International Journal of Grid and Distributed Computing
Vol. 10, No. 6 (2017), pp.55-70
http://dx.doi.org/10.14257/ijgdc.2017.10.6.05

Studies on Performance Analysis of Cloud Computing Based Data Centers with Queuing Models Using MATLAB

Nakka Thirupathi Rao[1], Pilla Srinivas[1], Ch. Rajkumar[1], Debnath Bhattacharyya[1] and Hye-jin Kim[2*]

[1]Department of Computer Science and Engineering,
Vignan's Institute of Information Technology,
Visakhapatnam-530049, AP, India
nakkathiru@gmail.com, debnathb@gmail.com
[2]Sungshin Women's University, 2, Bomun-ro 34da-gil,
Seongbuk-gu, Seoul, Korea
**hyejinaa@daum.net*

Abstract

Cloud computing was the technology developed to store the data and support the users with the access to the data stored by charging a minimal amount for the storage of data and for providing necessary steps for storing the data and fro providing security to the data that was stored. The content stored in various servers at various locations based on the type and size of the content. The content can be accessed to the users with valid registrations and a set of security verifications entered by the customers. The content that was hosted in the servers can also be used for hosting various applications and various other set of options of systems in various fields. It is one of the most famous and mostly used research areas in the recent years for further development in various set of applications and its usages related to several set of customers in the real time environment. Performance evaluation in cloud computing has been another major thrust area in the recent past, which is of crucial interest for both cloud providers and cloud customers. Only few notable works have been published with regards to performance evaluation in cloud computing. Generally analytical models established for assessing the working and the performance of cloud server farms can be studied under variety of configurations and assumptions are based on queuing theory and its accuracy is verified with numerical calculations and simulations. The problems at hand give rise to the task of evaluating the performance of data center with various queuing models to understand the distribution of the performance parameters with arrival and service rates, traffic intensity, number of servers and the associated probabilities. The goals of this thesis is to provide a framework through programs related to queuing models and evaluate the performance parameters, attempt validation, sensitivity analysis and make comparisons for data centers. Present thesis 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. Service rates have a wider range of distributions including exponential, generalize and Erlang type.

Keywords: *Cloud computing, queuing models, exponential distribution*

* Corresponding Author

ISSN: 2005-4262 IJGDC

1. Introduction

The cloud is where one can use technology when needed, as long as one needs it. The cloud can be both software and infrastructure service. In terms of maturity, software is much more evolved than hardware in the cloud. The cloud can be an application one can access through the web or a server. With cloud computing, the major advantage for users was that the programs that might be related to software programs related to several applications stored in the machine could not be executed in the local computers or laptops. All the programs related to the applications that were being connected to the cloud mechanism were being accessed by the internet to work. The internet connection was must in the field of cloud data processing and its related application areas. The major advantage for the customers and the users was the crash report. Whenever a crash occurs in the system or the system connected to the cloud, the data can be still available as the actual data was stored in the servers of the cloud not in the actual systems that were being connected to the cloud environment. The data will not be disturbed or any damage to the actual data as it was stored at various servers connected each other and located at various locations. The users might be customers from various numerous companies, numerous servers and their related applications and the numerous networks which could be used to connect all these servers and the machines. Here, the data was stored in various servers with various configurations and various operating systems and all these servers were located at various locations, so that the data can be secured even though a major damage or loss occurs for servers located at one place on the country or on earth.

1.1. Cloud Data Centers

Cloud computing was the technology and model which was aimed to provide the services to various set of customers by the model of paying some certain amount for utilizing certain set of operations and tasks from the system model. By using this model, the users can use the resources from various set of operations like networks, servers, applications and various services related to the system. The user can utilize the services and to download the data or the programs or the set of codes that were stored in various servers and located at various locations. The data can be downloaded or the utilized by locating at various places by simply having an authorized access to the set of customers. The customers were based on the set of applications they were likely to use or they wish to use the other set of data in the coming future also. The effort kept by users for providing data to the servers, maintaining the data in the servers and providing security to the data servers were very minimum. He most of the tasks and the security step for providing the system and the machines and servers will be taken care by the service providers and the people whoever maintaining the services.

Data centers were being maintained at various locations of the earth at various countries in the continents. The users related to that particular center can able to download the data at any point of time. The servers and the machines at each data center were internally connected to each other such that to maintain the data uniquely and to provide best service to the customers. The data that was stored at various data centers were being shared by various research organizations, remote processing applications and other related applications. Some organizations need to be used mostly for various applications and the data should be secured such that the more security should be provided to the data that was being stored by various research organizations and other multinational companies who will have a huge set of transactions and the customers related data.

1.2. Queuing Systems

The queuing systems which were in the form of theoretical model were intended to develop and provide the various set of models for predicting and estimating the performances of various systems subject to the random in nature of the systems. The history

of the queuing theory was first started in the years of 1908. The Copenhagen Telephone Company asked the Agner K. Erlang to study and identify the

Figure 1. Cloud Data Center

holding times of a telephone switch. Within short span of time he identified a great discovery of the queuing systems. He identified that the number of telephone conversations and the holding time of the each telephone unit can be able to fit into the Poisson and exponential distributions namely. This discovery had made a great change in the next future for the vast development of latest technologies based on the present queuing theory models. The analysis of the waiting lines can be analyzed easily with the help of queuing theory models. It also can be treated as the analysis of the waiting lines or queues at various places can be easily analyzed with the help of these queuing theory models. These models are highly famous and useful in predicting the performance of the various systems almost exactly the same condition of their actual behaviors. The research on various operations in a system called the operations research is also one of the parts of the queuing systems or queuing theory models. The long term average values for a system can be predicted or can be analyzed by these queuing systems. The input data that we are going to supply to a systems or a machine was measured for an extended period of time. The queuing systems will assume that the arrival times and service times are always random in nature.

