Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 6, June-2016

A Comprehensive Analysis of Response Time of Cloud

Sachin Annigeri[1], Prof. S.L Deshpande[2]

[1]*Department of Computer Network Engineering,VTU PG Centre,Belagavi*

[2]*Department of Computer Network Engineering,VTU PG Centre,Belagavi*

Abstract-*Cloud computing has got extremely large popularity due to its unique features such as Scalability, Reliability, dynamic and Economic computing solutions. Apart from these advantages many challenges are associated with it such as resource management, security, QoS. Services provided by the cloud have direct influence on user's satisfaction and service provider's profit. Therefore Performance analysis is one of the most important aspects of cloud computing. The factors that can affect the performance of the cloud computing are subject to the QoS metrics such as Processing time taken by the data center to process the user request, The response time and total cost evolved in designing the cloud environment. According to the analysis, as many as Data centers results greater reduction in overall response time and processing time however the corresponding cost is higher. Distribution of users and data centers over the different regions is optimal.*

Key words*-Cloud computing,QoS metrics,response time,processing time ,cost.*

I.INTRODUCTION

Cloud computing is a technology which provides the services or resources on-demand. In cloud computing the information stored is accessed through the internet. Cloud computing has got extremely large popularity due to its unique features such as Scalability, Reliability, dynamic and Economic computing solutions. Apart from these advantages many challenges are associated with it such as resource management, security, QoS. Services provided by the cloud have direct influence on user's satisfaction and service provider's profit. Therefore Performance analysis is one of the most important aspects of cloud computing. The factors that can affect the performance of the cloud computing are subject to the QoS metrics such as Processing time taken by the data center to process the user request, The response time and total cost evolved in designing the cloud environment. The main aim of the paper is to analyse the performance of the cloud computing based on the overall response time of request, Processing time at the data centers and Total cost of the data center using CloudAnalyst tool.

The high performance is the one of the big advantage of the cloud, the user expects satisfactory service in each request. These high performance services have direct influence on user satisfaction and commercial benefits. Commercial benefits are always depending on the ability to deliver Guaranteed Quality of Service.

The performance of the cloud computing can be analysed in many different ways:

- ***Performance can be analysed based on different metrics***–load balancing techniques, the number of incoming requests from the user, workload on the data center, cost and throughput.
- ***Performance can be analysed based on methods-*** such as parto traffic methods, fuzzy systems.
- ***Performance can be analysed based on Applications and services provided by the cloud-*** Cloud computing provides the variety of services where each of service is the simulation component to measure the performance.
- ***Performance can be analysed based on Environments-*** The cloud service providers like Google, Amazon have their own cloud environment with own characteristics.
- ***Performance can be analysed based on part of cloud-*** Sometimes the performance can be analysed based on only specific part of the cloud.

Performance, in simple terms, how faster an application can respond to a given user request. Therefore response time is the important measure which is the time taken to process the client request. In this paper, Performance of the cloud computing is evaluated based on Overall response time, Processing time of workload and cost of the data center.

B.Queuing Models

Queuing models are used to investigate the behaviour of buffers used in telecommunication system, Traffic Engineering etc. Queuing system consists of one or more buffers to store the arrival requests for a time period. It either rejects or pre-emits the request if time stamp expires.

