Journal of Mathematics and Computer Science 7 (2013) 23 – 32

Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Simulation and Optimization of Affective Causes on Quality of Electronic Services

Mohammadali Pirayesh Neghab

Iran University of Ferdowsi , *Faculty of Industrial Engineering*

pirayesh@ferdowsi.um.ac.ir

Shahrzad Mohsenian Heravi

Iran University of Ferdowsi , *Faculty of Industrial Engineering*

shmoseniyan@gmail.com

Mohsen Kahani

Iran University of Ferdowsi , Faculty of Computer Engineering

kahani@ferdowsi.um.ac.ir

Article history:
Received February 2013
Accepted March 2013
Available online April 2013

Abstract

By expanding bank services and increment in access requests for information resources, web servers perform weak about response time. So it is necessary to use simulation and mathematical models in order to analyzing complicated systems, optimizing and managing web server systems.

In this paper, services which are offered through internet are introduced and analyzed as a queue. Web servers are one of the main and effective parts in quality of internet services. Therefore, the operation of web servers is analyzed and optimized by concept of queue and simulation. One of the significant points in this paper is analyzing a problem taken from real world in internet field and also performing a new analysis from user's requests in web servers. The main purpose in this paper is to persuade users and answering theme as soon as possible. In this paper, after introducing the problem's structure, simulation and queue concept are used for analyzing, optimizing and managing web servers. Finally, a summery from results of simulation's calculation is declared.

Keywords: Simulation, Optimization, Queuing theory, Quality of service, Web server, Response time.

1. Introduction

Today, with the development of information technology and offering Internet services, network management issues and providing optimal service quality are considered as important issues. In addition, the Internet is a global network that millions of people throughout the world exchange of information via it, and this in turn will impact on web traffic and demand. Therefore, one of the most important issues in this area is providing quality service required for users, and appropriate performance in this field. Criteria for performance evaluation in this regard include the response time (waiting time in system), the efficiency of Web servers, the percentage of failed login of users and so on.
On the other hand, one of the most important and effective factors in profitable business is using information and communication technology tools, which leads the organizations to make some competitive advantage by creating web pages and providing electronic services. Thus, for the benefits of electronic services, the provider companies are required to provide a good level of service. Accordingly, one of the important topics in the field of electronic services is electronic service quality that today many studies are being conducted around [1].

Quality concept has different meanings according to the field it will be applied. Therefore, from the manufacturer perspective, it means the ability of products in terms of function, and from the customer perspective, it means to satisfy the demands of the customer. Electronic service quality has different dimensions such as accountability, user satisfaction, efficiency, innovation, security and support, and achievement and content quality [3]. In this paper, we focus on the dimensions of accountability and user satisfaction.

So to achieve this goal, we will seek solutions that improve the performance of server systems in order to reduce response time and increase user satisfaction. In general, these solutions are divided into software and hardware groups. The software solutions to improve the quality of Internet services include routing and scheduling algorithms in the allocation of the input stream to the web servers and load balancing in them, their expiration deadline, and so on, while the hardware solutions encompass adding a web server or its promotion, increased bandwidth, infrastructure design and network topology.

The purpose of this paper is to use software solutions, simulations and mathematical models for performance analysis of web servers in the scope of network management and optimization. To do so, we will determine the average waiting time for users in the system through queue concepts, and identifying the involved factors in this regard.

Then, after reviewing the literature, model structure and hypotheses of the web server will be discussed, then in Section 4, simulation model and verification of model by the existing queuing models will be explained, and finally, in Section 5, computational results will be presented.

2. Literature Review

In order to increase efficiency and improve system performance, there are many solutions in this respect which lead to reduced response time and increased user satisfaction.

In summary, previous research studies discussed in this regard can be divided into system simulation and comparison of scheduling algorithms ([4]-[6]), the quality of electronic services [7], analyzing and optimizing server model with queuing models ([8], [9]) and cancellation of the entry and service [10].

Kingman (1961), Koenigsberg (1966), Flatto (1977) and Halfin (1985) studied a model with implementation of queue models in two servers in parallel with unlimited capacity ([11]-[14]). Zheng and Zipkin (1990) analyzed timing model with two parallel queues with unlimited capacity in a web server [15].

