Received: November 24, 2016

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Performance Analysis of Cloud Computing in Healthcare System Using Tandem Queues

Santhi Kannan[1]* **Saravanan Ramakrishnan[1]**

[1] *School of Information Technology and Engineering,*
Vellore Institute of Technology University, Vellore, India
* Corresponding author's Email: ksanthi@vit.ac.in

Abstract: Cloud technology is broadly supported by healthcare organizations worldwide. Throughout the previous few years, healthcare industries have recognized the potential of cloud computing architecture and how it can assistance them to give quality services to patients. The cloud technology has faster the way healthcare industry can use or share information crossways a network. This paper offering an e-health solution founded on healthcare information systems joined with cloud computing concept emphasis on medical information processing. Our proposed work 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 service of the public cloud has been modelled using queueing theory as two serially connected queues (Tandem Queues) M/M/s and single server retrial queue (SSRQ). The results obtained in our proposed model are considerably reducing the total waiting time of different class of units in the cloud system than the existing system and more utilization of service in cloud.

Keywords: M/M/s queue, orbit, retrial queue, cloud computing, e-Healthcare, Quality of Service, Waiting time.

1. Introduction

Cloud computing is an advanced model for the conveyance of computing structure, which aims to change the setting of the computing infrastructure to the grid to diminution the costs of administration and conservation of hardware and software assets [1].Cloud computing has been used to permit communication anytime and everywhere. This new pattern forms new chances to share information that can be continuously available. There is an opportunity to exchange information between the medical devices inside the institutions with another device located in another institution. By moving the infrastructure to the cloud, valued data extracted from the dissimilar databases of treatment, patients, diseases, and so on, it will be convenient to doctors to accomplish analytical studies and realize statistical outcomes. Second system provides service access to a database server. They offered that to provide the most precarious area that needs a lot of information, data and computing power is the healthcare field. Doctors need, in critical moments, the medical history of patients in real time. Patients are sent to various investigations, supposing a high rate exchange of data between departments of medical units. Doctors need complete medical information of the patients to provide a complete and accurate treatment.

Healthcare industry has come an extended method from Hospital Information Systems (HIS), Electronic Medical Records (EMR) to computer aided surgeries and remote patient care, since the advent of information technology into the healthcare domain. They examine brief some of the digital data challenges that the healthcare industry is facing and also they have shown the system that capable of offering various healthcare services that utilize cloud computing [2].

Cloud computing permits self-motivated sharing of the computing resources among the users. A service level agreement (SLA) agrees the quality of service (QoS) be providing to the user concerning various performance parameters such as throughput,

consistency, blocking probability and response time Cloud computing has increased worldwide attention from the various researchers, however, only a small portion of them have instructed the QoS performance problem.

We use cloud computing and queueing theory to model the problem of cloud resource utilization, reliability and scaling. By modelling the queueing system, we aim to provide scalability to the cloud infrastructure running on a given virtualized platform. Our work focussed on modelling QoS performance based on more resources utilisation service, Data/Service Reliability, Scalability (QoS), Flexibility and Interoperability in the cloud system in e-Health cloud platforms. The proposed work we concentrate on study the cloud architecture using queueing model is to analyse the performance measure such as waiting time of different class of units from the public cloud until they leave the system after service completion. Arrivals from the public cloud can get service(registration, checking, consultation etc.,) from the first queue called M/M/s queue and enter into the retrial queue (SSRQ) for accessing from the cloud data base if the server is free and leave the system after service completion. Otherwise, those arrivals enter into the orbit called retrial group with probability ϕ for accessing cloud database after some random amount of time or leave the system with probability $(1 - \phi)$.The main advantage of this paper is to reducing waiting time of different class of units, increase resource utilization services in the healthcare cloud system over the other existing method.

The rest of this paper is organised as follows. In section 2, we confer the related work. The cloud architecture of the system with stability conditions, notations and the operating characteristics (waiting time of each class in the first queue and in the orbit) are obtained in section 3 and in section 4, we study numerical results for a particular situation and give comparison between existing work and proposed work both numerically and graphically. Finally, we present the conclusion in section 5.

