International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-4 Issue-5, November 2014

A Queuing Model To Improve Quality of Service by Reducing Waiting Time in Cloud Computing

G. Vijaya Lakshmi, C. Shoba Bindhu

ABSTRACT- Cloud computing is an emerging technology to provide cost effective and to deliver the business applications, services in an adaptable way. In cloud computing, multi resources such as processing, bandwidth and storage, need to be allocated simultaneously to multiple users. The When cloud computing users(CCU'S) requests for the service to the cloud computing service providers (CCSP) at the same time but while at a moment, if cloud computing server is busy CCU's needs to enter into the waiting line until CCSP completes its service to the previous CCU . So this may leads to bottleneck in the network. . Therefore cloud computing users neither utilize the resources nor waits in the queue. Cloud Computing service providers use multiple servers to reduce the waiting time .Therefore, it is necessary to consider a measure for congestion control in cloud computing environment. This paper proposes a (M/M/C):(∞/FIFO) Queuing model which is applied at multiple servers inorder to reduce waiting time , queue length also improving the network performance and QOS effectively in cloud computing environment.

Keyword: Cloud Computing, waiting time, Queuing Theory, QOS.

I. INTRODUCTION

Cloud computing[1][5][3] often referred as 'cloud' is the delivery of ondemand computing resources everything from applications to data centers over the internet on a pay for use basis. There are most significant components of cloud computing architecture which are known as frontend and backend which they connect to each other over the internet. The front end is where the clients/users sees. The backend is the 'cloud' itself comprising various computers , servers and data storage devices.

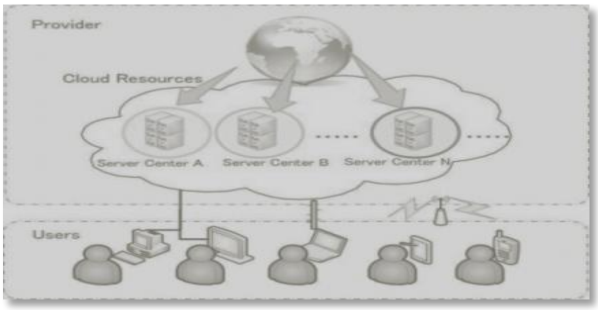

Fig1.Cloud Computing

Cloud computing providers offer their services [2] over the internet. These services are broadly divided into three categories namely:

Infrastructure as a software (Iaas): for instance webservices provides for the clients /users with virtual server instances and storages as well as application program interface that allows the user to access, to configure their virtual servers and storage.

- Platform as a service (Paas) : provides cloud based environment to access the resources such as physical machines, virtual machines etc
- Software as a service(Sass): allows to provide software applications as a services to the end users. It refers to software that is deployed on a hosted service and is accessible via internet.

Cloud computing can be used as software as a service (Saas) [6] .The main aim of the cloud service providers is to administer the system , monitor traffic flow to ensure maximum usage of the resources[5] in minimum waiting time. When Multiple users enters into the Cloud for sharing the resources or data at a time, when server is busy the CCU forms queue or may enter into reneging state which degrades the network performance . Therefore, in this paper (M/M/C): (∞/FIFO) Queuing model is applied at multiple servers to reduce mean waiting time which results in decrease in queue length and improving QoS in cloud computing environment.

II. RELATED WORK

Queuing theory has been applied to develop analytical methods for evaluating Cloud service performance. Xiong and Perros[8] modeled a Cloud computing system as an open queue network consisting of two tandem servers with finite buffer space, where both interarrival and service times are assumed to have exponential distributions . T.Saisowjanya et al.,[7] have shown M/M/S model for two servers which increases the performance over using one server by reducing the queue length and waiting time.. In order to study resource allocation for meeting performance requirements of clients with different priority levels [10] modelled a Cloud centre as an M/M/C/C queuing system, which has C servers with no buffer space and Markov processes for both arrival and departure. Yang et al [9] developed an queuing model for Cloud data centres . In this model both arrival and service times are assumed to be exponentially distributed and service response time is broken into three independent parts: waiting, service, and execution periods. In [11] they employ the queuing model to investigate resource allocation problems in both single-class service case and multiple-class service case. Furthermore, they optimize the resource allocation to minimize the mean response time.

