ISSN: 0975-766X
CODEN: IJPTFI
Available through Online *Review Article*
www.ijptonline.com

A SURVEYON QUEUEING MODELS FOR CLOUD COMPUTING

K.Santhi*, Dr. R. Saravanan
VIT University, Vellore, Tamilnadu, IndiaVIT University, Vellore, Tamilnadu, India.
Emails: ksanthi@vit.ac.in

Received on 12-05-2016 *Accepted on 02-06-2016*

Abstract

Cloud computing has developed as best way of sharing and providing resources over the internet. The capability to deliver guaranteed Quality of Service is vital for commercial success of cloud platforms. The process of entering into the cloud is usually in the form of queue, so each user need to wait until the current user is being served. In this paper, we have discussed about several queueing model for cloud computing. These models are used to reduce waiting time of customer(calls) and increase performance of the system. Furthermore, we have presented comparison of several queueing models results which are used for cloud computing environment.

Keywords: QoS, SLA, Queueing models, Erlang distribution, resources pool(orbit, buffer),Stochastic Reward Nets, Sojourn time, Rejection probability.

1. Introduction

Cloud computing well-defined as a pool of abstracted, highly scalable, and managed complete infrastructure capable of hosting end-customer applications and billed by consumption. Virtualization is a technique, which allows sharing single physical instance of an application or resource among multiple organizations or tenants (customers). It does so by assigning a logical name to a physical resource and providing a pointer to that physical resource when demanded [7]. Queueing theory is the mathematical study of waiting lines and other operating characteristic of queues. In queuing theory, a model is constructed so that queue lengths and waiting times can be predicted (Wikipedia, 2013). The work in (Leung, 2000) sees queueing theory as the study of queueing systems, where some customers get some services from some servers. Queueing theory is useful in computer and communication system for determining throughput, response time, utilization, lost call probability and resource requirements. In this paper, we focus on several queueing models for cloud computing and also present the analysis and performance of queueing models used in cloud computing.

2. Queueing Models for Cloud Computing

Rahul Ghoshet, al.,[1] author discussed about Handling diverse client demands and managing unexpected failures without degrading performance are two key possibilities of a cloud delivered service. Still, the evaluation of a cloud service quality becomes tough as the scale and complexity of a cloud system increases. In a cloud environment, service request from a user goes through a variety of provider specific processing steps from the instant it is submitted until the service is fully delivered. Measurement-based evaluation of cloud service quality is expensive especially if many configurations, workload scenarios, and management methods are to be analysed. To overcome these difficulties, this paper they propose a general analytic model based approach for an end-to end perform ability analysis of a cloud service. They exemplify our approach using Infrastructure-as-a-Service (IaaS) cloud, where service availability and provisioning response delays are two key QoS metrics. A novelty of approach is in reducing the complexity of analysis by dividing the overall model into sub models and then obtaining the overall solution by iteration over individual sub-model solutions.

Hamzeh Khazaei et, al.,[2] author determined novel approximate analytical model for performance evaluation of cloud server farms and solve it to obtain accurate estimation of the complete probability distribution of the request response time and other important performance indicators. The model permits cloud operators to determine the relationship between the number of servers and input buffer size, on one side, and the performance indicators such as mean number of tasks in the system, blocking probability and probability that a task will obtain immediate service, on the other. They also discuss results indicate that a cloud centre accommodates heterogeneous services may impose longer waiting time for its clients compared to its homogeneous equivalent with the same traffic intensity.

Goswami et, al.,[3] author determined several characteristics of cloud computing. The process of entering into the cloud is generally in the form of a queue, so that each user needs to wait until the current user is being served. In the system, each Cloud Computing User (CCU) requests Cloud Computing Service Provider (CCSP) for use of resources. If CCU finds the server busy, then the user has to wait till the current user completes the job. This might result in increase of queue length as well as waiting time, which may lead to request drop. To hold this difficult, CCSP needs to find ways to reduce waiting time. They proposed a finite multiserver queueing model with queue dependent heterogeneous servers where the web applications are modelled as queues and the virtual machines are modelled as service providers. CCSP's can use multiple servers and the number of busy servers changes depending on the queue length for reducing queue length and waiting time. This helps them to dynamically create and remove

virtual machines in order to scaling up and down. They develop a recursive method to obtain the system steady-state probabilities. Various performance measures of the proposed scheme have been described and evaluated.