2. Related Works

A cloud structure is involved in a network of computer servers that are obtainable in demand as a service, and they are deliberated to be scalable and flexible. H. Khazaei et al. [3] have proposed an innovative estimated analytical model for performance evaluation of cloud server farms using queueing model and solve it to obtain an accurate estimation of the complete probability distribution of the request response time and additional significant performance indicators. Furthermore, they have discussed the model in which cloud operators to conclude the relationship between quantity of servers and input buffer size. They have obtained the performance measures such as a mean number of jobs in the system, blocking probability, and the probability that a job will find immediate service. Finally, they showed the results that a cloud center accommodates heterogeneous services may execute longer waiting time for its clients compared to its homogeneous equal with the same traffic intensity.

K. Xiong et. al [4] have focussed on queueing network model for learning the performance of computer services in cloud computing and then established an approximation method for computing a response time distribution using Laplace transform in the cloud computing system. They have shown the calculation of cumulative distributions of the response time, and the maximal number of customers for given computer service resources in cloud computing in which customer services can be sure in the term of the percentile of response time.

O. Sorina Lupş et al. [5] have focussed on cloud computing technology in medical act may considerably improve the contact information, which can be done much easier. Using current cloud technology the availability of a physician, a medical specialist, a product or a service at dissimilar times and in diverse cases can be checked. Patients can be directed to fit persons or units can discover what they need. This is a huge benefit for patients and health professionals for increasing the quality of the medical service. They have constructed private cloud solution ensured the security of data and communication among departments and messaging through in a secure way. Umamakeswari et al. [6] have described healthcare organisation petabytes of data in terminologies of the patient record, medical images and lab outcomes. The main importance of the healthcare organisation is about the spending of an onsite medical image database. They formed up a medical image archive result using window Azure cloud and SQL Azure. They have shown results as decreases the management cost and also defends the data from disaster recovery. Furthermore, they also concentrate on providing security for the healthcare IT sector which has been attained by storing the images on the digital imaging and communications in medicine server. N. Ani Brown Mary et al. [7] have described about cloud data center is modelled [(M/G/1) : (∞/ GDMODEL)] as queueing system for the single task arrivals and task request buffer of infinite capacity. Based on file size that was selected in byte value, the waiting time and response time of

the user were calculated. They also have shown that the response time is more when compared with the waiting time. N. Ma et al. [8] have focused on G/M/m, M/G/m and G/G/m queueing models where the time between arrivals and service time do not follow an exponential distribution which are much more complex. Many theoretical studies have been based on widespread research in performance evaluation, including those that investigated the M/G/m model. Furthermore, they found the complexity in these cases originates from the impossibility of finding a closed formula to represent the probability distributions of the response or waiting time of customers in the queue, and so requires outcome approximate models.

R.Marcu et al. [9] considered the model M/M/s and M/M/1 queue in series in which different priority clients / patients / users obtained service from the first queue and enter into the second queue to get service from the server with probability p or directly access the database with probability (1– p). Furthermore, the authors have considered an e-health solution based on healthcare information systems combination and cloud computing idea with a focus on medical imaging department. They also considered Hybrid cloud architecture and request management application which have been analysed in terms of waiting time well-defined as Quality of Service criteria (QoS). The service has been demonstrated by means of queueing theory as two serially connected queues M/M/s and M/M/1. Furthermore, they have obtained waiting time of a class of units or patients in both M/M/s queue and M/M/1 queue and total waiting time for this model.

J. Vilaplana et al. [10] have discussed the queueing model consisting two systems M/M/m in series for the cloud e-Health platform. The first system offers compute services, and the good QoS, in terms of average waiting times. The proposed system can develop the e-Health field by providing a model to support medical software, saving resources and improving the control and management of the patients and also shown the results significant performance improvement when the number of servers increases. Several authors studied cloud computing using queueing models for more details one can refer [11, 12, 13].

J. Vilaplana et al. [14] have described the model for planning cloud computing architectures with good QoS based on queuing theory and the open Jackson's networks were designated as the basic means to promise a certain level of performance in line with the waiting and response times of such networks.

S. Addamani et al. [15] have focused on a closed queuing model for performance analysis of cloud computing platforms. They considered model as multiple queues and the virtual machines VMs are modelled as service centers. They have used software monitoring to measure model parameters on the two web applications and the model was studied by using JMT tool. The difference between the response time with the number of users and with the number of instances (or Virtual Machines) were calculated and the results can be used to determine the optimal number of users and the number of VMs for a desirable value of the response time.

