Volume 4, Issue 9, September 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Resource Allocation in Cloud Computing with M/G/s-Queueing System

R. Murugesan
Dept. of Computer Science,
C.P.A College,
Bodinayakanur, TN, India

C. Elango
Dept. of Mathematical Sciences,
C. P.A. College,
Bodinayakanur, TN, India

S. Kannan
Dept. of Computer Applications
Madurai Kamaraj University
Madurai, TN, India

Abstract-Cloud computing is a new trend for computing resource allocation. Successful development of cloud computing paradigm necessitates accurate performance evaluation of cloud data centers. The computing resource allocation and performance managing have been one of the most important aspects of cloud computing. In this model, the resource allocation are modeled as queues and the virtual machines are modeled as service centers. We considered, M/G/s queue as a tool regulate task's arrivals, and general service time for requests with single server and infinite waiting space. We used this model in order to evaluate the performance analysis of cloud server farms and we solved it to obtain accurate estimation of the complete probability distribution of the request response time and other important performance indicators.

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

I. 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 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

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'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]. 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 M/G/s queuing system with single task arrivals 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.

II. 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. We focused on the how the computer resource can efficiently allocate different tasks in the cloud centers.

III. 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 exponentially distributed with parameter $\lambda_1 > 0$ and the task service times are iid random variable with mean $\tau_1 > 0$. 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', according to a Poisson process with rate $\lambda_1 > 0$.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 stations i has s_i independent servers, and the queueing model at station i is M/G/s_i type. The cloud computing network diagram is described in figure 1

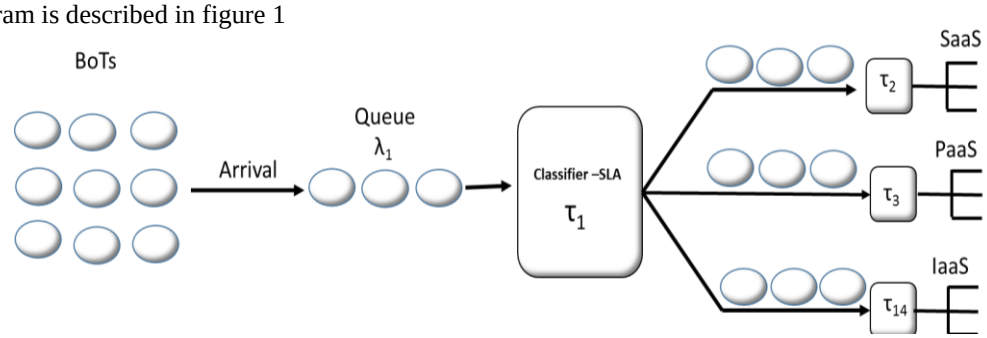


Fig. 1

A. 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 i from outside the network according to a Poisson process with parameters λ_i (i = 1, 2, …, N) and $\lambda_i > 0$.
- All arrival process is independent of each other.
- Service times for customers (service requests) of station i are independent and identically distributed (iid) random variables with mean τ_i.
- 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 Є S ={1, 2, … , N} is called the routing probabilities and the matrix P = (p_{ij}) i,j Є 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 server 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 capacity 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 a single station that 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 nature of the exponential distribution, a number of quite simple relationships can be derived for several performance measures based on the arrival rate and service rate are obtained. M/G/s system has a s servers 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 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]

Consider a single-station queueing system where customers arrive according to a PP (λ) and require iid service times with mean τ, variance σ^2 and second moment $s^2 = \sigma^2 + \tau^2$. The service times may not be exponentially distributed. The queue is serviced by a s_i servers at each stations and has infinite waiting room. Such a system is called a network of M/G/s queues.

Let X(t) be the number of customers in the system at time t. $\{X(t).t \geq 0\}$ is a continuous-time stochastic process with state space {0, 1, 2, . . .}. Knowing the current state X(t) does not provide enough information about the remaining service time of the customer in service (unless the service times are exponentially distributed), and hence we cannot predict the future based solely on X(t). Hence $\{X(t).t \geq 0\}$ is not a CTMC. Hence we will not be able to study an M/G/s queue in as much detail as the M/M/1 queue. Instead we shall satisfy ourselves with results about the expected number and expected waiting time of customers in the M/G/s queue in steady state. Since the arrival process to our Open Jackson networks Poisson, the whole networks (CCN) can be viewed as 4 independent M/G/s queue. [6]

B. Stability

For the CCN, we considered,

Let $\rho_i = \lambda_i \tau_i$, be the traffic intensity.

Theorem 1.(Condition of Stability). Consider a single-station queue with s servers and infinite capacity. Suppose the customers enter at rate λ and the mean service time is τ. The queue is stable if

$$\lambda_i \tau_i \leq s_i .$$

Proof:

We have, B = expected number of busy servers = $\lambda \tau_i$,

Where, λ is the arrival rate of entering customers. However, the number of busy servers cannot exceeds, the total number of servers. Hence, for the argument to be valid, we must have

$$\lambda_i \tau_i \leq s_i .$$

It follows that the M/G/s queue is stable if

$\rho_i < 1$, for each i, where $\rho_i = \dfrac{\lambda \tau_i}{s_i}$, where s_i is the number of servers at station i.

