PERFORMANCE ANALYSIS OF CLOUD COMPUTING CENTER WITH QUEUEING MODEL

S. Anand Gnana Selvam[1]*, S. Rita[2], M. Reni Sagayaraj[3]

[1]Department of Mathematics, AET College, Salem, Tamilnadu.
[2]Department of Statistics, Periyar University, Salem, Tamilnadu
[3]Principal, Holy Cross College, Tirupattur, Vellore, Tamilnadu.
Corresponding author: (e-mail: anandjuslin@gmail.com).

ABSTRACT

Cloud computing is a novel raising computing resource allocation. Successful development of cloud computing paradigm necessitates explicit performance evaluation of cloud data centers. The computing resource allocation and performance managing have been one of the most important septets of cloud computing. In this paper, we consider the cloud center as a queuing system with single task arrivals and a task request buffer of infinite capacity. We assess the performance of queuing system by using an analytical model and solve it to obtain important performance factors like mean number of tasks in the system. Using this model in order to evaluate the performance analysis of cloud server farms and obtained solved it to obtain accurate estimation of complete probability distribution of the request response time and other paramount performance indicators.

Keywords: Cloud computing, performance analysis, response time, queuing theory, Markov chain process

I. INTRODUCTION

Cloud computing is the Internet-based expansion and use of computer knowledge. It has become an IT buzzword for the past a few years. Cloud computing has been often used with synonymous terms such as software as a service, grid computing, cluster computing, autonomic computing, and utility computing [1]. Cloud computing is a novel paradigm for the provision of computing infrastructure, which aims to shift the location of the computing infrastructure to the network in order to reduce the costs of management and maintenance of hardware and software resources. This cloud concept emphasizes the transfers of management, maintenance and investment from the customer to the provider[7]. Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources Networks, servers, storage, applications, and services. Cloud Computing has become one of the most talked about technologies in recent times and has got lots of attention from media as well as analysts because of the opportunities it is offering. Cloud Computing encompasses different types of services. The cloud has a service-oriented architecture, and there are three classes of technology capabilities that are being offered as a service[8]. Queuing theory is a collection of mathematical models of various queuing systems. Queues or waiting lines arise when demand for a service facility exceeds the capacity of that facility i.e. the customers do not get service immediately upon request but must wait or the service facilities stand idle and waiting for customers. The basic queuing process consists of customers arriving at a queuing system to receive some service. In [10] the servers are busy, they join the queue in a waiting in line. They are then served according to a prescribed However, cloud centers differ from traditional queuing systems in a number of important aspects. A cloud center can have a large number of facility server, nodes, typically of the order of hundreds or thousands; traditional queuing analysis rarely considers systems of this size. Task service times must be modeled by a general, rather than the more convenient exponential, probability distribution. Moreover, the coefficient of variation of task service time may be high well over the value of 1. Due to the dynamic nature of cloud environments, diversity of user requests and time dependency of load, cloud centers must provide expected quality of service at widely varying loads. In [2] the cloud center as an M/G/m/m+r queueing system with single task arrivals and a task request buffer of finite capacity. The performance using analytical model and solve it to obtain important performance factors like mean number of tasks in the system. In [5] cloud environment as an M/G/m queuing system which indicates that inter-arrival time of requests is exponentially distributed, the service time is generally distributed and the number of facility nodes is m, without any restrictions on the number of facility nodes. the system performance increases efficiently by reducing the mean queue length and waiting time than compared to the conventional approach of having only single server so that the consumers need not wait for a long period of time and also queue length need not be bulky.

II. MODEL DESCRIPTION

We can define ergodicity of a Markov chain as follows: A Markov chain is called ergodic if it is irreducible, recurrent non-null, and a periodic. We define communicability as follows, State I communicates with j, written in $i \to j$, if the chain may ever visit state j with positive probability, starting from i. That is, $i \to j$ if pij(n) > 0 for some $n \geq 0$. We say i and j inter communicate if $i \to j$ and $j \to i$, in which case we write $i \leftrightarrow j$. It can be seen that $\leftrightarrow$ is an equivalence relation, hence the state space S can be partitioned into the equivalence classes of $\leftrightarrow$; within each equivalence class all states are of the same type.

A set C of states is called

(a) Closed, if pij = 0 for all i ∈C, j /∈C.

(b) Irreducible, if i_j for all i, j ∈C.

The kendall's classification of queuing systems exists in several modifications. Queuing models are generally constructed to represent the steady state of a queuing system, that is, the typical, long run or average state of the system. As a consequence, these are stochastic models that represent the probability that a queuing system will be found in a particular configuration or state.

