Sensitivity Analysis of Cloud under Transient Conditions

Satyanarayana.A[1], Dr. P. Suresh Varma[2], Dr.M.V.Rama Sundari[3], Dr.P Sarada Varma[4]

**[1] Computer Science Department, Adikavi Nannaya University,
Rajahmundry, Andhra Pradesh, India**

**[2] Computer Science Department, Adikavi Nannaya University,
Rajahmundry, Andhra Pradesh, India**

**[3] Information Technology Department, GIET,
Rajahmundry, Andhra Pradesh, India**

**[4] Mathematics Department, GMRIT,
Rajam, Andhra Pradesh, India**

Abstract

Cloud computing is a new economic computing paradigm in which information and shared resources can be accessed from a Web browser by customers. Clouds consists of a collection virtualized resources which includes both computational software and hardware facilities that can be provided to the customers based on demand. This on demand model creates a flexible and cost effective means to access compute resources. In cloud computing it is customary to consider that the job arrivals are characterized by Poisson process. This is assumption holds good if the job arrivals are homogeneous and independent of time. In this paper, we develop and analyze a cloud computing model with the assumption that the job arrivals are characterized by homogeneous Poisson Process. It is further assumed that allocation of Resources (Virtual Machines (VMs)) for each job dependent on number of jobs in the buffer at that instant using dynamic allocation strategy. The transient behavior of the cloud also analyzed by deriving various measures like mean number of jobs in the buffer, throughput of the cloud, utilization of the cloud and mean delay in allocation of resources etc,. The sensitivity analysis of the cloud reveals that the homogeneous Poisson arrivals and dynamic resource allocation strategy can reduce burstness in buffer and improve quality of service.

Keywords: Cloud Computing, Dynamic Resource Allocation, Performance Evaluation, Poisson Process, Sensitivity Analysis.

1. Introduction

Cloud computing has been often used with synonymous terms such as Platform as a service (PaaS), software as a service (SaaS), grid computing, cluster computing, and utility computing[1]. Services offered by web technology are becoming increasingly powerful in the information technology world. Cloud computing is completely new technology put forward from industry circle, it is the development of parallel computing, distributed computing, and grid computing cost efficient computing paradigm in which information and computational power can be accessed from the web browser easily by customers over internet. A service provider (such as Google or Amazon) provides hardware and virtualization software, and sometimes also applications. Instead of us, as users, hosting our own servers, their computers run virtual machines where our server can reside. The service provider's machine has a certain type of software that does virtualization (VM ware), so that a single machine in the provider can run many virtual machines. Multiple jobs enter into the Cloud requesting for various resources or services. In earlier approach of Cloud computing, which provided shared pool of resources on-demand over network on pay-per use[2], so the overload on that machine increases which affects the system performance. To increase performance, the Cloud requires efficient allocation of resourced virtual machines is essential in Cloud environment for increasing resource virtualization and efficient deployment of applications in virtual machines. The main advantage of having multiple virtual machines in Cloud computing is the system performance increases effectively by reducing the mean job queue length and waiting time of the job in the queue compared to other traditional approach of having only single virtual machine on each server so that the jobs need not wait for a long period of time and also queue length need not be large. Due to the unpredicted nature of demands of Virtual machines, waiting time of the jobs increases. Generally performance is mainly concerned with the problems of allocation and distribution of virtual machines to the jobs[3]. Recently P.Suresh varma & A. satyanarayana[4] has developed and analyzed a Cloud Computing model using dynamic allocation of virtual machines based on job queue length using queuing analogy.

In this paper we are considered dynamic allocation of virtual machines depending upon the requests in the buffer. Very little work is reported regarding sensitivity analysis of Cloud [5,6]. It is basically the study of how the uncertainty in the output of Cloud model that can be apportioned to different parameters of uncertainty in its inputs. Sensitivity analysis is very useful when attempting to determine the impact the actual outcome of a particular variable will have if it differs from what was previously assumed. Sensitivity Analysis is used to determine how "sensitive" a model is to changes in the value of the performance measures of the model and to changes in the structure of the model [7,8]. Parameter sensitivity is usually performed as a series of tests in which the modeler sets different parameter values to see how a change in the parameter causes a change in the dynamic behavior of the Cloud. By showing how the model behavior responds to changes in parameter values, sensitivity analysis is a useful tool in model building as well as in model evaluation [9,10].

