International Journal of Software Engineering and Its Applications
Vol.13, No.2 (2019), pp.25-30
http://dx.doi.org/10.21742/ijseia.2019.13.2.04

Analyzing the Performance of M/G/1 and M/Er/1 Queuing Models on Data Centers

N. Thirupathi Rao[1] and Debnath Bhattacharyya[2]

[1]*Department of Computer Science & Engineering Vignan's Institute of Information Technology Visakhapatnam, AP, India*
[1]*nakkathiru@gmail.com*

Abstract

Cloud computing is the process of allocating the network access admission to a group of selected users having advanced and smart pattern of computing facilities on the plan of usefulness of the network permission for accessing the network resources whenever there's a demand for the facility to be provided from the cloud. It may be a customary term and thus the regular service that was delivering the required services to the hosts among net. Here, the cloud computing mechanism is employed for describing each the list of platforms that were out there to the users for operating and additionally the many styles of applications which will be processed. The current technique was being thought about by most of the analyzers because of the most potential and therefore the most helpful space for the analysis and also for analysis in academe like universities and major research laboratories. Solely few notable works are revealed with regards to performance analysis in cloud computing. Generally the analytical models were geared toward coming up with the models that use the cloud and its services through that the performance of the model was analyzed and evaluated below numerous configurations and assumptions. These assumptions were based on the queuing theory and its accuracy is verified with numerical calculations and simulations. Present paper deals with the performance evaluation in-terms of steady-state parameters of a small cloud server farm using single and multi-server queuing models. Single server model includes M/G/1 and M/Er/1.

Keywords: *Cloud computing, Queuing models, Performance parameters, Traffic intensity*

1. Introduction

Presently it is seen that new trends related to computer technology emerge on a daily basis and one of those new trends is cloud computing, which is anticipated to bring an enormous change in the way one uses computers and internet. A total software and its related applications required environment was being provided under cloud computing [1][2][3]. Primary uses of cloud computing is the cloud service in terms of data storage and web applications. Cloud service developments tools by Amazon, Google app engine and IBM etc. are well accepted and utilized.

An important component of cloud computing is Infrastructure-as-a-service (*IaaS*), which is the ability to remotely access computing resources. The remote access of the services provided by the cloud computing environment was the access to the network, services related to the routing and the storage-related issues and their applications. The major work and the

Article history:
Received (August 13, 2019), Review Result (September 24, 2019), Accepted (November 15, 2019)

Print ISSN: 1738-9984, eISSN: 2208-9802 IJSEIA

application of the IaaS provider will be in supplying the basic services and the services related to the hardware and other services related to the administrative services which were required to accumulate the several list of applications and a stage for the management of several set of applications which requires the services from the cloud and its related areas [4][5]. Typical examples of *IaaS* include computer cycles, servers, storage, network and backup etc. One of the advantages of *IaaS* is that one can access very expensive data center resources through a rental means. The most and the very important point to be considered was that cloud computing can be taken as the important aspect for both the providers of the cloud and the customers of the cloud computing.

The current research paper evaluates the performance parameters of cloud-based data centers based on queuing theory models for both single server and multi-server models. The steady-state performance parameter formulations identified are programmed in MATLAB® environment. The present models selected for evaluation for single servers include *M/M/1*, *M/G/1*, *M/Er/1*. Interarrival rates for all the above models have exponential distribution. Service rates have a wider range of distributions including exponential, generalized, deterministic and Erlang type.

2. Queuing system - overview

Queuing theory is the subject of mathematics and its applied areas like the applied mathematics and the subject of statistics. It mainly deals with the waiting lines [6]. It is tremendously functional in predicting, identifying and evaluating the performance of the system. Operations research is one of the applicable and mostly used subject and subject strategy of the queuing systems. Customary queuing theory problems that were being observed by several users and the researchers are the customers going and visiting a store, corresponding to requirements incoming at a device. Queuing theory provide extended period of regular values. It is not possible or being considered in the queuing system procedure to identify or observe the occurrence of the whether the next event will occur or not occur. In queuing models or the queuing theory the arrival times to be considered as the random and similarly the service times are also random in nature. A queuing system [Figure 1] can be mentioned as “the customers resolve to appear for a specified check, wait if the check cannot begin right away and go away following being offered” The term “customer” can be men, products, machines etc.