Wendy Ellenset al., [4] considered the general problem of resource provisioning within cloud computing. They analysed the problem of how to allocate resources to different clients such that the service level agreements (SLAs) for all of these clients are met. A model with multiple service request classes made by different clients is proposed to appraise the performance of a cloud computing centre when multiple SLAs are negotiated between the service provider and its customers. They have modelled a cloud centre using the *M/M/C/C* queueing system with different priority classes. SLA is specified by the request for each class, rejection probabilities of the clients in that class. They proposed solution supports cloud service providers in the decision making about 1) defining realistic SLAs,2) the dimensioning of data centres, 3) whether to accept new clients, and 4) the amount of resources to be reserved for high priority clients. Finally they demonstrate the potential of the solution by a number of experiments conducted for a large and therefore realistic number of resources.

Preeti kambleet al., [5] authors have studied performance analysis of the cloud using M/G/m/m+r queueing system till date, which gives the novel and approximate analytical model. It provides the relationship between the input buffer size and number of servers available. It also stretches the performance indicators like mean number of tasks in the system, task blocking probability and immediate service probability.

N.Ani Brown Mary et al., [6] author focussed on a novel paradigm for the provision of computing infrastructure in cloud computing 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 really provide access to large pools of resources. Resources provided by cloud computing systems hide a great deal of services from the user through virtualization. In this paper, the cloud data centre is modelled as [(M/G/1): (∞/GDMODEL)] queueing system with a single task arrivals and a task request buffer of infinite capacity. The results can be analysed using simulation. The mean as well as standard deviation can be computed. The blocking probability and probability of immediate service can be computed.

Ayad Ghany Ismaeel[7] used effective technique for allocating servers to support cloud, i.e. determine the available servers which relatively have a higher idle (not busy) to support source servers using queue model at the same time employs GIS and GPS techniques through algorithm of Haversine equation to select the idle server which closer to satisfy lowest cost and reach the optimal throughput (performance) and also using Google Maps find more fast in

finding the distances between idle server and source server than GIS packages. He also discussed higher relative to other techniques, because they used Haversine equation to compute the nearest idle server.

Khaled Salah [8] determined achieving proper elasticity for cloud jobs is a challenging research problem. In this paper, hefocusedon how to achieve proper elasticity for highly parallelized jobs which run on cloud clusters. They have presented an analytical model based finite queueing systems that can be used to determine at any given instance of time and under current workload conditions the minimal number of cloud resources needed to satisfy the SLO response time.

Jordi Vilaplana,etal., [9] presented the ability to provide guaranteed Quality of service is vital for the commercial success of cloud platforms. Using queueing theory to study computer service QoS in cloud computing. By applying open Jackson network to determine and measure the QoS assurances in the cloud platform in order to get response time. They considered different parameters such as arrival rate of customer service and service rate of processing servers among others. This model provided us the best option to guarantee QoS in the real cloud computing systems. They proposed a multi-server system with the queueing model using open Jackson network. This paper concluded that the combination of *M/M/*1 and *M/M/m* in sequence was proposed to model the cloud platform. They also showed that to provide good QoSin terms of response time, they have to determine where the system has a bottleneck and then improve the corresponding parameter. They conclude that our modelcan be very useful for tuning service performance, i.e., QoS [response time (*T)*], thus guaranteeing the SLA contract between the client and the service provider.

G. Vijaya Lakshmi,et al., [10] have studied about cloud computing, multi resources such as processing, bandwidth and storage need to be allocated instantaneously to multiple users. When the cloud computing users requests (CCU) for the service to the cloud computing service providers (CCSP) at the same time but while at a instant, if cloud computing server is busy CCU's needs to enter into the waiting line until CCSP completes its previous CCU . This might be leads to bottleneck so cloud computing users neither utilize the resources norwaits in the queue. Cloud Computing service providers use multiple servers to reduce the waiting time. This paper proposed a(∞/FIFO) Queuing model which is applied atin order to reduce waiting time, queue length the network performance and QOS effectively in cloud computing environment.