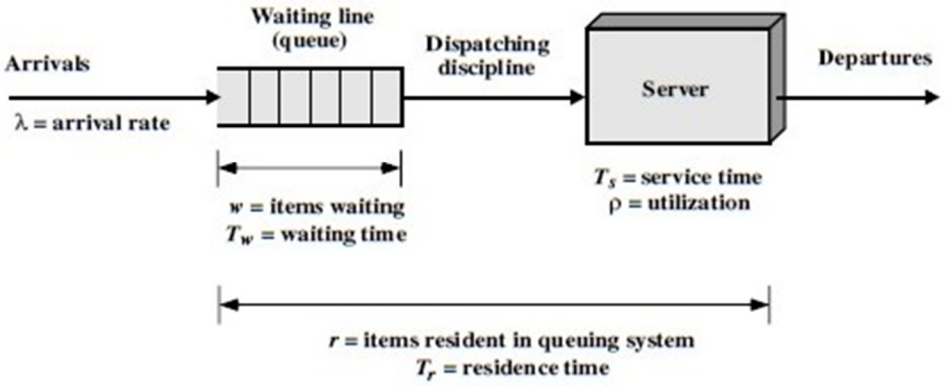

Figure 1. Queuing Model

Queuing system consists of three parts:

- Request arrival
- Service Mechanism
- Queuing Discipline

The request arrivals may be from one sources or from many sources,called population. If A*i* is the arrival time between (i-1)th and *i*'th requests, then inter-arrival time λ is given by = 1/(E(A) arrival rate.The service mechanism defines number of servers to be used to serve the incoming request. Each server in the system has a queue and probability distribution of service time. Let S_i be the service time of the ith customer, then mean service time μ is given by = 1/(E(S) service rate of the server.Queuing discipline is the rule that is applied to choose the next customer from the queue. Frequently used queuing techniques are: First In First Out (FIFO), Last In Last Out (LIFO), Priority – based on the importance of request .

C.Response Time

It refers to the time, how faster an application can respond to a given user request. Therefore response time is the important measure. User access the services of the cloud through requests, the cloud schedules these requests in the form of queue, each user added to queue needs to wait till the prior requests gets processed. Whenever new request enters into the cloud, it gets added to the buffer, so the length of the queue increases and hence the response time.

D.Processing Time

Processing time is the time taken to process the client request. The processing time is depend upon the data center configurations and number of requests to be processed. To get the service from the cloud, user has to wait in the queue. User requests the cloud to use resources. If resource is available, seize and hold it for length of time. Which leads to the queue length of new arriving requests and hence waiting time From the users satisfaction view the requests need to be served immediately. Therefore to reduce the processing time, service providers should implement as many as data centers. However the associated cost is higher.

Objective of the Paper

The main objective of the paper is to analyse the performance of the cloud computing based on the Overall response time, Processing time of data centers and Total cost of the data center.

II.LITERATURE REVIEW

The performance of the cloud computing can be evaluated based on the queuing theory[1].The requests from the user to the cloud are queued in buffer so the user needs to wait until the prior requests in the queue gets processed. As the new requests arrives, the waiting length increases so the waiting time also increases. It is observed that as number of arrival requests to the buffer increases the use of the cloud also increases which in turn increases the latency in the buffer so the response time. P. Suresh Varm, Satyanarayana A, M V Rama Sundari proposed a request dependent model where resource allocation is done linearly depending on the number of requests in the buffer. It is assumed that the arrival requests to the buffer follow the Poisson distribution λ and resource allocation also follows Poisson distribution μ. The performance is evaluated by

