IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-8 E-ISSN: 2347-2693

Optimization of Multi-server Configuration for Profit Maximization using M/M/m Queuing Model

M.G.Madhusudhan[1*] and K.Delhi Babu[2]

[1*] *Dept. of CSE, Sree Vidyanikethan Engg. College, India.*
[2] *Dept. of CSE, Sree Vidyanikethan Engg.College, India.*
mmadu512@gmail.com, Kdb_babu@yahoo.com

www.ijcaonline.org

Received: 16/07/ 2014 Revised: 28/07/ 2014 Accepted: 24 /08/ 2014 Published: 31 /08/ 2014

Abstract— Cloud computing is an emerging technology of business computing and it is becoming a development trend. The process of entering into the cloud is generally in the form of queue, so that each user needs to wait until the current user is being served. Cloud Computing User requests Cloud Computing Service Provider to use the resources, if Cloud Computing User finds that the server is busy then the user has to wait till the current user complete the job which leads to more queue length and increased of waiting time. So to solve this problem it is the work of Cloud Computing Service Providers to provide service to users with less waiting time otherwise there is a chance that the user might be leaving from queue. Cloud Computing Service Providers takes such factors into considerations as the amount of service, the workload of an application environment, the configuration of a multi-server system, the service-level agreement, the satisfaction of a consumer, the quality of a service, the quality of a service, the penalty of a low-quality service, the cost of renting and a service providers margin and profit. Cloud Computing Service Providers can use multiple servers for reducing queue length and waiting time. This project shows how the multiple servers can reduce the mean queue length and waiting time. The project approach is to treat a multi-server system as an M/M/m queuing model.

Keywords— Cloud Computing;Multi-server System; Queuing Model; Waiting time; Service-level agreement

I. INTRODUCTION

Cloud computing can be defined as the delivery of hosted services over the Internet, such that accesses to shared hardware, software, databases, information, and all resources are provided to consumers on-demand by centralized management of resources and services. A cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements[1] established through negotiation between the service provider and the consumers.

Cloud services include Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (Paas). The aim of cloud computing is to allocate virtual resources that enables computing and storage data access on demand basis. For allowing more requests, cloud services has the capacity of multiplexing the physical resources among requested resources. Cloud computing and networking are the two key functionalities that are involved in the distributed clouds. Convergence between cloud and networking is more important for QOS delivery and for creation of networked cloud environments.

Cloud computing is able to provide the most cost-effective and energy-efficient way of computing resources management. Cloud computing turn's information technology into ordinary commodities and utilities by using the pay-per-use pricing model [2], [3], [4]. However, cloud computing will never be free [5] and understanding the economics of cloud computing becomes critically important.

Like all business, the pricing model of a service provider in cloud computing is based on two components, namely, the income and the cost. For a service provider, the income (i.e., the revenue) is the service charge to users, and the cost is the renting cost plus the utility cost paid to infrastructure vendors. A pricing model in cloud computing includes many considerations, such as the amount of a service (the requirement of a service), the workload of an application environment, the configuration (the size and the speed) of a multi-server system, the service-level agreement, the satisfaction of a consumer (the expected service time), the quality of a service (the task waiting time and the task response time), the penalty of a low-quality service, the cost of renting, the cost of energy consumption, and a service provider's margin and profit.

The profit (i.e., the net business gain) is the income minus the cost. To maximize the profit, a service provider should

understand both service charges and business costs, and in particular, how they are determined by the characteristics of the applications and the configuration of a multi-server system.

II. RELATED WORK

Cloud service differs from traditional hosting in three principal aspects. First, it is provided on demand; second, it is elastic since users that use the service have as much or as little as they want at any given time (typically by the minute or the hour); and third, the service is fully managed by the provider. Due to dynamic nature of cloud environments, diversity of user requests, and time dependency of load, providing agreed quality of service (QoS) while avoiding over provisioning is a difficult task. Since many of the large cloud centers employ virtualization to provide the required resources such as PMs, we consider PMs with a high degree of virtualization. Real cloud providers offer complex requests for their users. For instance, in Amazon EC2, the user is allowed to run up to On-Demand or Reserved Instances, and up to 100 Spot Instances per region. We examined the effects of various parameters including ST arrival rate, task service time, the virtualization degree, and ST size on task rejection probability and total response delay. The stable, transient and unstable regimes of operation for given configurations have been identified so that capacity Planning is going to be a less challenging task for cloud providers.

