Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Analysis of Cloud Computing under Non Homogeneous Conditions

Satyanarayana .A	Dr. P. Suresh Varma	Dr. M.V.Rama Sundari	Dr. P Sarada Varma
Dept of Computer Science	Dept of Computer Science	Department of Information Technology	Dept of Mathematics
Adikavi Nannaya University	Adikavi Nannaya University	Godavari Institute of Engg. & Tech	GMRIT, Rajam
Rajahmundry, AP, India	Rajahmundry, AP, India	Rajahmundry, AP, India	AP, India

Abstract-Cloud computing is an evolutionary technology that differs from traditional parallel computing, distributed computing and grid computing as it provides solutions to many problems related to web based resource allocation by efficient and economical way of pay-per-use services. With tremendous increase in cloud services utilization, the allocation of resources to the job requests that enter the cloud has become a challenging task that arrives in non-homogeneous fashion. In this paper we developed and employed a cloud computing model for allocation of resources to the jobs that enter into the cloud by using queuing models. In this we also considered that arrival of jobs follows non-homogeneous Poisson process. We studied about the transient analysis of the model by using various performance measures such as Mean number of job requests in the cloud, Utilization, Throughput and Mean Delay in the cloud. It is observed that using queuing theory and by assuming arrival of jobs follow non-homogeneous Poisson process has tremendous influence on performance measures Mean number of job requests, Throughput, Utilization and Mean Delay in the Cloud

Keywords: Cloud computing, Grid computing, queuing model, Non-homogeneous Poisson Process, Transient analysis.

I. INTRODUCTION

Cloud Computing enables the massive scale service sharing, which allows users to access technology enabled service without knowledge of expertise of the system. Cloud computing has been often used with synonymous terms such as software as service, Grid Computing, Cluster computing, autonomic computing and utility computing. As stated in the majority of current cloud computing infrastructure as of 2009 consists of services that are offered up and delivered through a service centre such as a data centre that can be accessed from a web browser anywhere in the world [1], [2]. Hao-peng CHEN, Shao-chong Li.A [3] and R.D Mei, and H.B Meeuwissen [4] mentioned that the users do not care too much about the resources of the grid system but are more concerned with the services they are using. Hence, the function of service sharing enabled by cloud computing will be more interesting to general users than the resources sharing of the grid computing. A variety of Cloud Services are provided by the cloud system. The Cloud System could become very large even all over the whole internet. Users can request cloud services from any corner of the world. Some examples of commercial cloud services include Amazon EC2, Xen, Google Cloud, IBM Cloud and Microsoft Cloud. A little work is reported regarding services in the cloud by applying queuing theory models. Kaiqi Xiong and Harry Perros [5] studied service performance and analysis in Cloud Computing by applying queuing theory model at service centre.

In the modern competitive business environment providing Quality of Service is prime requisition for any service provider. With high exposition of technological innovation and developments the Cloud Computing is changing evolutionary in the modern era. The dynamic allocation of resources has emerged as promising technology to provide cost effectiveness in high performance cloud computing system for solving many complex problems in commercial application.

Recently P.Suresh Varma, A. Satyanarayana [6] has reported regarding cloud computing with request depend resource allocation for improving quality of service by utilizing idle resources by considered the arrival of jobs on service are homogeneous, but in many practical situations cloud service like Amazon, Google apps, Microsoft azure, the arrival of jobs are to be considered on time dependent in order to have accurate prediction of the performance measure of the cloud computing. In addition to this a number of measurements studies have reveal that the traffic generated by many real world application exhibit a high degree of burstness (time varying arrival rates) and posses correlation in the number of request arrivals [7], therefore the traditional models with simplified assumption regarding request arrivals modelled by poison process cannot capture the burst nature of request arrival process. Hence it is needed to develop queuing model with generalized Poisson Process for characterizing the arrival process. The time dependent nature of arrival rate of requests can be well captured by Non Homogeneous Poisson Process. Very little work has been reported regarding Non Homogeneous Process arrivals [8], [9]. We develop and analyzed a cloud computing model with Non Homogeneous Process arrival having request dependent resource Allocation. In request dependent resource Allocation the process rate

of each job request is adjusted depending on the content of the buffer at that instant. The Non Homogeneous Poisson Process is capable of portion the time dependent nature of the arrival process.