A general procedure for constructing and analyzing such queuing models is:

1. Identify the parameters of the system, such as the arrival rate, service time, queue capacity, and perhaps draw a diagram of the system,

2. Identify the system states.

3. Draw a state transition diagram that represents the possible system states and identify the rates to enter and leave each state. This diagram is a representation of a

Markov chain,

4. Because the state transition diagram represents the steady state situation between states there is a balanced flow between states so the probabilities of being in adjacent

states can be related mathematically in terms of the arrival and service rates and state Probabilities,

5. Express all the state probabilities in terms of the empty state probability, using the inter-state transition relationships, 6. Determine the empty state probability by using the fact that all state probabilities always sum to 1.

M/M/1 represents a single server that has unlimited queue capacity and infinite calling population, both arrivals and service are Poisson (or random) processes, meaning the statistical distribution of both the inter-arrival times and the service times follow the exponential distribution. Because of the mathematical nature of the exponential distribution, a number of quite simple relationships can be derived for several performance measures based on knowing the arrival rate and service rate. M/G/1 represents a single server that has unlimited queue capacity and infinite calling population, while the arrival is still Poisson process, meaning the statistical distribution of the inter-arrival times still follow the exponential distribution, the distribution ofthe service time does not. The distribution of the service time may follow any general statistical distribution, not just exponential. Relationships can still be derived for a number of performance measures if one knows the arrival rate and the mean and variance of the service rate. However the derivations are generally more complex and difficult. As most of these results relyon some approximation(s) to obtain a closed-form solution, they are not universally applicable.

1. Approximations are reasonably accurate only when the number of servers is comparatively small, typically below 10 or so, which makes them unsuitable for performance analysis of cloud computing data centers.

2. Approximations are very sensitive to the probability distribution of task service times, and they become increasingly inaccurate when the coefficient of variation of the service time, CoV, increases toward and above the value of one.

3. Finally, approximation errors are particularly pronounced when the traffic intensity is small. As a result, the results mentioned above are not directly applicable to performance analysis of cloud computing server farms where one or more of the following holds: the number of servers is huge; the distribution of service times is unknown and does not, in general, follow any of the well behaved probability distributions such as exponential distribution; finally, the traffic intensity can vary in an extremely wide range.

Let us assume that the arrivals follow a Poisson process with rate of arrival. We also assume that provision times are independently and identically distributed random variables with an arbitrary probability distribution. Let b(t) be the probability density function of provision time T between 2 departures. Let N(t) be the number of consumers in the system at time $t \geq 0$. Let N(t) be the number of consumers in the system at time . Let be the time instant at which the nth consumer completes service and departs. Let represents the number of consumers in the system when the nth customer departs. Also, the sequence of

Random variables $\{X_n : n = 1, 2, 3, ...\}$ is a Markov chain. Hence we have,

$$X_{n+1} = \begin{cases} X_n - 1 + A, ifX_n < 0i, eX_n \geq 1 \\ A, ifX_n = 0 \end{cases}$$

where A is the number of customers arriving during the provision time "T" of the (n+1)th customer. We know that, if U(Xn) denotes the unit step function , then we can write,

$$U(X_n) = \begin{cases} 1, ifX_n < 0(or)X_n \geq 1 \\ 0, ifX_n = 0 \end{cases}$$

Therefore $X_n + 1$ can be written as

$$X_{n+1} = X_n - U\left(X_n\right) + A \qquad (1)$$

Suppose the system is in steady state, then the probability of the number of consumers in the system is independent of time and hence is a constant.

That is, the average size of the system at departure is

$$E\left(X_{n+1}\right) = E\left(X_n\right)$$

Taking expectation on both sides of (1), we get

$$E\left(X_{n+1}\right) = E\left(X_n - U\left(X_n\right) + A\right)$$

$$E\left(X_{n+1}\right) = E\left(X_n\right) = E\left(U\left(X_n\right)\right) + E\left(A\right) \qquad (2)$$

$$E\left(X_{n+1}\right) = E\left(X_n\right)$$

$$E\left(X_n\right) = E\left(X_n\right) - E\left(U\left(X_n\right)\right) + E\left(A\right) \qquad (3)$$

$$E\left(U\left(X_n\right)\right) = E\left(A\right)$$

Squaring equation (1), we have

$$X_{n+1}^2 = \left(X_n - U\left(X_n\right) + A\right)^2$$

$$= X_n^2 + U^2\left(X_n\right) + A^2 - 2X_nU$$

$$\left(X_n\right) + 2AX_n - 2AU\left(X_n\right) \qquad (4)$$

But

$$U^2\left(X_n\right) = \begin{cases} 1, if X_n^2 < 0 \\ 0, if X_n^2 = 0 \end{cases}$$

$$= \begin{cases} 1, if X_n < 0 \\ 0, if X_n = 0 \end{cases}$$

Therefore X_n denotes the number of consumers and hence X_n cannot be negative.

$$U^2\left(X_n\right) = U\left(X_n\right)\left[U\left(X_n\right) = 1(or)0\right]$$