determining the various equations for mean queue length(*L*), throughput(*thp*) and waiting time(*W*). From derived equations, request dependent model yields better performance.
The advantage of having numerous data center leads to increased system performance but is associated with higher cost. So to solve, N Ani Browm Mary, K Sarvan modelled the cloud data center as M/G/1 queuing system [2] where the incoming requests from the user are queued with single arrival task and request buffer is of infinite capacity. It is more difficult to determine queue length and response time, if inter arrival time and service time are not exponential. Probability distribution also cannot be determined in M/G/m systems. They prposed the M/G/1 model where the distribution of inter arrival time and service time is depend on the number of requests to be processed in the system. In this model, the overall service time is distributed among the inter arrival requests. It is observed that it is possible to improve the performance of the cloud by modifying the Data center configurations such as number of processors, buffer size etc.
Swapana addamani, Anirban Basu[3] modelled the platform as multiple queues and the virtual machines (VM) . When an application is installed on the cloud, the request from the client to the application are queued before dispatching it to the machine.
The number of VM's to be created to handle the requests is decided by the application developer. It can be specified by SLA or by default empirical value is used. The VM's can run the application either on single node or multiple nodes. As the number of VM's increases the task to be processed is divided into many and each of gets executed on separate machine, which greatly results in reduction of the response time and processing time as well. The number of VM's to be created is need to be specified by the web application developer at the time of deploying the web application The incoming requests are served in many ways like First Come First Serve (FCFS),Last Come First Serve(LCLS),random order, round robin. The modelled system was analysed using JMT with distribution of service time over multiple servers.
Performance is exemplary idea that incorporates various measures like Reliability, Scalability, Efficiency, cost etc. Due to tremendous growth in cloud environment numerous elements influence the execution of cloud such as Number of clients, Number of Data Centers, Location, Latency, Number of virtual machines, workload on the processor, Network Bandwidth. Apart from these the metrics the factors which have direct impact on performance are Response time, Processing time, Cost [4].
T. Sai Sowjanya*, D.Praveen,K.Satish, A.Rahiman elaborated [8] how to reduce the waiting time to improve the cloud computing performance. Waiting lines emerges when there is need for service and the service mechanism is busy in serving the pat requests.In single server system, one process is processed at time. When request arrives to the clouds, if server is available the request directly enters into the server and processed immediately.
The factors that can affect the performance of the cloud computing are subject to the QoS metrics such as Processing time taken by the data center to process the user request, The response time and total cost evolved in designing the cloud environment have been studied in the literature.

III.PROPOSED WORK

Due to it's dynamic and Economic computing solutions, day by day the cloud computing has got enormous popularity. Since high performance is the one of the big advantage of the cloud, the user expects satisfactory service in each request. These high performance services have direct influence on user satisfaction and commercial benefits. Commercial benefits are always depending on the ability to deliver Guaranteed Quality of Service. As per the literature review some of the QoS parameters which affects the performance of the cloud computing are Overall Response time, Processing time of request and total cost.
The objective of this paper is to evaluate the performance of cloud computing based on the QoS metrics viz Overall Response time, processing time at the data center and total cost evolved in designing data center using CloudAnalyst tool. There are many QoS metrics which can be used to measure the performance of cloud computing. viz Overall Response time, processing time at the data center and total cost evolved in designing data center are measured in this paper using CloudAnalyst.The analysis of wide environments such as cloud is usually associated with simulation because to measure in real context is too difficult as there are many elements in cloud, which may not be predictable and controllable. There are two components to be design:
1) Data centers
2) Users
Data center is a component which includes various configuration parameters such as number of processors, the bandwidth, memory etc. It can be used in the different regions according to the requirements.
User is a customer who consumes the service provided by the cloud. Like Data Center User is a component, can be configure in terms location, the amount of work to be send etc.
The following three different scenarios are considered to measure the performance.
1) Simulation and analysis by modifying the Data Center configurations.
2) Simulation and analysis by modifying the User Configurations.
3) Simulation and Analysis by considering the geographical region of Data center and User.

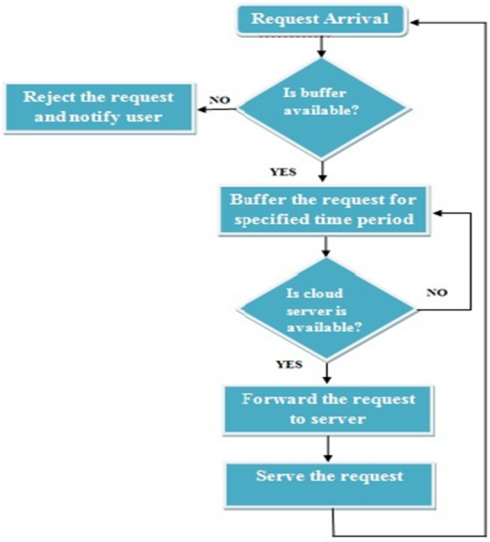

Figure 2 Process of request in cloud.