The following figure shows the queuing system model. The contents of the system are the customer arrivals i.e., the number of users or the customers who were using or entering into the queue for getting the service or to see something or get set of taking something from a point. When the customers enters, if the line is free they can easily get their service, if the line us busy then certain time will be taken for each user to get their application or their task will be completed.

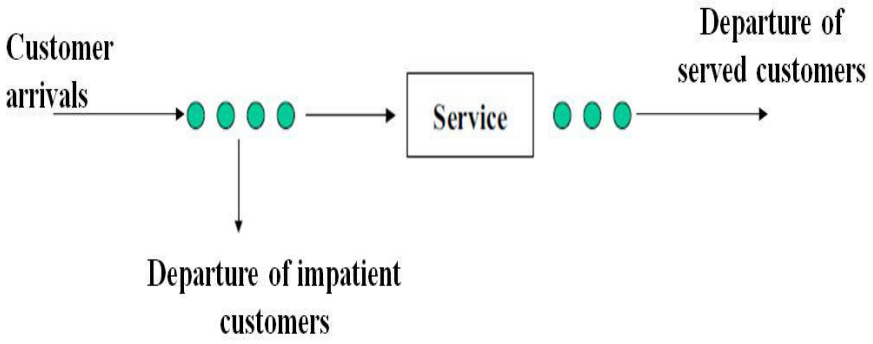

Figure 3. Queuing System

Various set of assumptions were involved in queuing systems whenever we are trying to solve the issues and the problems related to the queuing systems and its related applications. The arrivals to a queuing system or the system that was being solved by using the queuing system model are independent distribution, exponential distribution. The queues which were in heavy lines or small lines will not discourage the customers.

2. Problem Description and Solution

The problems at hand give rise to the task of evaluating the performance of data center with various queuing models to understand the distribution of the performance parameters with arrival and service rates, traffic intensity, number of servers and the associated probabilities. The goal of this thesis is to provide a framework through programs related to queuing models and evaluate the performance parameters, attempt validation, sensitivity analysis and make comparisons for data centers. Present thesis 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 considered 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.

Major problem in cloud computing is understanding the nature of cloud server performance, which depend on the evaluation and utilization of optimum performance parameters. Hence there is a necessity to analytically model the cloud servers/data centers using queuing systems and estimate the performance parameters. Generally various analytical models designed and developed for analyzing the performance of various cloud server applications with various set of assumptions and configurations. These assumptions are based on queuing theory and its accuracy can be verified with numerical calculations and simulations (which is out the scope of the present thesis).

2.1. Existing System

Performance parameters evaluated for single and multi server models especially with exponential arrival and service rate distributions have been popular and well validated. Few research publications focused on the analytical solutions when the distributions are not exponential. Sensitivity analysis of data centers with varying arrival and service rates has not been a major focus in the existing system, which provides overall insight in the data center behavior.

2.2. Proposed System

Present system 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 along with sensitivity analysis is also proposed.

2.3. Software and Hardware Requirements

The steady state performance parameter formulations were identified from literature and subsequently MATLAB programs were developed. Program for evaluating performance parameters was carried out in MATLAB® 7.60 (R2008a) software environment developed by MathWorks, Inc., USA. The MATLAB is a high-performance language for technical computing, integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Matlab is compatible on Windows, Unix and Mac OS platforms. Typical uses of the software include:

3. System Design

System design includes parameters, performance measures, stability and properties considered for the data center performance evaluation using queuing models. System parameters are,

a. It is customary to introduce some notation for the performance measures of interest in queuing systems.

b. ***Number of customers in the system (L_S)*:** In steady-state, the expected value of the state distribution gives the mean number of customers in the system.

c. ***Number of customers in the queue (L_Q)*:** In steady-state, the expected value of the state distribution in a queue gives the mean number of customers in the queue.

d. ***Utilization (or) Traffic intensity* (ρ):** For a queuing system with a single server, utilization ρ is the fraction of time the server is busy. When there is no limit on the capacity of the system, then

ρ = mean arrival rate/mean service rate $=\lambda/\mu$

e. The utilization when there are multiple servers (c), is the mean fraction of busy servers. Since ***c***is the overall service rate, in this case $\rho =$. For a stable (ergodic) system, the condition for stability is $\rho < 1$.

f. ***Throughput ( )***: The throughput for a queuing system with infinite capacity is the mean number of customers processed in a unit of time, i.e. the departure rate. Since the departure rate is equal to the arrival rate (and assuming $\rho < 1$), the throughput is $=c\rho$. For a queuing system with finite capacity, there can be loss in the systems, and so the throughput can be less than the arrival rate. In this case, throughput is often denoted differently (e.g. as *S*) **to distinguish it from the arrival rate.**

g. ***Response Time* (W_S):** (or sojourn time) It is the total time a customer spends in the system.

h. ***Waiting Time*** **(W_Q):** It is the time a job spends in the queue waiting to be serviced. Therefore, response time is the sum of the waiting time (W_Q) and the service time(***1/)*** for a customer i.e. $W_S = W_Q + (1/\mu)$

