International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 6487-6492

Performance Analysis of Cloud Computing Bulk Service Using Queueing Models

K. Santhi [1,*] and R. Saravanan [2]

[1,2] School of Information Technology and Engineering,
[1, 2] Vellore Institute of Technology University, Vellore, Tamil Nadu-632014, India.
** Corresponding author*

[1]Orcid ID: 0000-0002- 8038-9604

Abstract

In this paper, we have considered a Markovian bulk-service queue for cloud computing architecture. The multiple class i users from the public cloud entering into the queue are naturally based on the FCFS discipline. Partial batch service mode and full batch service mode have considered for accessing cloud database. The requesting and service of the public cloud has been modelled using queueing theory as a single server bulk service model. If the server is free, each user from the public cloud can enter into the $M/M^{[Y]}/1$ with probability ϕ or leave the system with probability $(1-\phi)$ without accessing the cloud database. Our proposed work is based on public cloud and request management application has been proposed and analysed in terms of waiting time defined as Quality of Service criteria (QoS). The results obtained in our proposed model are calculating the performance measures of different class i of units in the cloud system and also compare the results of both the partial and full batch service model.

Keyword Cloud Computing, $M/M^{[Y]}/1$queue, Quality of Service, Waiting Time

INTRODUCTION

Cloud computing is a novel paradigm for the provision of computing infrastructure, which aims to shift the location of the computing infrastructure to the network in order to reduce the maintenance costs of hardware and software resources. Cloud computing systems vitally provide access to large pools of resources. Resources provided by cloud computing systems hide a great deal of services from the user through virtualization.

Queueing theory is the mathematical understanding of queues or waiting lines. An old-style queueing system may be defined as one having a service ability, at which clients / users arrive for service and when there are more clients / users in the system than the service facility can handle immediately, a queue or waiting line is recognized. The waiting clients / users take their turn for service giving to a pre-assigned rule and leave the system after availing service. Thus, the input to the system consists of the clients demanding service and the output refers to the served clients. Therefore, we model the cloud center as $M/M^{[Y]}/1$ queueing system which indicate that clients are serviced in batches or groups with Poisson arrivals and service are done by exponentially distributed. Thus, we have analysed performance measures such as average waiting time of a client / user waiting for service in both queue and system and the average number of users / clients in the system and in the queue for arbitrary batch size and for constant batch size. Finally, the performance measures of different class i of users / clients of both partial and full batch service model have been shown graphically.

RELATED WORKS

Aattar et al., (2014) have proposed using the model of GE/G/m/m+r based on an analytical model for performance estimation of a cloud computing data centre. They proposed a model by using a generalized exponential arrival process that exposed the nature of arrivals in the cloud with batch Poisson with geometrically scattered batch sizes, a common service time, number of servers and a finite buffer capacity. By using this model they have calculated analytically the performance indicators such as the probability of immediate service, the average of response time, average number of jobs in the system and blocking probability. Mary et al., (2013) have described the model M/G/1:∞/GD for cloud data center as queueing system with only a task arrivals and a task requests buffer of infinite size. Performance of server is calculated by mean as well as standard deviation. The probability of instant service and blocking probability have also been computed.

Sowjanya et al., (2011) have deliberated a model M/M/s model for two servers which increase the performance over using one server by decreasing the queue length and also waiting time. They showed results using M/M/2 reduces queue length and waiting time when compared to M/M/1 and also assurance the QoS requirements of the cloud computing unit's jobs, and also can make the maximum incomes for the cloud computing service provider. Ellens et al.,(2012) have modelled a cloud center by means of the *M/M/c/c* queueing system with dissimilar priority classes. The foremost

performance principle in their analysis is the rejection probability for dissimilar customer classes, which can be analytically determined. Sahoo et al., (2016) have considered cloud centers where tasks attain in batches or groups of arbitrary size and task service times are predictable to follow an exponential distribution. This paper also perceives the cases where the arrival group size has a geometric distribution or a deterministic distribution. They have examined new analytical model for estimation of performance of such large scale systems and calculate the performance standards such as mean waiting time in the queue, mean system length and the mean number of busy servers mean request response time in the system.