It is assumed that, if CCU arrives at an average rate λ and server has service mean rate μ and finds the server in busy state then CCU has to wait till the server completes its job or CCU may enter into Balking or Reneging state .This

results increasing in waiting time and queue length. Therefore in order to overcome this problem (M/M/C): (∞/FIFO) queueing model is applied. In this model, the arrival process assumed to be Poisson each server has an independent identical exponential service time distribution.

Chandan Banerjee [11] et al., have studied a priority based service time distribution method using Erlang distribution for K-phases. The process of entering into the cloud is classically based on the priority of the service and form a queue. Each user needs to wait until the current user is being served if priority of the service is same, as per the First Come First Served rule. If the server is busy, then the user request has to be waited in the queue until the current user receives the result of the prescribed tasks. They also considered a priority class mentions to a collection of all customers having the same priority. They also consider $M/E_K/1$model for a priority based single server and $M/E_K/2$ model for priority based two servers in each class in cloud computing scenario to reduce overall mean queue length and waiting time and also considered a single server and two server retrial queueing systems in which considered a single server and two server queues with three classes of customers.

Multiple servers based scenario has improved the performance based on our system, based on queue length reduction and waiting time optimization over single server based scenario. Numerical results have been confirmed exhibiting higher performance in case of our priority based approach for $M/E_K/2$ which reduces the queue length and waiting time compared to $M/E_K/1$. The low priority of service classes 2 and 3 increases by 1 after every 60 seconds. This approach avoids starvation. $M/E_K/1$ has produced better result compared to M/M/1 and $M/E_K/2$ has shown better throughput compared to M/M/2 in terms of average queue length and average waiting time.

Xiaodong Liu et al., [12] have examined the use of virtualization; physical resources are converted into "resources pool" which delivers service on demand. The model considers the resources sharing among VMs. In addition, various types of failures, such as VMs failures, physical server's failures and network failures are also considered. A service request is divided into many subtasks and each subtask consists of a series of data processing and transmission. The average service time of service requests is obtained. They apply on Markov process using transition probability calculate processing time, transmission time, VMs failure and recovery time, server failure and recovery time, network failure and recovery time.

Mohamed Eisaet al., [13] focussed on improve the quality of service by minimize execution time per jobs, waiting time and the cost of resources to satisfy user's requirements. By queueing theory scheduling algorithm is suggested to improve scheduling process. They also showed experimental results indicate that our model increases utilization of

global scheduler and reduce waiting time. They proposed model performance increased by reducing the mean queue length and waiting time. They observe that the proposed model is convenient for global scheduler in which they need to maximize the use, reducing waiting time and deal with an infinite number of applications for cloud resources. And also showed, load balancing in global scheduler cannot be achieved by M/M/1 model that introduced a single channel for all requests.

A.Anupama [14] studied on each arriving Cloud Computing User (CCU) requests Cloud Computing Service Provider (CCSP) to use the resources, if server is available, the arriving user will seize and hold for a length of time, which leads to queue length and more waiting time. A new arrival leaves the queue with no service. After service completion the server is made instantly available to others. From the user's point of view needs to be served immediately and to prevent waiting the CCSP's can use infinite servers to reduce waiting time and queue length. The arrival pattern is often Poisson in queuing theory. In this paper they analysed the dynamic behaviour of the system with infinite servers by finding various effective measures like response time, average time spend in the system, utilization and throughput. They applied queueing theory to show performance analysis between M/M/1 and M/M/∞can reduce queue length and increase throughput and utilization.

Mohamed Benel aattar [15] discussed a model taking into assumption of batch arrivals in the center has never been studied before. In this paper, he studied by modelling the cloud data center as a (GE/G/m/k) queueing system with GE distribution task arrivals, a general service time for requests as well as large number of physical servers and a task buffer of finite capacity. Heused this model to evaluate the performance analysis of cloud server frames and get an accurate estimate of the complete probability distribution of the request response time and other important performance indicators such as the mean number of tasks in the system, the distribution of waiting time, the probability of immediate service and the blocking probability.

