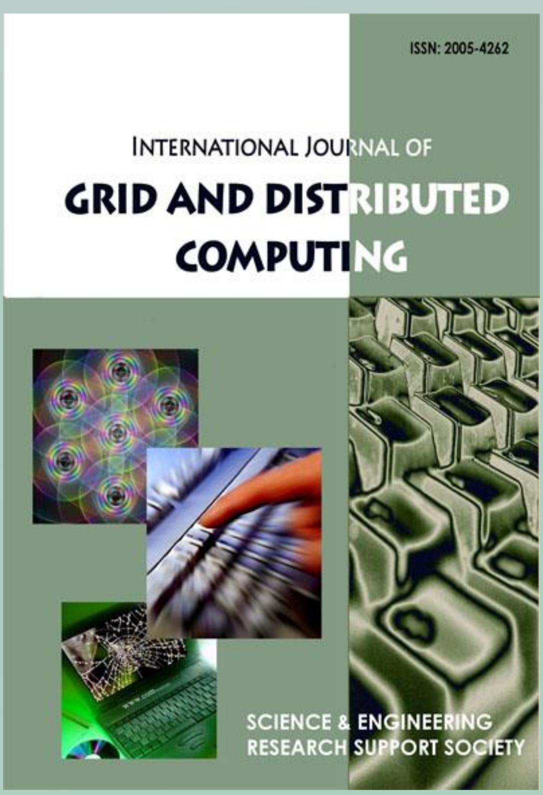
Performance Evaluation in Cloud Computing Model using Queuing Models

Amreen, Pilla Srinivas, Nakka Thirupathi Rao, Debnath Bhattacharyya and Hye-jin Kim



ISSN 2005-4262
Volume 10, Number 3, 2017

International Journal of Grid and Distributed Computing
Vol. 10, No. 3 (2017), pp.15-24
http://dx.doi.org/10.14257/ijgdc.2017.10.3.02

International Journal of Grid and Distributed Computing
Vol. 10, No. 3 (2017), pp.15-24
http://dx.doi.org/10.14257/ijgdc.2017.10.3.02

Performance Evaluation in Cloud Computing Model using Queuing Models

Amreen[1], Pilla Srinivas[1], Nakka Thirupathi Rao[1], Debnath Bhattacharyya[1], Hye-jin Kim[2]

Department of Computer Science and Engineering,
Vignan's Institute of Information Technology
Visakhapatnam-530049, India
{mallikamreen,srinivasp3,nakkathiru,debnathb}@gmail.com
[2]Sungshin Women's University,
2, Bomun-ro 34da-gil,
Seongbuk-gu, Seoul, Korea
hyejinaa@daum.net
(Corresponding Author)

Abstract

Cloud computing is the process of enabling the network access to a set of selected users with good configuration of computing resources on the basis of availability of the network access whenever there is an demand for the services from the cloud. Cloud computing is a regular term and the regular service that was being delivering the required services to the hosts in the internet. Here, the cloud computing mechanism is used for describing both the list of platforms that were available to the users for working and also the several types of applications that can be processed. The present technique was being considered by most of the researchers and the research institutes as the most potential and the most useful area for the research and also for research in academia like universities and major research laboratories. Performance evaluation of several applications the related applications and their sub parts were being considered as one of the useful and mostly used research area in the recent years. This technique and its services were being used mostly for the providers of the cloud and its related areas and the beneficiaries of this technique of area were both the providers of the cloud and the customers related to the cloud. Only few notable works have been published with regards to performance evaluation in cloud computing. In general the analytical models were aimed at designing the models which use the cloud and its services through which the performance of the model was analyzed and evaluated under various 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 models include M/M/1, M/G/1, M/D/1 and M/Er/1. Multi-server model considered include M/M/c, M/M/c/c, M/M/c/K and M/M/c/c+r. A comparison among the steady state parameters evaluated for the above queuing models with respect to traffic intensity is also presented.

***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 computer and internet. A total software and its related applications required environment was being provided under cloud computing. 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 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. 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 the cloud computing can be taken as the important aspect for both the providers of the cloud and the customers of the cloud computing.

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

2. Related Work

Extensive survey on cloud computing was highlighted by N.Ani Brown Mary et al. [2] where a widespread survey on quality of service with respect to their implementation details, strong points and limitations were presented. William Stallings [14] provided a practical guide to queuing analysis and also reviewed some elementary concepts in probability and statistics. Presentation points of cloud computing information stations using *[(M/G/1): (inf/GD model)*] queuing system were arrived in [1] particularly for mean number of tasks and waiting time in the system.