Kalyanaraman et al., (2016) have described about Poisson queue with batch arrival and two stages of batch service. They also have considered the server may breakdown, server take Bernoulli vacation and the services are assumed in two stages. For this model, by means of supplementary variable technique, the probability generating function of numbers of customers in the queue at various server states have been attained. Parveen et al., (2013) have considered a general arrival pattern and multiple working vacation period for single server bulk service queue. This model is analysed by using Embedded Markov Chain method. The steady state probability distribution at pre arrival period and arbitrary period are derived and measures similar by mean queue length are calculated.

PROPOSED CLOUD ARCHITECTURE

In this paper, we have considered $M/M^{[Y]}/1$ queueing model for designing cloud infrastructure for different types of clients / users from the public cloud who enter into the system for service using FCFS discipline. If server is free, different class i of clients / users enter with probability ϕ for getting service in batches of size K (access data from the cloud database) and leave the system after service completion. Otherwise, they are to be stayed in queue until their service request is fulfilled or leave the system with probability $(1 - \phi)$ without accessing cloud database. We consider arrivals are different class i each of which follows Poisson distribution with rate λ_i *(i = 1, 2…n)* and single server provide service to batch of users / clients consisting different class *i* of clients / users. The service time of each class i follow exponential distribution with rate μ_i. In service, the clients / users are served (access the cloud database) in batches of size K by two methods one of the methods is called partial batch service model and another method is called full-batch service model.

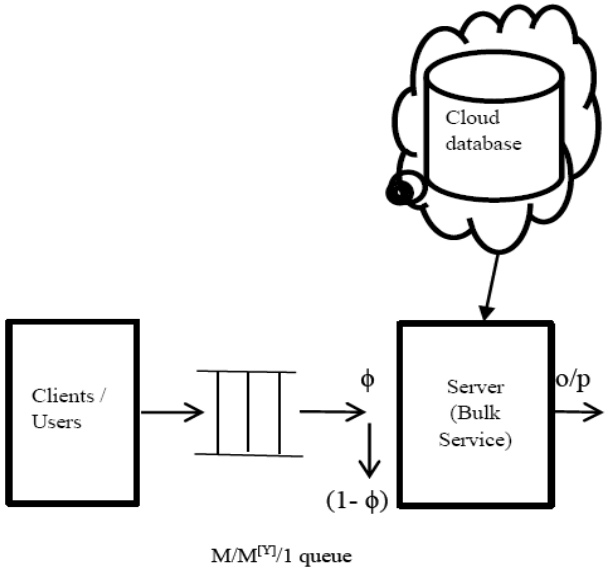

Figure 1: Batch service model in cloud computing architecture

Bulk Service ($M/M^{[Y]}/1$)

We consider a single-server bulk service queue for the cloud environment. We have accepted that clients / users arrive according to Poisson process with rate λ and the service is provided to batch of size K of clients / users whose service times are exponential with mean $1/\mu$ under FCFS criteria, there is no waiting-capacity restriction. We also have considered two kinds of bulk-service model, the full-batch service model and partial-batch service model which enable us how the system operates when the batch size is less than K and the constant batch size K in the system.