Hyacinth, C [16] discussed the model of a job traffic queue of a cloud based research collaboration platform. The impartial is to compute the size of those queues, the time that jobs spend in them consider the number of simultaneous HTTP GET file requests handled by the server, and the total time required to service a request. This system captures efficiency in service time, server utilization, queueing response times, queueing workload and contents for a cloud based research collaboration platform. They proposed queueing-model-based Adaptive Control approach combines both the modelling power of queuing theory and self-tuning power of adaptive control. Hence, it can handle both modelling inaccuracies and load disturbances in a better way. Using the model G/G/1 queue which is

usually referred to as generalized single-server queue with First-in-First-out discipline with a general distribution of the sequences of inter-arrival and service times was observed. The deductions from the model for the nominal throughput of the Queuing workload, Queuing waiting time, Server Utilization and service time shows that the model will facilitate service efficiency and optimal performance for the CRCM.

Mohamed Ben el aattar [17] discussed about problem in network modelling and find the suitable model that best reflects the reality of the studied network; where the traffic incoming in a cloud computing center has specific proprieties this problem becomes more important in the case of cloud computing. In this paper they examined this traffic as an overflow traffic and propose an analytical network queuing model using an IPP/G/m/k queuing system which indicates that the arrival process is an Interrupted Poisson Process (IPP), while the service time is usually distributed and the system under consideration is multi-servers and has a finite capacity. They proposed analytical model determined significant performance parameters for overflow traffic, such as the average number of tasks in the system, blocking and immediate service probabilities and the average of response time.

Jing LI [18] discussed about Cloud computing delivers geographically distributed resources. Replication is adopted to address these problems, but leads to energy increase. They introduce a model to decide replication opportune moment, which is based on queueing theory. Through statistical record of the arrival rate of uses, service period, and replicas already created, calculating length of queue, waiting duration time, busy period, service strength and a reasonable creation opportune moment of replica can be obtained. Analysis and experiments the results prove that the replication strategy is predictable, saving energy and improving the service rate of users 'requests.

P.Diamantopoulos [19] determined the measure of effectiveness for cloud services, reliability indicators are used by queueing theory. Cloud service is separated into two stages. The first one is that of generating request. To model this stage, they adopt a batch arrival queue and retrial. For the batch arrival model, they are involved in the distribution of the sojourn time of a request into the system, through for both models then compute the probability that a request will be served which is actually the reliability of this stage. The second stage is execution stage of a request which is common for both batch arrival and retrial models and in which they have concerned in the probability of successfully executing a request. Finally they calculated the overall reliability of the complete cloud service.

Chunling Cheng [20] focussed on incoming jobs in cloud computing environments have the nature of randomness and compute nodes have to be powered on all the time to await incoming tasks, these outcomes in a great waste of energy. An energy-saving task scheduling algorithm established on the vacation queueing model for cloud computing

systems is presented in this paper. The vacation queueing model with exhaustive service to model the task schedule of a heterogeneous cloud computing system. Furthermore, based on the busy period and busy cycle under steady state, author analyse the expectations of task sojourn time and energy consumption of compute nodes in the heterogeneous cloud computing system. Next, they proposed a task scheduling algorithm based on similar tasks to decrease the energy consumption. By using simulation results show that the proposed algorithm can reduce the energy consumption of the cloud computing system effectively while meeting the task performance.

A. D. Banik [21] described about an infinite-buffer single server queue where arrivals occur according to a batch Markovian arrival process. The server serves customers in batches of maximum size '*b*' with a minimum threshold size '*a*'. The service time of each batch follows general distribution independent of each other as well as the arrival process. They proposed analysis is based on the use of matrix-analytic procedure to obtain queue-length distribution at a post-departure epoch. They considered queue-length distributions at various other epochs such as, pre-arrival, arbitrary and pre-service using relations with post-departure epoch.