On the other hand, on the Web servers, usually there is a limit in the capacity of queue. Considering these hypotheses, the results of the investigation can not be directly applied in this field. Down and

Lewis (2006) have considered parallel queues, with the aim of minimizing the average cost and by delegation to users in transferring the users in the queues [16].

Another important issue in this field is the phenomenon of cancellation of entry and service by users that has a variety of conditions. One of the first things in cancellation of service was introduced by Barrer (1957) [17]. He considered definite cancellation in the single service provider model by Markov entry and service rate while the customers were selected for offering service randomly. Other attempts in modeling of cancellation of service are conducted by Baccelli (1984), Artalejo (1995), El-Paoumy (2009) and Choudhury (2009) [18].

Also, the first research on cancellation of entry was performed by Haight (1957). He introduced a logic on the behavior of the person cancellation of the cancellation of entry. This may result in understanding of the importance of not joining to the certain queue length, or indifference to joining the queue with non-zero length. Furthermore, both cancellation of entry and service are discussed in the articles by people like Haghighe (1986), Zheng (2005), Paoumy (2008), Sherbiny (2008), has been presented [19].

3. Defining the problem

In this paper, we have considered a queue system in which unlimited amount of the users resourse to a website with different requests, and give their requests in the form of a certain number of requests to the system. Then, each user as soon as login, sends his/her request to the Web server a random interval, and will remain in the system until receiving the response to all requests from the web server, and in this situation, between two consecutive requests from a single user, other requests from other users will be placed in the queue of the Web server.

In this model, we define the system with a Web server (service provider) with system limited capacity for users. Thus, the user requests will be placed in the queue of web server requests in order, and will leave the queue with the FIFO order.

3.1. Hypotheses of the Model

In this model, the simulation model inputs (such as inter-arrival time, service providing time) are modeled as random variables with exponentially distribution. Also, by sending the user request we mean entering any request by the user which is confirmed by the Enter key.

In this research, a user requests are assumed independently and with the same distribution function of service.

3.2. Parameters of Model

In order to determine the model, notation and decision variables are defined as below:

λ: rate of users refers to the system;

$\bar{\lambda}$: rate of effective logins of the users to the system;

$\acute{\lambda}$: rate of requests users entry in to the system;

μ: rate of service requests;

P (i = K): the percentage of failed login of user;

P (B): percentage of being busy of the Web Server;

K: system capacity restriction for the number of users;

L: the average number of users in the system;

W: the average user waiting time in the system;

L_re: average number of requests in the system;

W_re: average waiting time for requests in the system;

LQ_re: average number of requests in the queue;
WQ_re: average waiting time for requests in the queue;
r: is a numerical constant and equal to the number of requests of a single user.

Given the conditions and laws governing the Web pages, the question would have different modes, which we will discuss in the following.

Mode 1: After the web server processes every request of the user, and gives the response, the next request of the given user with the rate of $\acute{\lambda}$ would be placed in the queue of web server. This mode is used in registration websites and so on in which the phases are designed step by step.

Mode 2: After the web server processes the first request of the user, and gives the response, the next request of the given user with the rate of $\acute{\lambda}$ would be placed in the queue of web server. This mode is used in entering the user name, and then sending other requests of the user.

Mode 3: it is the same as the mode 2, except in that the user would not wait to receive response for his/her first request.

4. Methodology and Analysis

In the situation of market competition, the main effort of the manufacturing and service institutions is to provide users' satisfaction and reduce response time in the system; Hence for the economic survival, it is essential that the mathematical models and simulation be used in optimization and system analysis. Thus, we use queuing theory in order to mathematically study the queue of users request in a web server, and analyze parameters such as the average number of users in the system, the average user response time, users failing rate and so on.

On the other hand, since the main purpose of the study of queuing systems is to review and understand the real environment and optimal management of the factors affect it, we optimize the desired system performance by determining the important involved factors affect the model. To do so, we determine the average waiting time for the user in the system through simulation models and identify factors that affect it.

