Survey on Optimization of Resource Cost and Service Price Scheme in Cloud Computing

Kshitija S. Sontakke[1], Karan S. Mashal[2]

[1] Research Scholar, Dept .of Computer Engineering, RMD Sinhgad School of Engineering, Warje, Pune - 411058, Maharashtra, India

[2] Assistant Professor, Dept of Computer Engineering, RMD Sinhgad School of Engineering, Warje, Pune - 411058, Maharashtra, India

Abstract: *Cloud computing is a kind of internet-based computing and on demand computing where shared resources and information are provided to the customers on-demand. Profit is the most important factor from the cloud service providers' point of view and it is mainly determined by the configuration of a cloud service platform under given market demand. A single long-term renting scheme is usually used to configure a cloud platform, which cannot guarantee the quality of service but leads to serious resource waste. To overcome the drawbacks of single renting scheme, Double resource RR Renting scheme is designed which is the combination of both short term and long term renting. Double resource renting scheme not only guarantees the quality of service but also reduce the resource waste. In which M/M/m+D queuing model is used for job scheduling. Double resource renting RR scheme not only provides the Qos to the customers by using load balancing round robin algorithm but also maximize profit than single renting scheme*

Keywords: cloud computing, profit, pricing model. Service level agreement, queuing model, service charge, load balancing

1. Introduction

Cloud computing refers to both the applications delivered as services over the Internet, and the hardware and system software in the datacenters that provide those services. Cloud computing provides computing resources and services to the user on demand in pay per use manner anywhere and anytime, but for that user must connected to the high speed internet to access these services. Due to its advantages cloud has been more and more adopted in many areas such as banking, e-commerce etc. It provides secure, fast and convenient data storage. Cloud computing delivers three types of services as: Infrastructure as a Service, Platform as a Service, Software as a Service [3].

A service provider can build different multi server systems for different application domains, such that service requests of different nature are sent to different multi server systems. In cloud computing security is tremendously improved because of a superior technology security system, which is now easily available and affordable. Applications no longer run on the desktop Personal Computer but run in the cloud. This means that the PC does not need the processing power or hard disk space as demanded by traditional desktop software. Powerful servers and the like are no longer required. The computing power of the cloud can be used to replace or supplement internal computing resources.

Organizations no longer have to purchase computing resources to handle the capacity peaks. In the cloud computing for maximizing the profit first should understand the cost and revenue. Profit maximization must consider the user satisfaction also the cost of the cloud includes the renting cost and power consumption cost. For maximizing, must reduce the cost. For this it will configure the server perfectly. For configuring the server, calculate the expected waiting time and service charge is calculated.

2. Literature Survey

In the paper [2] they have discussed about pricing in cloud, aspects of pricing models and different types of pricing models such as pay-as-you go model, genetic model for pricing in cloud computing markets etc.

Pricing in cloud

Service provider provides requested services to the customer. Customer pays to the service provider based on the amount and the quality of the provided service. Pricing process can be of two types : Fixed and Dynamic In Fixed pricing mechanism same amount is charged to the customer all the time. Fixed pricing mechanism include pay per-use pricing model in which i.e customers charged based on their usage and consumption of a service.[2] Subscription is another type of fixed pricing, Subscription is another type of fixed pricing, in which the customer pays a fixed amount of money to use the service for longer periods at any convenient time or amount. In dynamic pricing mechanism customer is charged based on real time market conditions and the price of service is market dependent. Factors such as initial cost, lease period, quality of service, age of resources, maintenance cost which influence pricing in cloud computing is also discussed.

Aspects of pricing models in cloud
There are three main parameters of the pricing model [2] : Quality of Service, utilization period, pricing approach. The pricing approach describes the process by which the price is determined.

- **The pricing approach** could be one of the following:

The fixed price regardless of volume : In this fixed price is charged to the customer regardless of product utilization or volume of service.

The Fixed Price plus per-unit : In this fixed price plus a unit rate is charged to the customer

assured purchase volume plus per-unit price rate : In this customer pays fixed charge for certain quantity

per-unit rate with a ceiling : In this approach customer pays per unit rate up to a certain limit. The service provider will not charge the customer beyond that limit

Price per-unit approach: In this approach different price per unit is charged to the customer.

- **Quality of service (Qos)**
Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance. Quality of service include on time delivery of service, providing security and privacy, scalability and integrity of the service provider. If service provider maintains all these requirements at high level, no. of customers as well as customers loyalty towards service provider will increase.

- **The utilization period**
In the utilization period customer has the right to use the provider services based on service Level Agreements. It could change, based on the subscription period, or a pay-per-use model.

Cloud computing is a next generation technology. Cloud computing is on demand service because it provides flexible dynamic resource allocation for guaranteed and reliable services in pay as you-use manner to the customers. SaaS is a very important layer in the cloud computing because all the resource allocation is done through SaaS providers. This paper focused on for the SaaS provider customer request should minimize the cost or increase the profit with dynamic demand handling. For achieving user satisfaction and maximizing the profit of cloud service providers an efficient strategy should be required. This [3] paper focuses only on quality of service parameter and service level agreement of SaaS layer in cloud computing.

