International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 11 (2018) pp. 9485-9492

Performance Analysis of web Application deployment on cloud using M/M/S/K Queueing model

[1]N.Neelima, [2]B.Basaveswar Rao, [3]K.Gangadhara Rao, [4]K.Chandan

[1] *Assistant Professor, Velagapudi Ramakrishna Siddhartha Engineering College,A.P, India.*

[2] *Computer Centre, Acharya Nagarjuna University Guntur, A.P, India.*

[3] *Professor, Department of CSE, Acharya Nagarjuna University Guntur, A.P, India.*

[4] *Professor, Department of Statistics, Acharya Nagarjuna University Guntur, A.P, India.*

Abstract

This paper studies the performance analysis of web application deployment on cloud environment using M/M/S/K analytical queueing model. The performance measures end-to-end response time, utilization of computing resources consumed, blocking probability, average time spent in the system and average waiting time in the queue are considered to evaluate the dynamic nature of the model. The cost equation is derived with respect to virtual machine setup cost and web request holding cost .The numerical results are illustrated.

Keywords: Cloud computing, Queueing Theory, Performance Measure, Cost, SLA, QoS

INTRODUCTION

Cloud Computing is a unique pattern for the availability of computing infrastructure, which targets to shift the vicinity of the computing infrastructure to the network as a way to reduce the charges for the management and protection of hardware and software resources [3]. Cloud computing is an evolving profit-making infrastructure set up that promises to remove the need for investment and maintaining expensive computing facilities by the customer. It is a version for permitting ubiquitous, handy, on call for community to get entry to a shared pool of configurable computing assets (e.g., networks, servers, storage, applications, and offerings) which can be reconfigured dynamically to regulate to a variable load for improved performance and for optimum utilization with minimum control attempt or carrier company interplay [1].

Generally cloud computing presents the functionality through which regular and actual-time scalable resources likes applications, documents, packages, records, hardware, software and other computing amenities are available through the network to customers without the users being aware about the details of execution environment. These customers get the computing resources and services by entering in to a customized service level agreement (SLA) with the cloud provider. This SLA serves as the base for the expected level of service between the consumer and the provider, they effectively pay the rate according to the usage of time, the usage of mode, or the quantity of information transferred. Accordingly, SLA management includes the SLA agreement description, main schema with the Quality of Services (QoS) parameters which need to be closely monitored by both the parties [2]. The success of any cloud computing platform service providers depends upon its capability to deliver guaranteed QoS with 24 x 7 availability. The most important service parameters of QoS consists of availability, throughput, reliability, and security, in addition to many different other parameters.

Besides the QoS parameters there are other performance indicators like response time, job blocking probability, possibility of on the spot provision, and imply wide variety of obligations inside the cloud [5].All these performance measures can be computed by applying varied queuing models[4] ,which has been proved successfully in various walks of life where the issue of service provider and consumer comes into the picture .In order to accept the QoS of the cloud setup , it is equally important to enhance the QoS. As cloud computing dynamically gives computing resources to meet the requirements of QoS up on solicitation from exclusive client, orchestrating suitable resources could be a hard mission. On the other hand, a data centre has a huge variety of bodily computing nodes [6]; conventional queuing assessment rarely worries about structures of its length. Queueing models provide the analyst with a powerful tool for designing and evaluating the performance of cloud computing. When a user sends a request to a web application on cloud, the request join the queue if the servers are busy and waits for the service , then it is forwarded on to VMs which act as servers to process the requests in the queue. A Cloud Computing platform can be modelled as M/M/S/K queueing system consisting of single finite queue and multiple servers for different applications. This follows Poisson process, for which the inter-arrival time of requests is exponentially distributed. Here K represents the finite capacity of the queue and S represents the number of finite servers.