To evaluate the performance parameters of data center in cloud architecture, the programs consider the following input values:

i.Interarrival rates ,
ii. Service rates ,
iii. Number of servers , *c*
iv. Maximum number of customers allowed , *K*
v.Waiting capacity of customers , *r*
vi. Coefficient of variance , C_{oV}
vii. Erlang parameter , *Er*

The list of output parameters of the programs are highlighted below:

i.Length of customers in a system , L_S
ii. Length of customers in queue , L_Q
iii. Waiting time of customers in a system , W_S
iv. Waiting time of customers in queue , W_Q
v.Associated probabilities

4. System Implementation

Whenever if there is any uncertainty in arrival and service times of any system or application, the queuing models are the best fitted application or the model which could be used to estimate the presentation of the systems for various services to the users or customers. The simplest possible (single stage) 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. For the present work it is assumed that customers are served in the order in which they arrive in the system (First-Come-First-Served or FCFS). The MATLAB programs and formulations are given at Appendix 1.

4.1. Queuing Models

a) Queuing Model - M/M/1

The M/M/1 queue (Figure.7) has Interarrival times, which are exponentially distributed with parameter and also service times with exponential distribution with parameter. The system has only a single server (c=1) and uses the FIFO service discipline. The exponential distribution has a squared coefficient of variation of 1. The waiting line is of infinite size. The M/M/1 system is a pure birth-/death system, where at any point in time at most one event occurs, with an event either being the arrival of a new customeror the completion of a customer's service. What makes the M/M/1 system really simple is that the arrival rate and the service rate are not state dependent.

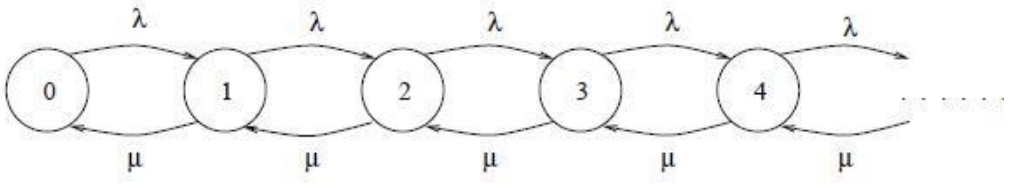

Figure 7. Markov Chain for M/M/1 Queue

Steady state performance measures for the above model are:

$$L_S = \frac{\rho}{1-\rho} \quad ; \quad L_Q = \frac{\rho^2}{1-\rho}$$

$$W_S = \frac{1}{\mu(1-\rho)} \quad ; \quad W_Q = \frac{\rho}{\mu(1-\rho)}$$

$$P(W_S < t) = 1 - e^{-(1-\rho)\mu t} \quad ; \quad P(W_S > t) = e^{-(1-\rho)\mu t}$$

$$P(W_Q <= t) = 1 - \rho e^{-(1-\rho)\mu t} \quad ; \quad P(W_Q >= t) = \rho e^{-(1-\rho)\mu t}$$

Note: P is the notation for probability

b). Queuing Model - M/D/1

The M/D/1 queue has Interarrival times, which are exponentially distributed with parameter and also service times with constant distribution with parameter. The system has only a single server (c=1) and uses the FIFO service discipline. The waiting line is of infinite size. Deterministic distribution (with constant service times) has a zero variance for this distribution. For this reason one can achieve always the highest throughput (lowest delays) for deterministic service times. The simplest important result is that the average number waiting is half that waiting with exponentially distributed service.

Steady state performance measures for the above model are:

$$L_S = \rho + \frac{\rho^2}{2(1-\rho)} \quad ; \quad L_Q = \frac{\rho^2}{2(1-\rho)}$$

$$W_S = \frac{2-\rho}{2\mu(1-\rho)} \quad ; \quad W_Q = \frac{\rho}{2\mu(1-\rho)}$$

$$Var(Ls) = \frac{\rho^3}{3(1-\rho)} + \frac{\rho^2}{2(1-\rho)} + 2\rho^2(3-2\rho)(1-\rho) + \rho(1-\rho)$$

$$Var(Ws) = 2Wq^2 + \frac{2\rho}{\mu^2 3(1-\rho)} + \frac{1}{\mu^2(1-\rho)} - Ws^2$$

$$Var(Lq) = \frac{\rho^3}{3(1-\rho)} + \frac{\rho^2}{2(1-\rho)} + 2\rho^2(1-\rho)$$

$$Var(Wq) = Wq^2 + \frac{2\rho}{\mu^2 3(1-\rho)}$$

c). Validation

Validation studies for the present MATLAB codes generated was undertaken using the existing queuing software's and published data. It was considered to validate the program first and then evaluate performance parameters for the input data of interest. The above approach was followed in the present work to prove the validity and reliability of the results generated. The results of validation are given at Appendix 2. A list of validation software's, published data utilized is highlighted below:

d). QtsPlus4Calc Software (*http://qtsplus4calc.sourceforge.net*):

QtsPlus4Calc, release 2006 is a freeware developed by Donald gross and Carl M. Harris of George Mason University. This software provides a platform to evaluate performance of various queuing models. A sample screenshot of the software environment is given at Figure.12. The calculator includes single server, multi-server, priority, bulk and network models.

Numerical Solutions to Queuing Systems *(http://queueing-systems.ens-lyon.fr):*Numerical solutions to queuing systems software provide a platform to evaluate performance of various queuing models with generalized service distributions. A sample screenshot of the software environment is given at Figure.13. The calculator includes single server models.