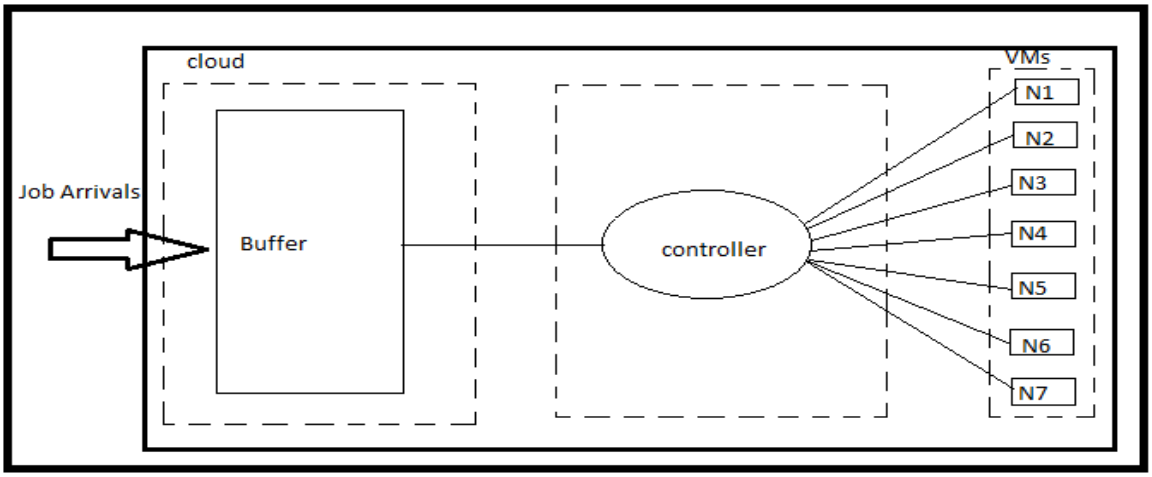
Using the difference differential equation the probability generating function of the number of requests in the each buffer are derived the transient behaviour of the cloud environment is analyzed [10] by deriving the system performance measures like the mean content of the buffer, mean delay, throughput, utilization explicitly.

II. CLOUD COMPUTING MODEL UNDER NON-HOMOGENEOUS CONDITIONS

In this paper a cloud system in which the arrival of job requests follows non-homogeneous Poisson process which is stored in buffer that scheduled for resource allocation using queuing models. The scheduling is carried with request dependent strategy. In request dependent strategy the resource allocation (virtual machines (VMs)) rate is a linear function of the number of request jobs in the buffer depending on the buffer content, the allocation time of the request packet is fixed with dynamic allocation of resources in the cloud.

Here, we assume that the arrival of request jobs follows a non-homogeneous Poisson process with parameter λ and αt where λ arrival rate of jobs, α is a constant, the number of job completion in the cloud follow Poisson processes with parameters μ. The mean service rate of the cloud is linearly dependent on the number of job requests in the buffer. The job arrivals are scheduled through the buffer by first in first out principle.

Fig.1 The schematic diagram representing the job requests in buffer to cloud system



The requests from various clients are stored in buffer which is connected to controller for allocation of resources in the cloud. The allocation of resources (VMs) is carried with request dependent strategy. The controller allocates resources (virtual machines) to the jobs based on job requests in the buffer. In request dependent strategy the resource allocation is a linear function of the number of requests in the buffer depending on the buffer content.

The schematic diagram representing the one buffer and one controller are in series with load dependent resource allocation. The mean service rate of the cloud is linearly dependent on the number of requests in the buffer. The jobs are services through the cloud by allocating virtual machines follows first in first out principle. With this structure the postulates of the cloud are

1. The arrival of the jobs in non-overlapping time intervals of time is statistically independent.
2. The probability that there is arrival of one job during small interval of time h is [λh+o(h)]
3. The probability that there is one job serviced through the cloud when there are n jobs in the buffer during small interval of time h is [n μ h + o(h)].
4. The probability that other than above jobs during small interval of time h is [o(h)]
5. The probability that there is no arrival of job in the buffer and no service completion of jobs during small interval of time h when there are n jobs in the buffer is [1- λh- n μ h+o(h)]

The Chapman Kolmogrov difference differential equations of the above postulates are

$$\frac{\partial P_n(t)}{\partial t} = -(\lambda(t)+n\mu)P_n(t)+\lambda(t)P_{n-1}(t)+(n+1)\mu P_{n-1}(t) \qquad n > 0$$

$$\frac{\partial P_0(t)}{\partial t} = -\lambda(t)P_0(t)+\mu P_1(t) \qquad n = 0$$

Where $\lambda(t) = \lambda+\alpha t$

By solving above difference differential equations

Let $P_n(t)$ denote the probability that there are n jobs in the buffer at time t.