Many researchers have done research on the analysis on infinite queues [12], [13], [15], [16], [17], [18], [19] which in reality most of the networks follows finite queues like M/M/S/K. This paper proposes an analytical model to check the performance analysis of web application hosted in a typical cloud setup. This cloud architecture is modelled as M/M/S/K queuing model and its arrival rate follow Poisson distribution whereas service rate exponential for given time interval, thus set of analytical equations and formulas for key performance measures are provided [26].The significant performance metrics calculated in this paper are end-to-end response time, utilization of computing resources consumed, blocking probability, average time spent in the system, average waiting time in the queue. The total Cost based on the performance metrics is derived and analysed with four different scenarios.

The rest of the paper is organised as follows section 2 discusses about related work, section 3 describes the finite queuing model, and section 4 elaborated the experimental results in tabular form and finally conclusions are provided in section 5.

RELATED WORK

Cloud computing delivers user a wide-range computing environment. Resource provisioning in cloud computing is still a challenging issue. In this section we discuss some related work for resource allocation in cloud computing using Queuing theory .In [7] when the number of facility nodes is m, without any restrictions on the number of facility nodes, a cloud environment is modelled as an M/G/m queuing system which indicates the inter-arrival time of requests when exponentially distributed, the service time when generally distributed. In [8] an approximate Markov Chain model for performance evaluation of a cloud centre using analytical technique which makes the model flexible in terms of scalability and diversity of service time and provides results with high degree of accuracy for the mean number of tasks in the system.

However in [9] a performance model suitable for analysing the service quality of large sized IaaS clouds, using interacting stochastic models were proposed. The properties of several bounds including arrival rate, task service time, the virtualization degree, and service task size on task rejection probability and total response delay was scrutinized. . In [10] an analytical model for performance evaluation of a cloud computing data centre using the queue GE/G/m/m+r was proposed for measuring the performance indicators analytically like average number of tasks in the system, blocking probability, probability of immediate service and the average of response time.

In addition in [11] the virtual machines are taken as service centres and the web applications are modelled as queues. The author hired the finite buffer multi server queuing system with queue dependent heterogeneous servers and the number of server's changes depending on the queue length. In [12] The performance of queuing system is evaluated analytically to obtain performance factors like mean number of tasks in the system with single task arrivals and a task request buffer of infinite capacity using Cloud centre as an [(M/G/1) : (∞/GD)] queuing system. In [13] a different cloud computing model in this paper emphasis on the improvement of allocation of resources dynamically following request dependent strategy under non homogeneous condition with time dependent arrival of jobs which is much useful for analysing the cloud more effectively and efficiently to increase performance measures of cloud. It demonstrates that the dynamic allocation of resources can reduce mean delay and mean service time.

In [14] the routing of incoming requests to the queue with reduced workload, response time and the average length of the queue was studied, by proposing the cloud computing model based on queuing system and obtained results indicate the increased utilization of global scheduler and decreased waiting time in cloud architecture. In [15] the author used a stochastic process to evaluate the dynamic conduct of infinite severs over single server and studied the utilization factor, throughput, length of server, and waiting time of infinite server system and hence found that it acquires service immediately. In [16] In order to analyse the performance of services in cloud computing they proposed a queuing model M/M/m and studied the performance of the optimization and the parameter for evaluating the service in cloud computing and developed a synthetically optimization method to optimize the performance which results in less wait time, queue length and more customers gaining the service. In [17-19] the author proposed a model to evaluate the performance of cloud computing centre using G/M/s queuing model in which the resource allocation is modelled as queues with the virtual machines as service centres and which reflects general nature of BoT's arrivals in the cloud. The author detected that when the arrival rate grows, the length of queue also rises, and the waiting time of a customer increases linearly with the arrival rate.

FINITE QUEUEING MODEL

In cloud computing, there are lot of users generating web resources that run applications remotely. The model of cloud computing is as in Figure 1[16].Cloud architecture of this model serves as a service center which consists of various resources and acts as a single point of access for all kinds of users around the world. Each user can access the service according to different kinds of needs and pays certain amount of money to the provider of the service. Cloud computing providers build the service center like Amazon to be used by clients, which provides several on-demand service instances. On-demand instances lets the user pay for compute capacity by the hour with no long-term commitments which frees the user from the costs and difficulties of planning, procuring, and maintaining hardware and converts what are usually large fixed costs into much smaller flexible costs [22].

