ijecs open access www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 8905-8910

Resource Allocation in Cloud Computing with General Classification Time and Exponential Service (G/M/s)

R.Murugesan[1], C.Elango[2], S.Kannan[3]

[1]Department of Computer Science,
Cardamom Planters' Association College,
Bodinayakanur, Tamilnadu,,India
rmncpa90@gmail.com

[2]Department of Mathematical Sciences,
Cardamom Planters' Association College,
Bodinayakanur, Tamilnadu,India
chellaelango@gmail.com

[3]Department of Computer Applications,
Madurai Kamaraj University,
Madurai, Tamilnadu, India
skannanmku@gmail.com

Abstract: **In this article we considered a Cloud Computing Network (CCN) with four nodes, classifier, SaaS, PaaS and IaaS. The classifier node serve as agent for the Service Level Agreement (SLA). It is a routing server which takes a random time which is independent identically and distributed. The other service stations took an exponentially distributed service time. Thus the CCN became a Network of G/M/s queues. The G/M/s type queue is justified because the arrival of service request to CCNs are in general not follow Poisson Process System perform measures are obtained to compute the total expected cost for the CCN.**

Keywords*:* Cloud Computing, Resource Allocation, Queueing Theory, Performance Measures**.**

1. Introduction

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. Cloud computing is a model for enabling ubiquitous, 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. Users get the computing resources and services by means of customized service level agreement (SLA); they only pay the fee according to the using time, using manner or the amount of data transferring, [6]. The main focuses on the SLA it emphasis the QoS of services, it includes availability, throughput, reliability, security, and many other parameters, but performance indicators are such as response time, task blocking probability, probability of immediate service, and mean number of tasks in the system, all of which may be determined by using the tool of Markov chain in queuing theory.

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: Infrastructure-as-a-Service (IaaS), where equipment such as hardware, storage, servers and network components are accessible via the Internet, the platform-as-a-Service (PaaS), which is a central component of the Cloud: the PaaS is responsible for developing applications for the cloud. It includes hardware with operating systems, virtualized servers, etc; and finally the Software-as-a-Service (SaaS) (resources software), which includes applications and other hosted services, [3].

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. If the servers are busy, they join the queue in a waiting room (i.e., wait in line). They are then served according to a prescribed discipline. 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 arrival process must be modeled by a general, rather than the more convenient exponential probability distribution. Moreover, the co-efficient of variation of task inter arrival time may be high (well over the value 1).
- Due to the dynamic nature of cloud environments, diversity of user's requests and time dependency of load, cloud centers must provide expected quality of service at widely varying loads. [4]

The authors already developed a CCN model, which has M/M/s type service stations [8] and M/G/service station [9]. In this paper, we study the resource allocation techniques to minimize the resource cost and minimize the service response time for cloud service providers. We model the cloud center as G/M/s queuing system with single task general process arrivals, with exponential service time and a task buffer of infinite capacity. We evaluate its performance using a combination of a transform-based analytical model and an approximate Markov chain model, which allows us to obtain a complete probability distribution of response time and number of tasks in the system

2. Related Work

Although cloud computing has attracted research attention, only a small portion of the work has addressed performance optimization question so far. In [4] an analytical technique based on an approximate Markov Chain model for performance evaluation of a cloud computing center. Due to the nature of the cloud environment, general service time for requests as well as large number of servers, which makes the model flexible in terms of scalability and diversity of service time. Numerical results showed that the proposed approximate method provides results with high degree of accuracy for the mean number of tasks in the system, blocking probability, probability of immediate service.

In [1] Cloud center as an [(M/G/1) : (∞/GD)] queuing system with single task arrivals and a task request buffer of infinite capacity. Evaluate the performance of queuing system using an analytical model and solve it to obtain important performance factors like mean number of tasks in the system. 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. There are lot works are carried out in this fields. In [3] Fatima Oumellal, Mohamed Hanini and Abdelkrim Haqiq, MMPP/G/m/m+r, the Markov modulated Poisson Process and the performance measures such as average number of tasks in the system blocking probability, probability of immediate serve the average response time. Recently the authors [9] studied a CCN system with M/G/s type servers and its performance. We focused on the how the computer resource can efficiently allocate different tasks in the cloud centers.