Manuscript Received on November 2014.

G. Vijaya Lakshmi, Asst. Pro. Department of Computer Science, Vikrama Simhapuri University Dargamitta, Potti Sreeramulu, Nellore, Andhra Pradesh, India .
C. Shoba Bindhu, Asst. Pro. Department of Computer Science, Vikrama Simhapuri University Dargamitta, Potti Sreeramulu, Nellore, Andhra Pradesh, India

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

III. QUEUEING THEORY

Queuing Theory [2] is a collection of mathematical models of various systems of queues. It is widely used to analyze the arrival rate and service time. Formation of queues arises when demand for a service exceeds the limited capacity of the system. To analyse the arrival rate & service rate and to deliver the packet to the destination a Queuing model[4] which is a Mathematical, Probalistic and Markovian model is applied at routing stages.

Queuing system is characterized by the components namely:

a) Arrival rate: describes the way the population arrives either static or dynamically..
b) Service rate: describes how many customers can be served when the service is available .
c) No of service channels: Service channel contains single or multiple. Customers enter one of the parallel service channels and is served by the customer
d) Queue discipline: describes the manner in which customers choose for the service like First in First out(FIFO), Last in First Out(FIFO).

Customer behaviour generally be in four states. They are:

- Balking : when the Queue is too long customer decides to enter or not in the queue .
- Reneging: The customer leaves from the queue if he has impatience to wait.
- Jockeying :when there are two or more parallel queues the customer move from one queue to other.

KENDELL'S NOTATION

A Queuing system can be described based on their notations:

A/S/M/B/K/D where
A : probability distribution of the arrival rate
S : service time distribution
M : number of servers
B : system capacity
K: population size
D : service discipline

Key notations:

λ: Mean arrival rate
μ:Mean Service rate
p= λ/ μ: server utilization

Steady state distribution: the system is in steady state when the behaviour of the system becomes independent of time.

IV. (M/M/C):(∞/FIFO) QUEUING MODEL

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 inorder to overcome this problem **(**M/M/C): (∞/FIFO) Queuing model is applied when there are multiple servers , C and each server has an independent identical exponential service time distribution. The arrival process assumed to be poisson .and ∞ indicates CCU. T he mean service rate will be Cμ .

The steady state probabilities are:

$$p_n = \begin{cases} \dfrac{\rho^n}{n!} p_0 & n = 1,2,\cdots,c \quad \text{where} \quad p_0 = \dfrac{1}{\sum \dfrac{\rho^n}{n!} + \dfrac{\rho^c}{c!\left(1-\rho/c\right)}} \\[2ex] \dfrac{\rho^n}{c!c^{n-c}} p_0 & n = c, c+1, \cdots \end{cases}$$

Measures of effectiveess :

$$L_q = \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} p_0$$

$$L = L_q + \rho$$

$$W_q = \frac{L_q}{\lambda}$$

$$W = W_q + \frac{1}{\mu}$$

Where ,

(Lq): Avg no of customers in queue.

L : Expected number of customers in the system.

Wq: expected waiting time per customers in the queue.

W: expected waiting time per customers in the system.

V. RESULT

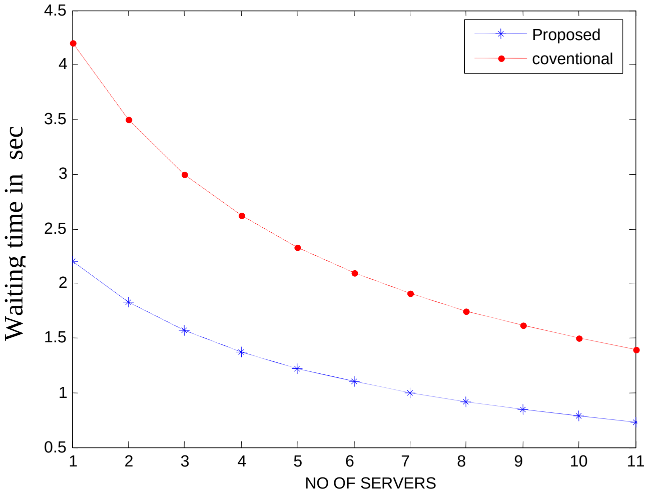

Fig 2. No of servers Vs waiting time

VI. CONCLUSION

When there are more number of cloud computing users in the queue, while the server is busy then there will be a formation of queue where the resources will not be obtained to users. This paper proposes (M/M/C): (∞/FIFO) Queuing model for multiple servers. It is observed from the result that when there will be more number of servers, waiting time reduces when compared to conventional method. Therefore Quality of service

(QOS) can be achieved effectively as CCSP’s provides the resources for CCU’s in Cloud Computing.