4.1. Simulation Model

Due to the specific nature of the problem, none of the queuing models in queuing theory solved the problem; therefore, we use the simulation as one of the most useful tools for functional analysis of complex systems [2].

4.1.1. The parameters of simulation model

In the simulation model that we have programmed by Matlab R2010a software, by changing the input parameters (λ, λ', μ, r, K), Average system performance criteria (L, L_re, W, W_re, P (i = K), P (B)) can be calculated in a stable period.

Also, the results of the simulation showed that Little rules (Formula 1) is correct for requests and users.

$$L = (\text{RATE OF ENTER}) \,.\, W \qquad (1)$$

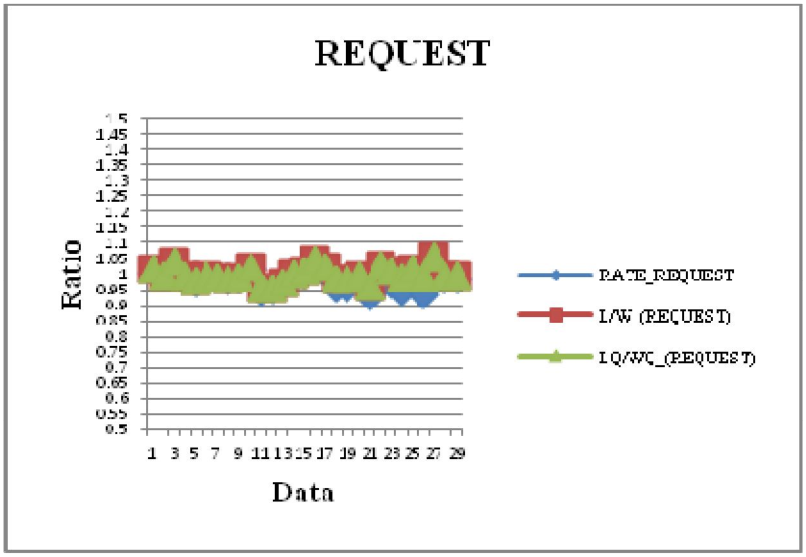

Figure 1. Little rule for requests

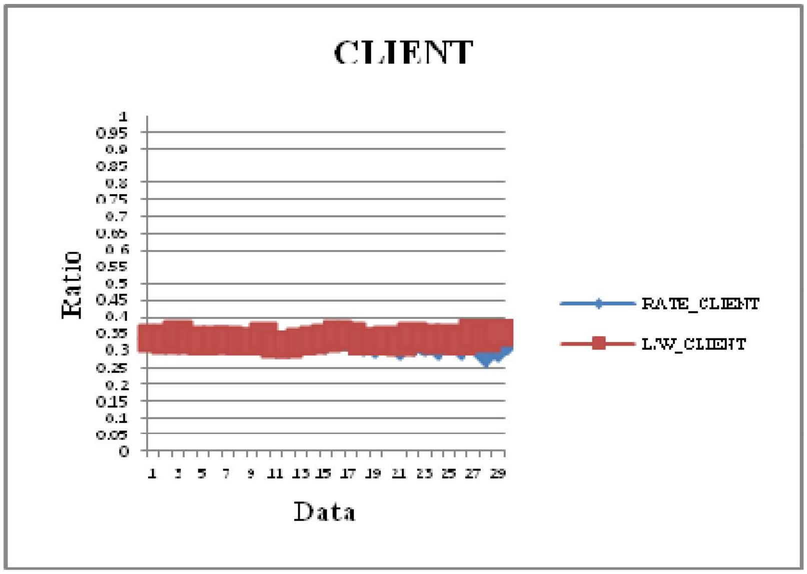

Figure 2. Little rule for users

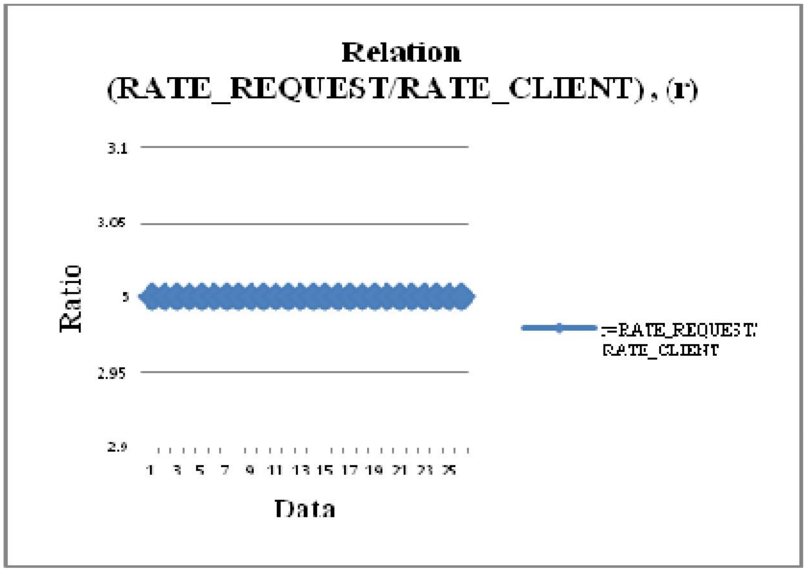

Figure 3. the relation between the effective rate of users and requests in the mode r=3.

4.1.2. Verification of the Simulation Model

In order to check the validity of the simulation model, considering some hypotheses to be used in existing queuing model, we analyze the model. Some of the results are shown in tables (1 and 2).

Confirmation Hypotheses in Simulation Model

For this purpose, in order to use M/M/1/rK queuing model to analyze the performance criteria of user requests, we consider system capacity equal to rK. Also, in order to use M/M/1/K queuing model to analyze the performance criteria of users, we consider service providing rate for each user equal to $\frac{\mu}{r}$, and regard request entry rate of each user a huge number ($\acute{\lambda} \ggg \lambda$), so that a given user's requests are placed in the queue of the Web server consecutively.

