Improving QOS of Cloud Service Provider Using Multi Server Configuration

[1]Arathi John, [2]P.Rajasekar

Department of Information Technology, SRM University, Kattankulathur-603203
[1]arathijohn.p@gmail.com@gmail.com , [2]rajasekar.p@ktr.srmuniv.ac.in

***Abstract* - Cloud computing is a recently developed new technology for complex systems that provides multiple features that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. Typical cloud computing providers deliver common business applications online that are accessed from another Web service. Most cloud computing infrastructures consist of services delivered through common centers and built-on servers. Commercial offerings are generally expected to meet quality of service (QoS) requirements of customers, and service level agreements (SLAs). It is difficult to build a profitable pricing function for service providers, because there are several factors that can influence price.This paper deals with the problem of offering profitable prices for the cloud service provider in Cloud Computing markets by showing how the multiple servers can reduce the mean queue length and waiting time. Our approach is to treat a multi-server system as an M/M/m queuing type, a new technique to improve profit for the cloud providers.**

Keywords - Quality of Service, Cloud Profit maximization Model, Cloud Multi-Server Model, probability density function

1. INTRODUCTION

A computing Cloud is a set of network enabled services, providing reliable, QoS guaranteed, less cost on demand computing infrastructures, which could be accessed in an easier and pervasive way. Cloud computing is becoming one of the next IT industry buzz words: by computing resources and computing services. By centralized management of resources and services, cloud computing delivers a combination of traditional IT functions, such as infrastructure, applications, databases, information, and all resources are provided to consumer's on-demand over the internet. Cloud Computing offers the opportunity to access IT resources and services with appreciable convenience and speed. Behind this primarily, is a solution that provides users with services that can be drawn upon on demand and invoiced as and when used. Suppliers of cloud services, in turn, get benefit as their IT resources are used completely, achieve additional economies of scale. However, cloud computing will never be free, and understanding the economics of cloud computing becomes critically important. The problem of configuring multiple servers for increasing the profit in a cloud computing environment is studied. Pricing model takes such factors into considerations as the amount of a service, the amount of work of an application environment to do at a given time, the configuration of a multi-server system, the SLAs, the satisfaction of a final user, the quality of a service, the extra charge of a poor-quality service, the amount of renting, the cost of power consumption, and a cloud service provider's profit. Here a multi-server system is treated as an M/M/m queuing type, such that the problem of optimization can be framed and worked out analytically. Two means of server speeds and power consumption models are considered, namely, the inactive-speed model and the persistent-speed model. The probability density function (pdf) of the waiting time of a recently arrived service request is derived. The expected charge for a service request is calculated. The expected over all improvement of business, after all positive and negative influences have been accounted, in one unit of time is obtained. The calculations of the optimal server size and the optimal server speed are demonstrated numerically.

2. SCOPE

The scope is configuration of multiple servers for increasing the profit. In this multi server configuration system, the systems are configured based on the size and speed of the multi-server system .That is the number of servers and the execution speed of the servers.Two server speed and power consumption models are considered, namely, the inactive-speed model and the persistent-speed model. The pdf of the waiting time of a recently arrived service request is derived. The expected charge for a service request is calculated.

3. RELATED WORK

With cloud computing becomes popular, more and more cloud services are offered to clients in a pay-as-you-go manner. Therefore scheduling the dynamic user service requests more cost-effectively with less SLA violations is one of the most important problem of service providers.[5] Queuing models are generally constructed to represent the steady state of a queuing system. As a result, these are stochastic models that are having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely [10]. In a real world market, there exist various economic models for setting the price of services based on *supply-and-demand* and their value to the user [3]. A service provider rents resources from the infrastructure vendors, makes appropriate multi-server systems, and provides different services to end users [11] .An end user sends

a service request to a service provider, receives the desired response from the service provider with certain SLA [5], and pays for the service based on the amount of the service and the quality of the service. There is an algorithm exists called M/M/1 Queuing Model [10] which represents a single server that has unlimited queue capacity and infinite calling population, where arrivals and service are Poisson (or random) processes. A number of quite simple relationships can be derived for different performance measures based on knowing the arrival rate and service rate.