Partial Batch Service Model

In this mode of service, the server can start service with at least one client / user and allow up to a maximum size K. Similarly, when server provided service to less than K clients / users, new arrivals immediately enter for service up to the limit K and complete service with the others who are receiving service already, irrespective of the entry time into service. The service time of any batch follow an exponential distribution with mean $1/\mu$. If p_n is the probability that there are n clients / users in the system in steady state and the corresponding probability generating function is

$$P(z) = \sum_{n=0}^{\infty} p_n z^n , \qquad (1)$$

Then we have the following difference equation in steady state:

$$(\lambda\phi+\mu)\,(P(z)-p_0) = \mu z^{-K}\sum_{n=1}^{\infty} p_{n+K}\, z^{n+K} + \lambda\phi z P(z) \quad (2)$$

$$\lambda\phi p_0 = \mu\sum_{n=1}^{K} p_n \quad (3)$$

After simplifying, we have

$$P(z) = \frac{\mu\sum_{n=0}^{K} p_n(1-z^{n-K})}{\lambda\phi(1-z)+\mu(1-z^{-K})} = \frac{\mu\sum_{n=0}^{K} p_n(z^n-z^K)}{\lambda\phi z^{K+1}-(\lambda+\mu)z^K+\mu} \quad (4)$$

If $a_1, a_2, \ldots, a_{K+1}$ are the roots of the characteristic equation $\mu a^{K+1}+(\lambda\phi+\mu)a+\lambda=0$, then

$$p_n = \sum_{i=1}^{K+1} k_i\, a_i^n,\ n\geq 0 \quad (5)$$

Since the total probability is one, $k_i = 0$ for all roots greater than one. So, we have exactly one root a_0 in (0, 1) by using Rouche's theorem. Therefore,

$$p_n = (1-a_0)\,a_0^n,\ n\geq 0,\ \ 0<a_0<1 \quad (6)$$

Thus,

(i) The expected number of clients / users in the system is $L_{PBSS} = \sum_{n=1}^{\infty} n\,p_n = \frac{a_0}{1-a_0}$ (7)

(ii) The expected waiting time of a client / user in the queue is $W_{PBSS} = \frac{L_{PBSS}}{\lambda\phi} = \frac{a_0}{\lambda\phi(1-a_0)}$ (8)

(iii) The expected number of clients / users in the system is $L_{PBSQ} = \frac{a_0}{1-a_0} - \frac{\lambda\phi}{\mu}$ (9)

(iv) The expected waiting time of a client / user in the queue is $W_{PBSQ} = \frac{a_0}{\lambda\phi(1-a_0)} - \frac{1}{\mu}$ (10)

Full Batch Service Model

In this mode of service, if there is exactly K clients / users request service and if the server is free, then they take up for service at a time for accessing cloud database where as if there is less than K clients / users in the system, then the server remains idle till there are 'K' clients / users request service and start the service process at which K clients / users available in the system. The service times is identical for all clients / users in a batch, and follow exponential distribution with mean 1/μ. The steady state balance equations for this model are as follows.

$$(\lambda\phi+\mu)p_n = \lambda\phi p_{n-1}+\mu p_{n+K},\quad n\geq K \quad (11)$$

$$\lambda\phi p_n = \lambda\phi\, p_{n-1}+\mu\, p_{n+K},\ 1\leq n<K \quad (12)$$

$$\lambda\phi p_0 = \mu p_K \quad (13)$$

Multiply equations (8) - (9) by appropriate powers of ' z ' and summing over 1 to ∞, we have

$$(\lambda\phi+\mu)\sum_{n=K}^{\infty} p_n z^n = \lambda\phi\sum_{n=K}^{\infty} p_{n-1}z^n + \mu\sum_{n=K}^{\infty} p_{n+K}z^n$$

$$\lambda\phi\sum_{n=1}^{K-1} p_n z^n = \lambda\phi\sum_{n=1}^{K-1} p_{n-1}z^n + \mu\sum_{n=1}^{K-1} p_{n+K}z^n$$

