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
www.iosrjournals.org

Cloud Computing Networks with Poisson Arrival Process-Dynamic Resource Allocation

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

[1]*Department of Computer Science, Cardamom Planters' Association College, Bodinayakanur, Tamilnadu*
[2]*Department of Mathematical Sciences, Cardamom Planters' Association College, Bodinayakanur, Tamilnadu*
[3]*Department of Computer Applications, Madurai Kamaraj University, Madurai, Tamilnadu*

Abstract: *In this article we considered a cloud computing network model which acts as a tool for dynamic resource allocation problem in cloud computing. Every successful servicing of cloud paradigm need an optimal resource allocation in turn to handle the problem of Bag of tasks (BoT). We assume that the allocation of BoT's are done by two stages of process namely classification according to service level agreement (SLA) of BoT and service providing (SaaS, PaaS and IaaS). After classification the request is routed to any one of the service provider with corresponding probability, Thus the Cloud Computing Networks (CCN) become a general open Jackson Queuing Network system performance measures are obtained to study the efficiency of the CCN.*

Keywords: *Cloud Computing, Resource Allocation, Jackson Queuing Network*

I. Introduction

Cloud Computing enables the massive scale resource sharing, which allows users to access technology enabled service without knowledge of expertise of the system. It also refers to the provision of computational resources on demand via a computer network. In the modern competitive business environment, providing Quality of Service (QoS) is prime requisition for any service provider. With high exposition of technological innovation and developments the Cloud Computing is changing evolutionary in the modern era. The dynamic allocation of resources has emerged as promising technology to provide cost effectiveness in high performance cloud computing system for solving many complex problems in commercial application.[7]

Resource allocation in a cloud computing environment can be modeled as allocating the required amount of multiple types of resource simultaneously from a common resource pool for a certain period of time for each request [8].There are three kinds of cloud services model, namely, Software as a Service (SaaS), Platform as a Service (PaaS) and Cloud Infrastructure as a Service (laaS) [4].

Software-as-a-Service (SaaS) is a software distribution model in which applications are accessible through a single interface, like a web browser over the Internet. Users do not have to consider the underlying cloud infrastructure including servers, storage, platforms, etc.

Platform-as-a-Service (PaaS) provides a high level of integrated applications that control of distributed applications and their hosting environment configurations. In general, developers accept all instructions on the type of software that can be written to change built-in scalability.

Infrastructure-as-a-Service (IaaS) provides users with computation processing, storage, networks and computing resources. IaaS users can implement an arbitrary application which is able to grow up and down dynamically. Also, IaaS sends programs and related data, while the cloud provider does the computation processing and returns the result [4].

Resource allocation in cloud computing is still a challenging issue. Due to existence of different workload types with various requirements that should be supported by cloud computing, no any single hardware or software solution can allocate resources to all imaginable types efficiently. Also, each type has its specific nature properties and a single solution cannot deal with in that regard optimally. Thus, it is a need to provide specific solutions for different workload types such as Bag of Tasks (BoT) and message passing applications, and provision resources in such a way that customers be able to just concentrate on demanded requests' results [5].The main advantage of having multiple servers in data center is, the increment in performance by reducing the mean queue length and waiting time (response time) than compared to the traditional approach of having only single server. [1]

In this paper we model the dynamic resource allocation of bag of task to cloud computing network cloud center at Poisson rate. We evaluate its performance using a novel analytical model and solve it to obtain important performance factors like mean number of tasks in the system and mean number of busy servers.

II. Model Description

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 distributed with parameter $\lambda_1 > 0$ and the task service times are also exponentially distributed with parameter $\mu_1 > 0$. 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 cloud computing network diagram is described in fig.1. The bag of task are arriving the first stations namely 'Classifier', according to a Poisson process with rate $\lambda_1 > 0$.The BoTs are taken for classification in FCFS discipline. After classification the BoT moves to any one of the stations which provides SaaS, PaaS and IaaS. Each stations i has s_i independent servers.