.

IV.RESULTS

A.Simulation-1:

In this, the number of users are assumed to be fixed equal to 80 and are equally distributed over the different regions of the world. Data centers are changed from 1 to 35 which are also distributed over the different regions and corresponding Response time, Processing time and associated cost of designing the data centers are shown in Table 1.

Table 1: Tabular column for Simulation-1.

Sl No.	No. of Users	No. of Data Centers	Response time (ms)	Processing time(ms)	Total Cost($)
1	80	1	280.93	0.62	1.01
2	80	5	69.29	0.55	3.02
3	80	10	50.1	0.49	5.53
4	80	15	50.1	0.45	8.04
5	80	20	50.1	0.39	11.01
6	80	25	50.1	0.39	13.85
7	80	30	50.1	0.39	15.57
8	80	35	50.1	0.38	18.08

Figure 3 shows the overall response time associated to scenario-1. As we can observe from the graph, initially to serve the number of users requests there was single data center so time taken to serve the user requests was more, hence the associated response time. In following simulations, the users are of fixed numbers, say 80, data centers are increased from 10,15,20 so on. Since there as many as data centers to serve the user requests there is reduction in response time as number of data centers increase.

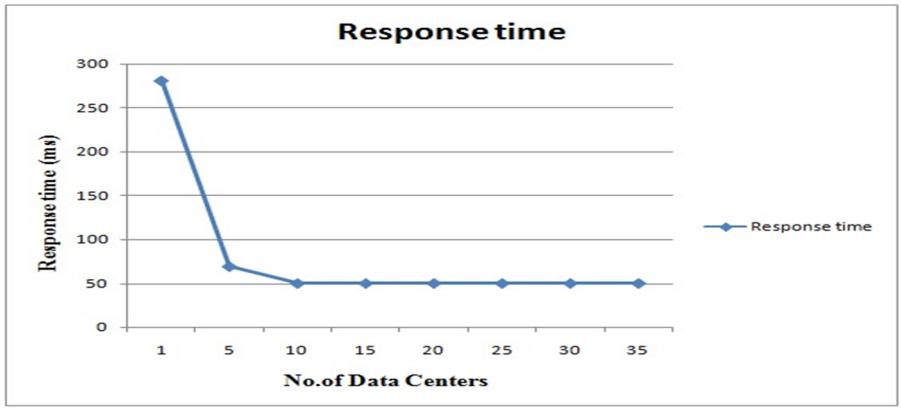

Figure 3: Overall response time in Simulation-1.

Figure 4 shows the processing time in data centers. The processing time is time taken by the data center to process the user request. The processing time is directly propotional to number of data centers. It can be observed that overall processing time is decreases as the number of data centers increases