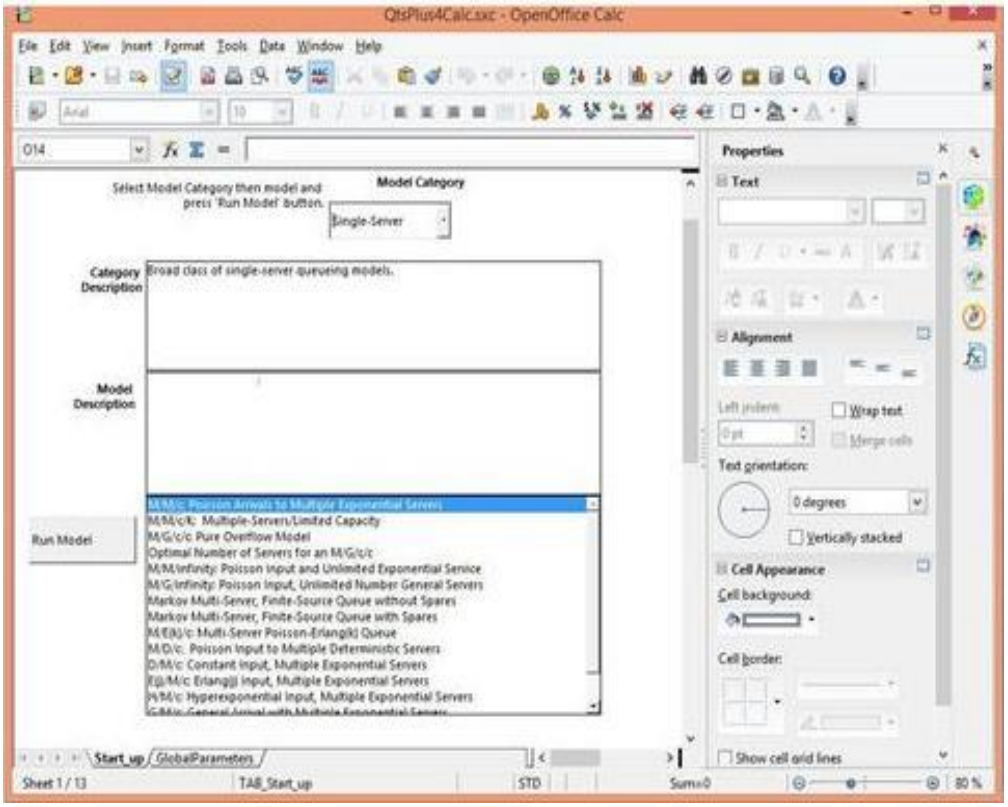

Figure 7. QtsPlus4Calc Environment Screenshot

6. Results and Discussion

Based on the queuing models discussed in the previous chapters (4 and 5), input parameters limits were identified from literature for performance evaluation of small cloud computing data center farm (*i.e.*, number of servers, *c* were limited to 2 and 4). The inter arrival and service rates are chosen for a range of traffic intensity (or utilization) varying from 0 - 1. Each time service rate was varied to values 1, 2 and the respective performance was evaluated. Input parameter limits are given at Table 3.

Table 3. Input Parameter Limits

Parameter	Chosen Limits
	0- 2
	1, 2
ρ	0- 1
c	2, 4
K	4, 8
r	4, 8
C_{ov}	1.5
Er	2, 4
t	0- 1

Evaluation of Cloud Computing Data Center Performance with M/M/1 Queuing Model:

Table 4. Performance of Data Center - M/M/1, Model = 1

	ρ	Ls	Lq	Ws	Wq	t	$P(Ws<t)$	$P(Ws>t)$	$P(Wq<=t)$	$P(Wq>=t)$
0	0	0.00	0.00	1.00	0.00	0	0.00	1.00	1.00	0.00
0.1	0.1	0.11	0.01	1.11	0.11	0.1	0.09	0.91	0.91	0.09
0.2	0.2	0.25	0.05	1.25	0.25	0.2	0.15	0.85	0.83	0.17
0.3	0.3	0.43	0.13	1.43	0.43	0.3	0.19	0.81	0.76	0.24
0.4	0.4	0.67	0.27	1.67	0.67	0.4	0.21	0.79	0.69	0.31
0.5	0.5	1.00	0.50	2.00	1.00	0.5	0.22	0.78	0.61	0.39
0.6	0.6	1.50	0.90	2.50	1.50	0.6	0.21	0.79	0.53	0.47
0.7	0.7	2.33	1.63	3.33	2.33	0.7	0.19	0.81	0.43	0.57
0.8	0.8	4.00	3.20	5.00	4.00	0.8	0.15	0.85	0.32	0.68
0.9	0.9	9.00	8.10	10.00	9.00	0.9	0.09	0.91	0.18	0.82
1	1	Inf	Inf	Inf	Inf	1	0.00	1.00	0.00	1.00

Table 5. Performance of Data Center - M/M/1 Model= 2

	ρ	Ls	Lq	Ws	Wq	t	P(Ws<t)	P(Ws>t)	P(Wq<=t)	P(Wq>=t)
0	0	0.00	0.00	0.50	0.00	0	0.00	1.00	1.00	0.00
0.2	0.1	0.11	0.01	0.56	0.06	0.1	0.16	0.84	0.92	0.08
0.4	0.2	0.25	0.05	0.63	0.13	0.2	0.27	0.73	0.85	0.15
0.6	0.3	0.43	0.13	0.71	0.21	0.3	0.34	0.66	0.80	0.20
0.8	0.4	0.67	0.27	0.83	0.33	0.4	0.38	0.62	0.75	0.25
1	0.5	1.00	0.50	1.00	0.50	0.5	0.39	0.61	0.70	0.30
1.2	0.6	1.50	0.90	1.25	0.75	0.6	0.38	0.62	0.63	0.37
1.4	0.7	2.33	1.63	1.67	1.17	0.7	0.34	0.66	0.54	0.46
1.6	0.8	4.00	3.20	2.50	2.00	0.8	0.27	0.73	0.42	0.58
1.8	0.9	9.00	8.10	5.00	4.50	0.9	0.16	0.84	0.25	0.75
2	1	Inf	Inf	Inf	Inf	1	0.00	1.00	0.00	1.00

