RESEARCH ARTICLE OPEN ACCESS

# The application of queue theory in cloud computing to reduce the waiting time

N.N. Bharkad*, Dr. M.H. Durge**
*Department of Mathematics, Institute of Technology & Management, Nanded-431 602, MS, India.
**Department of Mathematics, A.N. College of Arts, Commerce & Science,
Anandwan, Warora, Dist. Chandrapur, MS, India.

**ABSTRACT**
Cloud computing is a new technology in computer field to provide on line service to the customers. -Cloud computing has got enormous popularity as it offers dynamic, low-cost computing solutions. To get the service of cloud the user has to be in queue until he is served. 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 it 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 immediately available to others. From the user's point of view he needs to be served immediately and to prevent waiting the CCSP's can use infinite servers to reduce waiting time & queue length. The arrival pattern is often Poisson in queuing theory. In this article we analyzed the dynamic behavior of the system with infinite servers by finding various effective measures like response time, average time spend in the system, utilization and throughput.
***Keywords-*** Cloud computing, Poisson process, Queueing theory, Waiting time.

## I. INTRODUCTION

When you store your data online instead of on your home computer, or use web mail or a social networking site, you are using a "cloud computing" service. If you are an organization, and you want to use, for example, an online invoicing service instead of updating the in-house one you have been using for many years, that online invoicing service is a "cloud computing" service.Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on your own hard drive or updating applications for your needs, you use a service over the Internet, at another location, to store your information or use its applications. Doing so may give rise to certain privacy implications.

## II. CLOUD COMPUTING

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, web mail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications. [1]

The following definition of cloud computing has been developed by the U.S. National Institute of Standards and Technology (NIST):
Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.1

### A) Characteristics

The characteristics of cloud computing include on-demand self service, broad network access, resource pooling, rapid elasticity and measured service. On-demand self service means that customers (usually organizations) can request and manage their own computing resources. Broad network access allows services to be offered over the Internet or private networks. Pooled resources means that customers draw from a pool of computing resources, usually in remote data centres. Services can be scaled larger or smaller; and use of a service is measured and customers are billed accordingly.

### B) Service models

The cloud computing service models are Software as a Service (SaaS), Platform as a Service

(PaaS) and Infrastructure as a Service (IaaS). In Software as a Service model, a pre-made application, along with any required software, operating system, hardware, and network are provided. In PaaS, an operating system, hardware, and network are provided, and the customer installs or develops its own software and applications. The IaaS model provides just the hardware and network; the customer installs or develops its own operating systems, software and applications.

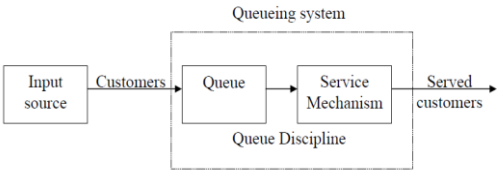
### C) Deployment of cloud services:

Cloud services are typically made available via a private cloud, community cloud, public cloud or hybrid cloud. Generally speaking, services provided by a public cloud are offered over the Internet and are owned and operated by a cloud provider. Some examples include services aimed at the general public, such as online photo storage services, e-mail services, or social networking sites. However, services for enterprises can also be offered in a public cloud. In a private cloud, the cloud infrastructure is operated solely for a specific organization, and is managed by the organization or a third party. In a community cloud, the service is shared by several organizations and made available only to those groups. The infrastructure may be owned and operated by the organizations or by a cloud service provider.

## III. QUEUEING THEORY

A queuing system consists of one or more servers that provide service of some sort to arriving customers. Customers who arrive to find all servers busy generally join one or more queues (lines) in front of the servers, hence the name **queuing systems**. There are several everyday examples that can be described as queuing systems, such as bank-teller service, computer systems, manufacturing systems, maintenance systems, communications systems and so on. [2]

### A) STUCTURE OF QUEUEING SYSTEM:




### B) COMPONENTS OF QUEUE SYSTEM:

A queuing system can be completely described by
a) Arrival pattern (input) ii) Service pattern (service mechanism) iii) Queue discipline iv) customer behavior