The cluster RMS supports four main functionalities: resource management; job queuing; job scheduling; and job execution. It manages and maintains status information of the resources such as processors and disk storage in the cluster system. Jobs submitted into the cluster system are initially placed into queues until there are available resources to execute the jobs. The cluster RMS then invokes a scheduler to determine how resources are assigned to jobs. After that, the cluster RMS dispatches the jobs to the assigned nodes and manages the job execution processes before returning the results to the users upon job completion.

III. M/M/M QUEUING MODEL PROCESS

A cloud computing service provider serves users' service requests by using a multi-server system, which is constructed and maintained by an infrastructure vendor and rented by the service provider. Multi-server model is treated as an M/M/m queuing model [6], for this queuing system, it is assumed that the arrivals follow a Poisson probability distribution at an average of λ customers per unit of time. The queue discipline is First-Come, First Served (FCFS) or Shortest Processing First (SPF) basis by any of the servers.

Service times are distributed exponentially, with an average of μ customers per unit of time. There is no limit to the number of the queue (infinite). The service providers are working at their full capacity. The average arrival rate is greater than average service rate. Service rate is independent of line length; service providers do not go faster because the line is longer.

In the Queuing model or waiting lines can be mainly three parts

1. Arrivals or inputs to the system. These have characteristics such as population size, behavior, and a statistical distribution.
2. Queue discipline, or the waiting line itself. Characteristics of the queue include whether it is limited or unlimited in length and the discipline of people or items in it.
3. The service facility. Its characteristics include its design and the statistical distribution of service times.

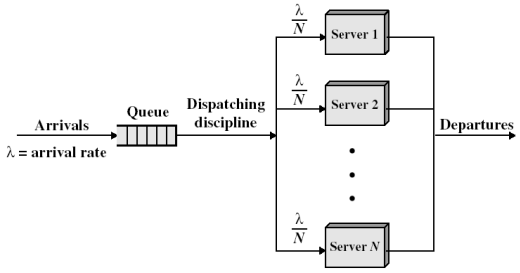

Fig1: Queuing Model Process

Let P_K denote the probability that there are K service requests (waiting or being processed) in the M/M/m queuing system for S. Then

$$P_K = \begin{cases} P_0 \dfrac{(m\rho)^K}{K!} & , K \le m; \\ P_0 \dfrac{m^m \rho^K}{m!} & , K \ge m; \end{cases}$$

Where

$$P_0 = \left(\sum_{K=0}^{m-1} \frac{(m\rho)^K}{K!} + \frac{(m\rho)^m}{m!} \cdot \frac{1}{1-\rho} \right)^{-1}$$

We now proceed to compute the performance measures of the queuing system.

The average waiting time in the system W

$$W = W_q + 1/\mu$$

The average waiting time in the queue W_q

$$W_q = L_q/\lambda$$

Utilization factor i.e., the fraction of time servers is busy ρ

$$\rho = \lambda / m\mu$$

The average number of customers in the system L

$$L = L_q + \frac{\lambda}{\mu}$$

The average number of customers in the queue L_q

$$L_q = \frac{P_0 (\lambda/\mu)^m \rho}{m!(1-\rho)^2}$$

The average number of service requests is $\overline{N}$

$$\overline{N} = \sum_{k=0}^{\infty} KP_k$$

The average task response time is $\overline{T}$

$$\overline{T} = \frac{\overline{N}}{\lambda}$$

IV. COSTS INTO THE MODEL

In order to evaluate the costs must be considered the decisions:

(i) Service Costs

(ii) Waiting time costs of customers.

Economic analysis of these costs helps the management to make a trade-off between the increased costs of providing better service and the decreased waiting time costs of customers derived from providing that service.