After simplifying the above equations by using equations (1) and (13), we have

$$\left[\lambda\phi+\mu-\lambda\phi z-\frac{\mu}{z^K}\right]P(z) = \left[\mu-\frac{\mu}{z^K}\right]\sum_{n=0}^{K-1} p_n z^n \quad (14)$$

$$\left[\lambda\phi(1-z)z^K+\mu(z^K-1)\right]P(z) = \mu(z^K-1)\sum_{n=0}^{K-1} p_n z^n \quad (15)$$

$$P(z) = \frac{(1-z^K)\sum_{n=0}^{K-1} p_n z^n}{t\,z^{K+1}-(t+1)z^K+1},\quad t=\frac{\lambda\phi}{\mu} = \frac{(1-z^K)\,(z_0-1)}{K\,(z_0-z)\,(1-z)}, \quad (16)$$

Where z_0 ($z_0 > 1$) is one of the root among (K+1) roots of the characteristic equation $\lambda\phi z^{K+1}+(\lambda\phi+\mu)z^K+\mu=0$ which lies outside the unit circle.

Thus,

(i) The expected number of customers in the system,

$$L_{FBSS} = P'(z)\big|_{z=1} = \frac{2+(z_0-1)(K-1)}{2(z_0-1)} \quad (17)$$

(ii) The expected waiting time of a customer in the system, $W_{FBSS} = \frac{L_{FBSS}}{\lambda\phi}$ (18)

(iii) The expected number of customers in the queue,

$$L_{FBSQ} = L_{FBSS} - \frac{\lambda\phi}{K\mu} \quad (19)$$

(iv) The expected waiting time of a client / user in the queue is $W_{FBSQ} = W_{FBSS} - \frac{1}{\mu}$ (20)