The probability generating function of $P_n(t)$ is

$$P(s,t)=\exp\left[\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}(s-1)\left(1-e^{-\mu t}\right)\right] \quad (2.1)$$

III. PERFORMANCE MEASURE OF THE CLOUD

Expanding $P(s;t)$ given in equation (2.1) and collecting the constant terms, we get the probability that the network is empty as

$$P_0(t)=\exp\left(-\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}\left(1-e^{-\mu t}\right)\right) \quad (3.1)$$

The probability generating function of the buffer size distribution as

$$P(s,t)=\exp\left[\frac{\lambda+\alpha t}{\mu}(s-1)\left(1-e^{-\mu t}\right)\quad \lambda<\mu\right] \quad (3.2)$$

The mean number of requests in the buffer is

$$L(t)=\left|\frac{\partial P}{\partial s}\right|_{s=1}=\left[\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}\left(1-e^{-\mu t}\right)\right] \quad (3.3)$$

The utilization of the cloud is

$$U(t)=1-P_0(t)=\left[1-\exp\left\{-\left[\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}\left(1-e^{-\mu t}\right)\right]\right\}\right] \quad (3.4)$$

The variance of the number of requests in the buffer is

$$\mathrm{var}(N)=\left[\frac{\lambda+\alpha t}{\mu}\left(1-e^{-\mu t}\right)\right] \quad (3.5)$$

The throughput of the cloud is

$$\mu\left(1-P_0(t)\right)=\mu\left[1-\exp\left\{-\left[\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}\left(1-e^{-\mu t}\right)\right]\right\}\right] \quad (3.6)$$

The mean delay in the buffer is

$$W(t)=\frac{L(t)}{\mu\left(1-P_0(t)\right)}=\frac{\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}\left(1-e^{-\mu t}\right)}{\mu\left[1-\exp\left(-\frac{(\lambda-\alpha/\mu)+\alpha t}{\mu}\left(1-e^{-\mu t}\right)\right)\right]} \quad (3.7)$$

IV. PERFORMANCE EVALUATION OF THE CLOUD

In this section, the performance of the cloud is discussed through numerical illustration. Different values of parameters are considered for allocation of resources(VMs) to the job requests. After that we considered that job arrival parameter λ varies from 2 x 10^4 to 6 x10^4 Jobs/sec, α varies from 0.2 to 0.8 and μ varies from 3 x 10^4 to 8 x10^4.

Using equations (3.1), (3.3), (3.4), (3.6) and (3.7) are probabilities of the Cloud emptiness, mean number of jobs in the buffer, Utilization, Throughput, mean delay are computed for different values of t, λ, α and μ are presented in table 1

TABLE.1 Different parameter values for varying α, t, λ , μ

α	T	λ	μ	$P_0(t)$	L(t)	U(t)	Th(t)	W(t)
0.2	1	2	3	0.466648	0.675707	0.533352	1.600056	0.422302
0.2	2	2	3	0.458287	0.77585	0.541713	1.625139	0.477405
0.2	5	2	3	0.376146	0.977777	0.623854	1.871562	0.522439
0.2	10	2	3	0.269520	1.311111	0.730480	2.191439	0.598288
0.2	15	2	3	0.193120	1.644444	0.80688	2.420641	0.679343
0.2	5	2	8	0.689440	0.371875	0.31056	2.484477	0.149679
0.2	5	3	8	0.608429	0.496875	0.391571	3.132568	0.158616
0.2	5	4	8	0.536937	0.621875	0.463063	3.704506	0.167870
0.2	5	5	8	0.473845	0.746875	0.526155	4.209240	0.177437
0.2	5	6	8	0.418167	0.871875	0.581833	4.654666	0.187312

0.2	1	2	4	0.466648	0.675707	0.533352	1.600056	0.422302
0.2	1	2	5	0.573507	0.527655	0.426493	1.705973	0.309299
0.2	1	2	6	0.644835	0.429089	0.355165	1.775825	0.241628
0.2	1	2	7	0.695174	0.360216	0.304826	1.828955	0.196952
0.2	1	2	8	0.732629	0.309921	0.267371	1.871600	0.165592
0.2	1	2	3	0.466648	0.675707	0.533352	1.600056	0.422302
0.4	2	2	3	0.410093	0.886686	0.589907	1.769720	0.501031
0.6	5	2	3	0.201896	1.600000	0.798104	2.394311	0.668251
0.7	10	2	3	0.053814	2.922222	0.946186	2.838558	1.029474
0.8	15	2	3	0.010278	4.577778	0.989722	2.969167	1.541772