**i) Arrival pattern (input):** It represent the pattern in which the customers arrive and join the system. Usually, the customers arrive in a more and less random way which is not worth making prediction. Thus the pattern can be described in terms of probabilities and consequently the probability distribution of for inter arrival times must be defined.
The present paper is only dealt with those queuing systems in which the the customers arrive in **Poisson or completely random fashion**. Other types of arrival pattern may also be observed in practice that has been studied in queuing theory. Two such patterns **are arrivals are of regular intervals and there is general distribution of time between successive arrivals.**

**ii) Service pattern (service mechanism):** It is specified when it is known how many customers can served at a time, what the statistical distribution of service is, and when service is available. It is true in most situations that service time is a random variable with the same distribution for all arrivals, but cases occur where there are clearly two or more classes of customers each with a different service time distribution. Service time may be constant or a random variable. Distributions of service time which are important in practice are **negative exponential distribution and the related Erlang (Gamma) distribution.**
In the present paper only those queuing systems are discussed in which the service time follows the **exponential distribution.**

iii) **Queue discipline:** it is the rule by which customers selected from the queue for service. The most common discipline is FIFO(first in ,first out) or FCFFS( first come ,first out),according to which the customers are served in the order of their arrival and the other service disciplines are :
LIFO-last in, first out or FILO-fist in last out
SIRO- Service in random order and Priority.
In this paper we shall be concerned only with first come first serve discipline (FCFS).

iv) **Customer behavior:** The customers generally behave in the following three ways:
1) Balking: A customer may leave the queue because the queue is very long and s/he has no time to wait or there is no sufficient waiting space.
2) Reneging: This occurs when a waiting customer leaves the queue due to impatience.
3) Jockeying: Customers may jockey from one waiting line to another.[3]

### Transient and Steady states:

Queuing Theory analyses the study of systems behavior over time. A system is said to be transient state when its operating characteristics are dependent on time.
A system is said to be transient state when its operating characteristics are independent on time.
In some situations if the arrival rate of the system is larger than its service rate a steady state cannot be reached regardless of the length of the elapsed time. In fact in this case the queue length will increase with

time and theoretically it could build up to infinite such case is called explosive case.
In this paper only steady state analysis will be considered.

### C) Kendall Notation

D.G. Kendall developed a notation that has been widely accepted for specifying the pattern of arrivals, the service time distribution, and the number of channels in a queuing model. This notation is often seen in software for queuing model. The basic three-symbol Kendall notation is in the form: arrival distribution/service time distribution/number of service channels open. Where specific letters are used to represent probability distributions. An abridged version of this convention is based on the format A/B/C/D/E/F. These letters represent the following system characteristics:[4]
**A** = represents the inter arrival-time distribution.
**B** = represents the service-time distribution.
[Common symbols for **A** and **B** include **M** (exponential), **D** (constant or deterministic), **Ek** ( Erlang of order k), and **G**(arbitrary or general)]
**C** = represents the number of parallel servers.
**D** = represents the queue discipline.
**E** = represents the system capacity.
**F** = represents the size of the population.

### D) Single-Channel Queuing Model with Poisson Arrivals and Exponential Service Times (M/M/1):

We present an analytical approach to determine important measures of performance in a typical service system. After these numeric measures have been computed, it will be possible to add in cost data and begin to make decisions that balance desirable service levels with waiting line service costs.

**Assumptions of the Model**

The single-channel, single-phase model considered here is one of the most widely used and simplest queuing models. It involves assuming that seven conditions exist:

**1.** Arrivals are served on a FIFO basis.

**2.** Arrivals are described by a Poisson probability distribution and come from an infinite or very large population.

**3.** Service times also vary from one customer to the next and are independent of one another, but their average rate is known.

**4.** Service times occur according to the negative exponential probability distribution.

**5.** The average service rate is greater than the average arrival rate.

When these five conditions are met, we can develop a series of equations that define the queue's operating characteristics.