NUMERICAL RESULTS

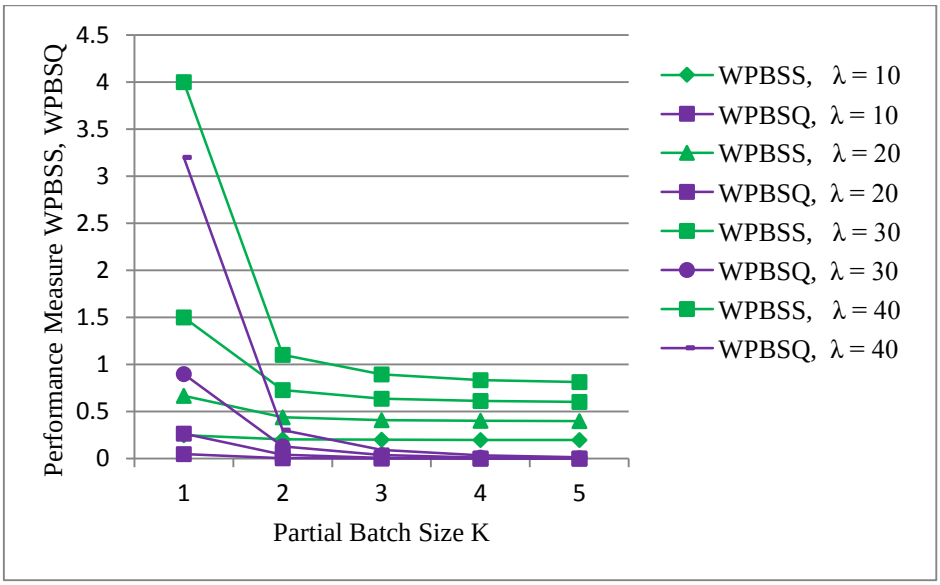

Figure 2: Partial Batch size K Vs Performance Measures L_{PBSS}, L_{PBSQ}

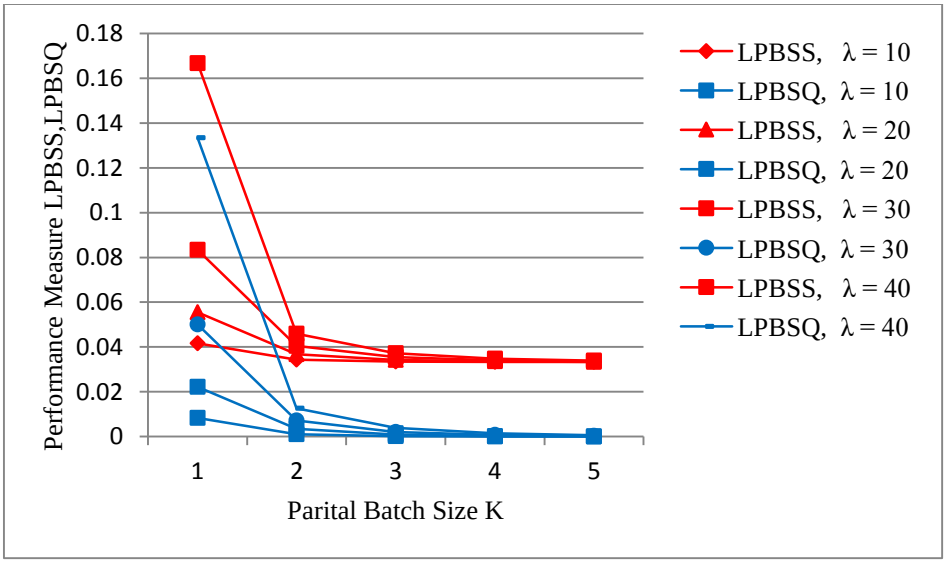

Figure 3: Partial Batch size K Vs Performance Measures W_{PBSS}, W_{PBSQ}

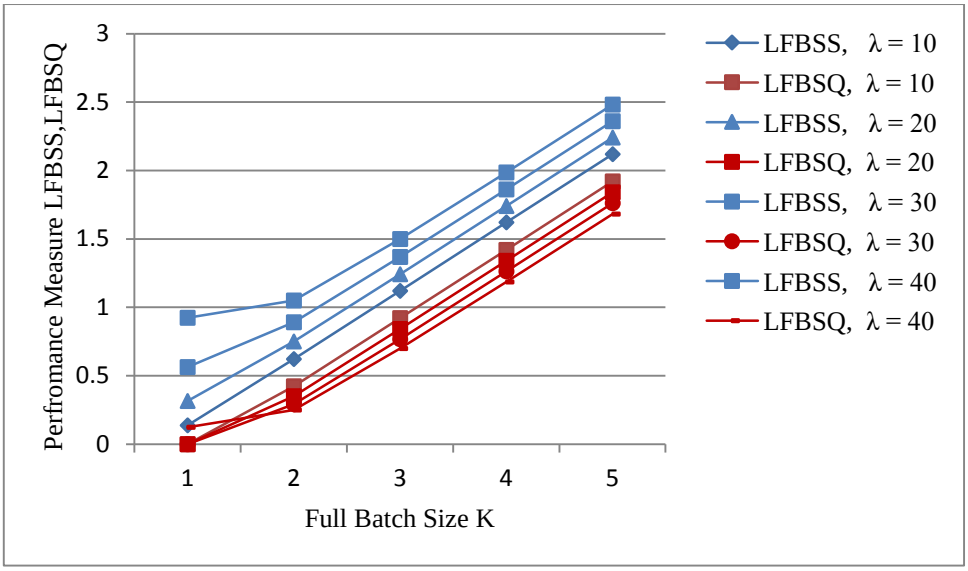

Figure 4: Full Batch Size K Vs Performance Measure L_{FBSS}, L_{FBSQ}

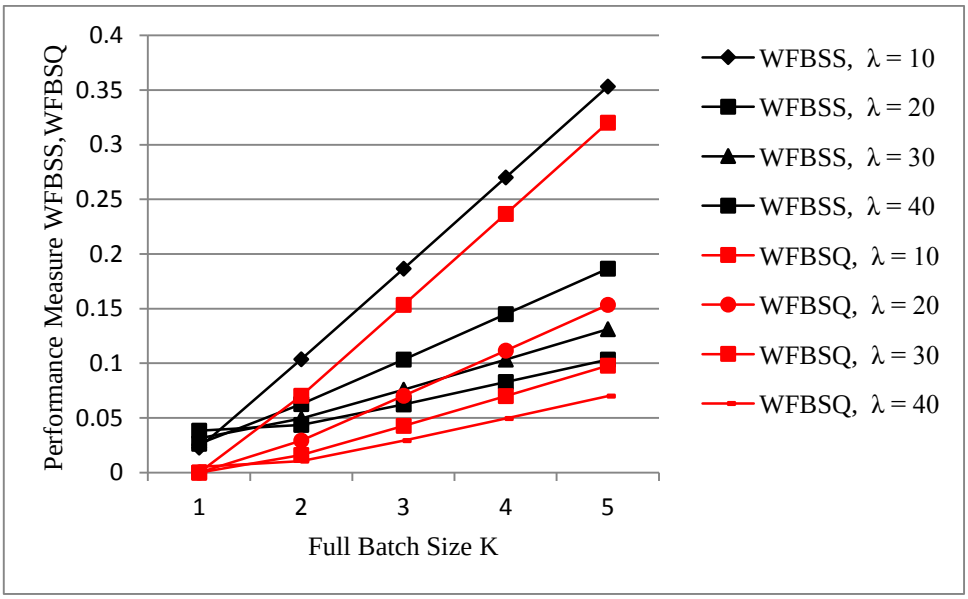

Figure 5: Full Batch Size K Vs Performance Measure W_{FBSS}, W_{FBSQ}

The simulation of the queueing system is done using SHARPE tool. The performance evaluation of the proposed single server bulk service model $M/M^{[Y]}/1$ is analysed for a particular situation based on numerical results in this section. The simulation of this kind is analysed to visualise the Quality of Service metrics such as utilisation of the server of the system and waiting time. We study numerical results for a particular situation in which clients / users arrive with rate λ = 10, 20, 30 and 40. The service rate of partial as well as full batch service model is $\mu = 50$ irrespective batch size K = 1,2,3,4, and 5 and $\phi = 0.6$. Note that during service period of first user, more than one client / user can enter into service station and get service along with the client /user who have already in service and leave the system after service completion simultaneously whereas in full batch service model, batch of size K clients / users can get service and leave the system after service completion. In this case, server can start the service only when there are K clients /users in a batch and will not serve one or two up to (K-1) clients / users. So, comparatively the waiting time of clients / users in full batch service model is larger than the waiting time of clients /users in the partial batch service model. In Figure 2, as the batch size K increases the average number of clients / users in the system (L_{PBSS}) and in the queue (L_{PBSQ}) for partial batch model decreases gradually and also the average waiting time of a client / user in the system (W_{PBSS}) and in the queue