From the equations (3.1) to (3.7), it is observed that as the time increases the probability of the cloud emptiness is decreasing, the mean number of requests in the buffer is increasing, the utilization of cloud is increasing, the variance of the number of requests in the buffer are increasing, the throughput of the buffer are increasing, the mean delay in the buffers are increasing, when other parameters are fixed.

It is also observed that as the mean arrival requests rate increases, the probability of the cloud emptiness is decreasing, the mean number of requests in the buffer is increasing, the utilization of cloud is increasing, the variance of the number of requests in the buffer are increasing, the throughput of the buffer are increasing, the mean delay in the buffers are increasing, when other parameters are fixed.

It is also observed that as the service rate of the cloud increases, the probability of the cloud emptiness is increasing, the mean number of requests in the buffer is decreasing, the utilization of cloud is decreasing, the variance of the number of requests in the buffer are decreasing, the throughput of the buffer are increasing, the mean delay in the buffers are decreasing, when other parameters are fixed. Few graphs representing different parameter values for the table are given below:

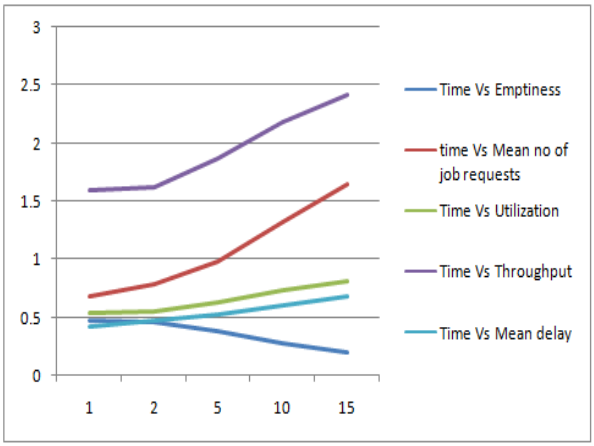


Fig.2 Analysis of parameters for constant λ , μ at varying t(α=0.2)

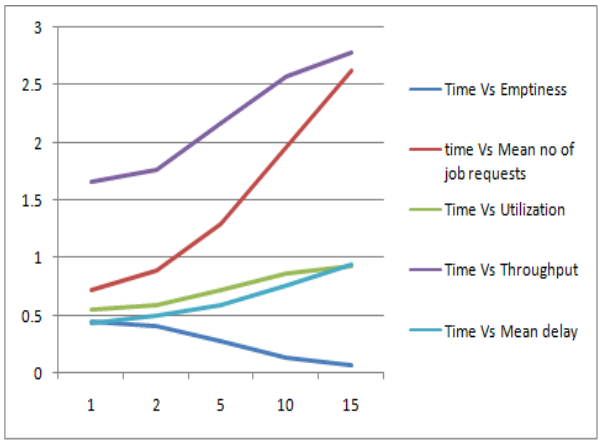


Fig.3 Analysis of parameters for constant λ , μ at varying t(α=0.4)

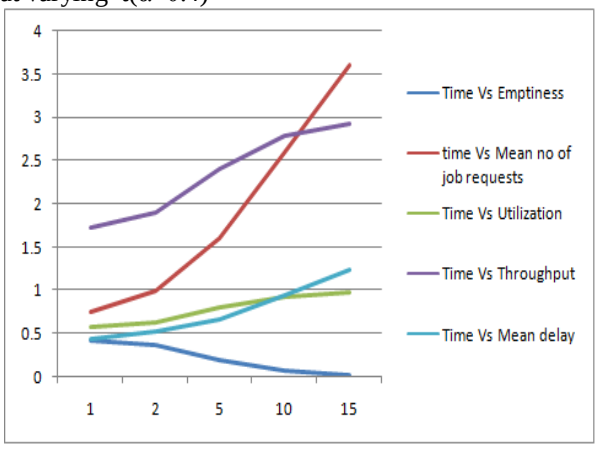


Fig.4 Analysis of parameters for constant λ , μ at varying t(α=0.6)