The problem of Cloud service and performance modeling subject to QoS metrics, such as mean queue length, throughput, utilization and mean delay time have been studied in the literature. For example R D Mei et al [11], H Karlapudi and J Martin[1] presented a model of web server performance in which an open queuing network was employed to model the behavior of Web servers on the internet. Based on the literature we developed and analyzed a model by considering service is depending on requests in the queue which is very useful in Cloud computing. It is generally difficult to perform laboratory experiment that capture dynamic allocation of virtual machines (i.e. changing virtual machine just before allocation of virtual machine) effect on jobs under wide variety of traffic conditions. In addition to these complexities in empirical analysis the virtual machines are connected to the buffer. Therefore to study the performance evaluation of dynamic virtual machine allocation through traffic dependent strategy, we develop markovian model (Using Queuing analogy). In this paper we study the performance of a Cloud with number of virtual machines connected to the queue under transient conditions. Using the difference differential equation, the joint probability distribution of the buffer size is derived. The performance measures mean queue length, throughput, utilization and mean delay in the Cloud are also analyzed.

In this paper we analyzed a Cloud, using queuing model which provide services by considering service depends on requests generated by web browsers. The sensitivity analysis of system has been studied by considering various parameters like mean numbers requests in the system, Mean Delay, Throughput and Utilization of the cloud. It is observed that load dependent service has tremendous influence on mean delay, throughput and congestion control. From sensitivity study a small change in the parameters λ, μ can bring drastic changes in the performance of the system.

2. Cloud Computing Model

Considering a Cloud computing system the requests from various clients are stored in buffer which is connected to dispatcher for allocation of resources in the Cloud. The computing is carried with request dependent strategy. The dispatcher allocates resources (Virtual Machines(VMs)) to the jobs based on 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 Dispatcher are in series with load dependent resource allocation is given in fig.1

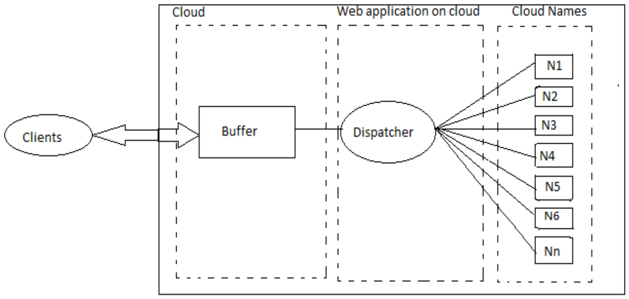

Fig. 1 Cloud computing model

Here, we assume that the arrival of requests follows a Poisson process with parameter λ and the number of virtual machine allocations in the Cloud follows Poisson processes with parameters μ. 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 occurrences of the jobs in non-overlapping time intervals of time are statistically independent.
2. The probability that there is arrival of one job in the first buffer 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 [nμh + 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 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:

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

$$\frac{\partial P_0(t)}{\partial t} = -\lambda P_0(t) + \mu P_1(t) \qquad \text{n} = 0 \qquad (2)$$

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:

$$P(s,t) = \exp\left[\frac{\lambda}{\mu}(s-1)\left(1-e^{-\mu t}\right)\right] \qquad (3)$$

3. Performance Analysis of Cloud computing under Transient Conditions

Consider a Cloud computing system the requests are packetized at source and stored in buffer which is connected to first transmitter for transmission in the Cloud. The transmission is carried with request dependent strategy. In request dependent strategy the transmission rate is a linear function of the number of request packets in the buffer depending on the buffer content, the transmission time of the request packet is fixed with dynamic allocation of virtual machines in the Cloud.
Here, we assume that the arrival of request packets follows a Poisson process with parameter λ and the number of transmission completion in the Cloud follow Poisson processes with parameters μ. The mean transmission rate of the Cloud is linearly dependent on the number of request packets in the buffers. The request packets are transmitted through the Cloud by first in first out principle. The schematic diagram representing the one buffer and one dispatcher in series with request dependent allocation of VMs is given in fig.1

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:

$$P(s,t) = \exp\left[\frac{\lambda}{\mu}(s-1)\left(1-e^{-\mu t}\right)\right] \qquad (4)$$

4. Performance measure of the cloud

Expanding $P(s;t)$ given in equation (4) and collecting the constant terms, we get the probability that the Cloud is empty as:

$$P_0 = \exp\left(-\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right) \qquad (5)$$

The probability generating function of the buffer size distribution is:

$$P(s,t) = \exp\left[\frac{\lambda}{\mu}(s-1)\left(1-e^{-\mu t}\right) \quad \lambda < \mu\right] \qquad (6)$$

The mean number of requests in the buffer is:

$$L(t) = \left|\frac{\partial P}{\partial s}\right|_{s=1} = \left[\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right] \qquad (7)$$

The utilization of the Cloud is:

$$U(t) = 1 - P_0(t) = \left[1 - \exp\left\{-\left[\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right]\right\}\right] \qquad (8)$$

The variance of the number of job requests in the buffer is:

$$\text{var}(N) = \left[\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right] \qquad (9)$$

The throughput of the Cloud is:

$$\mu\left(1-P_0(t)\right) = \mu\left[1 - \exp\left\{-\left[\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right]\right\}\right] \qquad (10)$$

The mean delay in the buffer is:

$$W(t) = \frac{L}{\mu\left(1-P_0(t)\right)} = \frac{\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)}{\mu\left[1-\exp\left(-\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right)\right]} \qquad (11)$$

5. Sensitivity analysis of the model

From the equations (5) to (11), 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 buffer 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.