Queuing models are very useful in the working and processing certain models and the data related to several applications like the intend of manufacture, shipping, transport and stocking systems in terms of capacities and control. Fundamental concepts in queuing theory are well documented in [7]. Both finite source and infinite source queuing models are dealt in the above reference. Chandrakala et al. [8] surveyed on a variety of mechanism and models those are used for studying and identifying the data center and its requirements and the performance, evaluation for providing the quality of service in IaaS cloud computing and it's related using systems. Transient probabilities of *M/M/1* queue were calculated by Myron Hlynka et al [9] based on task time, number of customers and traffic intensity. Niloofar Khanghahi et al. [10] brought out an on the whole standpoint on cloud estimation criteria and tinted it with assist of simulation.

Tom V. Mathew [11] proposed and analyzed a set of new queuing patterns which could be encountered while working on these models and the classification part of these patterns was also discussed. T.Sai Sowjanya et al [12] verified and proved that *M/M/c* model for two servers which increase the performance of the network model over using one server by tumbling the length of the queue and waiting time. Alenandre Brandwajn et al [13] measured an *M/G/1*-like system in which the service time distribution was being presented by a coxian series of several memory less stages. A new and novel approach based on conditional probabilities is used to obtain solution of such systems.

An exhibition of academic outcome for the *M/M/c/c+r* queuing model with eager clients is offered by Hideaki Takagi et al [15] with mathematical design as a fundamental form of call centers including derivations of combined sharing for the waiting time and probability of customer service and blocking. Ward Whitt [20] studied the *M/M/c* queue model with client desertion which was also treated as the Erlang-A model, which was sovereign and identically dispersed client discard period with an exponential distribution. Five statistical software packages for queuing theory were compared by Chinthanie Fernando Ramasundarahettige [22].

3. 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. 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 (Fig.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.

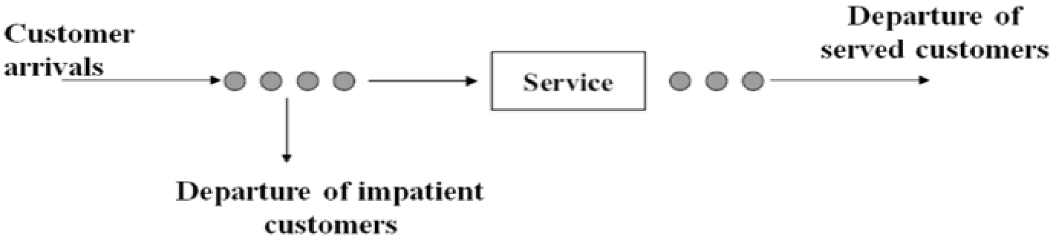

Customer arrivals
Service
Departure of served customers
Departure of impatient customers

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. 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). The service mechanism of a queuing system is precise in terms of number of servers (denoted by C) in which the each server comprises of its possess line or a ordinary line and the probability sharing of clients check moment denoted by $1/\mu$. Here ' T ' and ' X ' can take the following values:

M: Markovian (i.e. exponential)
G: General distribution
D: Deterministic
Er: Erlang distribution

Queue capacity (*K*) also denotes the loss of customers if queue is full. The size of the population (P) can be either finite or infinite. Service discipline (Z) basically can take the following values:

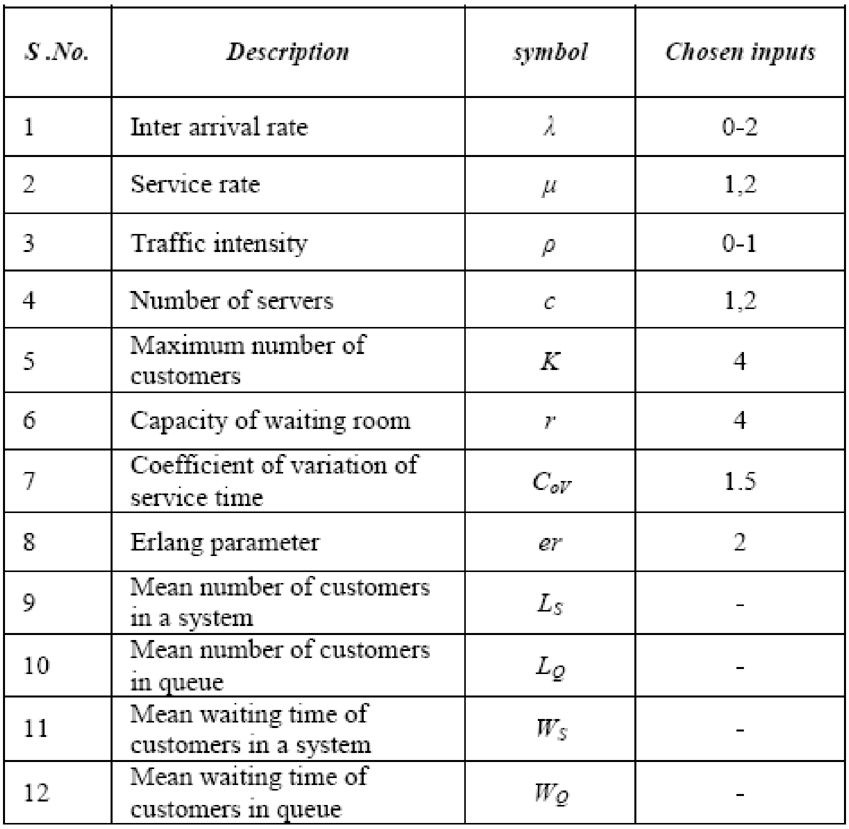
FCFS or FIFO : First Come First Served
LCFS or LIFO : Last Come First Served
RANDOM : service in random order
GD : General Discipline

The discipline of a queuing system explains in detain 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 an user.

4. Queuing Models & Formulations

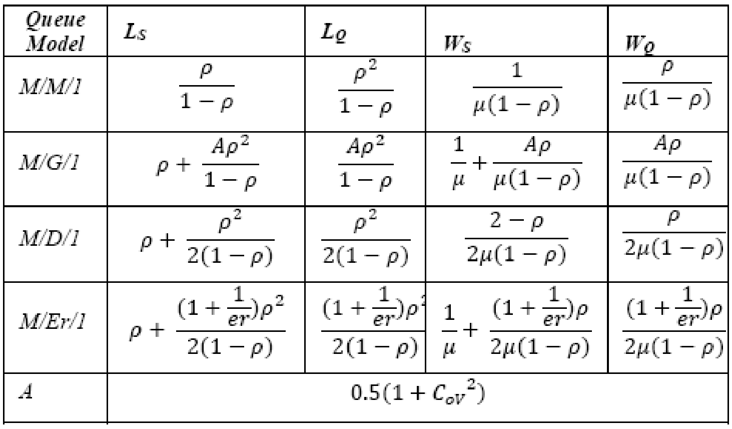
Queuing models (Table II & III) are very much helpful to the users in predicting or identifying the performance of the service systems whenever there is a chance of existence of uncertainty in arrivals and service times to the system. The simplest possible (single stage, Fig.1) queuing systems have the 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*). 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*	*Chosen inputs*
1	Inter arrival rate	λ	0-2
2	Service rate	μ	1,2
3	Traffic intensity	ρ	0-1
4	Number of servers	c	1,2
5	Maximum number of customers	K	4
6	Capacity of waiting room	r	4
7	Coefficient of variation of service time	C_{oV}	1.5
8	Erlang parameter	er	2
9	Mean number of customers in a system	L_S	-
10	Mean number of customers in queue	L_Q	-
11	Mean waiting time of customers in a system	W_S	-
12	Mean waiting time of customers in queue	W_Q	-