The drawback of the existing method is the patients / users can access the cloud database directly without medical staff (the server of the second queue) assistance. But, if the patients / users are unaware about how to access the cloud database, they expect someone assistance. In such case the personal information about the patients / users can be viewed by unknown accessor. To avoid this kind of critical situation, we allow the patients / users to access cloud database only through the authenticated medical staff (ie.,server in SSRQ) to secure information about the patients in our proposed method. Also, we observe that the resource utilisation (server utilisation) in the proposed method is more than the existing method. From the above analysis, the performance evaluation is carried out using queueing model in cloud computing. Several authors have addressed performance issues using a mathematical model in cloud computing among them there are few work which addresses the more utilisation service, Data/Service Reliability, Scalability (QoS), Flexibility and Interoperability in the cloud system. For more details one can refer [16, 17, and 18].

3. Proposed Work

Our proposed method refers figure 1 which shows multiuser from the public cloud enter into a healthcare system consisting two serially connected queue *M/M/s* queue and *M/M/1* retrial queue (SSRQ) in which we schedule using FCFS discipline and send to server for service in cloud computing infrastructure. We assume that after completing service from the first queue called M/M/s queue, users enter into second queue called *M/M/1* retrial queue (SSRQ). We consider arrivals are different class i each of which follows a Poisson distribution with rate λi *(i = 1,2…n)* and there are '*s'* parallel servers they provide service to each class *i*. The service time of each class follows an exponential distribution with rate μ_i. After