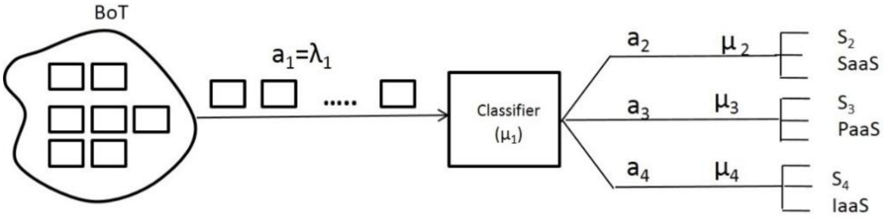


Fig. : 1

2.1 Analysis

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

- The network has N single stations with s_i servers at each station.
- There is an unlimited waiting space at each station (the classifications and service stations).
- The customers (BoT 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) exponential random variables with parameters μ_i (i = 1,2, …, N)
- Customers (service requests) finishing service at station i join the queue at station j with probability p_{ij} or leave the network altogether with probability r_i independently of each other.

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 a Open Jackson Queuing Network with N stations and s_i server at each stations [3].

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

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

,with the condition $\sum_{j=1}^{N} p_{ij} + r_i = 1, \quad 1 \le i \le N.$

We assumed in the CCN that each station has infinite capacity for waiting requests or jobs. This will lead to a new problem of stability analysis. Next we have to show that the CCN is stable in the long run. The next theorem proves the stability criteria for the proposed networks.

2.2 Theorem

The CCN, with external arrival rate vector $\bar{\lambda}$ and routing matrix P, is stable if the matrix I - P is invertible and $a_i < s_i\mu_i$, for all i=1,2,...,N, where

$$\overline{a} = [a_1, a_2, a_3, ..., a_N] \ \ with \ \ a_i = \lambda_i + \sum_{j=1}^{N} a_i p_{ij} \ \ i = 1,2,3,....N$$

Proof:

Since, I – P is invertible, the traffic equation has a unique solution $\bar{a} = \bar{\lambda}(I - P)^{-1}$. Here the ith service station has single queue having s_i servers with arrival rate a_i and mean service times $1/s_i\mu_i$ (i= 1, 2, …, N).

By the stability theorem of M/M/s queues it follow that the CCN is stable if $a_i < s_i\mu_i$, i = 1, 2,…, N. Hence the theorem □

This proves that our CCN model for Cloud Computing becomes Open Jackson Network and is also stable in the long run. This will induce us to compute the following performance measures:

1. Mean number of request waits in the network.
2. Probability that the network is busy.

III. 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 Xi (t) be the number of requests (BoTs) in the ith station i = 1, 2, 3, 4 at time t. The state of the system at time t be denoted as X(t) = (X1(t), X2(t), X3(t), X4(t)). Suppose that the CCN is stable, with the unique solution $\bar{a}$ to the traffic equation,

$$a_j = \lambda_j + \sum_{i=1}^{4} a_i p_{ij}$$

We can define the joint probability distribution in the long run as follows:

$$Let \quad p(n_1, n_2, n_3, n_4) = \lim_{t\to\infty} p_r\{\, X_1(t) = n_1, X_2(t) = n_2, X_3(t) = n_3, X_4(t) = n_4 \,\} = \prod_{i=1}^{4} p_i(n_i)$$

Since the stations are independent and p_i (n_i) denote the marginal probability that X_i (t) = n_i, that is there are n_i request wait in the queue M/M/s_i, where ith station's arrival rate a_i and service rate μ_i.

From limiting behavior of the queue M/M/s_i,we have

$$p_i(n) = p_i(0)\rho_i(n) \quad i = 1,2,3,4.$$

where

$$p_i(0) = \left[\sum_{n=0}^{s_i-1} \frac{1}{n!}\left(\frac{a_i}{\mu_i}\right)^n + \frac{\left(\frac{a_i}{\mu_i}\right)^{s_i}}{s_i!}\left(\frac{1}{1-\frac{a_i}{s_i\mu_i}}\right) \right]^{-1}$$

and

$$\rho_i(n) = \begin{cases} \frac{1}{n!}\left(\frac{a_i}{\mu_i}\right)^n & if \; 0 \le n \le s_i - 1 \\ \frac{S_i^{s_i}}{S_i!}\left(\frac{a_i}{S_{i_i}\mu_i}\right)^n & if \; n \ge s_i \end{cases}$$

Next we obtain the product form solution for the distribution of request in the CCN [2].

3.1 Theorem

The limiting behavior (steady state solution) of the simple CCN is given by

p (n_1, n_2, n_3, n_4) = $p_1(n_1)p_2(n_2)p_3(n_3)p_4(n_4)$, for n_i= 0, 1, 2, …. and i=1, 2, 3, 4.