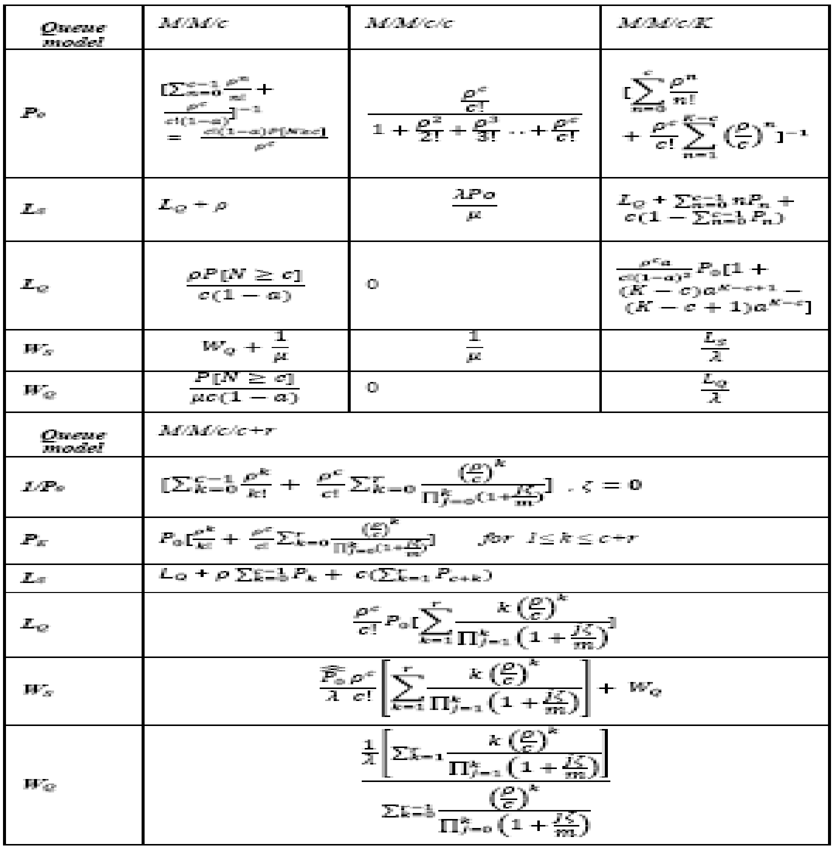
Table 2. Formulations- Single Server Models



Queue Model	L_S	L_Q	W_S	W_Q
M/M/1	$\frac{\rho}{1-\rho}$	$\frac{\rho^2}{1-\rho}$	$\frac{1}{\mu(1-\rho)}$	$\frac{\rho}{\mu(1-\rho)}$
M/G/1	$\rho + \frac{A\rho^2}{1-\rho}$	$\frac{A\rho^2}{1-\rho}$	$\frac{1}{\mu} + \frac{A\rho}{\mu(1-\rho)}$	$\frac{A\rho}{\mu(1-\rho)}$
M/D/1	$\rho + \frac{\rho^2}{2(1-\rho)}$	$\frac{\rho^2}{2(1-\rho)}$	$\frac{2-\rho}{2\mu(1-\rho)}$	$\frac{\rho}{2\mu(1-\rho)}$
M/Er/1	$\rho + \frac{(1+\frac{1}{er})\rho^2}{2(1-\rho)}$	$\frac{(1+\frac{1}{er})\rho^2}{2(1-\rho)}$	$\frac{1}{\mu} + \frac{(1+\frac{1}{er})\rho}{2\mu(1-\rho)}$	$\frac{(1+\frac{1}{er})\rho}{2\mu(1-\rho)}$
A	$0.5(1+{C_{oV}}^2)$			

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 I 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 and III.

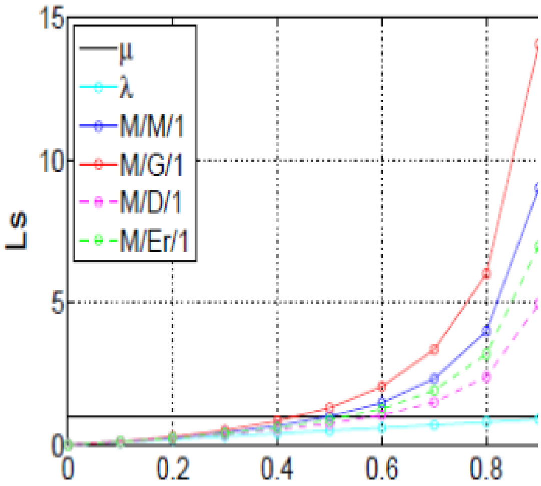
Table 3. Formulations- Multi Server Models