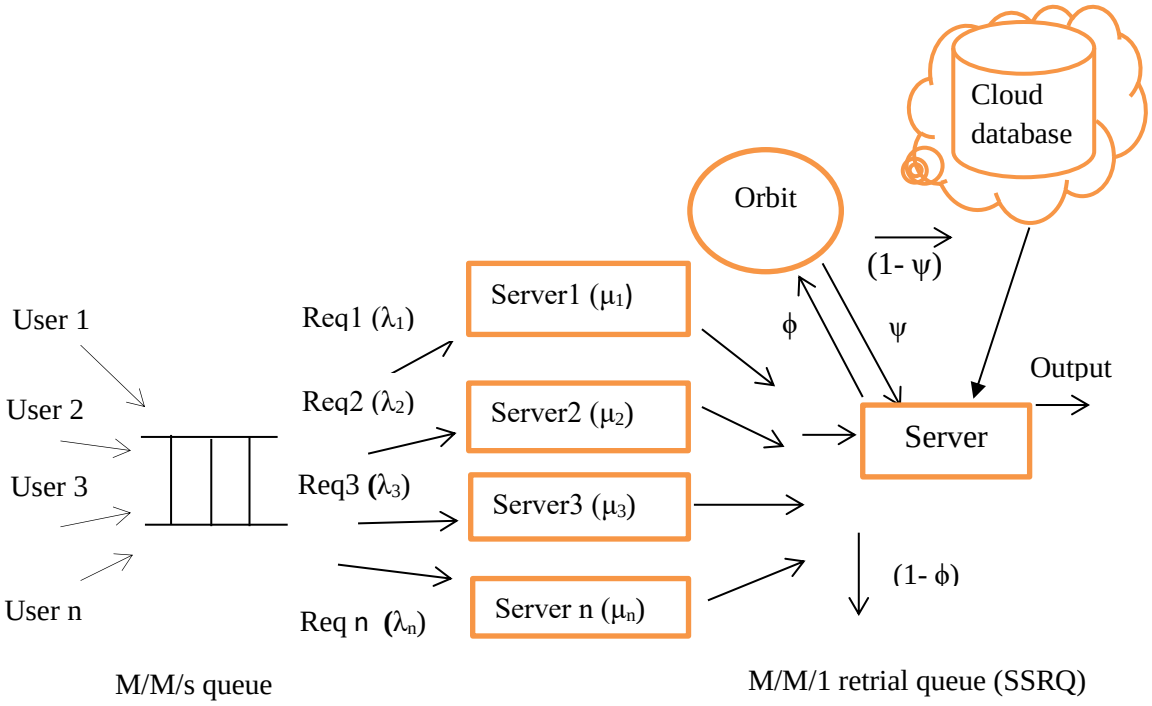

Figure.1 Queues in Series for Cloud Computing Architecture

completing service from *M/M/s* system, each class enters into SSRQ for service [19].If the server is free, arriving each class *i* occupies the server and leaves the system after completing service. If each class *i* finds server busy on its arrival, it leaves the system forever with probability *(1 - ϕ)* or enters into the orbit with probability *ϕ* to retry for service after a random amount of time. When the repeating class *i* from the orbit retries *j th* times and finds the server busy return to the orbit with probability *ψ* in order to retry *(j+1)th* times again for service after some random amount of time or leave the system always with probability *(1 - ψ)*. This type of retrial queue is called retrial queue with geometric loss [20].The retrial time, the time between two consecutive requests made by the class with rate *α* and independent of all previous retrial time and all another request in the system. In this paper, we use queueing models for cloud computing architecture which is applied in healthcare centres. After completed service (primary service such as appointment, consultation, physical test etc.,) from *M/M/s* queue usually, clients / patients / users request medical staff for accessing database from the cloud if the medical staff (server) is free. Otherwise, they enter into the orbit (buffer) with probability *ϕ* in order to seek service again after some random amount of time with retrial rate α or leave the system with *(1 - ϕ)*. When the repeating units from the orbit retries j^{th} times and finds the server busy return to the orbit with probability ψ in order to retry $(j+1)^{th}$ times again for service after some random amount of time or leave the system always with probability *(1 - ψ)*. In the existing model described in [10], after completing service from *M/M/s* queue usually, clients / patients / users request medical staff for accessing cloud database with probability *'p'* and leave the entire system after service completion or users / patients themselves directly access the cloud database with probability *(1-p)* and leave the entire system. But, the situations in which the users / clients / patients cannot access the database without medical staff assistance due to not knowing how to access directly or service not necessary from the second queue which were not considered. So, if the server is busy they enter into the orbit (buffer) with probability *ϕ* in order to request service once again after some random amount of time with retrial rate α or leave the system with probability *(1 - ϕ)* (this also considered for those patients who are not requiring service from SSRQ).

According to Kendall notation for queues, we study the model of a cloud system collected by two serially connected queues *M /M/s* entry queue and *M/M/1* retrial queue (SSRQ). We use the following notations for describing system.

For this model, we obtain waiting time of each class in *M/M/s* queue as well as in orbit for the case $0 \le \phi \le 1$ and $\psi = 1$. The system is stable if

$$\rho_1 + \rho_2 < 1 \qquad (1)$$

Table 1. Nomenclature

Notation	Meaning
λ_i	Arrival rate, i=1,2,…n
μ_i	Service rate, i= 1,2…n
α	Retrial rate
ϕ	Probability that units enter into the retrial queue
ψ	Probability that units from the orbit requesting service at the j^{th} attempt.
ρ	Utilization factor
ρ_1	M/M/s Utilization factor
ρ_2	SSRQ Utilization factor
π_0	Probability that the server is idle (free)
T_0	The amount of time until one of the 's' server
$W^{(i)}_{M/M/s}$	The average waiting time of units in the first queue
W_{SSRQ}	The average waiting time of units in the orbit
W_{TWT}	Expected mean total waiting time
W_{Q2}	The average waiting time units in the second queue

Where, $\rho_1 = \rho_k = \frac{\lambda_k}{s\mu}$ and $\rho_2 = \frac{\lambda_k}{\mu}$, $1 \le k \le r$

In the first queue, arrivals are Poisson with rate λ_i and multiple s servers. The service time for all r priority classes are independent and identically distributed with rate μ.

$$\mu = \mu_1 = \ldots = \mu_r$$

Let $$\sigma_k = \sum_{i=1}^{k} \rho_i \quad (2)$$

With $\sigma_0 = 0$ and $\sigma_r \equiv \rho = \lambda / s\mu$

And $$W^{(i)}_{M/M/s} = \sum_{k=1}^{i-1} E[T_k'] + \sum_{k=1}^{i} E[T_k] + E[T_0] \quad (3)$$

Where,

T_k - the required time to serve k^{th} priority patient / client / unit in the queue.