Also,

$$X_n U\left(X_n\right) = X_n$$

Hence (4) becomes

$$X_{n+1}^2 = X_n^2 + U\left(X_n\right) + A^2 - 2X_n + 2AX_n - 2AU\left(X_n\right)$$

$$2X_n - 2AX_n = X_n^2 - X_{n+1}^2 + U\left(X_n\right) + A^2 - 2AU\left(X_n\right)$$

$$2X_n\left(1-A\right) = X_n^2 - X_{n+1}^2 + U\left(X_n\right) + A^2 - 2AU\left(X_n\right)$$

Taking expectation on both sides, we get

$$2\left[E\left(X_n\right) - E\left(AX_n\right)\right] = E\left(X_n^2\right) - E\left(X_{n+1}^2\right) + E\left(U\left(X_n\right)\right) + E\left(A^2\right) - 2E\left(AU\left(X_n\right)\right)$$

$$2\left[E\left(X_n\right) - E\left(A\right)E\left(X_n\right)\right] = E\left(X_n^2\right) - E\left(X_{n+1}^2\right) + E\left(U\left(X_n\right)\right) + E\left(A^2\right) - 2E\left(AU\left(X_n\right)\right)$$

Therefore A and X_n are independent

$$2E(X_n)\left[1 - E(A)\right] = E(A^2) - E(A^2) + E(A) + E(A^2) - 2E(A)E(A)$$

$$2E(X_n)\left[1 - E(A)\right] = E(A^2) + E(A) - 2[E(A)]^2$$

$$E(X_n) = \frac{E(A) - 2[E(A)]^2 + E(A^2)}{2(1 - E(A))}$$

Since the arrivals during "T" is a Poison process with rate λ,

$$E\left(A/T\right) = \lambda T$$

$$E\left(A^2/T\right) = \lambda^2 T^2 + \lambda T \qquad (6)$$

This is obtained by mean and variance of the poison process,

$$E\left(X(t)\right) = \lambda t$$

$$E\left(X^2(t)\right) = \lambda^2 t^2 + \lambda t$$

Also,

$$E(A) = E\left(E\left(A/T\right)\right)$$

$$= E\left(\lambda T\right)$$

$$E(A) = \lambda E(T) \qquad (7)$$

Similarly,

$$E(A^2) = E\left(E\left(A^2/T\right)\right)$$

$$= E\left(\lambda^2 T^2 + \lambda T\right)$$

$$E(A^2) = \lambda^2 E(T^2) + \lambda E(T) \qquad (8)$$

Now equation (5) becomes,

$$E\left(X_n\right) = \frac{\lambda^2 E(T^2) + \lambda E(T) + \lambda E(T) - 2\left[\lambda E(T)\right]^2}{2\left(1 - \lambda E(T)\right)}$$

$$= \frac{\lambda^2 E(T^2) + 2\lambda E(T) - 2\left[\lambda E(T)\right]^2}{2\left(1 - \lambda E(T)\right)}$$

$$= \frac{2\lambda E(T)[1 - \lambda E(T)] + \lambda^2 E(T^2)}{2\left(1 - \lambda E(T)\right)}$$

$$E\left(X_n\right) = \frac{2\lambda E(T)\left[1 - \lambda E(T)\right]}{2\left(1 - \lambda E(T)\right)} + \frac{\lambda^2 E(T^2)}{2\left(1 - \lambda E(T)\right)} \quad (9)$$

The standard quantity of consumers in the system is obtained from the given equation. Notice that a multi server system with multiple identical servers has been configured to serve requests from certain application domain. Therefore, we will only focus on standard quantity of consumers in the system and do not consider other sources of delay, such as resource allocation and provision, virtual machine instantiation and deployment, and other overhead in a complex cloud computing environment.

III. Waiting Time Distribution

The waiting time of a consumer in the system is obtained with the help of the equation that has been already calculated as standard quantity of consumers in the system.

$$E\left(X_n\right) = \frac{2\lambda E(T)\left[1 - \lambda E(T)\right]}{2\left(1 - \lambda E(T)\right)} + \frac{\lambda^2 E(T^2)}{2\left(1 - \lambda E(T)\right)}$$

$$= \frac{2\lambda E(T)\left[1 - \lambda E(T)\right]}{2\left(1 - \lambda E(T)\right)\mu} + \frac{\lambda^2 E(T^2)}{2\left(1 - \lambda E(T)\right)\mu}$$

$$E\left(Xi_n\right) = \frac{2\rho E(T)\left[1 - \lambda E(T)\right]}{2\left(1 - \lambda E(T)\right)} + \frac{\rho\lambda E(T^2)}{2\left(1 - \lambda E(T)\right)} (10)$$

$$E\left(Xii_n\right) = \frac{2\lambda E(T)\left[1 - \lambda E(T)\right]}{2\left(1 - \lambda E(T)\right)} + \frac{\lambda^2 E(T^2)}{2\left(1 - \lambda E(T)\right)}$$

(11)

$$E\left(Xiii_n\right)=\left(\frac{2E(T)\left[1-\lambda E(T)\right]}{2\left(1-\lambda E(T)\right)}+\frac{\lambda^2 E(T^2)}{2\left(1-\lambda E(T)\right)}\right)-\left(\frac{1}{\mu}\right) \qquad (12)$$

With the help of waiting time distribution the delay and the queuing values are obtained by the consumers in the queue those who are waiting for the resources to be provided by the providers.