3. Proposed Model

Consider a cloud computing networks (CCN) which provides resources ranges from computing infrastructure and applications. The inter-arrival time of requests to the classifier node is independent identically distributed (iid) random variable with mean η and the task service times are exponentially distributed with parameters $\mu_i > 0$ (i=1, 2, 3, 4). Generally there are three kinds of requests. Depending on the type of clients request, three types of services are provided, namely Software (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The bag of task are arriving the first stations namely 'Classifier', with general inter arrival time distribution F(.) with mean η. The bag of tasks (BoTs) are taken for classification in FCFS discipline. After classification of BoTs according to SLA it moves to any one of the stations which provides SaaS, PaaS and IaaS. Each station i has s_i independent servers, and the queueing model at station i is of the type $G/M/s_i$. Here the input process in non-Markovians implies that output from each stations is non-Markovian. The cloud computing network diagram is described in figure 1.

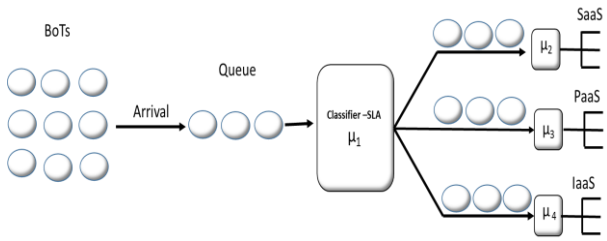


Fig. 1

3.1 Analysis

We model the Cloud Computing network, as a Open Jackson Queueing Network. Consider a general CCN with the following assumptions.

- The network has N single stations with s_i servers at each station i.
- There is an unlimited waiting space at each station (the classification and service stations).
- The customers (BoTs request) arrive at station 1 (classifier) from outside the network following a general distribution with inter arrival time distribution F (.), with mean rate η.
- Arrival process is independent of service process.
- Service times for customers (service requests) of station i are exponentially distributed with parameter $\mu_i > 0$.
- Customers (service requests) finishing service at station i proceeds to join the queue at station j with probability p_{ij} or leave the network altogether with probability r_i independently of each other.[8]

The probabilities p_{ij}, $i,j \in S = \{1, 2, \ldots, N\}$ is called the routing probabilities and the matrix $P = (p_{ij})$ $i,j \in S$ is called the routing probability matrix. By our assumption, the stochastic model of cloud computing network, we described becomes an Open Jackson Queuing Network with N stations and s_i servers at each station [4].

The routing matrix P can be expressed as a transition probability matrix of the form

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdot & \cdot & \cdot & p_{1N} \\ p_{21} & p_{22} & p_{23} & \cdot & \cdot & \cdot & p_{2N} \\ p_{31} & p_{32} & p_{33} & \cdot & \cdot & \cdot & p_{3N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{N1} & p_{N2} & p_{N3} & \cdot & \cdot & \cdot & p_{NN} \end{bmatrix}$$

We assumed in the CCN that each station has infinite space for waiting requests (jobs). Next we have to show that the CCN is stable in the long run.

Our previous work [8] is based on M/M/s represents that each station has unlimited queue capacity and infinite calling population, arrival process is Poisson and service time is exponentially distributed meaning the statistical distribution of both the inter-arrival times and the service times follow the exponential distribution. Because of the mathematical structure (members) of the exponential distribution, a number of quite simple relationships can be derived. Several performance measures based on the arrival rate and service rate are obtained. In the next model [9] we conduct a M/G/s system 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, but the distribution of the 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 (limited) number of performance measures if one knows the arrival rate and the mean and variance of the service times [6].

3.2 Model formulation:

Consider a single-station queueing system where customers arrive according to a renewal process with iid inter-arrival times having common cdf F (.) and mean η. The inter-arrival times may not be exponentially distributed. The queue is serviced by a s_i servers at node i and has an infinite waiting room. The service times are exponentially distributed random variables. Such a queue is called a G/M/s queue. If the inter-arrival times are exponentially distributed, it reduces to an M/M/s_i queue.

Let X(t)be the number of customers before the classification node l in the system at time t. {X(t),t ≥ 0} is a continuous-time stochastic process with state space {0, 1, 2,…}.Since the arrival process is not Poisson, and is of general nature, knowing the current state X(t) does not provide enough information about the time until the next arrival (unless the inter-arrival times are exponentially distributed), and hence we cannot predict the future based solely on X(t). Hence {X(t), t ≥ 0 }is not a CTMC. We will state some of the main results without proof and derive others from it. First define

$$\rho_i = \frac{\lambda_i}{\mu_i s_i},$$

as the traffic intensity of the G/M/s queue.