Expected Service Cost $(SC) = mC_m$

Where m = number of servers,

C_m = Service Cost of each server.

Expected Waiting time Costs $(WC) = (\lambda W) C_w$

Where λ = number of arrivals,

W = Average waiting time spends in the system,

C_w = Opportunity Cost of Waiting time by Customers

Expected Total Costs $(TC) = SC + WC$

Expected Total Costs $(TC) = mC_m + (\lambda W) C_w$

V. CASE STUDY

To compute the performance measures of the multi-server queuing system at the Banking System using the values of arrival rate (λ) =5 , Service rate(μ)=7 and the number of servers m=1,2,3,4,5.Then calculate the $\rho, P_0, L_q, L, W, W_q$ and finally calculate the total system cost.Here each server cost is 500 ($C_m = 500$) and waiting time cost is 800 ($C_w = 800$). Below table shows the Cloud Computing Service Providers waiting time and total system cost. The variations between the waiting time and total system cost should be shown in Figure 1,2,3.

Table 1: performance measures of multi-server queuing model at the Banking system

Performance servers (m)	**1**	**2**	**3**	**4**	**5**
Arrival rate (λ)	**5**	**5**	**5**	**5**	**5**
Service rate (μ)	**7**	**7**	**7**	**7**	**7**
Utilization (ρ)	**71%**	**35%**	**23%**	**17%**	**14%**

P_0	**0.2857**	**0.4737**	**0.4880**	**0.4893**	**0.4889**
L_q	**1.7864**	**0.1044**	**0.0122**	**0.0014**	**0.00014**
L	**2.5007**	**0.8186**	**0.7265**	**0.7157**	**0.7144**
W_q	**0.3572**	**0.0209**	**0.0024**	**0.0003**	**0.00003**
W	**0.5001**	**0.1637**	**0.1453**	**0.1432**	**0.1429**
Total system cost	**5502.40**	**1536.02**	**1922.24**	**2409.95**	**2908.38**