Proof:

Since {X (t); t ≥ 0} is a four dimensional Continuous Time Markov Chain (CTMC) with state space E = $N^0{}_X N^0{}_X N^0{}_X N^0$, where N^0= {0,1, 2, 3,…}and the limiting distribution obtained satisfy the balance equation, the product form solution of distribution of request in CCN is obtained easily □

IV. System Performance Analysis

As we obtained the steady state probabilities for the number of customers in the each of the stations as a product form solutions, we are able to find the mean number of BoTs waiting. The following system performance measures are crucial for our model.

1. Mean number of requests waits in ith station (L_i) $= \sum_{n=0}^{\infty} n\ p_i(n);\ i = 1,2,3,4.$

$$= \frac{\rho_i}{1-\rho_i},\ i = 1,2,3,4; \quad where\ \rho_i = \frac{a_i}{s_i \mu_i}, and\ p_i(0) = 1-\rho_i$$

2. The probability that the CCN system is busy $= \prod_{i=1}^{4}(1-\rho_i),\ i = 1,2,3,4.$

V. Cost Analysis

Let C_1 is the cost of waiting per unit time and C_2 be the service cost incurred per unit time. The total expected cost rate in the long run is given by

$$TC = \sum_{i=1}^{4}\left(\frac{\rho_i}{1-\rho_i}\right) C_1 + \prod_{i=1}^{4}(1-\rho_i)\ C_2$$

VI. Numerical Examples

6.1 Example 1:

Consider a simple CCN with Poisson arrival rate λ_1 at station 1and services rate µi at station i , (i=1, 2, 3, 4).

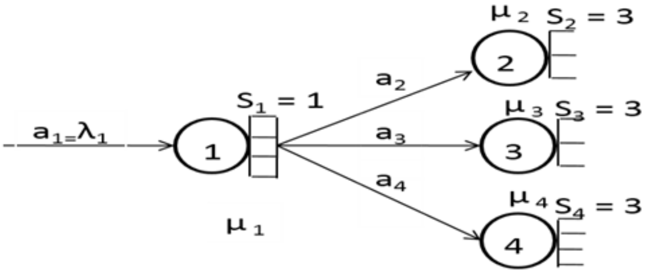


Fig.: 2

This CCN is an open Jackson queuing with the following parameters

N = 4, s_1 =1, s_2 = 3, s_3 = 3, s_4= 3 and routing probability matrix.

$$P = \begin{bmatrix} 0 & .2 & .3 & .5 \\ .2 & 0 & .4 & .4 \\ .1 & .2 & 0 & .7 \\ .1 & .4 & .5 & 0 \end{bmatrix}$$

For this CCN, $r_1 = 0$, $r_2= 1/3$, $r_3= 1/3$, $r_4= 1/3$. The arrival rate of each station i is given by a_i and a = (a1, a_2, a_3, a_4) and a = λ $(I - P)^{-1}$. In steady state $a_i < s_i\mu_i$ for i=1, 2, 3, 4

6.2 Cost analysis:

In this section we give a cost analysis for the CCN we considered in this paper. We emphasis on the convexity of the total cost functions TC, which varies with the average throughput of the network $\rho = \frac{1}{4}\left[\sum_{i=1}^{4}(\frac{a_i}{s_i\,\mu_i})\right]$ by imposing the cost structure, C_1: cost of waiting per customer C_2: cost of service.