Stability of the G/M/s Queue. We begin with the study of stability of the G/M/s queue. We assume that the G/M/s queue is stable if $\rho_i < 1$, i=1, 2, … , N.

Indeed, it is possible to show that the G/M/s queue is unstable if $\rho_i \geq 1$. Thus $\rho_i < 1$is a necessary and sufficient condition of stability. We shall assume that the queue is stable in the remaining analysis. Then the general the key functional equations of a G/M/s queue is given by

$$u = \overline{F}(\mu(1-u)) \quad where\ the Laplace\ transform \qquad (1)$$

The $\overline{F}(s) = \int_0^\infty e^{-st} f(x)dx$, ($f(x)$*is the density function*).

general statement of the theorem is as follows [6].

Theorem (Key Functional Equation). If $\rho_i \geq 1$, for all i there is no solution to the key functional equation in the interval (0, 1). If $\rho_i < 1$, there is a unique solution $\alpha \in (0, 1)$ to the key functional equation. ///

It is not always possible to solve the key function equation of the G/M/s queue analytically. In such cases, the solution can be obtained numerically by using the following recursive computation:

$$u_0 = 0,\ u_{n+1} = \overline{F}(\mu(1-u_n)),\ n \geq 0. \qquad (2)$$

Then, if ρ < 1,

$$\lim_{n \to \infty} (u_n)_i = \alpha_i$$

where, α_i is the unique solution in (0,1) to the key functional equation (2).

1) Exponential distribution

Suppose the inter-arrival times are iid random variables exponentially distributed with parameter λ > 0 (η = 1/ λ). Then, the key functional equation becomes

$$u = \frac{\lambda}{\lambda + \mu(1-u)}. \qquad (3)$$

This equations can be rearranged to get

$$(1-u)(u\mu - \lambda) = 0. \qquad (4)$$

The solution is given by

$$u = 1,\ u = \rho = \frac{\lambda}{\mu s},\quad where\ s\ denote\ the\ number\ of\ servers$$

2) Geometrical distribution

Suppose the inter-arrival time are iid geometric variables with parameter p, then,

From (1) we get

$$u_n = 0, \qquad u_{n+1} = \tilde{F}(\mu(1-u_n)$$

$$= \sum_{i=1}^{\infty} e^{-\mu(1-u_n)}(1-p)^{i-1} p$$

$$= e^{-\mu(1-u_n)} \sum_{i=1}^{\infty} p(1-p)^{i-1}$$

$$= e^{-\mu(1-u_n)}.$$

4. Steady State Analysis

Consider the CCN with 4 stations namely ‘classification’, ‘SaaS’, ‘PaaS’ and ‘IaaS’. The limiting behavior of the system in steady state can be studied as follows.

Let X_i (t) be the number of requests (BoTs) in the i^{th} station i = 1, 2, 3, 4 at time t. The n as X(t) = (X_1(t), X_2(t), X_3(t), X_4(t)) is a stochastic process with state space E = {(i, j, k, l) i, j, k, l = 0, 1, 2,….}.

Hence the CCN becomes an Open Jackson Queueing network with G/M/s_i, at each stations.

Embedded Distributions. Let X_n denote the number of customers in a G/M/s queue as seen by an arrival. Since the $\{X(t), t \geq 0\}$ process jumps by ±1.The arrival time distribution is the same as the departure time distribution,

$$\pi_j = \pi_j^*, j \geq 0,$$

using the solution to the key functional equation (2).

Theorem: (Arrival Time Distribution). In a stable G/M/s queue, the limiting distribution of the number of customers as seen by an arrival is given by

$$\pi_j^* = (1-\alpha_i)\alpha_i^j, \quad j \geq 0, \ i = 1, 2, 3, 4.$$

where α_j is the unique solution in (0, 1) to the key functional equation [6].