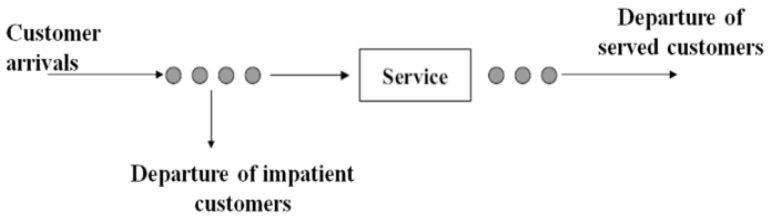

Figure 1. Queuing model

The characterization of the queuing model or the queuing system was very useful and very important for processing the several systems or several applications using these queuing systems [7][8]These systems can be characterized with several features or the factors like arrival processes of the customers, the time taken for providing the service, the discipline of the service, the capacity of the service and the number of service stages involved in finalizing

the completion of the service. The following is a standard notation system (Kendall's notation) of queuing systems T/X/C/*K*/P/Z with:

T: probability distribution of inter-arrival times

X: probability distribution of service times

C: Number of servers

K: Capacity of the Queue

P: Size of the population

Z: Discipline of the service

Arrivals might initiate as of single or several sources referred to as the population those were being called. The population that was being called might be either limited or 'unlimited'. The arrival process of the system comprises of explaining how the clients or the users turn up to the system which was denoted by λ (inter-arrival rate) [9]. The service mechanism of a queuing system is precise in terms of number of servers (denoted by C) in which each server comprises of its possess line or an ordinary line and the probability sharing of clients check moment denoted by $1/\mu$. The discipline of a queuing system explains in detail the process of whether a system follows a rule or a regulation to a server that how the server identifies or selects the next customer or the next item from the existing queue or the queue under process from which the server completes the task given by a customer or a user.

3. Queuing models & formulations

Queuing models [Table 2][Table 3] are very much helpful to the users in predicting or identifying the performance of the service systems whenever there is a chance of existence of uncertainty in arrivals and service times to the system [10][11]. The simplest possible (single-stage, [Figure 1] queuing systems have the following components: customers, servers, and a waiting area (queue). An arriving customer is placed in the queue until a server is available. To model such a system, we need to specify the characteristics of the arrival and service process; how (in what order) waiting customers are dispatched to available servers. For the present work it is assumed that clients are offered the services in which order they arrive in the system (First-Come-First-Served or *FCFS*). The mean value approach is used to determine mean performance measures, *LS* and *WS* directly by using Little's queuing formula and PASTA property.

Table 1. System parameters and performance measures

S.No	Description	Symbol	Inputs
1	Arrival Rate in middle	λ	0-4
2	Service Rates	μ	1,3
3	Intensity of the Traffic	€	0,1
4	Servers in number	C	1,2
5	Customers number in size	K	4
6	Capacity of the Buffers	R	4
7	Erlang Parameter	er	2
8	Mean number of customers in a model	Ls	-
9	Total number of customers in queue	Lq	-
10	Waiting time of the customers in the queue	Wc	

Table 2. Formulations- single server models

Queuing Model	*Ls*	*Lq*	*Ws*	*Wq*
M/G/1	$P + \frac{AP^2}{1-P}$	$\frac{AP^2}{1-P}$	$\frac{1}{\mu} + \frac{AP}{\mu(1-P)}$	$\frac{AP}{\mu(1-P)}$
M/Er/1	$P + \frac{(1+\frac{1}{er})P^2}{2(1-P)}$	$P + \frac{(1+\frac{1}{er})P^2}{2(1-P)}$	$\frac{1}{\mu} + \frac{\left(1+\frac{1}{er}\right)P}{2\mu(1-P)}$	$\frac{1}{\mu} + \frac{\left(1+\frac{1}{er}\right)P}{2\mu(1-P)}$

Nomenclature for system parameters, chosen input values (S. No.1-8) and performance measures (S. No.9-12) for the queuing models considered are highlighted at [Table 1] and throughout it is assumed that the system is in "steady-state", i.e., it has operated for a long time with the same values for all the parameters. The various formulations for the chosen queuing models are given at Table II.