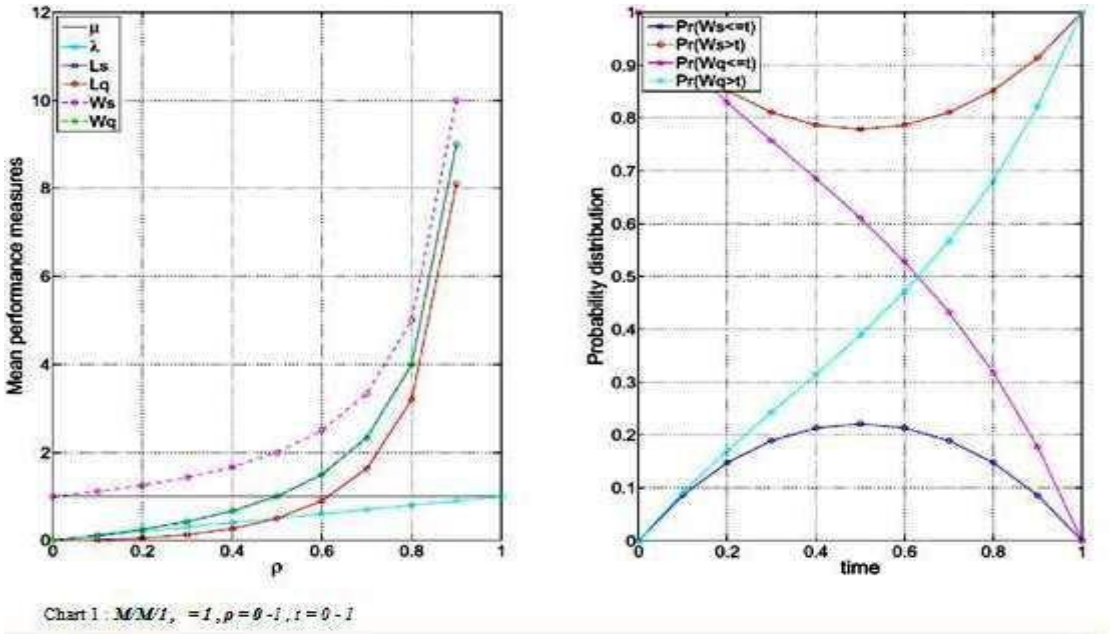

Figure 8. Graphical Representation of First Model M/M/1 for Probability

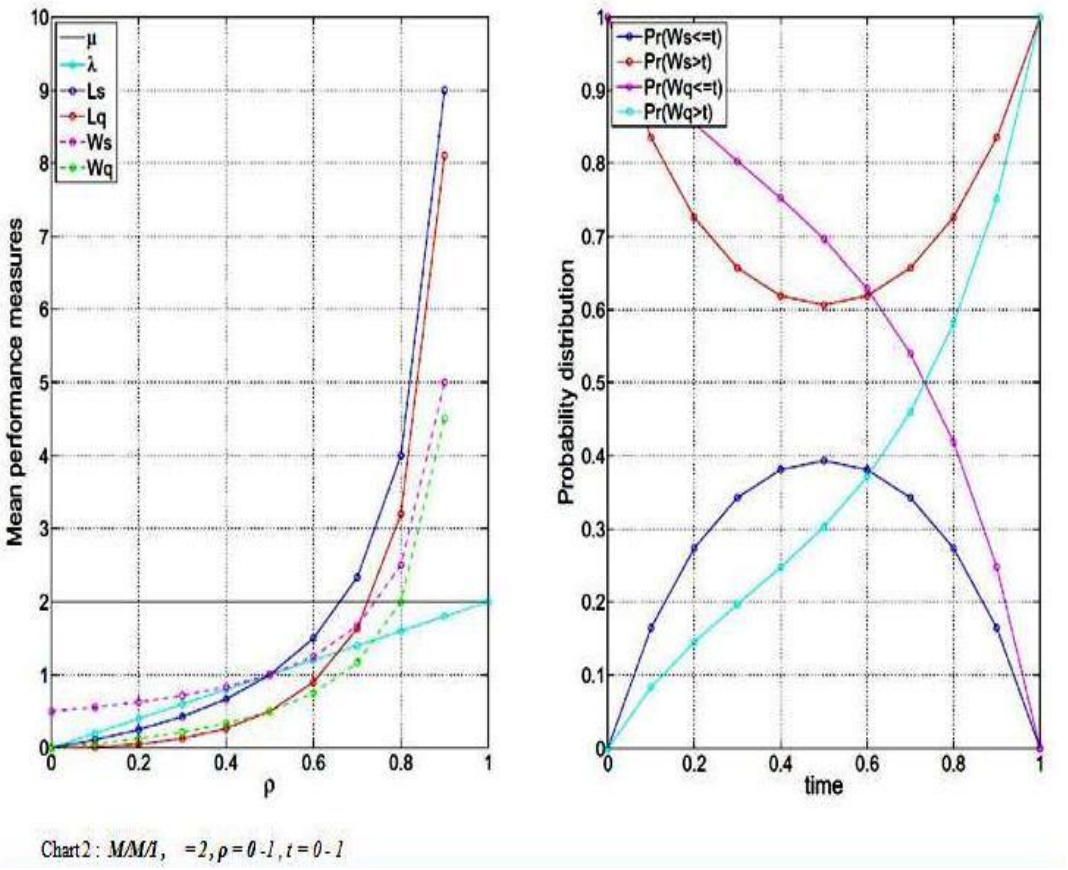

Figure 9. Graphical Representation of Second Model M/M/1 for Probability

Evaluation of Cloud Computing Data Center Performance with M/D/1 Queuing Model,

Table 8. Performance of Data Center - M/D/1 Model= 1

Variance										
		ρ	***Ls***	***Lq***	***Ws***	***Wq***	***Var(Ls)***	***Var(Ws)***	***Var(Lq)***	***Var(Wq)***
1	0	0	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
1	0.1	0.1	0.11	0.01	1.06	0.06	0.14	0.08	0.02	0.08
1	0.2	0.2	0.23	0.03	1.13	0.13	0.33	0.18	0.07	0.18
1	0.3	0.3	0.36	0.06	1.21	0.21	0.53	0.33	0.14	0.33
1	0.4	0.4	0.53	0.13	1.33	0.33	0.72	0.56	0.25	0.56
1	0.5	0.5	0.75	0.25	1.50	0.50	0.90	0.92	0.40	0.92
1	0.6	0.6	1.05	0.45	1.75	0.75	1.14	1.56	0.67	1.56
1	0.7	0.7	1.52	0.82	2.17	1.17	1.73	2.92	1.34	2.92
1	0.8	0.8	2.40	1.60	3.00	2.00	3.93	6.67	3.67	6.67