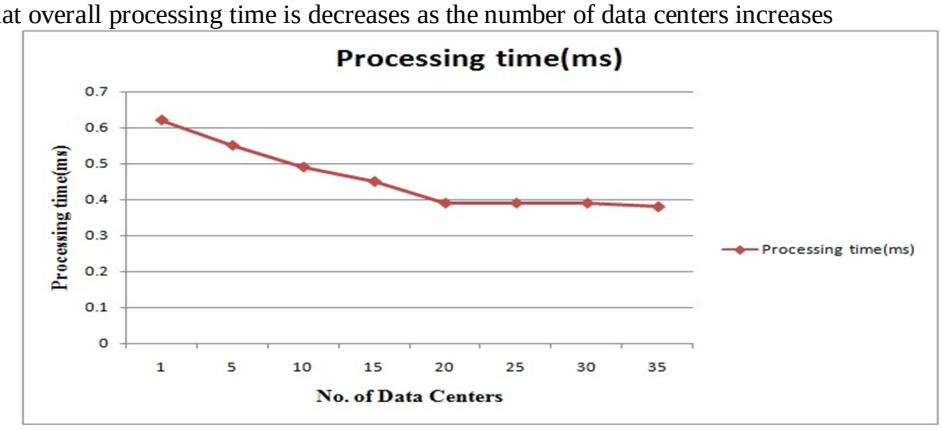

Figure 4: Overall processing time in Simulation1

Figure 5 shows the total cost associated designing the data centers. Cost increases and it affects on service providers income. If we try to optimize the response time and processing time of the cloud, obviously it leads to increase number of data centers hence the associated cost of designing the data centers.

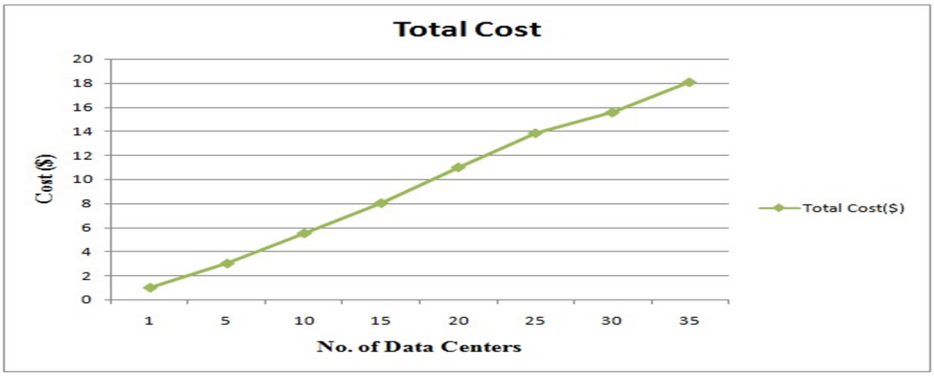

Figure 5: Total cost in Simulation1

B.Simulation -2:

In this scenario, number of Data centers to serve the client requests are constant, say 1 and the number of users are changed from 10,20,30,40.... The goal here is to measure over all processing time and to show how response time of individual gets affected as number of users being increased. Table 2 shows the variation of processing and response timing as the number of users added to the cloud.

Table 2: Tabular column for Simulation 2

Sl No.	No. of Users	No. of Data Centers	Response time (ms)	Processing time(ms)
1	10	1	369.12	0.61
2	30	1	379.62	0.63
3	60	1	383.12	0.69
4	90	1	388.6	0.8
5	120	1	392.2	0.81
6	150	1	397.36	0.83
7	180	1	399	0.88

Figure 6 shows the response time of data centers. Response time is time how quicker can application respond. Number of users in a particular geographical area also plays important role for measuring the performance of cloud. Note that whenever users get added to the cloud, the overall response time increases.

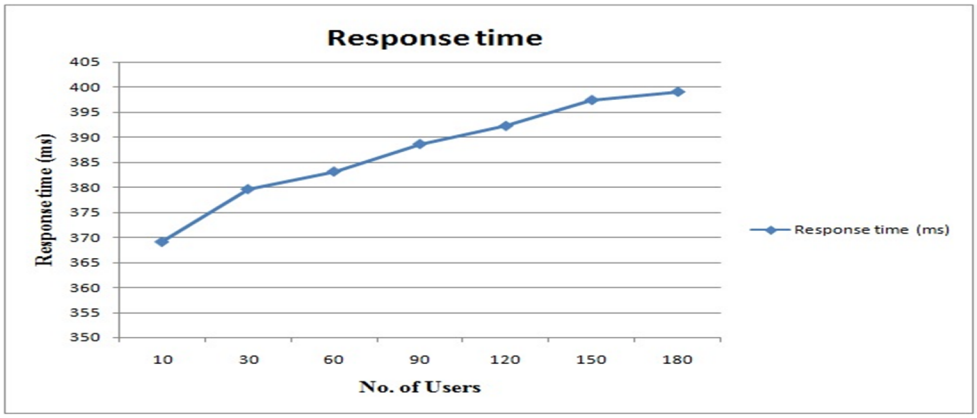

Figure 6 : Overall response time in Simulation 2

The processing time is shown in figure 7. As shown in figure the overloaded data center results increase in processing time. The hardware configuration of data centers also has the significant impact on the processing time.