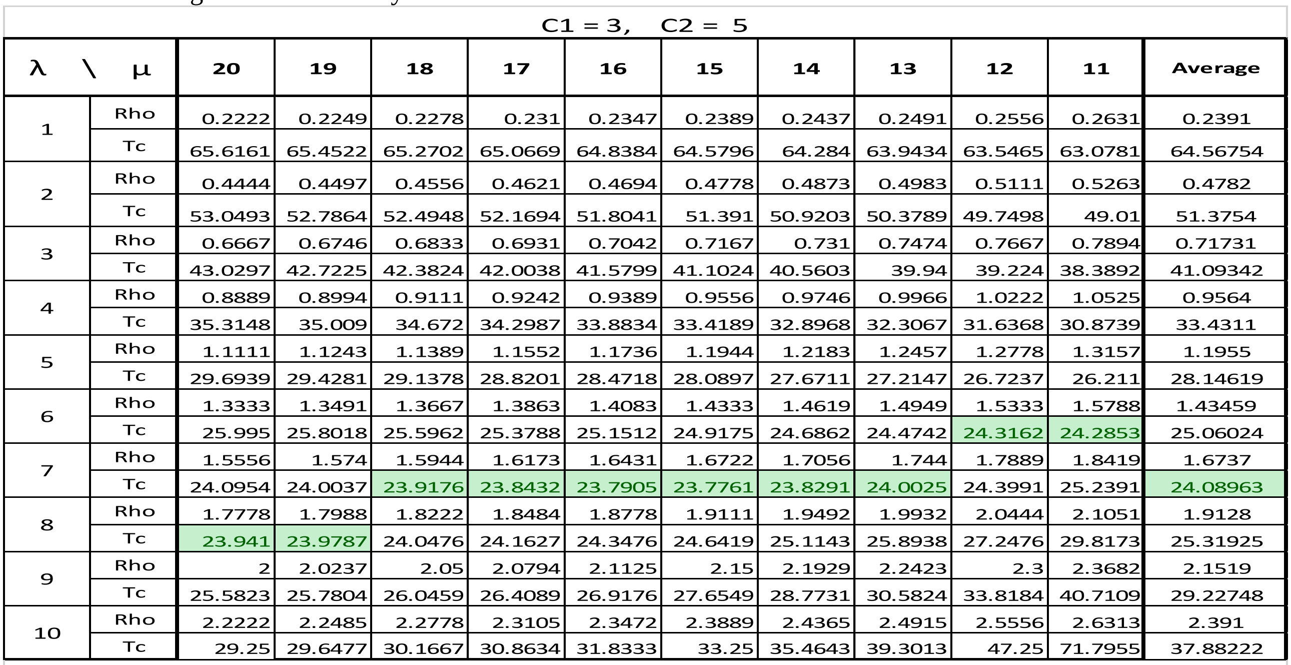
As we are unable to prove the convexity of the cost function TC analytically, we made numerical search by varying the parameters of the CCN system. We are able to get local optimum with convexity at specific intervals of permanent μ.

From table 1, we observe that whenever μ_1 the service rate at cloud lies in the interval (19, 20) and $\rho \in (1.59, 1.78)$ the optimal arrival rate $\lambda_1^* = 8$, with minimal total cost lies in $(23.94, 23.97)$.

Similarly whenever the service rate μ_1 lies in the interval (13, 18) and $\rho \in (1.59, 1.78)$ .The optimal arrival rate is $\lambda_1^* = 7$, and the whenever the service rate μ_1 , lies in the interval (11, 12) and $\rho \in (1.53, 1.57)$, the optimal arrival rate is $\lambda_1^* = 6$.

Thus the overall optimality of the system is obtained at the level $\lambda_1^* = 7$, with $\mu_1 \in (13, 18)$ and point $\rho \in (1.6, 1.8)$ (fig.: 3)

In this numerical example we assume that there is no feedback in the system and hence it becomes Tandem queue with Poisson arrival. The average throughput of the system $\rho = \frac{1}{4}\sum_{i=1}^{4}\rho_i$, is the indicator for the variation in average cost TC of the system.



C1 = 3, C2 = 5

λ \ μ		20	19	18	17	16	15	14	13	12	11	Average
1	Rho	0.2222	0.2249	0.2278	0.231	0.2347	0.2389	0.2437	0.2491	0.2556	0.2631	0.2391
	Tc	65.6161	65.4522	65.2702	65.0669	64.8384	64.5796	64.284	63.9434	63.5465	63.0781	64.56754
2	Rho	0.4444	0.4497	0.4556	0.4621	0.4694	0.4778	0.4873	0.4983	0.5111	0.5263	0.4782
	Tc	53.0493	52.7864	52.4948	52.1694	51.8041	51.391	50.9203	50.3789	49.7498	49.01	51.3754
3	Rho	0.6667	0.6746	0.6833	0.6931	0.7042	0.7167	0.731	0.7474	0.7667	0.7894	0.71731
	Tc	43.0297	42.7225	42.3824	42.0038	41.5799	41.1024	40.5603	39.94	39.224	38.3892	41.09342
4	Rho	0.8889	0.8994	0.9111	0.9242	0.9389	0.9556	0.9746	0.9966	1.0222	1.0525	0.9564
	Tc	35.3148	35.009	34.672	34.2987	33.8834	33.4189	32.8968	32.3067	31.6368	30.8739	33.4311
5	Rho	1.1111	1.1243	1.1389	1.1552	1.1736	1.1944	1.2183	1.2457	1.2778	1.3157	1.1955
	Tc	29.6939	29.4281	29.1378	28.8201	28.4718	28.0897	27.6711	27.2147	26.7237	26.211	28.14619
6	Rho	1.3333	1.3491	1.3667	1.3863	1.4083	1.4333	1.4619	1.4949	1.5333	1.5788	1.43459
	Tc	25.995	25.8018	25.5962	25.3788	25.1512	24.9175	24.6862	24.4742	24.3162	24.2853	25.06024
7	Rho	1.5556	1.574	1.5944	1.6173	1.6431	1.6722	1.7056	1.744	1.7889	1.8419	1.6737
	Tc	24.0954	24.0037	23.9176	23.8432	23.7905	23.7761	23.8291	24.0025	24.3991	25.2391	24.08963
8	Rho	1.7778	1.7988	1.8222	1.8484	1.8778	1.9111	1.9492	1.9932	2.0444	2.1051	1.9128
	Tc	23.941	23.9787	24.0476	24.1627	24.3476	24.6419	25.1143	25.8938	27.2476	29.8173	25.31925
9	Rho	2	2.0237	2.05	2.0794	2.1125	2.15	2.1929	2.2423	2.3	2.3682	2.1519
	Tc	25.5823	25.7804	26.0459	26.4089	26.9176	27.6549	28.7731	30.5824	33.8184	40.7109	29.22748
10	Rho	2.2222	2.2485	2.2778	2.3105	2.3472	2.3889	2.4365	2.4915	2.5556	2.6313	2.391
	Tc	29.25	29.6477	30.1667	30.8634	31.8333	33.25	35.4643	39.3013	47.25	71.7955	37.88222