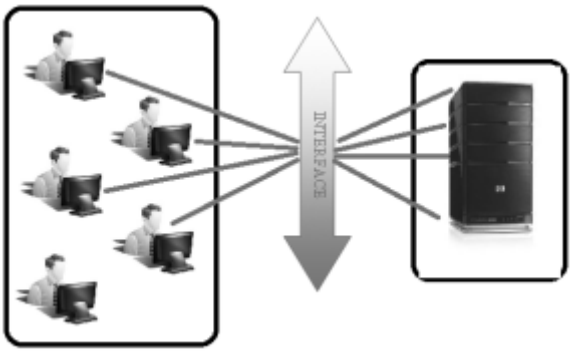

Fig 1: Bock diagram of the existing system.

But drawbacks of the system are low system performance by the increase of mean queue length, Increase of waiting time and response time, high Arrival rate and service rate, doesn't provide high quality service, traffic intensity can vary in an extremely wide range. There is also a difference between High Performance Computing (HPC) workload and Internet-based services workload.[1]

4. GENERAL STRUCTURE AND DESIGN

The problem of optimal multi-server configuration for increasing profit in a cloud computing environment is studied. A multi-server system is treated as an M/M/m queuing model, such that problem of optimization can be formulated and solved analytically. M/M/m queuing model is the only model that accommodates an analytical and closed form expression of the pdf of the waiting time of a newly arrived service request .The distribution of response time was obtained for a cloud center modeled as an M/M/m queuing system. Both inter-arrival and service times were assumed to be exponentially distributed, and the system maintains a finite buffer of size m.

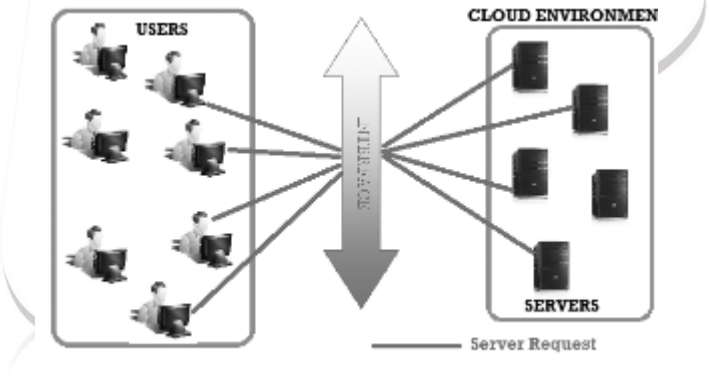

Fig 2: Bock diagram of the proposed system.

5. IMPLEMENTATION

The system which is implemented is divided into three parts. They are user multi-server, optimal size and speed. To improve the performance of the system a multi-server system can be considered as an M/M/m queuing model, such that optimization problem can be formulated and solved analytically.

Fig 3: Overall architecture

Two server speed and power consumption models, namely, the inactive-speed model and the persistent-speed model are considered. Main contributions are as follows. Deriving the pdf (pdf) of the waiting time of a recently arrived service request. This result is the base of this paper. Calculates the expected service charge to a service request. Based on these results, expected over all improvement of business, after all positive and negative influences have been accounted, in one unit of time can be obtained, and the optimal server size and the optimal server speed are calculated numerically.

6. SYSTEM MODEL

A.USER MODEL:

In this module, Users have authentication and security to access the detail. Before accessing or searching the details user should have the account in that otherwise they should register first. Once user gets validated they are allowed to communicate to server. In this module user views all the server uploaded files from cloud, then the user can give the request to server for the respective file. After sending request then calculation of the time period of getting response from server is calculated.