4. Results and discussion

Evaluation of performance parameters was carried out by programming in MATLAB® 7.60 (R2008a) environment developed by MathWorks, Inc., USA. The input to the program is according to Table 1 and the output results for performance parameters are given at [Figure 2] and [Figure 3].

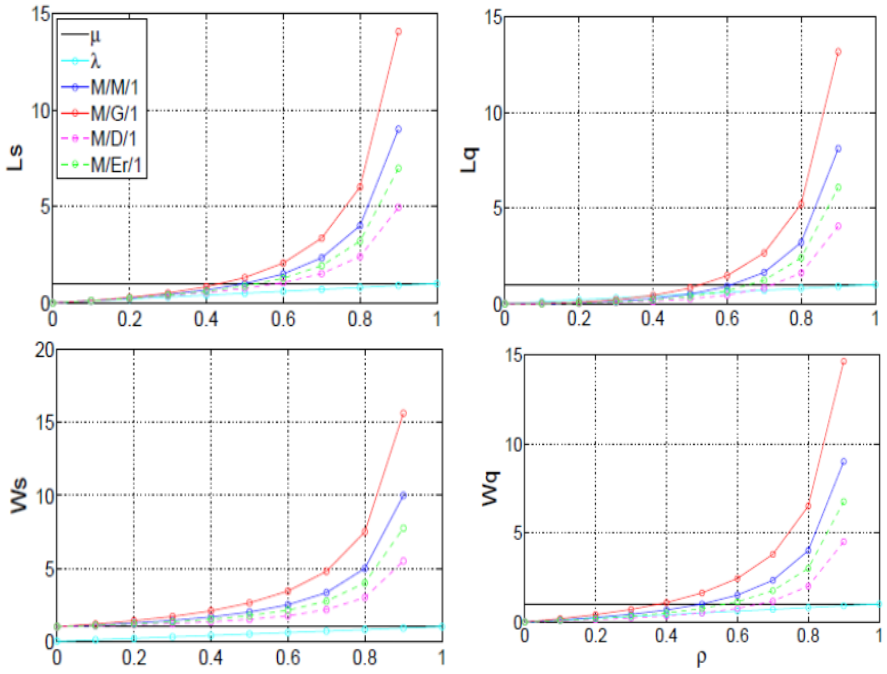

Figure 2. Single server performance parameters, μ=4

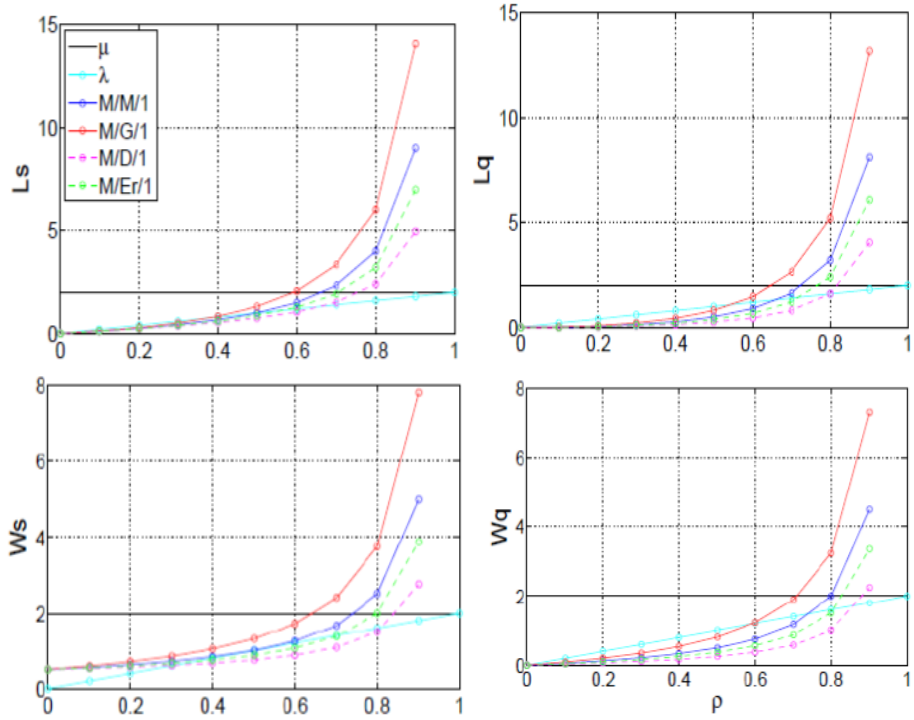

Figure 3. Single server performance parameters, μ=8

Performance evaluation for single servers indicate that as the service rate (μ) increases for a constant range of traffic intensity (ρ) only waiting times of customers in the system (W_S) and queue (W_Q) decreases, where as the length of customers in system (L_S) and queue (L_Q) remain unchanged as it is independent on μ. For the same input parameters *M/G/1* model shows optimum performance in terms of queue lengths and waiting times followed by *M/Er/1*. Performance of *M/G/1* shows detrimental nature when compared with other queuing models, which is attributed to higher value of *CoV*. For higher order Erlang parameters, *M/G/1* and *M/Er/1* models behave in close comparison.

5. Conclusions

Performance evaluation of small cloud computing data center is discussed with the theory based on queuing systems. Single server and multiple server models are presented along with their formulations for performance parameters. MATLAB programming/code generation and implementation for performance evaluation of cloud computing data server farm is accomplished. Comparisons among various models are attempted and relevant observations are highlighted.

References