Proof: Define π_j^* as the number of customers in the G/M/s queue as seen by the n^{th} arrival (not including the arrival itself). Then we first show that $\{X_n^*, n \geq 0\}$ is an irreducible and aperiodic DTMC on state space {0, 1, 2 …} and compute the transition probabilities

$$p_{ij} = P(X_{n+1}^* = j \mid X_n^* = i), i, j \geq 0,$$

in terms of μ and F(.). Then we show that $\pi_j^* = (1-\alpha_i)\alpha_i^j$ satisfies the balance equations

$$\pi_j^* = \sum_{i=0}^{4} \pi_i^* P_{ij}, j \geq 0. \qquad i = 1, 2, 3, 4$$

This, along with aperiodicity of the DTMC, shows that

$$\pi_j^* = \lim_{n \to \infty} P(X_n^* = j), \ j \geq 0.$$

This proves the theorem.□

Limiting Distribution: Next we study the limiting distribution $p_j = \lim_{t \to \infty} \left[P(X(t) = j) \right], \ j \geq 0.$

Theorem: (Limiting Distribution). In a stable G/M/s queue, the limiting distribution of the number of customers in the system is given by

$$p_0 = 1 - \rho, p_j = \rho \pi_{j-1}^*, \quad j \geq 1,$$

Proof. In steady state, the expected number of busy servers is ρ. Since the system is served by a single server, in the long run we must have

P(server is busy) = ρ,

and hence

P(server is idle) = $p_0 = 1 - \rho$.

Now, fix j ≥ 1. Consider the following cost structure. Whenever the number of customers in the system jumps from j – 1 to j, the system earns $1, and whenever it jumps from j to j - 1, it loses $1. Then R_u, the long-run rate at which it earns dollars, is given by the product of λ, the long-run arrival rate of customers, and π_{j-1}^*, the long-run fraction of customers that see j -1 customers ahead of them in the system. Thus

$$R_u = \lambda \pi_{j-1}^*.$$

On the other hand, R_d, the long-run rate at which it loses dollars, is given by the product of μ, the long-run departure rate of customers (when the system is in state j), and p_j, the long-run fraction of the time that the system is in state j. Thus

$$R_d = \mu p_j.$$

Since the system is stable, we expect these two rates to be the same. Hence we have

$$\lambda \pi_{j-1}^* = \mu \ p_j$$

This proves the theorem [6]. □

Examples: Consider a CCN, with the following parameters $\lambda_1 = \lambda$, $\lambda_i = 0$, i = 2, 3, 4. Mean service time at station i is Exp(μ_i) . The routing probability matrix is

$$\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and $r_i = 0$, $r_2 = r_3 = r_4 = 1$. From this system performance measures and total expected cost can be completed.

5. System Performance Analysis

As we obtained the steady state probabilities for the number of customers in the each of the stations. We are able to find the mean number of BoTs waiting. The following system performance measures are crucial for our model. [7]

The expected number of customers in the system is given by

$$L = \sum_{i=1}^{4} \left[\sum_{j=0}^{\infty} j p_j \right]$$

$$= \sum_{i=1}^{4} \left[\sum_{j=1}^{\infty} j \rho_i (1-\alpha_i) \alpha_i^{j-1} \right]$$

$$= \sum_{i=1}^{4} \left[\frac{\rho_i}{1-\alpha_i} \right].$$

1. Expected waiting time in the system

$$W = \frac{L}{\lambda} = \sum_{i=1}^{4} \frac{1}{\mu_i s_i} \left(\frac{1}{1-\alpha_i} \right).$$