$T_k^{'}$ - the service time of the patient /client / unit during the service time of k^{th} priority unit S_0 – the amount of time, remaining until the next server become available.

$$\sum_{n=s}^{\infty} \pi_n = \pi_0 \sum_{n=s}^{\infty} \frac{(s\rho)^n}{s^{n-s}\, s!} = \pi_0 \frac{(s\rho)^s}{s!(1-\rho)} \quad (4)$$

Where $$\pi_0 = \left(\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{(s\rho)^s}{s!(1-\rho)} \right)^{-1} \quad (5)$$

and $E[T_0|$ all channels busy] = $1/s\mu$. From the memoryless property of the exponential distribution

We have $$E[T_0] = \frac{(s\rho)^s}{s!(1-\rho)(s\mu)} \pi_0 \quad (6)$$

Therefore, the average waiting time i^{th} class patient / client / user in the first queue is

$$W^{(i)}_{M/M/s} = \frac{E[T_0]}{(1-\sigma_{i-1})(1-\sigma_i)} \quad (7)$$

Moreover, *M/M/1* retrial queue is considered as the second queue in which Poisson arrivals (λ) enter for an accessing database when the server is free, otherwise, arrivals enter into the orbit with probability ϕ and request for accessing the database after some random amount of time. Service time and retrial time follow an exponential distribution with rates μ and α respectively. By using the concept of retrial queueing system [15], we will derive the average time spent in the orbit W_0. Let $\pi_{i,j}$ be the probability that the system is in the state (i, j), Where $i \in \{0,1\}$ is the number of patients / clients / units in service and $n \in \{0, 1, 2, ...\}$ is the number of patients / clients / units in orbit, then the partial generating functions are defined as

$$\Pi_0(z) = \sum_{j=0}^{\infty} \pi_{0,j} z^j, \quad \Pi_1(z) = \sum_{j=0}^{\infty} \pi_{1,j} z^j \quad (8)$$

Using the above partial generating functions, the steady state differential equations are obtained as

$$\lambda_i \Pi_0(z) + \alpha\, z\, \Pi_0'(z) = \mu\, \Pi_1(z) \quad (9)$$

$$(\phi\lambda_i + \mu)\Pi_1(z) = \lambda_i \Pi_0(z) + \phi\lambda_i z\, \Pi_1(z) + \alpha\, \Pi_0'(z) \quad (10)$$

On solving (7) and (8), we have

$$\Pi_0(z) = \frac{(1-\phi\rho_2 z)}{(1+(1-\phi)\rho_2)} \left(\frac{1-\phi\rho_2}{1-\phi\rho_2 z} \right)^{\frac{\lambda_i}{\alpha}+1} \quad (11)$$

$$\Pi_1(z) = \frac{\rho_2}{(1+(1-\phi)\rho_2)}\left(\frac{1-\phi\rho_2}{1-\phi\rho_2 z}\right)^{\frac{\lambda_i}{\alpha}+1} \quad (12)$$

Let L_{SSRQ} be the expected number of patients / clients / units in the orbit, then

$$L_{SSRQ} = \Pi_0'(1) + \Pi_1'(1) \quad (13)$$

Therefore, the average time spent by the units in the orbit is

$$W_{SSRQ} = W_{Q2} = \frac{L_{SSRQ}}{\phi\lambda_i} = \frac{\lambda_i}{(\mu - \phi\lambda_i)(\mu + (1-\phi)\lambda_i)} X \left\{\frac{(\alpha + \mu + (1-\phi)\lambda_i)}{\alpha}\right\} \quad (14)$$

Thus, the total expected time wait in the queue is

$$W_{TWT} = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_{M/M/s}^{(i)} + W_{SSRQ} \quad (15)$$

Particular cases

(i) If $\phi = 1$ and $\lambda_i = \lambda$, W_{SSRQ} coincides with the time spent in orbit in classical *M/M/1* retrial queue considered by Falin and Templeton [19].

(ii) If the retrial rate is sufficiently large, the second factor in W_{SSRQ} tends to 1 and this model coincides with the model considered by [20].