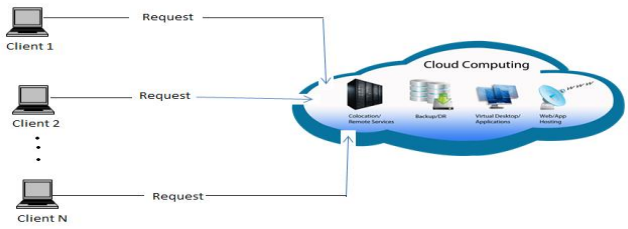

Figure 1. Cloud Computing Service Model [16]

The cloud computing service model showed in Figure 1 can be represented as a queueing model in Figure 2. Queuing theory has emerged as a scientific tool to deal with different types of queues for modelling and analysis of performance of computer systems and many other systems [23, 24]. An M/M/S/K queuing system model is a queue model where arrivals follow Poisson process that has limited queue capacity and finite requests and multiple servers. The Queuing discipline that is followed here if First-come-first-served (FCFS) .The mathematical analysis of the model gives the measurers system performance, average waiting time, server utilization, the blocking probability of exceeding buffer, the average busy servers, the period of activity server, etc. [25,26]

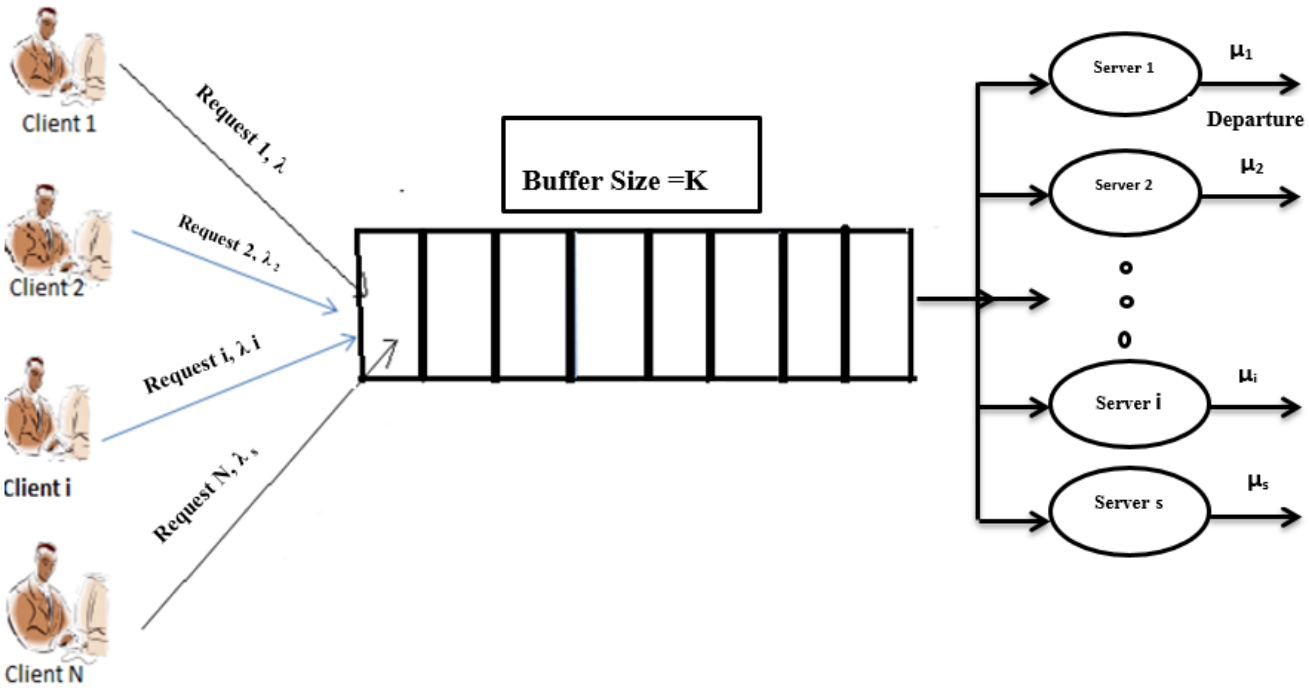

Figure 2. A queuing model for computing services in cloud computing services.