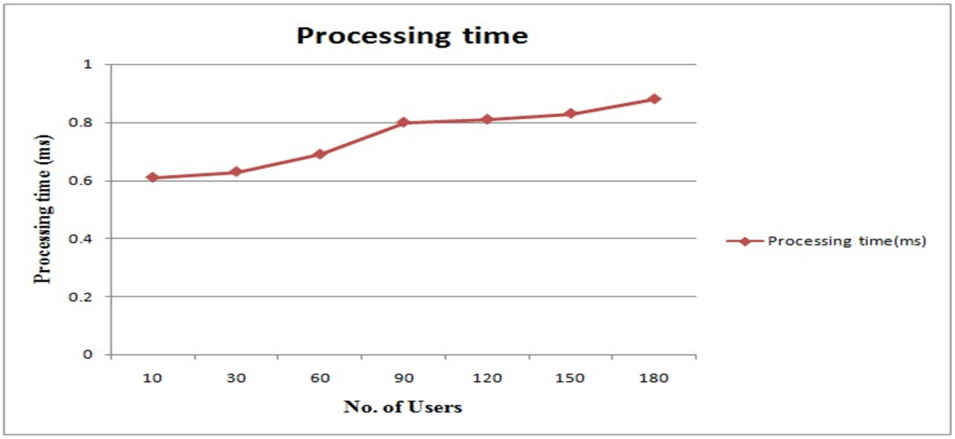

Figure 7: Overall processing time Simulation 2.

4.3 **Simulation -3:**

The geographical location of Data centers and individuals also plays very important for organizations to retain the data in specific locations. In this scenario, based upon the geographical location of Data Center and Users the performance will be evaluated.

There are three following cases.

1) Both server farm and all clients are put in same area (single region).
2) In second case, server farms are put in one locale and all clients are in another region.
3) In third case both data centers and users have been distributed.

Table 3: Tabular column for Simulation 3

Parameter	Same Region	Distributed Region	Separate Region
Processing time (ms)	0.88	0.22	0.64
Response time (ms)	62.63	360	375.11

Figure 8 shows the response time in different regions. The geographical location of data centers and users also plays a very important role in measuring the performance of cloud computing. Note that the response time is least in case where both data center and user are in same region and it is more in separate region. As data center is away from the user, the request from the user to data center needs a time to get back response.

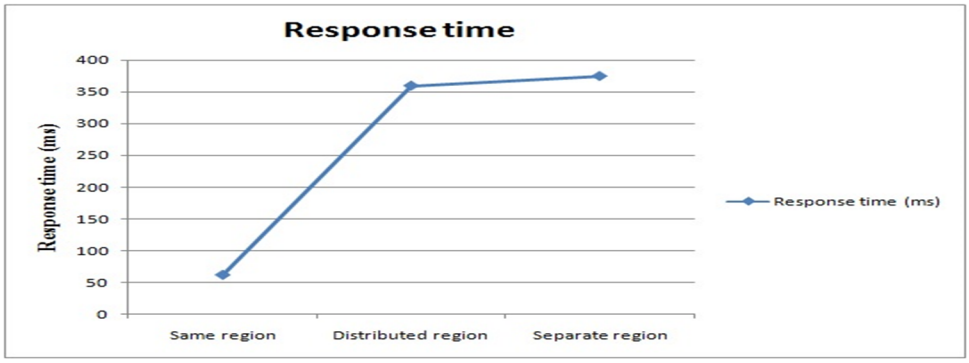

Figure 8: Response time in all three cases.