## IV. QUEUING EQUATIONS

$\lambda$ = mean number of arrivals per time period .
μ = mean number of people or items served per time period
When determining the arrival rate ( $\lambda$ ) and the service rate (μ), the same time period must be
Used. For example, if the is the average number of arrivals per hour, then must indicate the average number that could be served per hour.[5]

1) Average number of units the in the system : $L_s = \lambda / \mu - \lambda$
2) Average number of units the in the queue : $L_q = \lambda^2 / \mu(\mu - \lambda)$
3) Average waiting time of an arrival in the system : $W_s = 1 / \mu - \lambda$
4) Average waiting time of an arrival in the queue : $W_q = \lambda / \mu(\mu - \lambda)$
5) Variance of queue length = $\dfrac{\lambda/\mu}{(1-\lambda/\mu)^2}$
6) The utilization factor for the system : $\rho = \dfrac{\lambda}{\mu}$
7) Probability of queue being greater than or equal to k:P(n $\geq k$) = $(\dfrac{\lambda}{\mu})^k$
8) The present idle time, that is, the probability that no one is in the system:$P_0$=1-$\dfrac{\lambda}{\mu}$

### E) Multiple-Channel Queuing Model With Poisson Arrivals And Exponential service Times (M/M/S):

The next logical step is to look at a multiple-channel queuing system, in which two or more servers or channels are available to handle arriving passengers. Let us still assume that travelers awaiting service form one single line and then proceed to the first available server. Each of these channels has an independent and identical exponential service time distribution with mean 1/ μ .The arrival process is Poisson with rate $\lambda$ . Arrivals will join a single queue and enter the first available service channel.The multiple-channel system presented here again assumes that arrivals follow a Poisson probability distribution and that service times are distributed exponentially. Service is first come, first

served, and all servers are assumed to perform at the same rate.Other assumptions listed earlier for the single-channel model apply as well.[6,7]

**Equations for the Multichannel queuing Model:**

If we let

$\lambda$ = average arrival rate, and

μ = average service rate at each channel.

C= number of parallel service channels

$P_n$=probability of n customers in the system

n =number of customers in the system

The following formulas may be used in the waiting line analysis

1) Probability of the system shall be idle:

$$P_0=\frac{1}{\sum_{n=0}^{c-1}\frac{\left(\lambda/\mu\right)^n}{n!}+\frac{\left(\lambda/\mu\right)^c}{c!}\times\frac{c\mu}{c\mu-\lambda}}$$

2) average number of customers in the queue :

$$L_q=\frac{\lambda\cdot\mu\left(\lambda/\mu\right)^c}{(c-1)!(c\mu-\lambda)^2}\times p_0$$

3) average number of customers in the system: $L_s=L_q+\frac{\lambda}{\mu}$

4) average number of customers in the system: $W_q=\frac{L_q}{\lambda}$

5) average number of customers in the system: $W_s=\frac{L_s}{\lambda}$

6) probability of a customer has to wait: $P(n\geq c)=\frac{\mu\left(\frac{\lambda}{\mu}\right)^c}{(c-1)!(c\mu-\lambda)}\times P_0$

7) probability of a customer enters the service without waiting =1-P(n≥c)

8) Utilization factor : $\rho=\frac{\lambda}{c\mu}$

**F)Numerical example results and analysis**:

We are taken some examples and calculated there different parameters as given in the following Tables

For M/M/1 model

| Sr. no. | $\lambda$ | $\mu$ | $\rho$ | $L_s$ | $L_q$ | $W_s$ | $W_q$ |
|---|---|---|---|---|---|---|---|
| 1 | 12 | 5 | 2.4 | -1.71 | -4.10 | -0.14 | -0.34 |
| 2 | 15 | 6 | 2.5 | -1.67 | -4.17 | -0.11 | -0.27 |
| 3 | 17 | 6 | 2.8 | -1.54 | -4.31 | -0.9 | -0.25 |
| 4 | 20 | 7 | 2.9 | -1.54 | -4.46 | -0.7 | -0.20 |
| 5 | 23 | 8 | 2.9 | -1.53 | -4.48 | -0.7 | -0.20 |
| 6 | 22 | 8 | 2.7 | -1.57 | -4.28 | -0.7 | -0.19 |
| 7 | 26 | 9 | 2.9 | -1.53 | -4.48 | -0.6 | -0.17 |
| 8 | 21 | 8 | 2.6 | -1.61 | -4.18 | -0.8 | -0.21 |
| 9 | 20 | 8 | 2.5 | -1.67 | -4.17 | -0.8 | -0.20 |
| 10 | 17 | 6 | 2.8 | -1.54 | -4.31 | -0.9 | -0.25 |
| 11 | 20 | 7 | 2.9 | -1.54 | -4.47 | -0.8 | -0.23 |
| 12 | 20 | 8 | 2.5 | -1.67 | -4.17 | -0.8 | -0.20 |
| 13 | 17 | 6 | 2.8 | -1.54 | -4.31 | -0.9 | -0.23 |
| 14 | 15 | 7 | 2.1 | -1.88 | -3.95 | -0.12 | -0.25 |
| 15 | 13 | 6 | 2.2 | -1.86 | -4.9 | -0.14 | -0.30 |
| 16 | 10 | 6 | 1.7 | -2.5 | -4.25 | -0.25 | -0.42 |