The M/M/S/K Analytical Queuing model for deploying web applications on cloud has the following parameters:

λ = the arrival rate of user requests,

μ = the service rate for each virtual machine for completion of user request,

s = Number of Virtual machines

K = Buffer size; Total number in system ≤ k+s

Let P_n be the probability that there are n requests in the system in the steady-state. n = 0, 1, 2, 3… k. The balanced equations for performance measures in general are [26]

$$P_1 = (\lambda / \mu)\, P_o$$

$$P_2 = (\lambda^2 / 2!\, \mu^2)\, P_o$$

$$P_3 = (\lambda^3 / 3!\, \mu^3)\, P_o$$

. . . . . .

$P_n = (\lambda^n / n!\, \mu^n)\, Po$, when n<S (3.1)

$P_n = (\lambda^n / S!\, S^{n-s}\, \mu^n)\, P_o$, when s≤n≤k (3.2)

Where

$$P_o = \left[\sum_{n=0}^{s-1} \frac{\rho^n}{n!} + \sum_{n=s}^{k} \frac{\rho^n}{S!\, S^{n-s}}\right]^{-1} .$$

Blocking Probability **BP**:

BP =Pr {system is full} =P_k . (3.3)

Effective Arrival Rate λ_e :

$\lambda_e = \lambda (1\text{-}BP) = \lambda (1\text{-}P_k)$. (3.4)

Average Customers in System **Ls** :

$$L_s = \frac{P_o}{S!}\left[\sum_{n=0}^{s-1} n.\rho^n + S^s \sum_{n=s}^{k} n \frac{\rho^n}{S^n}\right] . \quad (3.5)$$

Average Busy Servers $\mathbf{L_B}$:

$L_B = \lambda_e / \mu$. (3.6)

Average Customers in Queue $\mathbf{L_q}$:

$$L_q = \sum_{n=s}^{k} (n-s).P_n \ . \quad (3.7)$$

Utilization of the System **U:**

$U = 1\text{-}P_o$. (3.8)

Utilization of the Service **SU**:

SU = 1-(P_o + P_1 + P_2 +…. +P_{s-1}). (3.9)

Average Time Spent in System **Ws**:

$W_s = L_s / \lambda_e$. (3.10)

Average Waiting Time in the Queue **Wq**:

$W_q = L_q / \lambda_e$. . (3.11)

Cost Model

After examining the above measures it is observed that the cost of web application deployment mainly depends on the number of virtual machines used, service rate at a given time and

average requests in the buffer. Based on the dynamic nature of these parameters the total Cost 'C' defined as [20-21]

$$C = \alpha\, s\, \mu + \beta\, L_q$$
$$= \alpha\, s\, \mu + \beta\, \lambda e\, Wq$$
$$= \alpha\, s\, \mu + \beta\, (1-P_K)\, \lambda\, Wq \qquad (3.12)$$

Where $L_q = \lambda e\, Wq$ & $\lambda e = (1-P_K)\, \lambda$

α is the price of request completion resources consuming cost and β is the price for Job waiting cost for requests waiting in the buffer respectively. The cost in cloud infrastructure is determined by the number of virtual machines, service rate for the requests, completed requests and number of jobs waiting in the queue .The linear cost model in Eq (3.12) has been justified by the numerical illustration with different scenarios, the scenarios are λ varying scenarios, μ varying scenario and S varying scenario and K varying scenario.

NUMERICAL ILLUSTRATION

In this section, to give some numerical examples for web application deployment on cloud environment through M/M/S/K queuing model, the parameters S, K, λ, μ are given values, and then found total Cost incurred for the Cost derived in section 3.