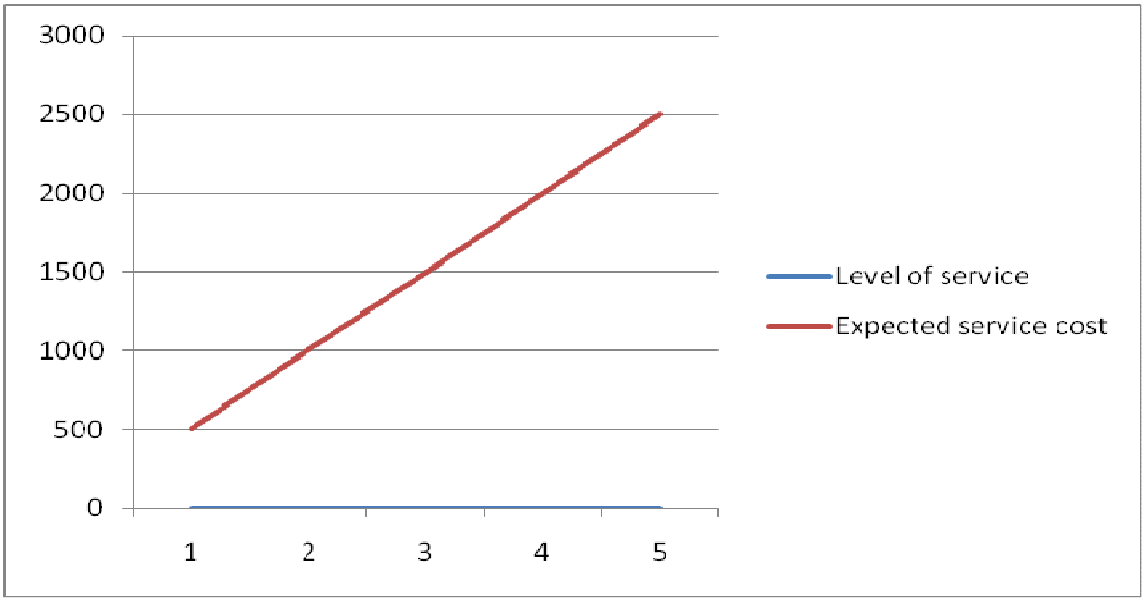

Fig1:Expected Service Cost against Level of service

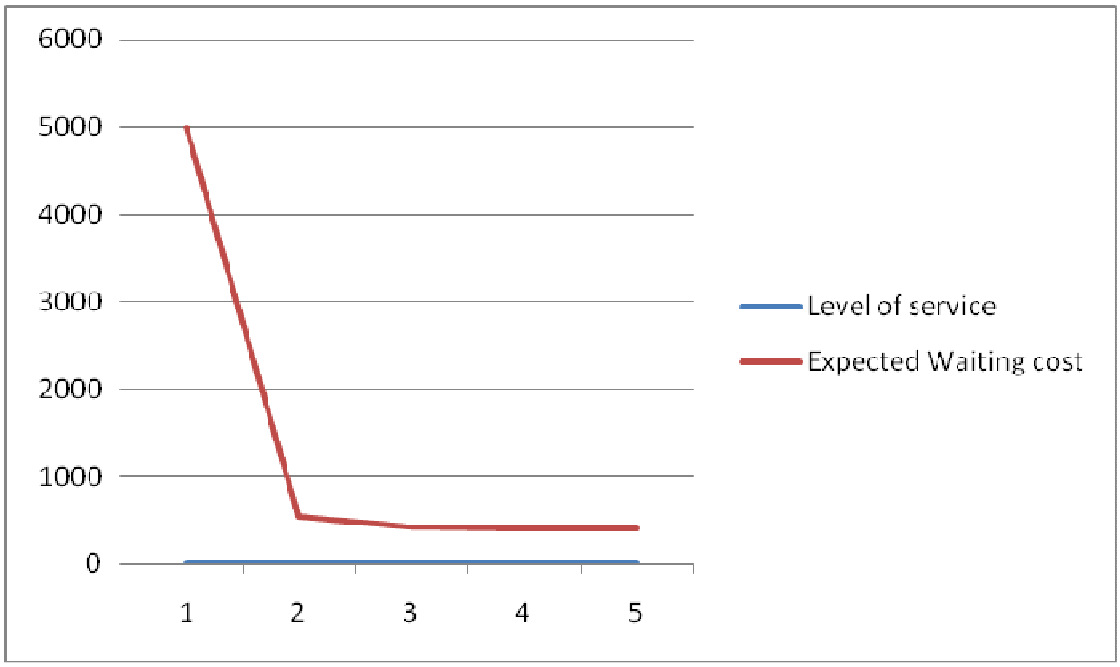

Fig2:Expected Waiting time Cost against Level of Service

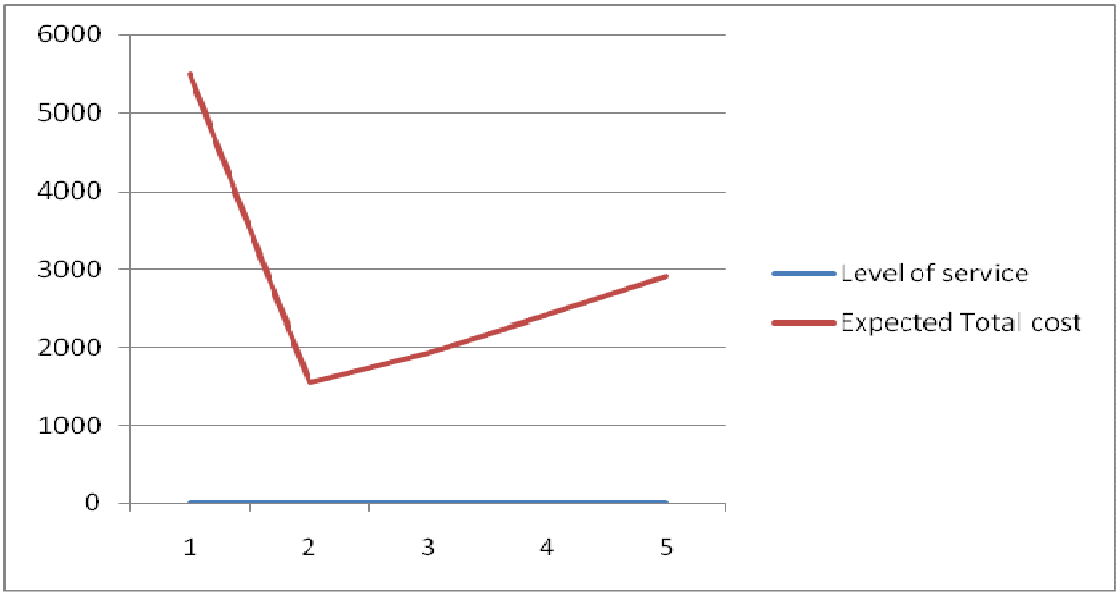

Fig3: Expected Total Cost against Level of Service

VI. CONCLUSION

By using M/M/m queuing model, the problem of optimal multi server configuration and minimization of cost in cloud computing environment can be achieved. We are considering the factors Number of servers (m), Cost, Waiting Time(W). From the above case study, it is observed that as the number of servers is increased the waiting time is reduced but the cost is increased.

If we consider Cost as the only criteria to be minimized then it is minimum for number of servers is 2 (m=2). If we want to minimize the waiting time only, then it is best when having maximum number of servers (here it is m=5). Finally if we want the optimal solution (i.e. by considering both waiting time and cost as optimal), then we have to identify the best combination of both (here it is for 2 servers usage with optimal cost and waiting time).

In this paper, we have used the single queue at the servers for processing the service requests which can be extended by using multi level queues, so that the cost and the waiting time can be reduced by using minimum number of servers.

REFERENCES

[1]http://en.wikipedia.org/wiki/Service_level_agreement, 2012.

[2] M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, Feb. 2009.

[3] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the Fifth Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.

[4] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," Nat'l Inst. of Standards and Technology, http://csrc.nist. gov/groups/SNS/cloud-computing/, 2009.