1	0.9	0.9	4.95	4.05	5.50	4.50	19.12	26.25	18.99	26.25
1	1	1	Inf	Inf	Inf	Inf	Inf	NaN	Inf	Inf

Table 9. Performance of Data Center - M/D/1 Model= 2

Variance										
		ρ	Ls	Lq	Ws	Wq	Var(Ls)	Var(Ws)	Var(Lq)	Var(Wq)
2	0	0	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
2	0.2	0.1	0.11	0.01	0.53	0.03	0.14	0.02	0.02	0.02
2	0.4	0.2	0.23	0.03	0.56	0.06	0.33	0.05	0.07	0.05
2	0.6	0.3	0.36	0.06	0.61	0.11	0.53	0.08	0.14	0.08
2	0.8	0.4	0.53	0.13	0.67	0.17	0.72	0.14	0.25	0.14
2	1	0.5	0.75	0.25	0.75	0.25	0.90	0.23	0.40	0.23
2	1.2	0.6	1.05	0.45	0.88	0.38	1.14	0.39	0.67	0.39
2	1.4	0.7	1.52	0.82	1.08	0.58	1.73	0.73	1.34	0.73
2	1.6	0.8	2.40	1.60	1.50	1.00	3.93	1.67	3.67	1.67
2	1.8	0.9	4.95	4.05	2.75	2.25	19.12	6.56	18.99	6.56
2	2	1	Inf	Inf	Inf	Inf	Inf	NaN	Inf	Inf

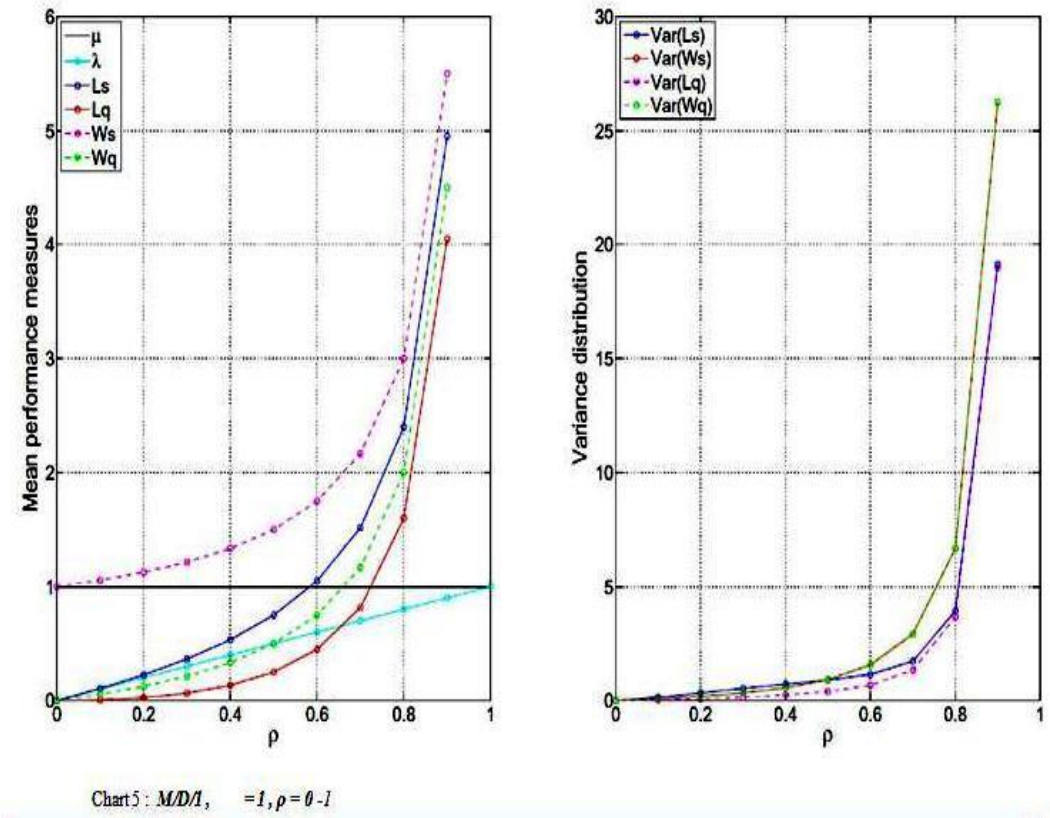

Figure 10. Graphical Representation of First Model M/D/1 for Variance and Performance Measures