Further they consider the system-length distributions at post-departure and arbitrary epochs using queue-length distribution at post-departure epoch. They analysis some important performance measures, like mean queue-lengths and mean waiting times have been obtained. Total expected cost function per unit time is also derived to determine the locally optimal values of *a* and *b*. Secondly, they perform similar analysis for the corresponding infinite-buffer single server queue where arrivals occur according to a BMAP and service process in this case follows a non-renewal one, namely, Markovian service process (MSP).

Jing Li [22] studied cloud-based message queueing services (CMQSs) with distributed computing and storage is generally adopted to improve availability, scalability, and reliability in virtualization technology. Yet, a critical issues its performance and the quality of service (QoS). In this paper, he employed both the analytical and simulation modelling to address the performance of CMQSs with reliability guarantee. He presented a visibility-based modelling approach (VMA) for simulation model using coloured Petri nets (CPN).

This model incorporates the important features of message queueing services in the cloud such as replication, message consistency, resource virtualization, and especially the mechanism named visibility timeout which is adopted in the services to guarantee system reliability.

Finally, this model estimate through different experiments under varied scenarios to obtain important performance metrics such as total message delivery time, waiting number, and components utilization. The results reveal

considerable insights into resource scheduling and system configuration for service providers to estimate and gain performance optimization.

Tuan Phung-Duc [23] studied about the processing unit (server) and the storage unit (buffer) are separated using retrial queueing model for cloud computing systems. Jobs that cannot inhabit the server upon arrival are stored in the buffer from which they are sent to the server after some random time. After completing a service the server stays idle for a while waiting for either a new job or a job from the buffer.

After the idle period, the server starts searching for a job from the buffer. Then they assumed that the search time cannot be disregarded during which the server cannot serve a job.

They have provided a model in this system using a retrial queue with search for customers from the orbit and obtain an obvious solution in term of partial generating functions. However they have presented a recursive scheme for computing the stationary probability of all the states.

B. Sundarraj [24] author discussed in this paper a key problem in Cloud data center management due to the numerous and heterogeneous strategies that can be applied, ranging from the VM placement to the federation with other clouds. Performance evaluation of Cloud Computing infrastructures is required to predict and quantify the cost-benefit of a strategy portfolio and the corresponding Quality of Service (QoS) experienced by users. In this paper, they present an analytical model, based on Stochastic Reward Nets (SRNs), that is both scalable to model systems composed of thousands of resources and flexible to represent different policies and cloud-specific strategies.

Furthermore several performance metrics are defined and evaluated to analyse the behaviour of a Cloud data center, utilization, availability, waiting time, and responsiveness. A resiliency analysis is also provided to take into account load bursts.

Gang Sunna [25] proposed an improved serial migration strategy and introduced the post-copy migration scheme into it. Then proposed mixed migration strategy that is based on the improved serial migration strategy and the parallel migration strategy.

Also, he developed queueing models (i.e., the M/M/C/C and the M/M/C queueing models) to quantify performance metrics, such as the blocking ratio and average waiting time of each migration request.

He evaluated the performance of the proposed migration strategy by conducting mathematical analysis, the numerical results showed that proposed strategy outperforms of the existing approach.

3. Comparison of Queueing models for Cloud Computing.

S. No	Authors	Queueing Models	Methods	Parameters	Computation
1	Rahul Ghosh	(M/M/C)	FCFS	Arrival rate, service rate	Job arrival rate, job service rate), fault load (e.g. machine failure rate) and system capacity (PMs in each pool).
2	Hamzeh Khazaei	M/M/1 (FIFO)	Poisson Arrival Process	Arrival rate, task service time, the virtualization degree, task rejection	Reliable response time and blocking probability avoidance.
3	V. Goswami, etal.,	M/M/s	Poisson Arrival Process	Arrival rate, service rate	Services performance prediction.
4	Wendy Ellens	*M/M/C/C*	Poisson Arrival Process	Arrival rate, service rate	Rejection probability for different customer classes, realistic SLAs, dimensioning of data centers, acceptance of new clients and reservation of resources
5	Preeti kamble	M/G/m/m+r	Poisson Arrival Process	Arrival rate, service rate	Mean number of tasks in the system, task blocking probability and immediate service probability.
6	N.Ani brown mary	[(M/G/1): (∞/GDMODEL)]	Poisson Arrival Process	Arrival rate, service rate	The blocking probability and probability of immediate service can be computed.
7	Ayad Ghany Ismaeel	Haversine equation	Poisson Arrival Process	Arrival rate, service rate	Finding the distances between idle server and source server
8	Khaled Salah	M/G/1/K	Poisson Arrival Process	Arrival rate, service rate	Proper elasticity for highly parallelized jobs
9	Jordi Vilaplana et al.,,	*M/M/*1 and *M/M/m*	Poisson Arrival Process	Arrival rate, service rate	Guarantee QoS in cloud