Table 1. Simulation model verification with M/M/1/rK queuing model and analysis of the performance criteria of requests in the mode r=3.

Data				model M/M/1/rK		model simulation	
K	λ	λ́	μ	L_re	W_re	L_re	W_re
50	2	5	0.5	149.889	299.778	148.493	311.299
50	2	5	20	0.333	0.067	0.569	0.095
50	2	5	8	1.667	0.333	5.248	0.847
90	2	5	1	269.750	269.750	268.128	265.704
20	2	5	1	59.750	59.750	58.318	60.277
40	2	5	1	119.750	119.750	118.272	116.531
50	2	12	1	149.909	149.909	148.315	141.955
50	2	5	1	149.750	149.750	148.123	144.493
50	2	8	1	149.857	149.857	148.203	146.084

Table 2. Simulation model verification with M/M/1/K queuing model and analysis of the performance criteria of users in the mode r=3.

Data				model M/M/1/K		model simulation	
K	λ	λ́	μ	L	W	L	W
50	20	2000	1	49.983	149.949	48.437	152.927
50	12	800	1	49.971	149.914	49.932	141.913
50	8	800	1	49.957	149.870	49.907	143.873
50	2	800	20	0.429	0.214	0.394	0.197
50	2	700	8	3.000	1.500	2.494	1.223
90	2	800	1	89.800	269.400	89.781	280.723
20	2	500	1	19.800	59.400	19.820	61.968
40	2	500	1	39.800	119.400	39.807	120.613
50	2	500	1	49.8	149.4	49.7972	142.2621

5. Computational results

In comparison between the modes according to the diagrams (7 and 8), the mode 3 is the lowest in the system response time, and so in the parametric analysis, we will consider this type.