Indeed, it is possible to show that the M/G/s queue is unstable if $\rho_i \geq 1$: Thus $\rho_i < 1$ for every i is a necessary and sufficient condition of stability. We shall assume that the queue is stable.□

IV. 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 state of the system at time t be denoted as $X(t) = (X_1(t), X_2(t), X_3(t), X_4(t))$. Hence the CCN becomes an Open Jackson Queueing

network with M/G/s_i, at each stations, where $\lambda_1 = \lambda$, $\lambda_i = 0$, i=2, 3, 4. Mean service time at station i is τ_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 $r_i = 0$, $r_2 = r_3 = r_4 = 1$

As each station i (i=1, 2, 3, 4) has a M/G/s_i queue, the expected number of customer in station i is given by

$$L_i = \rho_i + \frac{\lambda^2 S_i^2}{2(1-\rho_i)} ,$$

where, $S_i^2 = \sigma_i^2 + \tau_i^2$, and the expected waiting time in the queue is given by

$$W_i = \tau_i + \frac{\lambda S_i^2}{2(1-\rho_i)}$$. Where S_i^2 = is the sample variance for service time at station i.

V. 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 traffic intensity is given by

$$\rho_i = \frac{\lambda \tau_i}{s_i}$$

1. Expected waiting time in the system

$$W = \sum_{i=1}^{4} \left[\tau_i + \frac{\lambda S_i^2}{2(1-\rho_i)} \right]$$ Where, $S_i^2 = \sigma_i^2 + \tau_i^2$,

2. The expected number of tasks waiting for transmission is given by

$$L = \sum_{i=1}^{4} \left[\rho_i + \frac{\lambda^2 S_i^2}{2(1-\rho_i)} \right].$$

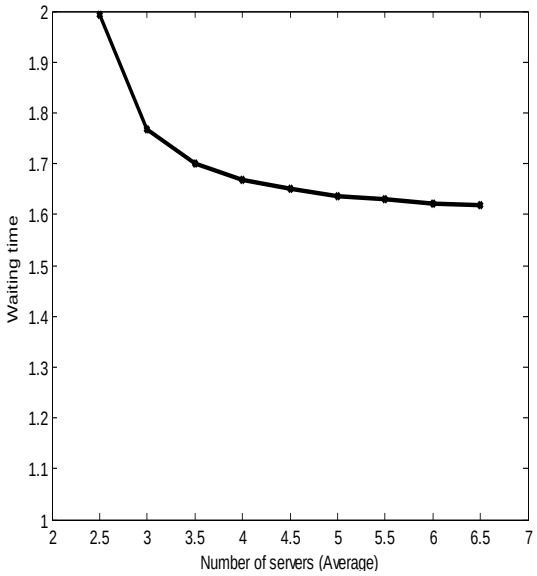
VI. NUMERICAL EXAMPLES

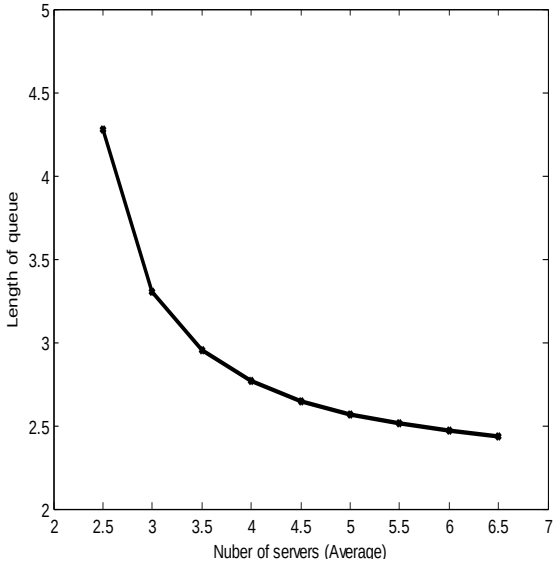
In this section, we evaluate the performance of a of cloud computing center as a network of M/G/s queue. The following numerical example illustrates the model. Consider the CCN with 4 stations,

$\tau = (0.1, 0.2, 0.3, 0.4)$, $\lambda = 3$, and $\sigma = (0.1, 0.1, 0.1, 0.1)$.

Table 1 for Average no. of server, waiting time and length of queue.

Number of servers s_1 s_2 s_3 s_4	Server Average	Waiting time (W)	Length of queue (L)
3 2 3 2	2.5	1.9923	4.2768
3 3 3 3	3.0	1.7664	3.2991
3 4 3 4	3.5	1.7001	2.9504
3 5 3 5	4.0	1.6684	2.7654
3 6 3 6	4.5	1.6497	2.6491
3 7 3 7	5.0	1.6374	2.5694
3 8 3 8	5.5	1.6287	2.5111
3 9 3 9	6.0	1.6222	2.4666
3 10 3 10	6.5	1.6172	2.4315





From table 1, we observe that the length of queue in the system decreases with the average number of servers and the waiting time of a customer in the system also decreases with the average number of servers. In general '*efficiency of a server*' is measured by the mean service time τ_i of a customer and its variance σ_i^2.

VII. CONCLUSION AND FUTURE DEVELOPMENT

In this paper, we proposed an approximate model to evaluate the performance of a center of cloud computing using the M/G/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 M/G/s queueing system that reflects the nature of BoTs arrivals in the cloud. This system has general service time, number of servers and a infinite buffer capacity. We described a new analytical approximation for performance evaluation of a center of cloud computing and resolved it to get a very decent estimate. This result may be extended to a general case when arrival process to the CCN is of arbitrary nature, but service time is exponential, that is G/M/s system.

REFERENCE

[1] Ani Brown Mary.N and K.Saravanan,"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., 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, Changyuan Jiang, "Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System",

[8] Murugesan R, C.Elango, S.Kannan," Cloud Computing Networks with Poisson Arrival Process-Dynamic Resource Allocation". (In print)