10	G. Vijaya Lakshmi and C.Shobana Bindhu	(M/M/C): (∞/FIFO)	Poisson Arrival Process	Arrival rate, service rate	Increasing in waiting time and queue length
11	Chandan Banerjee et al.,	$M/E_K/1$ and $M/E_K/2$	Batch arrivals	Arrival rate, service rate	Reduce overall mean queue length and waiting time
12	Xiaodong Liu et al.,	M/M/1	Poisson Arrival Process	Arrival rate, service rate	VMs failures, physical server's failures and network failures
13	Mohamed Eisa	M/M/1	Poisson Arrival Process	Arrival rate, service rate	Improve scheduling process
14	A.Anupama	M/M/1 and $M/M/\infty$	Poisson Arrival Process	Arrival rate, service rate	Response time, average time spend in the system, utilization and throughput
15	Mohamed Ben el aa	GE/G/m/k	Batch arrivals	Arrival rate, service rate	Mean number of tasks in the system, the distribution of waiting time, the probability of immediate service and the blocking probability.
16	Hyacinth C	Job Traffic Queue	Poisson Arrival Process	Arrival rate, service rate	Service time, server utilization, queuing response times, queuing workload.
17	Mohamed ben el aattar	IPP/G/m/k	Interrupted Poisson Process	Arrival rate, service rate	Overflow traffic in Networks.
18	Jing LI	*M/M/C*	Poisson Arrival Process	Arrival rate, service rate	Saving energy and improving the service rate of users 'requests.
19	P.Diamanto poulos	*M/M/C*	Batch arrival	Arrival rate, service rate	Overall reliability of the complete cloud service.
20	Chunling Cheng	M/G/1 vacation model	Poisson Arrival Process	Arrival rate, service rate	Reduce the energy consumption.
21	A. D. Banik	$BMAP/G^{(a,b)}/1/\infty$	Poisson Arrival Process	Arrival rate, service time	Mean queue-lengths and mean waiting times.
22	Jing Li	Markovian Model with service in two phases	Poisson Arrival Process	Arrival rate, service time	Total message delivery time, waiting number, and components utilization.
23	Tuan Phung-Duc	*M/M/1*	Poisson Arrival	Arrival rate, service rate	Mean number customer waiting orbit,mean number

			Process		calls in the system , recursive scheme probability of all the states.
24	B. Sundarraj	Stochastic Reward Nets	Poisson Arrival Process	Arrival rate, service rate	Cloud data center: utilization, availability, waiting time, and responsiveness.
25	Gang Suna	M/M/C/C and the M/M/C	Poisson Arrival Process	Arrival rate, service time	Average waiting queue length and the average waiting time of the arriving migration requests

4. Conclusion:

Cloud computing is an emerging technology in IT field .In this paper we focused on several queueing model for cloud computing. Furthermore we analysed models which shows the resource sharing, priority based distribution, scalability and also performance of cloud computing center in cloud computing environment.