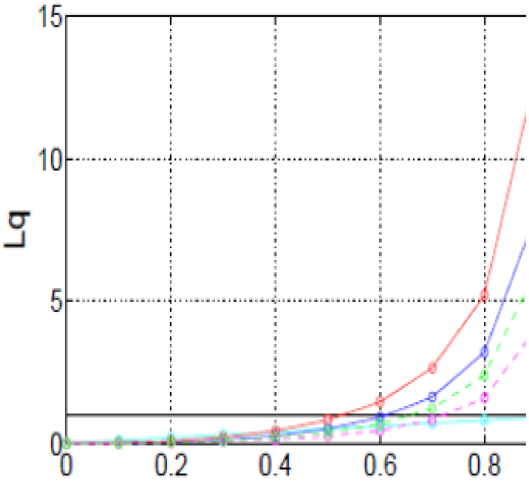


Queue model	M/M/c	M/M/c/c	M/M/c/K
P_o	$[\sum_{n=0}^{c-1}\frac{\rho^n}{n!}+\frac{\rho^c}{c!(1-a)}]^{-1} = \frac{c!(1-a)P[N\geq c]}{\rho^c}$	$\frac{\frac{\rho^c}{c!}}{1+\frac{\rho^2}{2!}+\frac{\rho^3}{3!}\cdots+\frac{\rho^c}{c!}}$	$[\sum_{n=0}^{c}\frac{\rho^n}{n!}+\frac{\rho^c}{c!}\sum_{n=1}^{K-c}(\frac{\rho}{c})^n]^{-1}$
L_s	$L_Q+\rho$	$\frac{\lambda Po}{\mu}$	$L_Q+\sum_{n=0}^{c-1}nP_n+c(1-\sum_{n=0}^{c-1}P_n)$
L_Q	$\frac{\rho P[N\geq c]}{c(1-a)}$	0	$\frac{\rho^c a}{c!(1-a)^2}P_o[1+(K-c)a^{K-c+1}-(K-c+1)a^{K-c}]$
W_s	$W_Q+\frac{1}{\mu}$	$\frac{1}{\mu}$	$\frac{L_s}{\lambda}$
W_Q	$\frac{P[N\geq c]}{\mu c(1-a)}$	0	$\frac{L_Q}{\lambda}$
Queue model	M/M/c/c+r		
$1/P_o$	$[\sum_{k=0}^{c-1}\frac{\rho^k}{k!}+\frac{\rho^c}{c!}\sum_{k=0}^{r}\frac{(\frac{\rho}{c})^k}{\prod_{j=0}^{k}(1+\frac{j\zeta}{m})}]$, $\zeta=0$		
P_K	$P_o[\frac{\rho^k}{k!}+\frac{\rho^c}{c!}\sum_{k=0}^{r}\frac{(\frac{\rho}{c})^k}{\prod_{j=0}^{k}(1+\frac{j\zeta}{m})}]$ for $1\leq k\leq c+r$		
L_s	$L_Q+\rho\sum_{k=0}^{c-1}P_k+c(\sum_{k=1}^{r}P_{c+k})$		
L_Q	$\frac{\rho^c}{c!}P_o[\sum_{k=1}^{r}\frac{k(\frac{\rho}{c})^k}{\prod_{j=1}^{k}(1+\frac{j\zeta}{m})}]$		
W_s	$\frac{\widehat{P_o}}{\lambda}\frac{\rho^c}{c!}\left[\sum_{k=1}^{r}\frac{k(\frac{\rho}{c})^k}{\prod_{j=1}^{k}(1+\frac{j\zeta}{m})}\right]+W_Q$		
W_Q	$\frac{\frac{1}{\lambda}\left[\sum_{k=1}^{r}\frac{k(\frac{\rho}{c})^k}{\prod_{j=1}^{k}(1+\frac{j\zeta}{m})}\right]}{\sum_{k=0}^{c-1}\frac{(\frac{\rho}{c})^k}{\prod_{j=0}^{k}(1+\frac{j\zeta}{m})}}$		

5. 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 I and the output results for performance parameters are given at Fig. 2-7.