[5] D. Durkee, "Why Cloud Computing Will Never be Free," Comm. ACM, vol. 53, no. 5, pp. 62-69, 2010.

[6] L. Kleinrock, Queueing Systems:Theory, vol. 1. John Wiley and Sons, 1975.

[7] Junwei Cao, Kai Hwang, Keqin Li, and Zomaya A.Y, "Optimal Multiserver Configuration for Profit Maximization in Cloud Computing," IEEE Trans. Parallel and Distributed Systems, vol. 24, no. 6, pp. 1087-1096, Jun. 2013, doi: 10.1109/TPDS.2012.203.

[8] K. Li, "Optimal Load Distribution for Multiple Heterogeneous Blade Servers in a Cloud Computing Environment," Proc. 25th IEEE Int'l Parallel and

Distributed Processing Symp. Workshops, pp. 943-952, May 2011.

[9] K. Li, "Optimal Configuration of a Multicore Server Processor for Managing the Power and Performance Tradeoff," J. Supercomputing, vol. 61, no. 1, pp. 189-214, 2012.

[10] F.I. Popovici and J. Wilkes, "Profitable Services in an Uncertain World," Proc. ACM/IEEE Conf. Supercomputing, 2005.

AUTHORS PROFILE

M.G.Madhusudhan received the B.Tech degree in Computer Science and Engineering, from Santhiram Engineering College affiliated to JNTUA, Nandyal, Andhra Pradesh, in 2011. He is currently doing M.Tech in Computer Science from the Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, during 2012-2014.His current research interests include network security and cloud computing.

K.Delhi Babu received the MS from BITS Pilani in Software Systems and PhD degree in Software Architecture from Sri Venkateswara University Tirupati, 2011. He is currently Working as Professor in Computer Science Department at Sree Vidyanikethan Engineering College, Tirupati , Andhra Pradesh. His current research interests include Software testing, Software Architecture and Software Engineering.