[1] K. Saravanan, et. al., "Performance factors of cloud computing data centers using [(M/G/1): (inf/GD model)] queuing systems," International Journal of Grid Computing & Applications, vol.4, pp.12-18, **(2013)**

[2] K. Jayapriya, et. al., "An extensive survey on QoS in cloud computing," International Journal of Computer Science and Information Technologies, vol.5, pp.15-25, **(2014)**

[3] Ivo Adan and Jacques Resing, "Queuing systems," Eindhoven University of Technology, The Netherlands, vol.10, pp.1-12, March, **(2015)**

[4] Dr. Janos Sztrik, et. al., "Basic queuing theory," University of Debrecen, Faculty of Informatics, **(2012)**.

[5] Chandrakala, et. al., "Survey on models to investigate data center performance and QoS in cloud computing infrastructure," First International Conference on Recent Advances in Science & Engineering, pp.36-49, **(2014)**

[6] Samantha Molinaro, et. al., "Comparing expected wait times of an M/M/1 queue," Department of Mathematics and Statistics, University of Winsor, pp.1-12, **(2010)**

[7] Reza Ravanmehr, et. al., "Cloud computing performance evaluation: Issues and challenges," International Journal on Cloud Computing Services and Architecture, vol.3, pp.18-27, **(2013)** DOI: 10.5121/ijccsa.2013.3503

[8] Tom V. Mathew, "Queuing analysis, transportation systems engineering," Indian Institute of Technology, Bombay, pp.56-64, Feb, **(2014)**

[9] D. Praveen and K. Satish, "The queuing theory in cloud computing to reduce the waiting time," IJCSET, vol.1, pp.25-31, **(2011)**

[10] Hongyun Wang, et. al., "A conditional probability approach to M/G/1 - like queues," Performance evaluation, vol.65, **(2008)**

[11] Arnika Tripathi, "Simulation of queuing models," International Journal of Engineering Science and Innovative Technology, vol.2, pp.25-32, **(2013)**