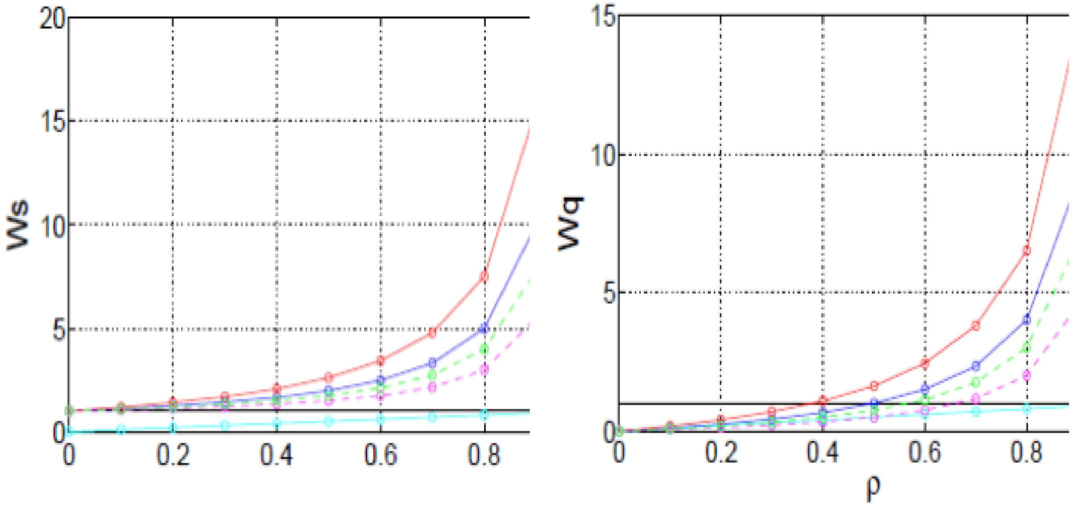


Figure 2. Single Server Performance Parameters, μ=1

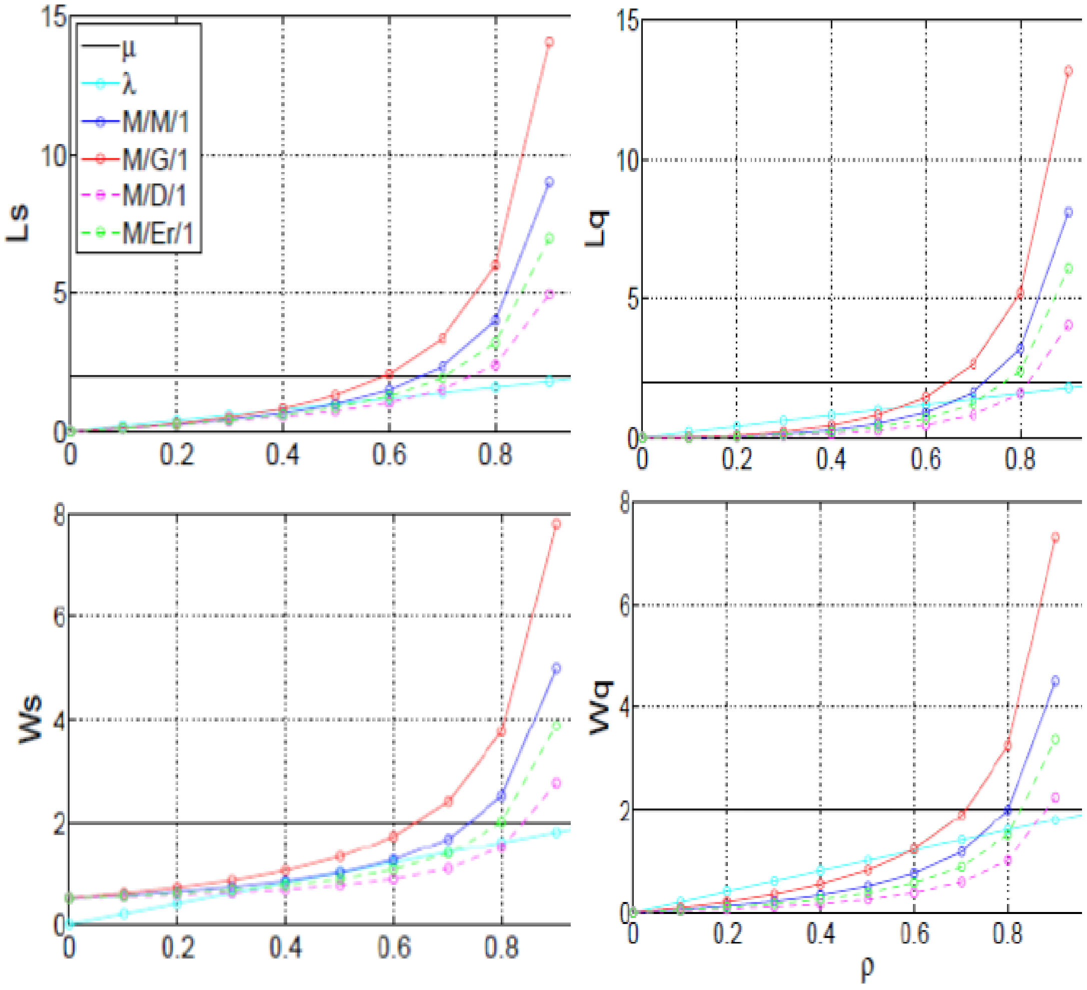


Figure 3. Single Server Performance Parameters, μ=2

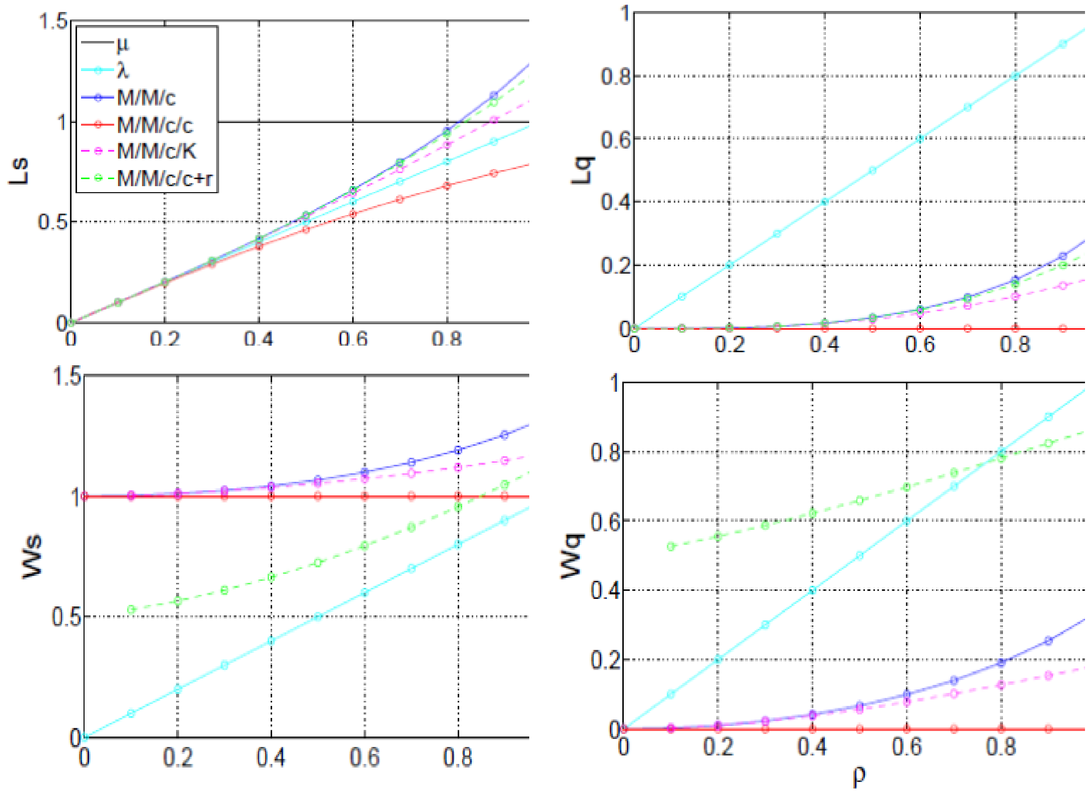


Figure 4. Multi Server Performance Parameters, μ=1

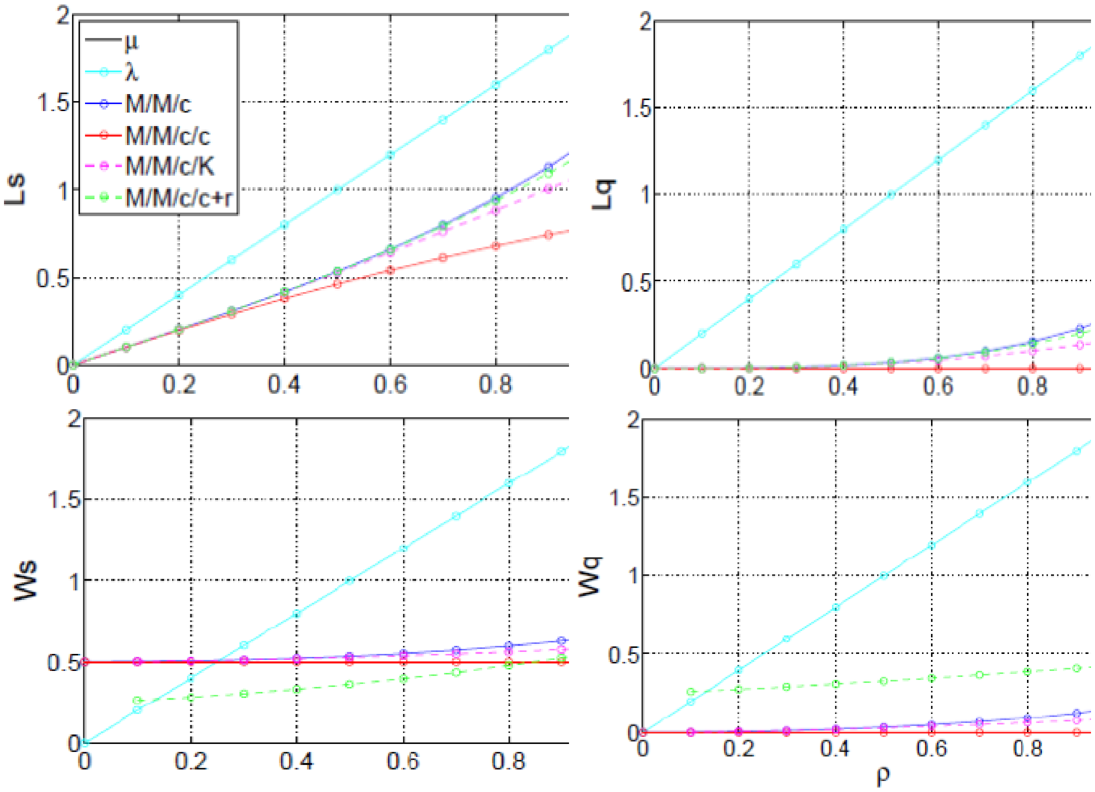


Figure 5. Multi Server Performance Parameters, μ=2

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/D/1* model shows optimum performance in terms of queue lengths and waiting times followed by *M/Er/1*, *M/M/1*. Performance of *M/G/1* shows detrimental nature when compared with other queueing 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.

Multiple server (c = 2) performance evaluation indicates that as the service rate (μ) increases for a constant range of traffic intensity (ρ) , similar nature as in the case of single servers is observed with regards to queue lengths and waiting times. For the same input parameters *M/M/c/c* possess least waiting times and customer lengths in queue, followed by *M/M/c/K* (K = 4) and *M/M/c*. Performance of *M/M/c/c+r* (r = 4) is detrimental in terms of waiting times in