REFERENCES

[1] Suneeta Mohanty, Prasant Kumar Pattnaik and Ganga Bishnu Mund “A Comparative Approach to Reduce the Waiting Time Using Queuing Theory in Cloud Computing Environment” International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 4, Number 5 (2014), pp. 469-474.

[2] N.Ani Brown Mary and K.Saravanan “ 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.

[3] Kusaka, T, Okuda, T, Ideguchi, T, Xuejun,Tian “ Queuing theoretic approach to server allocation problem in time-delay cloud computing systems “ Teletraffic congress (ITC), 2011, 23rd International publications , 2011 pp:310-311.

[4] C. Knessl, B. Matkowsky, Z. Schuss and C. Tier, “Asymptotic analysis of a state-dependent M/G/1queueing system,” SIAM J. Appl. Math. 46 (1986) 483–505.

[5] Souvik Pal and P. K. Pattnaik, “Efficient architectural Framework of Cloud Computing”, in “International Journal of Cloud Computing and Services Science (IJ-CLOSER)”, Vol.1, No.2, June 2012, pp. 66-73.

[6] P.Mell and T. Grance, “Definition of Cloud Computing” v15, National Institute of Standards and Technology (NIST), 2009.

[7] T. sai Sowjanya et al, “The Queuing Theory in cloud Computing to Reduce the Waiting Time”, International Journal of Computer Science and Engineering Technology, April 2011, Vol. 1, Issue 3, pp. 110-112.

[8] K. Xiong and H. Perros, “Service performance and analysis in cloud computing,” in Proceedings of the 5th World Congress on Services (SERVICES ’09), Los Angeles, Calif, USA, July 2009, pp. 693–700.

[9] B. Yang, F. Tan, Y.-S. Dai, and S. Guo, “Performance evaluation of cloud service considering fault recovery,” in Proceedings of the 1st International Conference on Cloud Computing (CloudCom '09), Beijing, China, December 2009, pp. 571–576.

[10] W. Ellens, M. Zivkovic, J. Akkerboom, R. Litijens, and H. Berg, “Performance of cloud computing centers with multiple priority classes,” in Proceedings of the 5th IEEE International Conferenceon Cloud Computing,Honolulu, Hawaii, USA, June 2012, pp. 245–252.

[11] X. M. Nan, Y. F. He, L. Guan. “Optimal Resource Allocation for Multimedia Cloud Based on Queuing Model”, Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on, 2011, pp.1-6 .

Dr. G.Vijaya Lakshmi, received her B.Tech Degree in Computer Science & Engineering from Intel College of Engineering, Affiliated to JNTU, Anantapur, India, in 2002. M.Tech in Software Engineering in JNTU College of Engineering, Anantapur, India, in 2005.Received Ph.D degree in Computer Science & Engineering from JawaharLal Nehru Technological University, Anantapur, A.P., India in February 2012. At Present she is Assistant Professor, Computer Science Dept in Vikrama Simhapuri University, Nellore, INDIA . Her current Research Interest includes computer networks , wireless communication network, cloud computing.

Dr. C. Shoba Bindu received her B.Tech Degree in Electronics & Comm. Engineering from JNTU College of Engineering, Anantapur, India, in 1997; M.Tech in Computer Science & Engineering from JNTU College of Engineering, Anantapur, India, in 2002. Received Ph.D degree in Computer Science from Jawahar Lal Nehru Technological University, Anantapur, A.P., India in May 2010. At present she is Associate Professor, Computer Science & Engg Dept in J.N.T.U.A, Anantapur, INDIA. Her Current Research Interest includes Network Security and Wireless Communication Systems, cloud computing.