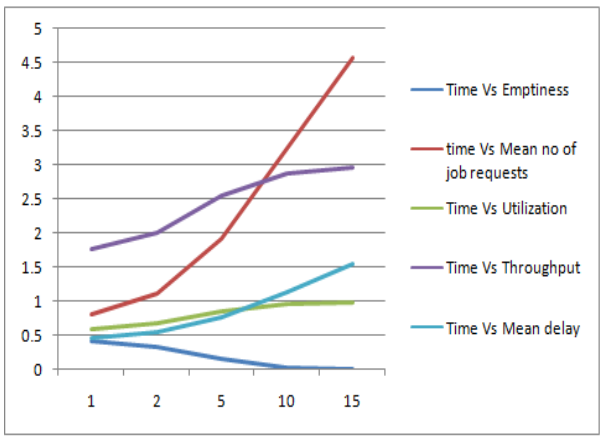


Fig.5 Analysis of parameters for constant λ , μ at varying t(α =0.8)

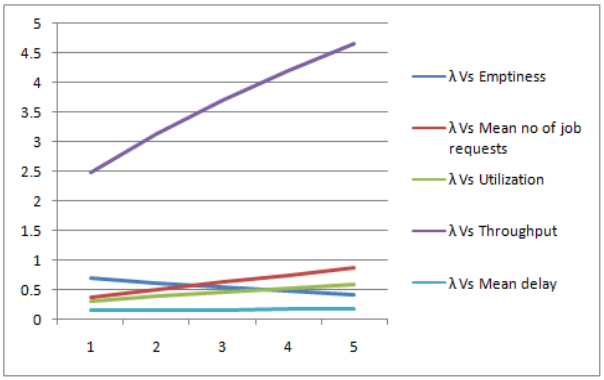


Fig.6 Analysis of parameters for constant t, μ at varying λ (α=0.2)

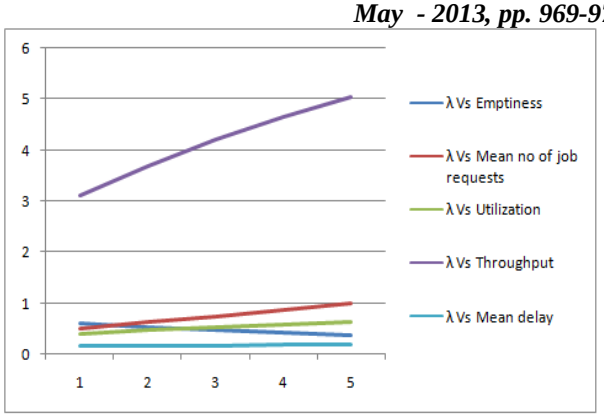


Fig.7 Analysis of parameters for constant t, μ at varying λ (α=0.4)

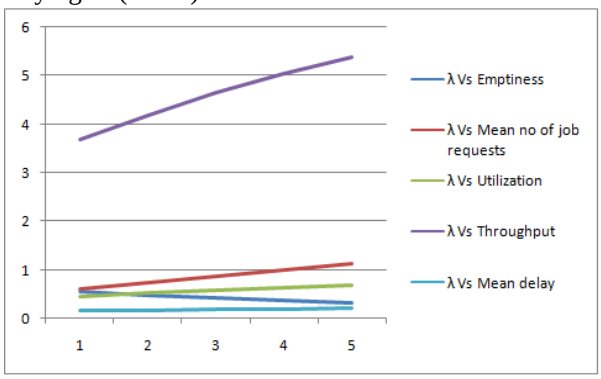


Fig.8 Analysis of parameters for constant t, μ at varying λ (α=0.6)

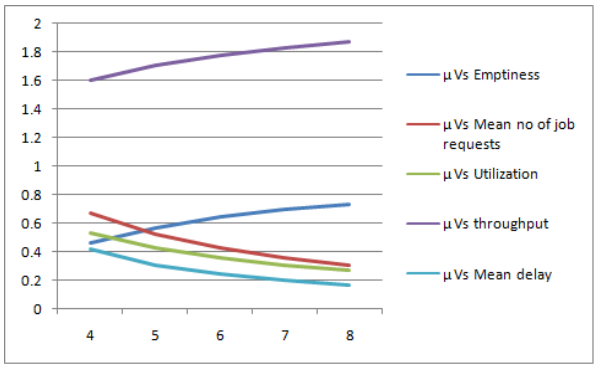


Fig.9 Analysis of parameters for constant t, λ at varying μ (α=0.2)