The sensitivity analysis of the cloud is performed with respect to the effect of the changes in the parameters t, λ, μ on mean number of jobs in the buffer, utilization, throughput and mean delay.

The following data has been considered for the sensitivity analysis, t = 1 sec, λ = 2 jobs/sec, μ = 3 jobs/sec. The mean number of jobs, the utilization of Cloud, the throughput of the cloud and mean delay are computed with variation of -15%, -10%, -5%, 0%, +5%, +10%, +15% on the cloud parameters and presented in Table 1, 2 and 3.

Fig.2 Sensitivity analysis of L(t), U(t), Th(t) and W(t) at varying values of λ

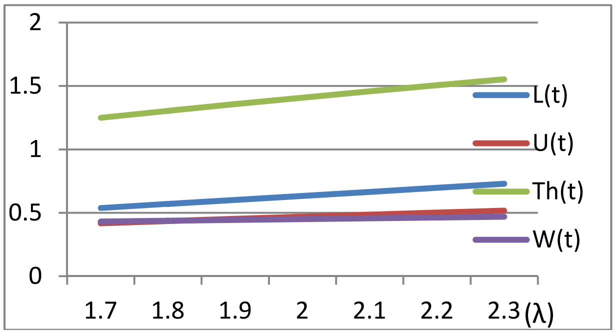

Fig.3 Sensitivity analysis of L(t), U(t), Th(t) and W(t) at varying values of μ

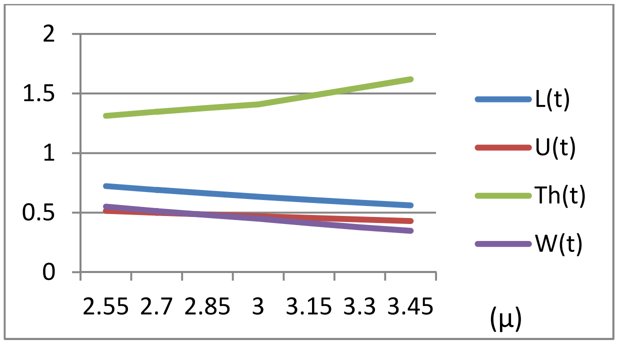

Fig.3 Sensitivity analysis of L(t), U(t), Th(t) and W(t) at varying values of t

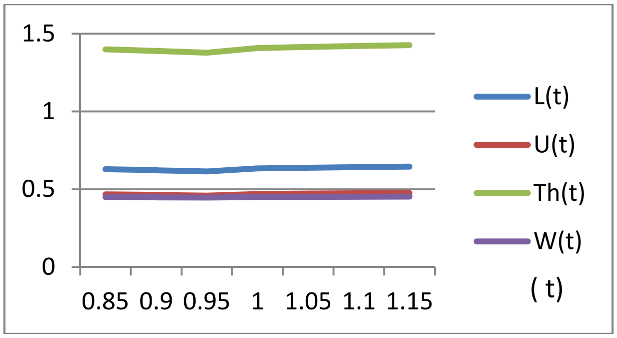

Table1. The values of L(t), U(t), Th(t), W(t) for different values of λ.

	λ-15%	λ-10%	λ-5%	λ	λ+5%	λ+10%	λ+15%
L(t)	0.538454	0.570128	0.601802	0.633475	0.665149	0.696823	0.728497
U(t)	0.41635	0.434547	0.452176	0.469256	0.485803	0.501834	0.517366
Th(t)	1.24905	1.30364	1.356529	1.407768	1.457409	1.505503	1.552098
W(t)	0.431091	0.437335	0.443634	0.449986	0.456391	0.46285	0.469363

Table2 . The values of L(t), U(t), Th(t), W(t) for different values of μ

	μ-15%	μ-10%	μ-5%	μ	μ+5%	μ+10%	μ+15%
L(t)	0.723073	0.690959	0.661162	0.633475	0.607713	0.583707	0.561307
U(t)	0.514741	0.498905	0.483749	0.469256	0.455405	0.442173	0.429537
Th(t)	1.31259	1.347043	1.378684	1.407768	1.478156	1.548544	1.618933
W(t)	0.550875	0.512945	0.47956	0.449986	0.411129	0.376939	0.346714