Scenarios: In Customer service rate increases scenario the service rate λ, the computation resource cost β, the number of servers S and k is kept fixed and cost is calculated for different arrival rates μ. In buffer capacity increase scenario, for fixed values of λ, μ, S and for different k values the cost is estimated. In virtual machines increases scenario, for fixed values of λ, k, μ and for different S values the cost is estimated. In customers' requests increases scenario, for fixed values of k, s, μ and for different λ values the cost is estimated as shown in table 1. For all the scenarios the price of request completion resources consuming cost α and the price for Job waiting cost for requests waiting in the buffer β are taken as 5 and 1 units respectively.

Table 1. Cost calculation for fixed values and varying values of λ, μ, s, k

Scenarios	Fixed Parameter values			Varying Parameter values	Cost
	λ	S	k	μ	C
Customer service rate increases	50	6	10	5	152.7567
				10	300.7547
				15	450.1373
				20	600.0305
				25	750.0086
				30	900.0029
				35	10500
				40	12000
				45	13500
				50	15000
				60	18000
				70	21000
				80	24000
				120	36000

Scenarios	Fixed Parameter values			Varying Parameter values	Cost
				180	5400
				250	7500
				300	9000
				400	12000
	λ	**S**	**μ**	**K**	**C**
When buffer capacity increase	20	2	25	2	250
				3	250.0569
				4	250.1002
				5	250.1258
				6	250.1394
				7	250.1463
				8	250.1496
				9	250.1511
				10	250.1518
	λ	**K**	**μ**	**S**	**C**
When virtual machines increases	20	50	25	5	625.0003
				6	750.0000
				7	875.0000
				8	1000
				9	1125
				10	1250
				15	1875
				20	2500
				25	3125
				30	3750

Scenarios	Fixed Parameter values			Varying Parameter values	Cost
				35	4375
				40	5000
	S	**k**	**μ**	**λ**	**C**
When customers' requests increases	2	10	25	20	250.1518
				25	250.3270
				30	250.6302
				35	251.1072
				40	251.7729
				45	252.5801
				50	253.4286
				55	254.2170
				60	254.8862
				65	255.4233
				70	255.8425
				75	256.1672
				80	256.4196
				85	256.6180
				90	256.7761
				95	256.9040
				100	257.0090

Observations: From different scenarios as shown in table 5, it is observed that

- As the Completion rate of the user increases the computing cost increases significantly,
- As the queue capacity increases the computing cost increases slowly and is insignificant
- due to loss probability (i.e. P_k is insignificant)
- The number of Virtual Machines increases, the computing cost also increases
- The User request rate increases, the computing cost increases slowly