B.MULTI- SERVER MODEL:

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. The architecture detail of the multi-server system can be quite flexible. In this module server has authentication, server can upload the files.When server chooses the upload files file gets stored into the cloud storage. Client uses this storage files to give the request to server. The server views all the requests and their details. Then, collects the performance of all servers by considering factors such as ram speed, CPU usage, processor and memory. Based on performance server can be allocated to user for reducing the response and request time.

C.OPTIMAL SIZE AND SPEED

Server size optimization has clear physical interpretation. When m is small such that _ is approximately to 1, service requests waiting times are very long, and the service charges and the business profit are low. As m increases, the waiting times are reduced, and the service charges and the net business profit are increased. However, as m again increases, there will be not be any increase in the expected services charge which has an upper bound a_r; on the other hand, the cost of a service provider (i.e., the rental cost and base power consumption) increases, so that the net business profit is actually reduced. Hence, there is an optimal choice of m which maximizes the profit.

D.PROFIT MAXIMIZATION

The status of multi-server such as complete request, pending request are to be calculated. Based on that cost can be calculated. If there is delay, or change in SLA, penalty can be set by setting conditions.

7. APPLICATION

Application can be used for server allocation. In this application multiple servers are configured based on the user request. If multiple users sends the request to the single server, it cannot mange those request. If multiple requests came on time, they can be allocated to multiple servers for user.

8. CONCLUSION

This paper proposes a novel pricing demand scheme designed for a cloud based environment that aims at the increasing of the cloud profit with predictive demand price solution on economic way of user profit. A powerful multi server system reduces the penalty of breaking a service-level agreement and increases the revenue. By using an M/M/ m queuing technique, the problem of

optimal multi-server configuration for profit maximization in a cloud computing environment can be designed and worked out. This discussion can be easily extended to other service charge functions. This methodology can be applied to other pricing models.

REFERENCES

[1] 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.

[2] A.P. Chandrakasan, S. Sheng, and R.W. Brodersen, "Low-Power CMOS Digital Design," IEEE J. Solid-State Circuits, vol. 27, no. 4, pp. 473-484, Apr. 1992.

[3] R. Buyya, D. Abramson, J. Giddy, and H. Stockinger, "Economic Models for Resource Management and Scheduling in Grid Computing," Concurrency and Computation: Practice and Experience, vol. 14, pp. 1507-1542, 2007.

[4] D.E. Irwin, L.E. Grit, and J.S. Chase, "Balancing Risk and Reward in a Market-Based Task Service," Proc. 13th IEEE Int'l Symp. High Performance Distributed Computing, pp. 160-169, 2004.

[5] Y.C. Lee, C. Wang, A.Y. Zomaya, and B.B. Zhou, "Profit-Driven Service Request Scheduling in Clouds," Proc. 10th IEEE/ACM Int'l Conf. Cluster, Cloud and Grid Computing, pp. 15-24, 2010.

[6] R. Buyya, D. Abramson, J. Giddy, and H. Stockinger, "Economic Models for Resource Management and Scheduling in Grid Computing," Concurrency and Computation: Practice and Experience, vol. 14, pp. 1507-1542, 2007.

[7] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," Proc. 41st Design Automation Conf., pp. 868-873, 2004.

[8] Prof. Kishor S. Trivedi, " Availability, Performance and Cost Analysis for Large Scale Cloud" In Proc. RACOS workshop, held in conjuction with SRDS 2011

[9] H. Khaaei, J. Misic, and V.B. Misic, "Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 5, pp. 936-943, May 2012.

[10] N.Ani Brown Mary and K.Saravanan,"Performance Factors Of Cloud Computing Data Centers Using [(M/G/1) : (/GDmodel)] Queuing Systems" International Journal of Grid Computing & Applications (IJGCA) Vol.4, No.1, March 2012 Junwei Cao, Kai Hwang,

[11] Keqin Li and Albert Y. Zomaya, "Optimal mutiserver configuration for profit maximization in cloud computing",2013