queue, where as *M/M/c* has higher customer queue lengths. Waiting times in queue for *M/M/c/c+r* can be reduced by reducing the waiting capacity of customers in queue. Also the performance with *M/M/c/K* can be improved in terms of waiting times in queue by decreasing the maximum number of customers allowed in the system. Multiple servers always behave better in satisfying a queue than single servers.

6. 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.

Acknowledgment

Authors would like to express sincere gratitude to the Department of Information Technology (IT), Vignan's Institute of Information Technology (VIIT) for providing an opportunity to take up the above research work and publish the content. The authors also thank numerous people for their valuable support during the tenure of the research activity.

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, **(2013)**.

[2] K.Jayapriya et. al., “An Extensive Survey on QoS in Cloud computing”, International Journal of Computer Science and Information Technologies, Vol.5, **(2014)**.

[3] Ivo Adan and Jacques Resing, Queuing systems, Eindhoven University of Technology , The Netherlands , 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, **(2014)**.

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

[7] Reza Ravanmehr et. al., “Cloud Computing Performance Evaluation: Issues and Challenges”, International Journal on Cloud Computing Services and Architecture, Vol.3, **(2013)**.

[8] Tom V. Mathew, Queuing Analysis, Transportation Systems Engineering, Indian Institute of Technology , Bombay, Feb 2014

[9] D.Praveen and K.Satish, “The Queuing Theory in Cloud Computing to Reduce the Waiting Time”, International Journal of Computer Science and Engineering Technology, Vol.1, **(2011)**.

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

[11] William Stallings, Queuing Analysis, **(2000)**.

[12] Harry Groenvelt, “A Note on Queuing Models”, **(1996)**.

[13] Arnika Tripathi, “Simulation of Queuing Models”, International Journal of Engineering Science and Innovative Technology, Vol.2, **(2013)**.

[14] QtsPlus4CalcRelease0.7, http://qtsplus4calc.sourceforge.

[15] Chinthanie Fernando Ramasundarahettige, “A comparison of Queuing Software packages”, Department of Mathematics and Statistics, University of Windsor, Canada.