The pictorial representation for the observations are presented below

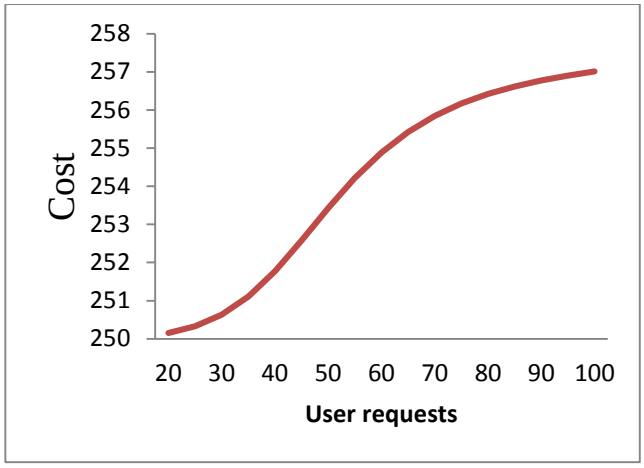

Figure 3. Cost estimation when the user requests increases

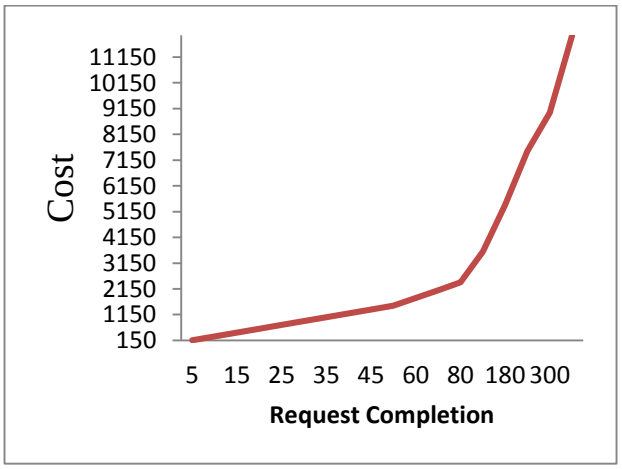

Figure 4. cost estimation when Request completion rate increases

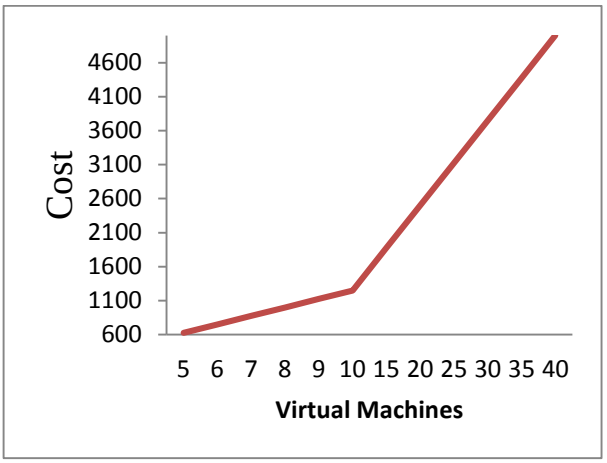

Figure 5. Cost estimation when Virtual Machines increases

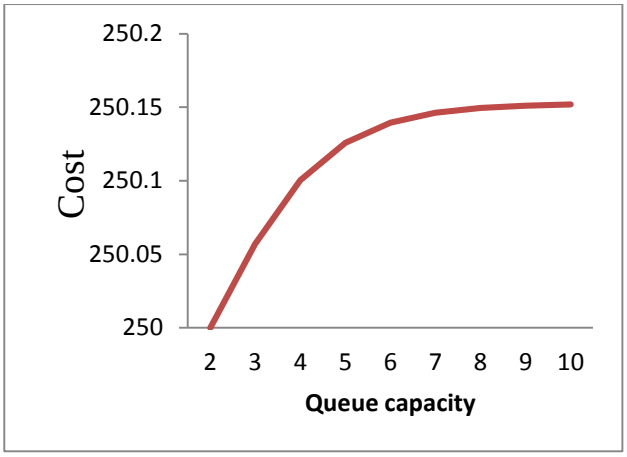

Figure 6. Cost estimation when Queue capacity increases

CONCLUSION

This paper investigated the analysis of web application deployment on cloud environment for studying the cost point of view. This analysis is useful for estimating the future requirement of the cloud resources based on number of user requests. The total Cost based on the performance metrics is derived and analysed with four different scenarios. It is observed that as the completion rate of the user request increases the computing cost also increases significantly. The queue capacity increases the computing cost increases slowly and insignificantly as loss probability tends to zero (i.e. P_k is insignificant).The number of virtual machines increases, the computing cost also increases. The user request rate increases, the computing cost increases slowly. The future scope of this paper is to conduct experiment in AWS or other cloud services.

REFERENCES

[1] "The NIST Definition of Cloud Computing," http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145. Pdf.

[2] A. Keller and H. Ludwig, "The wsla framework: Specifying and monitoring service level agreements for web services," J. Netw.Syst. Manage. vol. 11, pp. 57–81, March 2003. [Online]. vailable:http://dl.acm.org/citation.cfm?id=635430.635442

[3] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition", Computer Communication Review, vol. 39, no. 1, pp. 50–55, 2008.

[4] L. Kleinrock, Queuing Systems: Theory, vol. 1, Wiley-Interscience, New York, NY, USA, 1975

[5] L.Wang, G. vonLaszewski, A. Younge et al., "Cloud computing: a perspective study," New Generation Computing, vol. 28, no. 2, pp. 137–146, 2010

[6] "Amazon Elastic Compute Cloud, " User Guide, API Versioned. Amazon Web Service LLC or its

affiliate, 2010, http://aws.amazon.com/documentation/ec2.