References:

1. Rahul Ghosh" End-to-End Performability Analysis for Infrastructure-as-a-Service Cloud: An Interacting Stochastic Models Approach",Conference paper · December, 2010.
2. Hamzeh khazaei, "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.
3. V. Goswami, S. S. Patra, G. B. Mund "Performance Analysis of Cloud with Queue Dependent Virtual Machines", Int'l Conf. on Recent Advances in Information Technology RAIT-2012.
4. Wendy Ellens "Performance of Cloud Computing Centers with Multiple PriorityClasses", IEEE Fifth International Conference on Cloud Computing,2012.
5. Preeti kamble "Performance analysis of cloud computing centers by breaking-down response time", International Journal of Advanced Computational Engineering and Networking, Volume- 1, Issue- 8, Oct-2013.
6. N.ani brown mary "Performance factors of cloud computing data centers using [(m/g/1) : (∞/gdmodel)] queuing systems", International Journal of Grid Computing & Applications (IJGCA) Vol.4, No.1, March 2013.
7. Ayad Ghany Ismaeel" Effective Technique for Allocating Servers to Support Cloud using GPS and GIS", Science and Information Conference 2013.
8. Khaled Salah "A Queueing Model to Achieve Proper Elasticity for Cloud Cluster Jobs",IEEE Sixth International Conference on Cloud Computing,2013.

9. Jordi Vilaplana, Francesc SolsonaIvan, Teixidó Jordi MateoFrancesc and Abella Josep Rius,"A queuing theory model for cloud computing", J Supercomputer, © Springer Science Business Media New York 2014.
10. G. Vijaya Lakshmi , C. Shoba Bindhu "A Queuing Model To Improve Quality of Service by Reducing Waiting Time in Cloud", International Journal of Soft Computing and Engineering (IJSCE), Volume-4 Issue-5, November 2014.
11. Chandan Banerjee, Anirban Kundu,Ayush Agarwal1, Puja Singh, Sneha Bhattacharya and Rana Dattagupta "Priority based K-Erlang Distribution Method in Cloud Computing",Int. J. on Recent Trends in Engineering and Technology, Vol. 10, No. 1, Jan 2014.
12. Xiaodong Liu , Weiqin Tong , Xiaoli Zhi "Performance analysis of cloud computing services considering resources sharing among virtual machines",J Supercomputer (2014).
13. Mohamed Eisa "Enhancing Cloud Computing scheduling based on Queuing Models", International Journal of Computer Applications, Volume 85 – No 2, January 2014.
14. A.Anupama "Using Queuing theory the performance measures of cloud with infinite servers", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 01 Jan 2014.
15. Mohamed Ben el Aattar "Performance Modelling for a Cloud Computing Center Using GE/G/m/k Queuing System", International Journal of Science and Research (IJSR),Volume 3 Issue 5, May 2014.
16. Hyacinth C" Model of a Job Traffic Queue of a Cloud- Based Research collaboration Platform", International Journal of Current Engineering and Technology, 01 Oct 2014.
17. Mohamed Ben el Aattar "Analysis of Queuing system model for overflow tasks routed to a cloud computing center", International Journal of Applied Mathematics and Modelling, IJA2M, Vol.2, No. 3, 28-37, May, 2014.
18. Jing LI "A Model of Green Replication in Cloud Computing Environment " International Conference on Computer Science and Software Engineering, 2014.
19. P.Diamantopoulos "Cloud computing service reliability modelling with batcharrivals and retrial queues" , Safety, Reliablity and Risk Analysis:Beyond the Horizon 2014.
20. Chunling Cheng "An Energy-Saving Task Scheduling Strategy Based on Vacation Queuing Theory in Cloud Computing",Tsinghua science and technology,Volume 20, Number 1, February 2015.
21. Jing Li, "Modelling Message Queueing Services with Reliability Guarantee in Cloud Computing Environment Using Coloured Petri Nets", Mathematical Problems in Engineering Volume 2015.

22. Tuan Phung-Duc " Retrial Queue for cloud systems with separated processing and storage units ",Queueing Theory and Network Applications, pp 143-151, Volume 383,2015.
23. B. Sundarraj" A Stochastic Model to Investigate Data Center Performance and QoS in Iaas Cloud Computing Systems"., International Journal of Innovative Research in Computer and Communication Engineering,Vol. 3, Issue 3, March 2015.
24. Gang Suna"A new technique for efficient live migration of multiple virtual machines",Future Generation Computer Systems ,Volume 74–86 ,(2016).

Corresponding Author:
K.Santhi*,
Emails: *ksanthi@vit.ac.in*