(W_{PBSQ}) for partial batch service model are decreasing gradually as the batch size K increases gradually which is shown in figure 3. Moreover, from figure 4, we have observed that the average number of clients / users in the system (L_{FBSS}) and in the queue (L_{FBSQ}) is decreasing when the full batch size K increases gradually one by one. Also, the average waiting time of client / user in both the system and queue are increasing when full batch size K increases gradually which are shown in figure 5. Furthermore, when K=1 the average number of clients /users and average waiting time of a client / user are decreasing whereas the average number of clients / users and average waiting time of client / user are increasing gradually when $K \geq 2$ which can be observed from figures 2 - 5. Thus, the partial batch service model is reducing waiting time of clients / users and increases the quality of service than the full batch service model.

CONCLUSION

In this paper, we have proposed the model for cloud computing architecture to analyse performance measure such as waiting time of different class i of clients / units from the public cloud who access cloud database in batches of size K. We have considered partial batch service model and full batch service model for accessing cloud database. We simulated our proposed model using SHARPE tool. The results analysed by proposed model shows enhancement in the performance of the cloud queueing system. The performance of the cloud system is measured by using the total waiting time and utilisation of resources. Simulation results showed that the significant variation between the total waiting time and the total number of users / clients in the system for different arrival rates and for different batch sizes. We are also going to study the above model for the general service time distribution and server's vacation for our future work.