[7] Hamzeh Khazaei, Jelena Mi si C,Vojislav B. Mi si C,"Modelling of Cloud Computing Centres Using M/G/m Queues", 2011 31st International Conference on Distributed Computing Systems Workshops

[8] Hamzeh Khazaei, Jelena Misic, and Vojislav B. Misic, "Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems", IEEE transactions on parallel and distributed systems, VOL. 23, NO. 5, MAY 2012

[9] Hamzeh Khazaei, Jelena Mi ˇ si´ c, and Vojislav B. Mi ˇ si ´c, "A Fine-Grained Performance Model of Cloud Computing Centers", IEEE transaction on parallel and distributed systems, vol. X, no. Y, 2012

[10] Mohamed Ben el aattar, Abdelkrim Haqiq,"Performance Modelling for a Cloud Computing Centre Using GE/G/m/k Queuing System", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, 2012.

[11] Goswami V, Patra, S. S, Mund G. B, "Performance Analysis of Cloud with Queue Dependent Virtual Machines", 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012.

[12] Ani Brown Mary, N and Saravanan, K,"Performance factors of Cloud computing data centres using, [(M/G/1): (∞/GDMODEL)] Queuing system", International Journal of Grid Computing & Applications (IJGCA) Vol.4, No.1, March 2013

[13] Satyanarayana A, Dr. P. Suresh Varma, Dr. M.V.Rama Sundari, Dr. P Sarada Varma," Performance Analysis of Cloud Computing under Non Homogeneous Conditions", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X.

[14] Mohamed Eisa, E. I. Esedimy, M. Z. Rashad, "Enhancing Cloud Computing Scheduling based on Queuing Models", International Journal of Computer Applications (0975 – 8887) Volume 85 – No - 2, January 2014

[15] Anupama A, G.Satya Keerthi, "Using Queuing theory the performance measures of cloud with infinite servers," Using Queuing theory the performance measures of cloud with infinite servers, ISSN: 2229-3345 Vol. 5 No. 01 Jan 2014.

[16] Lizheng Guo, Tao Yan, Shuguang Zhao, Changyuan Jiang," Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System", Journal of Applied Mathematics, Volume 2014, Article ID 756592, 8 pages

[17] Murugesan, R, Elango C, and Kannan S, " Resource Allocation in Cloud Computing with M/G/s – Queuing Model", Volume 4, Issue 9, September 2014, ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering, PP 443-447.

[18] Murugesan R, Elango C, and Kannan S," Cloud Computing Networks with Poisson Arrival Process-Dynamic Resource Allocation", IOSR Journal of Computer Engineering (IOSR-JCE) e- ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 124-129

[19] Murugesan R, Elango C, and Kannan S, "Resource Allocation in Cloud Computing with General Classification Time and Exponential Service (G/M/s)", International Journal of Engineering And Computer Science ISSN: 2319-7242, Volume 3, Issue 10 October, 2014 Page No. 8905-8910.

[20] Hamid Reza Feili1, Mohsen Momeni Tabar, Navid Akar," Calculating the Number of Optimal Servers in Queue M/M/s/K", International Conference on Nonlinear Modelling & Optimization, 28-29 Aug. 2012, Shomal University, Amol, Iran

[21] Xiaoming Nan, Yifeng He, Ling Guan," Queuing model based resource optimization for multimedia Cloud", J. Vis. Commun. Image R. 25 (2014) 928–942

[22] Amazon EC2 Pricing, http://aws.amazon.com /ec2/pricing/.

[23] L. Breuer, D. Baum "An Introduction to Queueing Theory", Springer Verlag, 2005.

[24] Leonard Kleinrock "QUEUEING SYSTEMS "VOLUME J: THEORY, A Wiley-Interscience Publication

[25] Lec_14_MMsk_Queueing System.pdf.

[26] Donald Gross, John F. Shortle, James M. Thompson, Carl M. Harris "Fundamentals of Queueing Theory", 4th Edition