Table :1

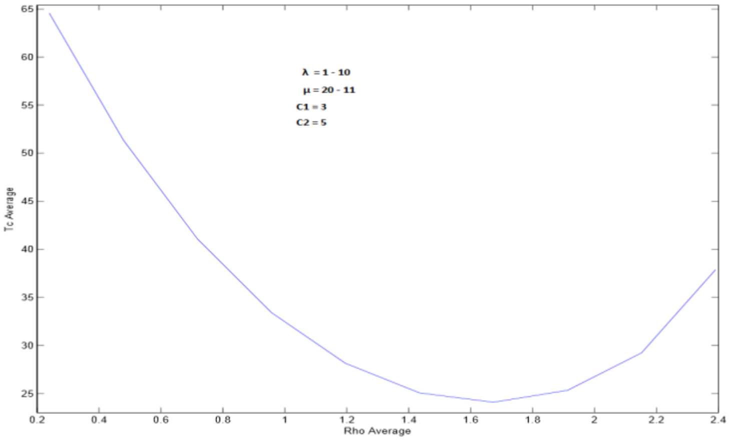


Fig.:3

VII. Conclusions And Future Directions

In this articles, we have studied the Cloud Computing Network (CCN) with Poisson arrival process and exponentially service times. Through the calculation, we are able to answer two questions 1.Mean number of requests waits and 2.Probability that the system is busy. This system performance measures and used to get the optimal resource allocation parameters. There are many opening for further investigations in the area of controlling cloud center (CCN) with Poisson arrival and general service times.

References

[1] L.M. Vaquero.L.M, L. Rodero-Merino, J. Caceres, and M. Lindner, “A Break in the Clouds: Towards a Cloud Definition,” ACM SIGCOMM Computer Comm. Rev., vol. 39, pp. 50-55, Dec. 2008.

[2] Shin-ichiKuribayashi, “Optimal Joint Multiple Resource Allocation Method for Cloud Computing Environments”, International Journal of Research and Reviews in Computer Science Vol. 2, March 2011.

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

[4] SaiSowjanya.T, D.Praveen. D, Satish.K, Rahiman. A,” The Queueing Theory in Cloud Computing to Reduce the Waiting Time”,IJCSET April 2011.

[5] MasoudSalehpour and AsadollahShahbahrami,”Alleviating Dynamic Resource Allocation for Bag of Tasks Applications in Cloud Computing”, International Journal of Grid and Distributed Computing Vol. 5, No. 3, September, 2012

[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, MAY 2012

[7] Satyanarayana .A , P. Suresh Varma, M.V.RamaSundari, P SaradaVarma“ Performance Analysis of Cloud Computing under Non Homogeneous Conditions”, Volume 3, Issue 5, May 2013,International Journal of Advanced Research in Computer Science and Software Engineering

[8] Mohamed Eisa , E. I. Esedimy M. Z. Rashad,” Enhancing Cloud Computing Scheduling based on Queuing Models”,International Journal of Computer Applications (0975 – 8887) Volume 85 – No 2, January 2014