4. Numerical results

The simulation of the queuing system is done using SHARPE tool. The performance evaluation the proposed model (SSRQ) is analysed. Based on numerical results are studied for a particular situation in this section. The simulation of this kind is analysed to visualise the Quality of Service metrics such as utilisation of the system and waiting time. The proposed model is compared with an existing model which is shown as result as more utilisation of service and reducing priority), $\lambda_2 = 2\lambda_1$ requests/s (medium level priority) and $\lambda_3 = 3\lambda_1$ request/s (low level priority) for class1, class2 and class3 priority units waiting e. We assume that arrival rate of the request in the s the system $\lambda_1 = 7$ requests/s (high level respectively and service time $\mu = 28$ for M/M/s queue and $\mu = 140$ for M/M/1 queue with retrial rates $\alpha = 5$ request/s for all class of patients/customers. Figure (2) shows that as the number servers (s) increase gradually, the total waiting time (W_{TWT}) of the system decreases gradually. Moreover, we observe that the considerable variation in the total waiting time between proposed method high level priority (class 1) units and the existing method high level priority (class 1) units whereas the total waiting time of medium level priority (class 2) in the proposed method is smaller than the total waiting time of medium level priority (class 2) in the existing method. Also, the total waiting time of low level priority (class 3) units in the existing method is higher than the waiting time of low level priority (class 3) units in the proposed method. So, comparatively the total waiting time of our proposed method (PM) is considerably lower than the existing method (EM) [19]. The comparison between the total waiting time of different class of patients both in existing work and in proposed work is given in the following table (3) for the particular parameters and servers s =1,2…8.

We observe that as the probability (ϕ) for the second queue (SSRQ) increases gradually, the waiting time in the orbit increases gradually from figure (3). Furthermore, there is no much more difference between high level priority class1 units in both proposed method and existing method. But, the increase in waiting time of medium level priority class 2 units in the proposed method after the probability $\phi = 0.5$ is smaller than the waiting time of medium level priority class 2 units in the existing method and the similar observation happened between the low level priority class 3 units in both proposed method and existing method after the probability $\phi = 0.6$ due to the fixed retrial rate $\alpha = 200$. Note that if the retrial rate α is keep on increasing, the waiting time of all class of units in the proposed method is considerably smaller than the waiting time of all class of units in the existing method which shown in figure 4 and figure 5 for $\alpha = 1000$ and $\alpha = 2000$.

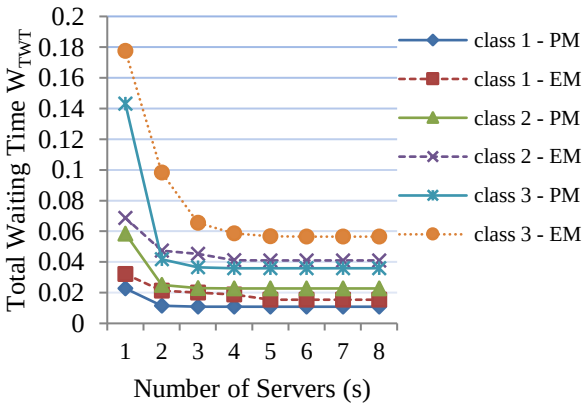

Figure.2 Number of servers Vs Total Waiting Time

Table 2. Comparison between Existing method and proposed method for scientific data

Work carried out	Model	Tools	Analytical model/ Numerical model	Quality of service factors	Metrics of QOS achieved
Performance Analysis of Cloud Computing in Healthcare Systems Using Tandem Queues (Proposed Method)	M/M/s and retrial M/M/1	SHARPE	Numerical model	Steady state	Efficiency, resource utilization, waiting time, security and lower cost.
Healthcare Integration based on Cloud Computing	M/M/s and M/M/1	DICOM and HL7	Numerical Model	Steady state	Efficiency, cost and security.
The cloud paradigm applied to e-Health	Two M/M/m	Open Stack	Analytical model	Steady state	Scalable, flexible, waiting time.
A Queuing theory model for cloud computing	M/M/1 and M/M/m	Sage 5.3 Mathematical software	Analytical Model	Steady state	Waiting time and response time.