Table3 . The values of L(t), U(t), Th(t), W(t) for different values of t

	t-15%	t-10%	t-5%	t	t+5%	t+10%	t+15%
L(t)	0.628104	0.621863	0.614612	0.633475	0.638099	0.642078	0.645503
U(t)	0.466397	0.463057	0.459149	0.469256	0.471704	0.473802	0.475601
Th(t)	1.399192	1.38917	1.377448	1.407768	1.415112	1.421406	1.426804
W(t)	0.448905	0.447651	0.446196	0.449986	0.450917	0.45172	0.452412

The performance measures are highly affected by the changing in the values of t and cloud parameters λ and μ. As t increases from -15% to +15%, the mean number of jobs in the buffer is increasing along with the utilization, throughput and mean delay in the cloud. As the parameter λ increases from -15% to +15%, the mean number of jobs in the buffer is increasing along with the utilization, throughput and mean delay in the cloud. Similarly for the parameter μ increases, the mean number of jobs in the buffer is increasing along with the utilization, throughput and mean delay in the cloud.

6. Conclusion

In this paper, we developed and analyzed cloud computing model by applying queuing theory to allocate resources(VMs) depending upon buffer size, which can increase the performance of cloud. The sensitivity analysis was also studied with respect to mean number of jobs in the buffer, utilization, throughput, mean delay. It is observed that using queuing theory in the allocation of resources (VMs) has tremendous influence on mean number of jobs in the buffer, utilization, throughput and mean delay.

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] 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.

[3] Rubinstein R. Y, “Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models”, Operations Research Vol.37, 1989, 72-81.

[4] In the proceeding of International conference on cloud computing technologies and management (ICCCTAM-12) organized by bits, dubai from 8th to 10th 2012

[5] 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

[6] T.Sai Sowjanya, D.Praveen, K.Satish, A. Rahmain, “The Queueing Theory in Cloud Computing to Reduce the waiting Time”, IJCSET ,April-2011.

[7] Arsham. H, A..Feuerverger, D. L. Mcleish, J. Kreimer and R.Y. Rubinstein,“Sensitivity Analysis and the What If Problem in Simulation Analysis”, Mathl. Comput. Modeling 12, 1989, pp.193-219.

[8] Glynn, P. W. A. Thesen, H. Grant and W D Kelton (eds.), “Likelihood Ratio Gradient Estimation: An Overview”, In Proceedings of the Winter Simulation Conference, IEEE Press, 1987, pp.366-375.

[9] Nakayama, M. K., A. Goyal, and P.W. Glynn, “Likelihood Ratio Sensitivity Analysis for Markovian Models of Highly Dependable Systems”, Operations Research 42, 1994, 137-157.

[10] Reiman. M. I, and A. Weiss, “Sensitivity Analysis for Simulations Via Likelihood Ratios”, Operations Research 37, 1989, pp.830-844.

[11] R.D Mei, H.B Meeuwissen, and F. Phillipson, “User perceived Quality-of-Service for voice-over-IP in a heterogeneous multi-domain network environment”, Proceedings of ICWS, Sept 2006.

Mr.A.Satyanarayana received the Bachelor’s degree in Computer Science and Engineering from NEC engineering college, A.P. Master’s degree in Computer Science and Engineering from RVR&JC College of Engineering and Technology, Guntur, Andhra Pradesh, India, in 2009. He Currently Pursuing PhD in Department of Computer Science at Adikavi Nannaya University, Rajahmundry, A.P, India. His area Research interests include Artificial Intelligence, Parallel processing, distributed computing and cloud computing.

Dr. P. Suresh Varma received M.Tech (CST) from Andhra University. He received Ph.D. (CSE) from Acharya Nagarjuna University. He is currently working as Professor in Department of Computer Science and Dean, College Development Council in Adikavi Nannaya University, Rajahmundry, A.P., India. He published several papers in National and International Journals. He is active life member of various professional bodies like ISTE, ORSI, ISCA, IE etc,. His current research is focused on Computer Networks, Cloud Computing and Data Mining.

Dr.M.V.Rama Sundari received the B.Tech(CS) from Acharya Nagarjuana University, M.Tech (CST) and Ph.D.(CSSE) from Andhra University. She is currently working as Associate Professor, Department of Information Technology, GIET, Rajahmundry. Her current research is foucussed on Computer Networks, Cloud Computing etc,.

Dr.P.Sarada Varma received the M.Sc Mathematics and Ph.D.in Mathematics from Andhra University. She is currently working as Assistant Professor, Department of Mathematics, GMRIT, Rajam. Her current research is System Modeling and Cloud Computingetc,.