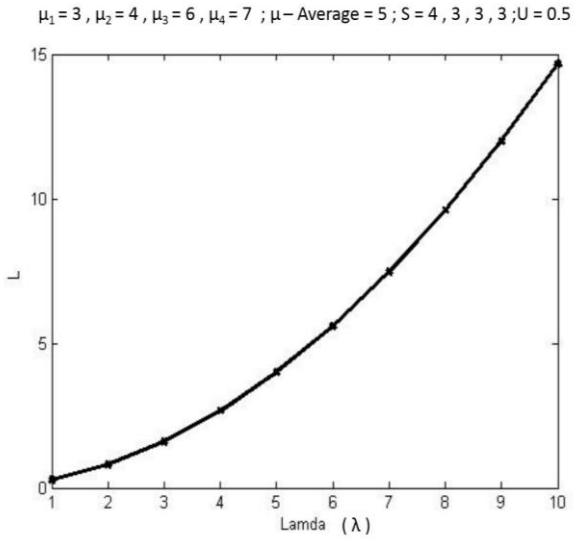
6. Results and Analysis

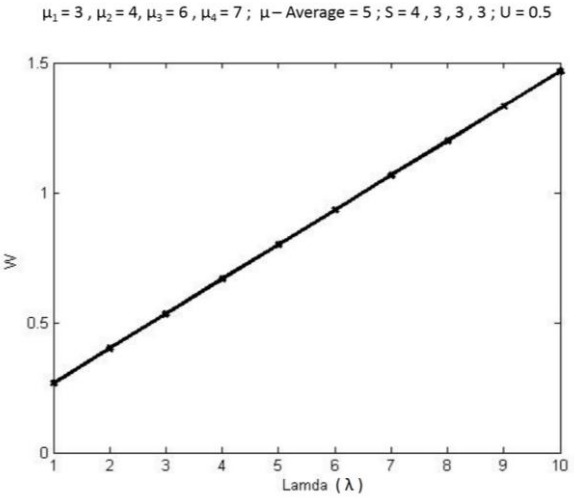
We observe that when the arrival rate λ increases, the length of queue (queue size) also increases, waiting time of a customer increases linearly with λ. (see the table 1)

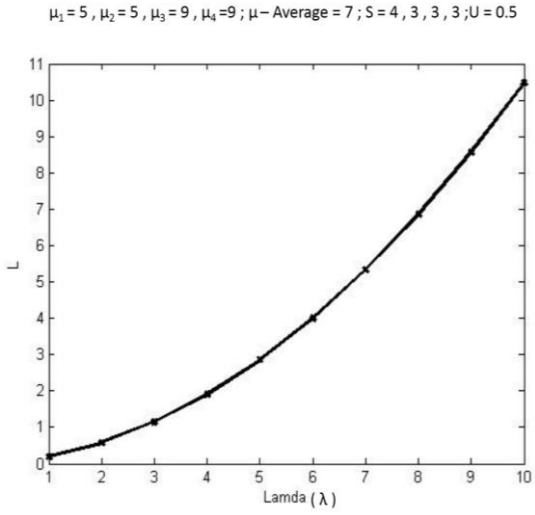
If the service rate μ is very high compared to λ, (ie. λ/μs < < 1). Then the queue length and waiting time became constant.

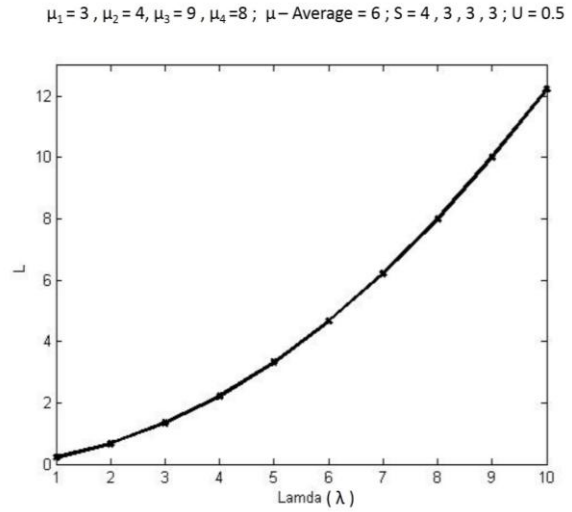
λ \ μ		5	6	7	8	9	10
1	L	0.2667	0.2222	0.1905	0.1667	0.1481	0.1333
	W	0.2667	0.2222	0.1905	0.1667	0.1481	0.1333
2	L	0.8000	0.6667	0.5714	0.5000	0.4444	0.4000
	W	0.4000	0.3333	0.2857	0.2500	0.2222	0.2000
3	L	1.6000	1.3333	1.1429	1.0000	0.8889	0.8000
	W	0.5333	0.4444	0.3810	0.3333	0.2963	0.2667
4	L	2.6667	2.2222	1.9048	1.6667	1.4815	1.3333
	W	0.6667	0.5556	0.4762	0.4167	0.3704	0.3333
5	L	4.0000	3.3333	2.8571	2.5000	2.2222	2.0000
	W	0.8000	0.6667	0.5714	0.5000	0.4444	0.4000
6	L	5.6000	4.6667	4.0000	3.5000	3.1111	2.8000
	W	0.9333	0.7778	0.6667	0.5833	0.5185	0.4667
7	L	7.4667	6.2222	5.3333	4.6667	4.1481	3.7333
	W	1.0667	0.8889	0.7619	0.6667	0.5926	0.5333
8	L	9.6000	8.0000	6.8571	6.0000	5.3333	4.8000
	W	1.2000	1.0000	0.8571	0.7500	0.6667	0.6000
9	L	12.0000	10.0000	8.5714	7.5000	6.6667	6.0000
	W	1.3333	1.1111	0.9524	0.8333	0.7407	0.6667
10	L	14.6667	12.2222	10.4762	9.1667	8.1481	7.3333
	W	1.4667	1.2222	1.0476	0.9167	0.8148	0.7333