Table 3. Comparison between Existing method and Proposed method

Parameters						Total Waiting Time (W_{TWT})	
λ_i	μ_1	μ_2	p	α	s	EM	PM
7	28	140	0.9	20	1	0.0321	0.0227
					2	0.0212	0.0114
					3	0.0200	0.0109
					4	0.0189	0.0108
					5	0.0154	0.0108
					6	0.0154	0.0108
					7	0.0154	0.0108
					8	0.0154	0.0108
14	28	140	0.9	20	1	0.0687	0.0585
					2	0.0473	0.0251
					3	0.0452	0.0230
					4	0.0411	0.0228
					5	0.041	0.0227
					6	0.041	0.0227
					7	0.041	0.0227
					8	0.041	0.0227
21	28	140	0.9	20	1	0.1775	0.1430
					2	0.0983	0.0417
					3	0.0654	0.0366
					4	0.0587	0.0360
					5	0.0567	0.0359
					6	0.0565	0.0359
					7	0.0565	0.0359
					8	0.0565	0.0359

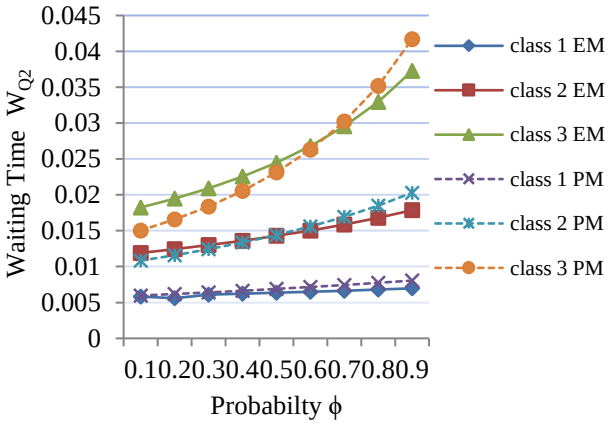

Figure.3 Waiting Time Comparison between Existing Method and Proposed Method for different Probability ϕ and α = 200

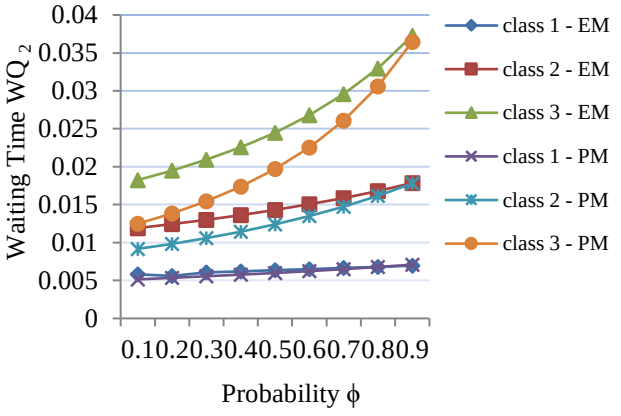

Figure.4 Waiting Time Comparison between Existing Method and Proposed Method for different Probability ϕ and α = 1000

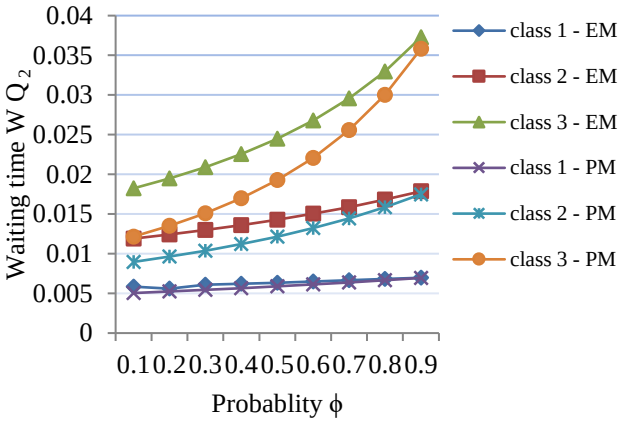

Figure.5 Waiting Time Comparison between Existing Method and Proposed Method for different Probability ϕ and α = 2000.

5. Conclusion

In this paper, we have proposed the model for the cloud computing architecture which is applied in the healthcare centres to analyse performance measure such as waiting time of different class of units from the public cloud who access 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 is measured by using the total waiting time, utilisation of resources and cost .Simulation results shows the fact that total waiting time rate has decreased by 25% compared to the existing model and increases the resource utilisation effectively. We are also going to study the above model for the general service time distribution and server's vacation for our future work.

Acknowledgments

Authors would like to thank the referees for their valuable comments and suggestions to improve the article.

References

[1] L. M. Vaquero, L. Rodero -Merino, J. Caceres and M. Lindner, "A break in the clouds: Towards a cloud definition, "*ACM SIGCOMM Computer Communication Review",* Vol. 39, No. 1, pp. 50–55, 2009.