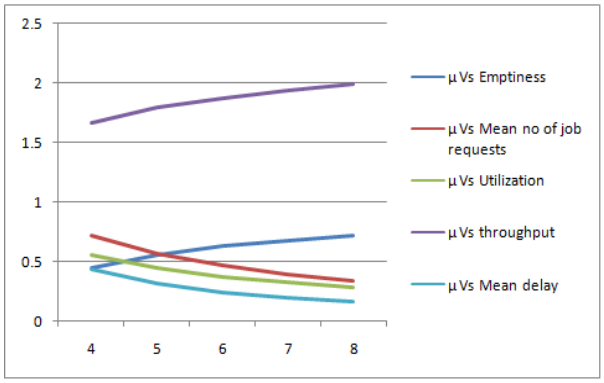


Fig.10 Analysis of parameters for constant t, λ at varying μ (α=0.4)

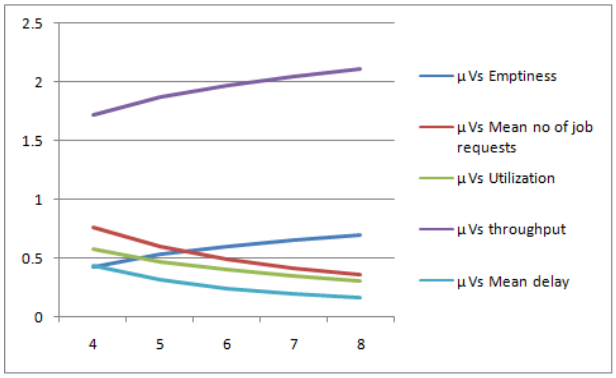


Fig.11 Analysis of parameters for constant t, λ at varying μ (α=0.6)

V. Conclusion

In this paper, we developed a novel cloud computing model which is much useful for analysing the cloud more effectively and efficiently to increase performance measures of cloud. The work presented in this paper focus on the improvement of allocation of resources dynamically following request dependent strategy under non homogeneous condition with time dependent arrival of jobs. It shows that dynamic allocation of resources can reduce mean delay and mean service time. The developed model is much useful for evaluating the performance of several clouds like google apps, amazon etc,. This model includes some of the earlier models as particular cases for specific values of the parameters.

References

1. H. Karlapudi, and J. Martin, "Web application performance prediction", Proceedings of the IASTED International Conference on Communication and Computer Networks, Boston, MA, Nov 2004 ,pp 281-286.
2. J. Lu, and J. Wang, "Performance modeling and analysis of Web Switch", In Proceedings of the 31st Annual *International conference on Computer Measurement* (CMG05), Orlando, FL Dec 2005.
3. Hao-peng CHEN, Shao-chong Li.A, "Queueing-based Model for Performance Management on Cloud", Proceedings of IEEE International conference on Advanced Information Managements and Services 2010, pp.83-88.

4. R.D Mei, and H.B Meeuwissen, "modeling end-to-end Quality-of-Service for transaction-based services in multi-domain environment", In performance Challenges for *Efficient Next Generation Networks* (Eds. X.J. Liang, Z.H. Xin, V.B Iversen. and G.S. Kuo), Proceedings of the 19th International Tele traffic Congress (ITC19), Beijing, China ,Aug 2005 pp. 1109-1121.
5. Kaiqi Xiong and Harry Perros, "Service Performance and Analysis in Cloud Computing", ICWS in Proc International workshop on Cloud Computing,july,6-10(2009),LA,2009.
6. P. Suresh varma, A. Satyanarayana, Rama sundari M.V, "Performance Analysis of Cloud Computing Using Queuing Models", International conference on cloud computing technologies and management (ICCCTAM-12), IEEE, 8-10 dec 2012, pp 12-15.
7. T.Sai Sowjanya, D.Praveen, K.Satish, A.Rahmain, "The Queueing Theory in Cloud Computing to Reduce the waiting Time", in *IJCSET* ,April-2011.
8. R.DMei,H.B Meeuwissen,andF.Phillipson, "User perceived Quality-of-Service for voice-over-IP in a heterogeneous multi-domain network environment", In Proceedings of *ICWS*,Sept 2006.
9. Sonam Rathore, "Efficient allocation of virtual machine in cloud computing environment", International journal of computer science and informatics, Vol.2, Issue 3, 2012, 59-62.
10. Rubinstein R. Y, "Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models", Operations Research Vol.37, 1989, 72-81.