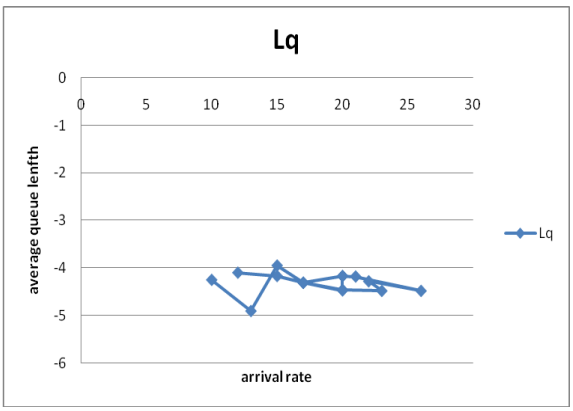
For M/M/2 model

| Sr. no. | $\lambda$ | $\mu$ | $\gamma$ | $\rho$ | $P_0$ | $L_q$ | $L_s$ | $W_q$ | $W_s$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 12 | 5 | 2.4 | 1.20 | -0.09 | -7.7 | -6.5 | -0.64 | -0.54 |
| 2 | 15 | 6 | 2.5 | 1.25 | -0.11 | -10.7 | -9.45 | -0.71 | -0.63 |
| 3 | 17 | 6 | 2.8 | 1.40 | -0.17 | -5.8 | -4.4 | -0.34 | -0.31 |
| 4 | 20 | 7 | 2.9 | 1.45 | -0.18 | -5.4 | -3.95 | -0.27 | -0.19 |
| 5 | 23 | 8 | 2.9 | 1.45 | -0.18 | -5.4 | -3.95 | -0.23 | -0.17 |
| 6 | 22 | 8 | 2.7 | 1.35 | -0.14 | -7.65 | -6.3 | -0.34 | -0.28 |
| 7 | 26 | 9 | 2.9 | 1.45 | -0.18 | -5.4 | -3.95 | -0.20 | -0.15 |
| 8 | 21 | 8 | 2.6 | 1.30 | -0.13 | -6.34 | -5 | -0.30 | -0.23 |
| 9 | 20 | 8 | 2.5 | 1.25 | -0.11 | -10.7 | -9.45 | -0.53 | -0.47 |
| 10 | 17 | 6 | 2.8 | 1.40 | -0.11 | -5.8 | -4.4 | -0.34 | -0.26 |
| 11 | 20 | 7 | 2.9 | 1.45 | -0.18 | -5.4 | -3.95 | -0.27 | -0.20 |
| 12 | 20 | 8 | 2.5 | 1.25 | -0.11 | -10.7 | -9.45 | -0.53 | -0.47 |
| 13 | 17 | 6 | 2.8 | 1.40 | -0.17 | -5.8 | -4.4 | -0.34 | -0.26 |
| 14 | 15 | 7 | 2.1 | 1.05 | -0.02 | -18.52 | -17.47 | -1.2 | -1.16 |
| 15 | 13 | 6 | 2.2 | 1.10 | -0.04 | -10.64 | -9.54 | -0.81 | -0.73 |
| 16 | 10 | 6 | 1.7 | 0.85 | 0.08 | 4.35 | 5.2 | 0.435 | 0.52l |