[2] N. John and S. Shenoy, "Health Cloud - Healthcare As service (HaaS)", In: *Proc. of International Conf. on Advances in Computing, Communications and Informatics (ICACCI),* pp.1963-1966, 2014.

[3] H. Khazaei, J. Misic, and V.B. Misic, "Performance Analysis of Cloud Computing Centers using M/G/m/m+r Queuing System", *IEEE transactions on parallel and distributed systems*, Vol. 23, no. 5, pp.936-943, 2012.

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

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

[6] A. Umamakeswari, N. Vijalakshmi and T. Renugadevi, "Storage and Retrieval of medical images using cloud computing", J*ournal of Artifical Intelligence*, Vol.5, No.4, pp.207-213, 2012.

[7] N. Ani Brown Mary and K. Saravanan "Performance Factors of Cloud Computing DataCentersUsing [(M/G/1) :(∞/GDModel)] Queuing Systems", *International Journal of Grid Computing & Applications (IJGCA)* ,Vol.4, No.1, pp.1-10, 2013.

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

[9] R. Marcu, D. Popescu, and J. Danila, "Healthcare Integration Based On Cloud Computing", *U.P.B. Sci. Bull.,* Vol.77, No. 2, pp.31-42, 2015.

[10] J. Vilaplana, F. Solsona, F. Abella, R. Filgueira and J. Rius, " The Cloud Paradigm Applied to e-Health", *BMC Medical Informatics and Decision Making* , pp.1-10, 2013.

[11] L. Wang, G. VonLaszewski and A. Younge , "Cloud computing: a perspective study," *New Generation Computing*, Vol. 28, No. 2, pp. 137–146, 2010.

[12] T. Anthony, J. Toby and R. Elsenpeter, "*Cloud Computing: A Practical Approach",* Tata McGraw-Hill, New York, 2010.

[13] V. Goswami and S. S. Patra and G. B. Mund, "Performance Analysis of Cloud with Queue-Dependent Virtual Machine", 1st*International Conference on Recent Advance in Information Technology | RAIT*, pp. 357-362, 2012

[14] J. Vilaplana, I. Teixido. J. Mateeo, F. Abella and J. Rius,"A Queuing theory model for cloud computing", *The Journal of Supercomputing*, Vol 69, pp 492-507, 2014.

[15] K.S. Addamani and A. Basu, "Performance Analysis of Cloud Computing Platform**",**

International Journal of Applied Information Systems (IJAIS), Vol.4, No.4, pp. 30-33, 2012.

[16] P. Suresh Varma. A. Satyanarayana and M. V. Rama Sundari, "Performance analysis of Cloud Computing using Queueing Models", In: *Proc. of International Conf. on Cloud Computing Technique Application and Management, IEEE*, pp.12-15, 2012.

[17] D. Mani and A. Mahendran, "Availability Modelling of Fault Tolerant Cloud Computing System", *International Journal of Intelligent Engineering and Systems*, Vol.10, No.1, pp. 154-165, 2017.

[18] F. Cătălin, "Stochastic Processes and Queueing Theory used in Cloud Computer Performance Simulations", *Database Systems* , Vol. 4, No. 2, pp. 56-62, 2015.

[19] G. I. Falin and J. G. C. Templeton, "*Retrial queues*", Chapman and Hall, London. 1997.

[20] B. D. Choi and Y. Chang, "Single Server Retrial Queues with Priority Calls", *Mathematical and computer modelling*, Vol.30, pp.7-32, 1999.

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

[22] M. LawanyaShi, B. Balusamy, S. Subha, "Threshold-Based Workload Control for an Under-Utilized Virtual Machine in Cloud Computing", *International Journal of Intelligent Engineering and Systems*, Vol.9, No.4, pp. 234-241, 2016.

[23] K. Santhi, R. Saravanan, "A survey on queueing models for cloud computing", *International Journal of Pharmacy & Technology",* Vol.8, No.2, pp.3964-3977, 2016.

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

[25] K.S. Trivedi and R. Sahner, " SHARPE at the age of twenty two", *ACM SIGMETRICS Performance Evaluation Review* , Vol. 36, No. 4, pp. 52-57, 2009.