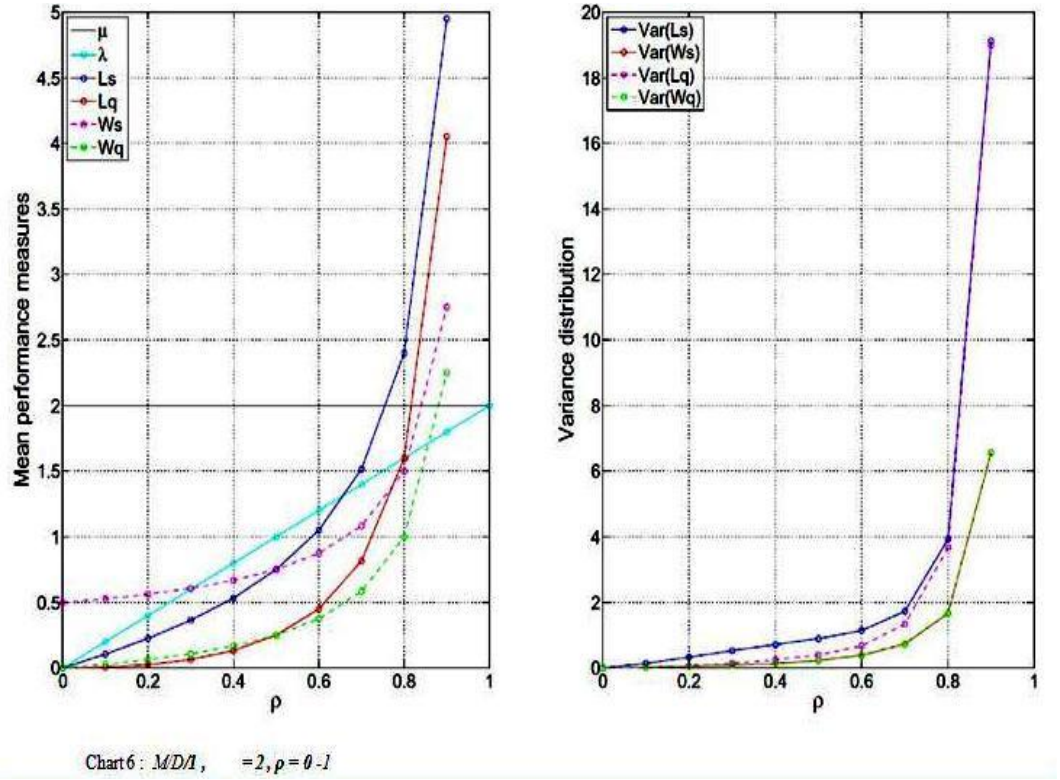

Figure 11. Graphical Representation of Second Model M/D/1 for Variance and Performance Measures