Figures and Tables (Table 1. Utility and Delay)

M/GD	Utility	Queue	Delay
1000	5.62037	0	0.00634
5000	21.16239	0	0.61171
10000	47.63793	1	5.96816
15000	86.85273	2	26.43108
20000	87.68589	2	29.94669
25000	88.572	2	48.10515
30000	89.55422	2	58.40535
35000	90.09729	3	51.01807
40000	113.27625	3	46.77912
45000	135.6864	4	44.72912
50000	157.6482	4	44.00284

Depending on the file sizes that is allotted in bytes the values are calculated for response time of user, and the users waiting in the queue and the waiting time is calculated. Here we can see clearly that the response time is more when compared with the waiting time. Figure 1. Utility and Delay

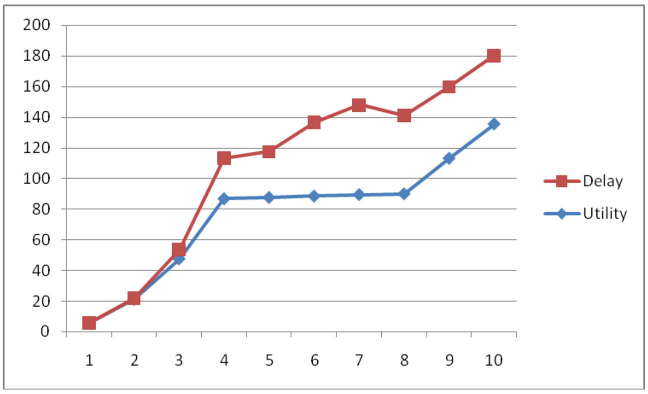

Figure 1. Utility and Delay

CONCLUSION

In this paper, we proposed model for performance evaluation of a cloud computing data centre with queue. We assessed the performance of queuing system by using an analytical model and solved it to obtain important performance factors like mean number of tasks in the system. Using this model in order to evaluate the performance analysis of cloud server farms and obtained solved it to obtained accurate estimation of complete probability distribution

REFERENCES

[1] W. Kim, “Cloud computing: Today and Tomorrow,” Journal of Object Technology, 8, 2009.

[2] L. Vaquero , L. Rodero-Merino , J. Caceres and m. Lindner, “A break in the clouds: towards a cloud definition,” ACM SIGCOMM Computer Communication Review, vol. 39, no.1, 2009.

[3] J. Zhang and N. Tansu, “Optical gain and laser characteristics of InGaN quantum wells on ternary InGaN substrates,” *IEEE Photon. J.*, vol. 5, no. 2, Apr. 2013, Art. no. 2600111.

[4] Bharathi, M., Sandeep Kumar, P., Poornima, G .V.”Performance factors of cloud computing data centers using M/G/m/m+r queuing systems”, IOSR Journal of Engineering (IOSRJEN) e-ISSN: 2250-3021, p-ISSN: 2278-8719.

[5] Fatima Oumellal, Mohamed Hanini and Abdelkrim Haqiq,”MMPP/G/m/m+r Queuing System Model to Analytically Evaluate Cloud Computing Center Performances ”, British Journal of Mathematics & Computer Science, 4(10): 1301, 1317, 2014.

[6] Hamzeh Khazaei, Jelena Misic,and Vojislav B. Misic,” Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems”, IEEE transactions on parallel and distributed systems, VOL. 23, NO. 5, 2012.

[7] Kulkarni, V.G, “Introduction to modeling and analysis of stochastic system” 2 nd edition, Springer text in statistic,2011.

[8] Wang L., Von Laszewski G., Younge A. , He X.,Kunze M., Tao J., and Fu C. (2010) Cloud Computing: A Perspective Study. New Generation Computing, vol. 28. 137-146.

[10]. Xiong K. and Perros H. (2009) Service Performance and Analysis in Cloud Computing. proceedings of the 2009 Congress on Services. Los Angeles, CA, 6-10 July 2009. IEEE. 693-700.

[11] Kimura T. (1996b) Optimal Buffer Design of an M/G/s Queue with Finite Capacity. Comm. In Statistics Stochastic Models, vol. 12, 165-180.