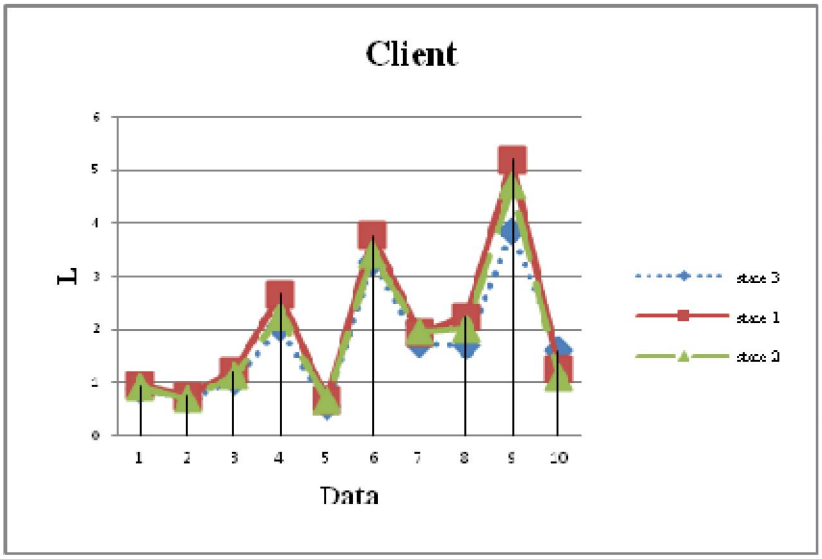

Figure 7. Comparison between the modes and analyzing the average number of user.

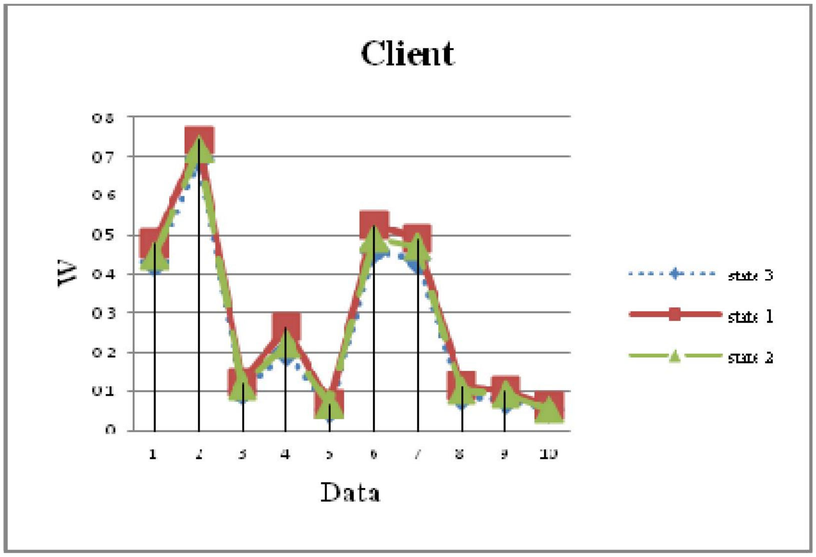

Figure 8. Comparison between the modes and analyzing the average time of the user.

Table (3), indicates that by lowering the systems restrictions, more users can access the system, which in turn will affect their expected waiting time on the system. Therefore, the network administrator makes a balance between the percentage of failed logins of the users and waiting time in the system.

Table 3. Computational results obtained from the modified K

$\lambda = 2\ \acute{\lambda} = 5\ \ \mu = 7\ \ r = 3$			
K	L	W	P(i = K)
2	1.050	0.775	0.328
5	2.273	1.236	0.096
7	2.705	1.445	0.047
10	3.356	1.731	0.016
20	4.239	2.128	0.002
100	4.381	2.179	0

Also, the results of Table 4 suggest that by increasing the service providing rate and speed of service providing, the average number of users and the waiting time in the system will decrease, and subsequently, the percentage of failed logins of the users and the percentage for the web server to be busy will reduce.

Table 4. Computational results obtained from the modified µ

$\lambda = 2$ $\acute{\lambda} = 5$ $K = 10$ $r = 3$				
µ	L	W	P(i = K)	P(B)
5	7.121	4.366	0.184	0.980
7	3.433	1.773	0.022	0.838
10	1.718	0.851	0.001	0.600
20	1.001	0.499	0	0.278
45	0.855	0.433	0	0.133

6. Conclusion

In this paper, the performance of Web servers as an effective component of quality of Web services have been analyzed by using queuing theory concepts and simulation principles. The results of simulation showed the effect of speed and capacity of the web server on the average waiting time of the users. As a result, Web site administrators can use the results of this research to make appropriate decisions to improve the quality of their service.

