 2013. Proceedings 6. Springer International Publishing, pp. 150-161, 2013.

[2] Özüdoğru G. Problems faced in distance education during the COVID-19 pandemic // Participatory Educational Research, Vol. 8, pp. 321-333, 2021.

[3] Abdullaev E., Zakirov V., Shukurov F. Assessment of the distance learning server's operation strategies and service capacity in advance //E3S Web of Conferences. – EDP Sciences, Vol. 420, pp. 06016, 2023.

[4] Barral H., Jaloyan G.A., Thomas-Brans F., Regnery M., Géraud-Stewart R., Heckmann T., Souvignet T., Naccache D. A forensic analysis of Google Home: Repairing compressed data without error correction //Forensic Science International: Digital Investigation, Vol. 42, pp. 301437, 2022.

[5] Schwarte A., Haase P., Hose K., Schenkel R., Schmidt M. Fedx: Optimisation techniques for federated query processing on linked data //The Semantic Web-ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23–27, 2011, Proceedings, Part I 10, pp. 601-616, Springer Berlin Heidelberg 2011.

[6] Lee M., Lee M., and Kim C.S., A JIT Compilation-Based Unified SQL Query Optimisation System, 6th International Conference on IT Convergence and Security (ICITCS), pp. 1-2, IEEE 2016.

[7] Vakhid Z., Eldor A., Farrukh S. System's load reduction by using asynchronous and synchronous service methods //Universum: technical science, Vol. 4-6 (109), pp. 65-70, 2023.

[8] Zakirov V., Abdullaev E., Determining the efficiency of service quality in the open loss and waiting methods of single-channel synchronous systems // Current issues in the development of innovative information technologies in transport, Vol. 1, No. 2, pp. 22–33, 2022.

[9] Vora M. N., Shah D., Estimating effective web server response time, 2017 Second International Conference on Information Systems Engineering (ICISE), pp. 37–44, IEEE 2017.

[10] Kurbanov F., Yaronova N.V., Kodirova L.A., "Remote Control and Monitoring of the Unguarded Railway Crossing System," 2023 International Russian Automation Conference (RusAutoCon), Sochi, Russian Federation, pp. 993–997, 2023. doi: 10.1109/RusAutoCon58002.2023.10272764.

[11] Khazaei, H., Misic, J., Misic, V.B. Performance analysis of cloud computing centres using m/g/m+r queuing systems IEEE Transactions on Parallel and Distributed Systems, Vol. 23(5), 936–943, 2011.

[12] Youcef, S., Bhatti, M. U., Mokdad, L., Monfort, V. Simulation-based response-time analysis of composite Web services. 2006 IEEE International Multitopic Conference, pp. 349–354, IEEE 2006.

[13] Almeida, L., Pedreiras, P. Scheduling within temporal partitions: response-time analysis and server design. In Proceedings of the 4th ACM International Conference on Embedded Software, pp. 95–103, 2004.

[14] Chiew T.K., Renaud K. Estimating web page response time based on server access log //2015 9th Malaysian Software Engineering Conference (MySEC), pp. 140–144, IEEE 2015.

[15] Lozhkovsky A.G. Theory of Queuing in Telecommunications: A Textbook //Odessa: ONAS im. AS Popova, 2012.

[16] Sharma D. Response time-based balancing of load in web server clusters. 7th International Conference on Reliability, Infocom Technologies, and Optimisation (Trends and Future Directions) (ICRITO), pp. 471-476, 2018.

[17] Zhang X., Zhang J., Peng C., Wang X. Multimodal optimisation of edge server placement considering system response time //ACM Transactions on Sensor Networks, Vol. 19, pp. 1–20, 2022.

[18] Tong Z., Deng X., Mei J., Liu B., Li K. Response time and energy consumption co-offloading with the SLRTA algorithm in cloud-edge collaborative computing // Future Generation Computer Systems, Vol. 129, pp. 64–76, 2022.

[19] Huang C., Huang G., Liu W., Wang R., Xie M. A parallel joint optimised relay selection protocol for wake-up radio-enabled WSNs //Physical Communication, Vol. 47, pp. 101320, 2021.

[20] Bocchi E., De Cicco L., Rossi D. Measuring the quality of experience of web users //ACM SIGCOMM Computer Communication Review, Vol. 46, pp. 8–13, 2016.

[21] Zhong H., Fang Y., Cui J. Reprint of "LBBSRT: An efficient SDN load balancing scheme based on server response time" Future Generation Computer Systems, Vol. 80, pp. 409–416, 2018.

[22] Cao K., Li L., Cui Y., Wei T., Hu S. Exploring placement of heterogeneous edge servers for response time minimization in mobile edge-cloud computing //IEEE Transactions on Industrial Informatics, Vol. 17, pp. 494-503, 2020.

[23] Tochukwu N.J., Mary O.E.C. Performance Evaluation of Web Servers Using Response Time and Bandwidth //Performance Evaluation, Vol. 9. pp. 133–138, 2020.

[24] Martinez J., Dasari D., Hamann A., Sañudo I., Bertogna M. Exact response time analysis of fixed priority systems based on sporadic servers, Journal of Systems Architecture, Vol. 110, pp. 101836, 2020.

[25] Ergüzen, A., Erdal, E., Ünver, M., Özcan. A. Improving the technological infrastructure of distance education through trustworthy platform-independent virtual software application pools //Applied Sciences, Vol. 11. – no. 3. – pp. 1214, 2021.

Information about the authors

Zakirov Vahid Maripovich / Vakhid Zakirov — Toshkent davlat transport universiteti “Radioelektron qurilmalar va tizimlar” kafedrasi professori v.b. t.f.n, E-mail: vakhidzakirov@mail.ru
Tel.:+ 998 97 780 03 21
https://orcid.org/0000-0002-2290-2625

Abdullayev Eldor Sa’dulla o‘g‘li / Eldor Abdullaev — Toshkent davlat transport universiteti “Radioelektron qurilmalar va tizimlar” kafedrasi doktaranti. E-mail: eldorabdullayev0223@gmail.com
Tel.:+99890 043 11 04
https://orcid.org/0000-0002-8954-9731