Figure 9 shows the processing time associated with three cases. It is least in distributed scenario.

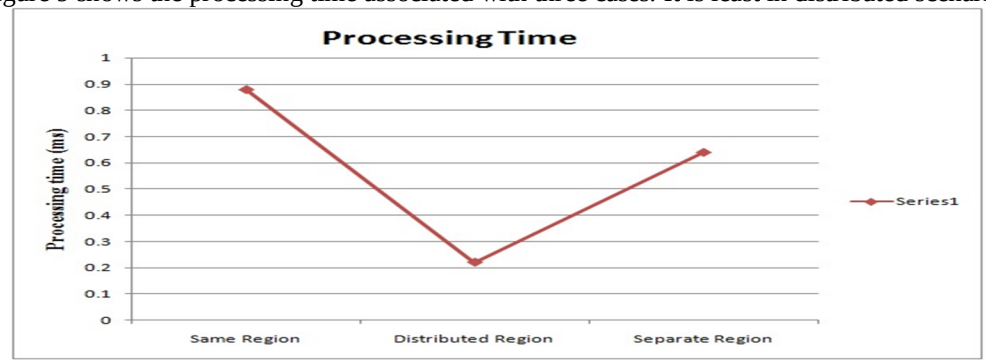

Figure 9: Overall processing time in all three regions.

According to the analysis, the various aspects are conclude as the service provider must have some standards to ensure the customers satisfaction in terms of Quality of services such as Reliability, availability etc.

To meet customer satisfaction, the service provider should enforces the various mechanisms to provide the quick and efficient responses. Because of the development in distributed computing, the capacity and computing is transferred into mists. Therefore architectural analysis of data centers should be performed to ensure the performance of the cloud.

V.CONCLUSION:

According to the analysis , as many as Data centers results greater reduction in overall response time and processing time however the corresponding cost is higher. For service providers benefits, if Data center is overloaded with workloads which degrades the efficiency. So, depending upon the number of users the data centers need to be increased to improve the efficiency and to achieve service satisfaction in each request. Distribution of users and data centers over the different regions is optimal.

REFERENCES

[1] P. Suresh Varm, Satyanarayana A, M V Rama Sundari. "*Performance analysis of cloud computing using Queuing models*" International conference on cloud computing technologies, applications and management In the year of 2012.

[2] N Ani Browm Mary, K Sarvan."*Performance factors of cloud computing data centers using M/G/1 Queuing system*" International Journal of grid computing and Applications. Vol.4 N0.1, March-2013.

[3] Swapana addamani, Anirban Basu."*Performance analysis of cloud computing platform*" International Journal of Applied Information System (IJAIS) vol-4,No.4 Oct-2012.

[4] Niloopar khanghahi, Reza Ravanmehr. *"Performance monitoring and Analysis of cloud computing environment"* Department of computer Engineering Tslamic azad university ,Central Tehran branch.
[5] A.Anupama, G.Satya Keerthi "*Using Queuing theory the performance measures of cloud with infinite servers*" International Journal of Computer Science & Engineering Technology (IJCSET).
[6] Assoc. Prof. Rajkumar Buyya "*CloudAnalyst: A CloudSim-based Tool for Modelling and Analysis of Large Scale Cloud Computing Environments"*- MEDC Project Report,.
[7] " Hao-peng CHEN, Shao-chong LI "*A Queueing-based Model for Performance Management on Cloud,*. School of Software, Shanghai Jiao Tong University, Shanghai, China, School of Computer Science, Georgia Institute of Technology, Atlanta, USA
[8] T. Sai Sowjanya*, D.Praveen,K.Satish, A.Rahiman "*The Queueing Theory in Cloud Computing to Reduce the Waiting Time*" Vol 1, Issue 3,110-112 111. IJCSET | April 2011 |