In comparison between the values obtained in putting some hypotheses in the simulation model, and results of the analysis of existing queuing models, the amounts of numbers and waiting time of the request and user did not show significant differences. Therefore, he results of the simulation model will be approved.

Further Suggestions

The following researches can be conducted in this respect:

- Entering the costs in decision-making models;
- Analyzing the effect of scheduling algorithms on how users enter their requests;
- Generalizing the model for multiple web servers;
- Using a service cancellation (Time Out) as a constant number;

REFERENCES

[1] Sh. Zahedi, J. Bibiyaz, Analyzing the Quality of Electronic Services in Raja Pa‹ssenger Trains Company, Information Technology Management journal, vol. 1, pp. 65-82 **(2009)**.

[2] M. A. Pirayesh Neghab, Sh. Mohsenian Heravi, Analyzing Internet-based Services Using Queuing Theory and Simulation Concepts, International Conference on Operation Research, Tabriz, Spring **(2012)**.

[3] Kerin, Hartly, Berkowitz and Rudelious, Marketing, USA: McGraw-Hill Irwin, pp. 25-55 **(2006)**.

[4] Y.M. Teo, R. Ayani, Comparison of load balancing strategies on cluster-based web servers, The International Journal of the Society for Modeling and Simulation, vol. 77, no. 6, pp. 185–195 **(2001)**.

[5] D.R.W. Holton, M. Younas, I.U. Awan, Priority scheduling of requests to web portals, The Journal of Systems and Software, vol. 84, pp. 1373–1378 **(2011)**.

[6] A. Wierman, Fairness and scheduling in single server queues, Operations Research and Management Science, vol. 16, pp. 39-48 **(2011)**.

[7] N. Ye, E.S. Gel, X. Li, T. Farley, Y.C. Lai, Web server QoS models: applying scheduling rules from production planning, Computers & Operations Research, vol. 32, pp. 1147– 1164 **(2005)**.

[8] Z. Zhang, W. Fan, Web server load balancing: A queueing analysis, European Journal of Operational Research, vol. 186, pp. 681–693 **(2008)**.

[9] N.Gautam, Performance analysis and optimization of web proxy servers and mirror sites, European Journal of Operational Research, vol. 142, pp. 396-418 **(2002)**.

[10] A. Choudhury, P. Medhi, A Simple Analysis Of Customers Impatience In Multiserver Queues, Journal Of Applied Quantitative Methods, vol. 5, no. 2, pp. 182-197 **(2010)**.

[11] J.F.C. Kingman, Two similar queues in parallel, Annals of Mathematical Statistics, vol. 32, no. 4, pp. 1314–1323 **(1961)**.

[12] E. Koenigsberg, On jockeying in queues, Management Science, vol. 12, no. 5, pp. 412–436 **(1966)**.

[13] L. Flatto, H.P. McKean, Two queues in parallel, Communications of Pure and Applied Mathematics, vol. 30, pp. 255–263 **(1977)**.

[14] S. Halfin, The shortest queue problem, Journal of Applied Probability, vol. 22, pp. 865–878 **(1985)**.

[15] Y.S. Zheng, P. Zipkin, A queueing model to analyze the value of centralized inventory information, Operations Research, vol. 38, no. 2, pp. 296–307 **(1990)**.

[16] D.G. Down, M.E. Lewis, Dynamic load balancing in parallel queueing systems: Stability and optimal control, European Journal of Operational Research, vol. 168, no. 2, pp. 509–519 **(2006)**.

[17] D.Y. Barrer, Queuing with impatient customers and ordered Service, Operations Research, vol. 5, vol. 5, pp. 650-656 **(1957)**.

[18] A. Choudhury, A few words on Reneging in M/M/1/K queues, Contributions to Applied and Mathematical Statistics, vol. 4, pp. 58-64 **(2004)**.

[19] A.M. Hagighi, J. Medhi, S. G. Mohanty, On a multi server Markovian queuing system with Balking and Reneging, Computer and Operational Research, vol. 13, no. 4, pp. 421-425 **(1986)**.