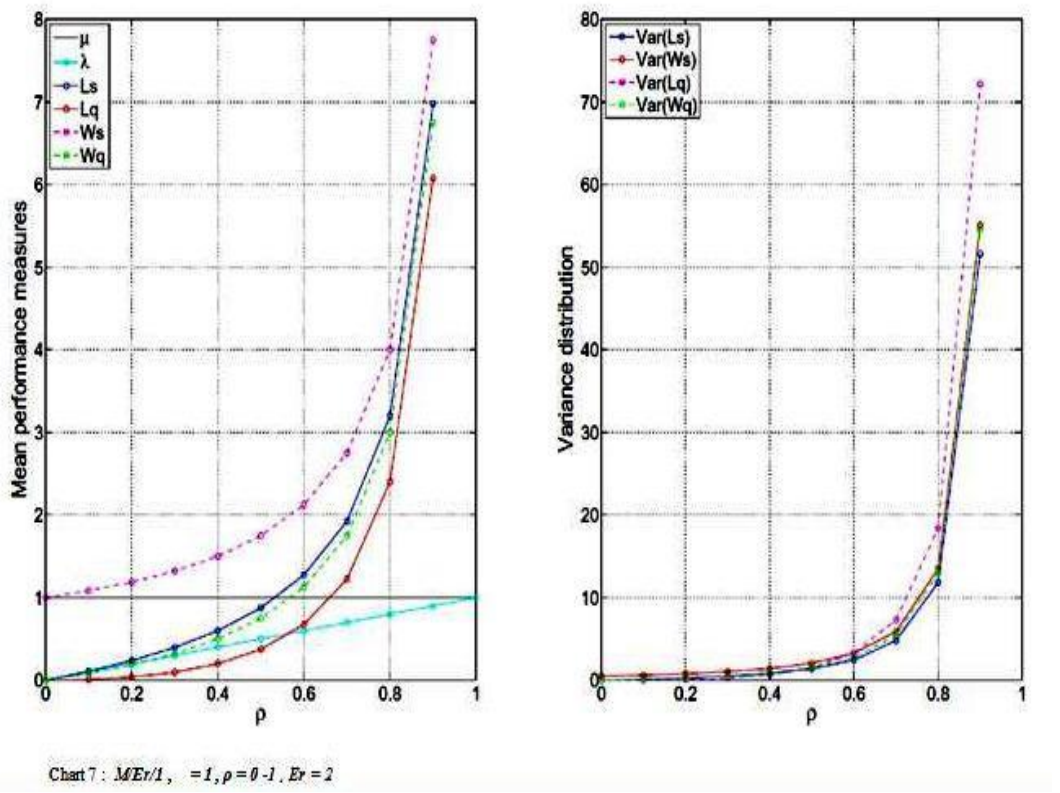

Figure 12. Graphical Representation of First Model M/D/1 for Variance and Mean Performance Measures

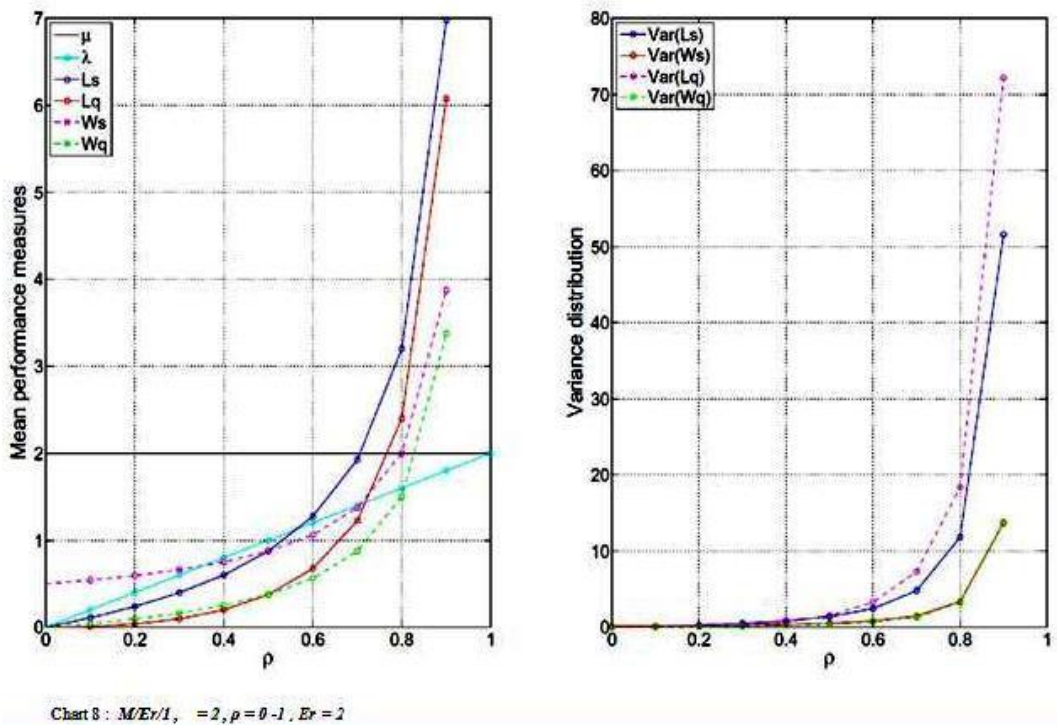

Figure 13. Graphical Representation of Second Model M/D/1 for Variance and Mean Performance Measures

7. Conclusions

Performance evaluation for single servers indicate that as the service rate (r) 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. 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/D/1* shows detrimental nature when compared with other queuing models, which is attributed to higher value of C_{ov}. 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 ( ) 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. 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] J. Hamzeh Khazaei and Vojislav, "Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems", IEEE Transactions on parallel and distributed systems, vol. 23, **(2012)** May.
[2] H. Khazaei, "Performance Modeling of Cloud Computing Centers", Doctoral dissertation, The University of Manitoba, Canada, **(2012)** October.

[3] B. Yang, F. Tan, Y. Dai, and S. Guo, "Performance evaluation of cloud service considering fault recovery", First International Conference on Cloud Computing (CloudCom), **(2009)** December.
[4] I. Adan and J. Resing, "Queuing systems", Eindhoven University of Technology, The Netherlands, **(2015)** March.
[5] J. Sztrik, "Basic Queuing Theory", University of Debrecen, Faculty of Informatics, **(2012)** December.
[6] J. Shetty Chandrakala, "Survey on Models to Investigate Data Center Performance and QoS in Cloud Computing Infrastructure", First International Conference on Recent Advances in Science & Engineering, **(2014)**.
[7] M. Hlynka and S. Molinaro, "Comparing expected wait times of a M/M/1queue", Department of Mathematics and Statistics, University of Winsor, **(2010)** June.
[8] N. Khanghahi and R. Ravanmehr, "Cloud Computing Performance Evaluation: Issues and Challenges", International Journal on Cloud Computing Services and Architecture, vol. 3, **(2013)** October.
[9] T. V. Mathew, "Queuing Analysis, Transportation Systems Engineering", Indian Institute of Technology, Bombay, (2014) February.
[10] T. Sai Sowjanya, D. Praveen, K. Satish and A. Rahiman, "The Queuing Theory in Cloud Computing to Reduce the Waiting Time", IJCSET, vol. 1, **(2011)** April.
[11] A. Brandwajn and H. Wang, "A conditional probability approach to M/G/1 – like queues", Performance evaluation, vol. 65, **(2008)**.