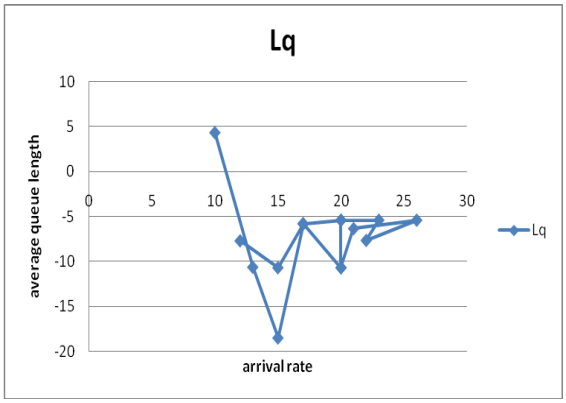
For M/M/3 model

| Sr. no. | $\lambda$ | $\mu$ | $\gamma$ | $\rho$ | $P_0$ | $L_q$ | $L_s$ | $W_q$ | $W_s$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 12 | 5 | 2.4 | 0.80 | 0.056 | 2.58 | 3.38 | 0.215 | 0.282 |
| 2 | 15 | 6 | 2.5 | 0.83 | 0.045 | 3.36 | 4.19 | 0.224 | 0.279 |
| 3 | 17 | 6 | 2.8 | 0.94 | 0.014 | 13.37 | 14.31 | 0.786 | 0.842 |
| 4 | 20 | 7 | 2.9 | 0.95 | 0.011 | 16.99 | 17.94 | 0.849 | 0.897 |
| 5 | 23 | 8 | 2.9 | 0.95 | 0.011 | 16.99 | 17.94 | 0.738 | 0.78 |
| 6 | 22 | 8 | 2.7 | 0.92 | 0.020 | 9.43 | 10.35 | 0.428 | 0.47 |
| 7 | 26 | 9 | 2.9 | 0.96 | 0.010 | 24.38 | 25.34 | 0.937 | 0.97 |
| 8 | 21 | 8 | 2.6 | 0.87 | 0.035 | 5.278 | 6.148 | 0.251 | 0.292 |
| 9 | 20 | 8 | 2.5 | 0.83 | 0.045 | 3.36 | 4.19 | 0.209 | 0.209 |
| 10 | 17 | 6 | 2.8 | 0.94 | 0.014 | 13.37 | 14.31 | 0.842 | 0.842 |
| 11 | 20 | 7 | 2.9 | 0.95 | 0.011 | 16.99 | 17.94 | 0.849 | 0.897 |
| 12 | 20 | 8 | 2.5 | 0.83 | 0.045 | 3.36 | 4.19 | 0.168 | 0.209 |
| 13 | 17 | 6 | 2.8 | 0.94 | 0.014 | 13.37 | 14.31 | 0.786 | 0.026 |
| 14 | 15 | 7 | 2.1 | 0.71 | 0.094 | 1.225 | 1.935 | 0.082 | 0.129 |
| 15 | 13 | 6 | 2.2 | 0.72 | 0.083 | 1.879 | 2.599 | 0.144 | 0.199 |
| 16 | 10 | 6 | 1.7 | 0.55 | 0.169 | 0.376 | 0.092 | 0.037 | 0.092 |

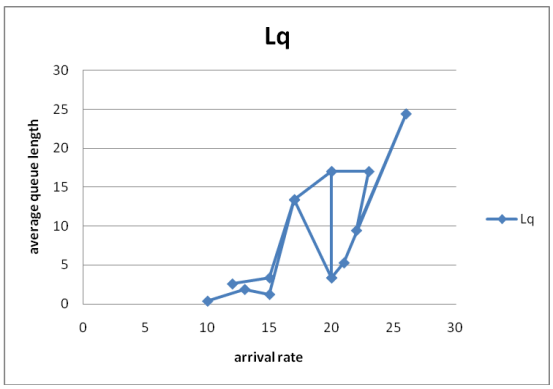
Graph of M/M/1




Graph of M/M/2




Graph of M/M/3




## V. CONCLUSION

The aim of this work is to develop an efficient procedure for cloud computing queuing. From the above study it is clear that to reduce the waiting time of customer, we have needed to increase to three number of service channels. This work do not include both waiting as well as service cost due to non availability of required data. In further study the cost factor will be considered to find better cloud computing queuing.

## REFERENCES

[1] Huimin Xiao, Guozheng Zhang, The Queuing Theory Application in Bank Service Optimization, *IEEE* 2010.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Luqun Li, "An Optimistic Differentiated Service Job Scheduling System for Cloud Computing Service Users and Providers", *Third International Conference on Multimedia and Ubiquitous Engineering*,2009.

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