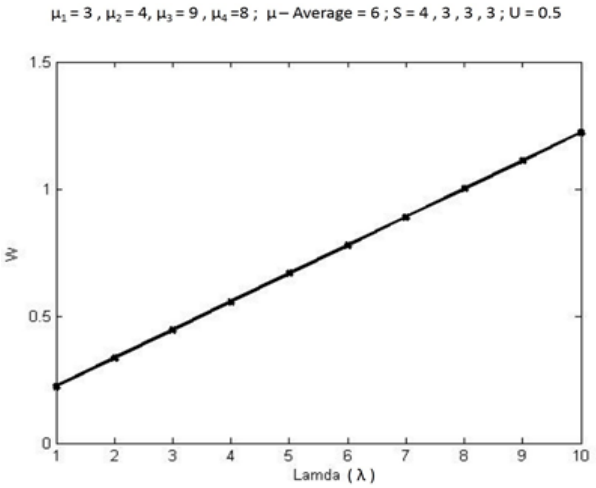
Table : 1

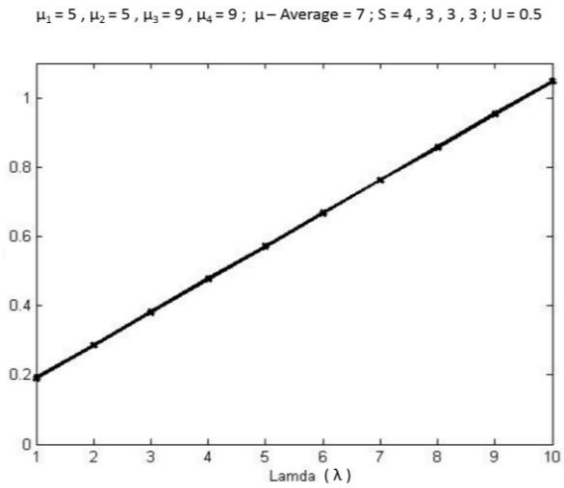












7. Conclusion and Future Development

In this paper, we proposed an approximate model to evaluate the performance of a cloud computing center using the G/M/s queue model method. Due to the nature of the environment of cloud computing and the diverse needs and demands of users, we considered a G/M/s queueing system that reflects the general nature of BoT's arrivals in the cloud. This system has general inter-arrival time, more number of servers and a infinite buffer capacity. In future we plan to extend the work to describe more general G/G/s Queueing system. Numerical examples are provided to illustrate the model.

Reference

[1] Ani Brown Mary,N and Saravanan, K,"Performance factors of Cloud computing data centers using, [(M/G/1): (∞/GDMODEL)] Queueing system", International Journal of Grid Computing & Applications (IJGCA) Vol.4, No.1, March 2013.

[2] Bharathi, M., Sandeep Kumar, P,and 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, www.iosrjen.org* Volume 2, Issue 9 (September 2012), PP 06-10 DOI: 10.5121

[3] 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

[4] 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, MAY 2012

[5] Hamzeh Khazaei,Jelena Mi si C,Vojislav B. Mi si C, " Modelling of Cloud Computing Centers Using M/G/m Queues", 2011 31st International Conference on Distributed Computing Systems Workshops.

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

[7] Lizheng Guo,Tao Yan, Shuguang Zhao, and Changyuan Jiang, "Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System",

[8] Murugesan R, Elango C, and Kannan S," Cloud Computing Networks with Poisson Arrival Process-Dynamic Resource Allocation", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 124-129.

[9] Murugesan, R, .Elango C, and Kannan S," Resource Allocation in Cloud Computing with M/G/s – queueing Model", Volume 4, Issue 9, September 2014 , ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering, PP 443-447.