REFERENCE

[1] Vaquero L.M, Merino R.L, Caceres. J and Lindner M. 2009. A break in the clouds: Towards a cloud definition. ACM SIGCOMM Computer Communication Review. Vol.39, pp 50–55.

[2] Mohamed Ben aattar. 2014. Performance Modeling for a Cloud Computing Center Using GE/G/m/k Queuing System International Journal of Science and Research (IJSR),Vol 3, pp 783-789.

[3] Ani Brown Mary. N and Saravanan .K. 2013. Performance Factors of ClouComputing Data Centers Using [(M/G/1):(∞/GDModel)] Queuing Systems International Journal of Grid Computing & Applications (IJGCA). Vol.4, pp 1-10

[4] Sai Sowjanya .T, Praveen .D, Satish.K , Rahiman. 2011. A The Queueing Theory in Cloud Computing to Reduce the Waiting Time IJCSET, Vol 1,pp.110-112

[5] Ellens. W, Zivkovi .M, Akkerboom .J, Remco Litjens, 2012. Performance of Cloud Computing Centers with Multiple Priority Classes 2012 IEEE Fifth International Conference on Cloud Computing.

[6] Sahoo, C.N. and Goswami, V. 2016. Performance Evaluation of Cloud Centers with High Degree of Virtualization to provide MapReduce as Service. International Journal of Advances in Soft Computing & Its Applications, Vol.8, pp 193-203.

[7] Kalyanaraman. R, Nagarajan. P. 2016. Bulk arrival, fixed batch service queue with unreliable server, Bernoulli vacation, Two stages of service and with Delay time, International Journal of Mathematics Trends and Technology (IJMTT) Vol.38, pp 1-8

[8] Jemila Parveen .M and Afthab Begum I. 2013. General Bulk Service Queueing ystem with Multiple Working Vacation, International Journal of Mathematics Trends and Technology- Vol.4, pp.163-173

[9] Umamakeswari. A, Vijalakshmi. N and Renugadevi .T.2012. Storage and Retrieval of medical images using cloud computing Journal of Artifical Intelligence. Vol.5, pp. 207-213.

[10] Ma .N, Mark. J.1998. Approximation of the mean queue length of an M/G/c queueing system. Oper Res, Vol 43, pp. 158–165.

[11] Gross . D, Harris. C, 2014. Fundamental of queueing theory, John Wiley & Sons, Fourth edition.

[12] Santhi. K, Saravanan. R. 2016. A survey on queueing models for cloud computing International Journal of Pharmacy & Technology Vol. 8, pp. 3964-3977.

[13] Santhi K, Saravanan R, 2017.Performance Analysis of Cloud Computing in Healthcare System Using Tandem Queues International Journal of Intelligent Engineering and Systems, Vol.10, pp. 256-264.

[14] Kleinrock. L.1975. Queueing Systems: Theory. Vol.1. A Wiley- Interscience, New York.

[15] Trivedi K.S and Sahner. R. 2009. SHARPE at the age of twenty two ACM SIGMETRICS Performance Evaluation Review., Vol.36, pp 52-57.

[16] Xiong, and Perros. H. 2009. Service performance and analysis in cloud computing. IEEE Computer Society, pp 693–700.

[17] Arokia Muthu. 2016. Performance Analysis of Cloud Computing Centers using M/G/m=m+ r Queuing Systems. International Journal of Research in engineering, science and technologies. Vol 2, pp 1-16.

[18] Sorina Lupşe .Marcella.O. Vida .M and Stoicu-Tivadar. L. 2012. Cloud Computing and Interoperability in Healthcare Information System. The First International Conference on Intelligent Systems and Applications. Vol. 2, pp. 81-85.