In the [1] paper pricing model is proposed for cloud computing which include the factors , such as the service requirement, application environments workload, the configuration of a multi-server system, SLA, consumer satisfaction, Qos, the penalty of a low-quality service, renting cost, energy consumption cost, and margin of a service provider and profit. M/M/m queuing model is used to formulate and solve problem of the optimal multi-server configuration for profit maximization in a cloud computing environment.

Multi objective genetic algorithm is proposed in the [4] for solving multivariable optimization problem. This system provides a new cloud brokering mechanism with cloud service discovery using this optimization technique. This paper mainly focus the IaaS layer of the cloud computing. In this system user submits job to the service provider and service provider provides services to the user. Here the user

requirement is to obtain services within deadline and within minimum cost. At the same time service providers aim is to increase the income for that he should give the quality of service to the user by minimizing customers cost , customers waiting time and providers idle time and maximize service providers benefit.

3. Proposed Work

3.1 Service Level Agreement (SLA)

A service-level agreement (SLA) is a contract between customers and its service providers. A consumer sends a service request to a service provider and receives the desired result from the service provider with certain service-level agreement, then pays for the service based on the amount of the service and the quality of the service. If the service provider violates this SLA no charge is provided for the particular service. So there will be the loss of the profit. Many types of businesses use SLA. SLA adopts a price compensation mechanism to guarantee service quality and customer satisfaction. For example, China Post gives a service time commitment for domestic express mails. It promises that if a domestic express mail does not arrive within a deadline, the mailing charge will be refunded. The SLA is also adopted by many real world cloud service providers such as Microsoft Azure , Rackspace, Joyent and so on. [5]

3.2 Queuing Model

The cloud service platform is a multi-server system with a service request queue. The clouds provide resources for jobs in the form of virtual machine (VM). In addition, the users submit their jobs to the cloud in which a job queuing system such as SGE, PBS, or Condor is used.[5] In the centralized manner job scheduler schedules all jobs and assigned it to different Virtual Machines. Hence, it is one type of service request queue. For example, Condor is a specialized workload management system for compute intensive jobs and it provides a job queueing mechanism, scheduling policy, priority scheme, monitoring of resources, and management of resources. Users submit their jobs to Condor, and Condor put them into a queue, chooses when and where to run them based upon a policy. An M/M/m+D queueing model [1] is build for our multi-server system with varying system size. And then, an optimal configuration problem of profit maximization is formulated in which many factors are taken into considerations, such as the market demand, the workload of requests, the server-level agreement, the rental cost of servers, the cost of energy consumption, and so forth. The optimal solutions are solved for two different situations, which are the ideal optimal solutions and the actual optimal solutions.
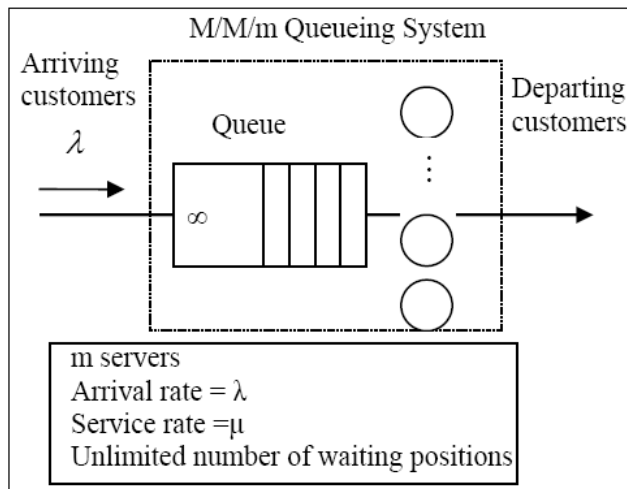
Figure 1: Queuing Model

3.3   Load Balancing

Load balancing is used to distributing a larger processing load to smaller processing nodes for enhancing the overall performance of system. In cloud computing environment load balancing is required distribute the dynamic local workload evenly between all the nodes.[8]

Load balancing is used to allocate computing   resource to achieve a high satisfaction of user and proper utilization of resources. High utilization of resources and Proper balancing of load helps in reducing resource consumption. It helps in implementing fail over, scalability, and avoids bottlenecks.

- Load balancing is a technique that helped networks and resources by providing a Maximum throughput with minimum response time. Load balancing[8] is divide the traffic between all servers, so data can be sent and received without any delay with load balancing.
- In cloud environment many algorithms are available that helps in proper traffic Loaded between all available servers. Most of them can be applied in the cloud environment with suitable verifications. In cloud computing environment load balancing algorithms is split into two main groups : first algorithm type is Batch mode heuristic scheduling algorithms (BMHA) and second is online mode heuristic algorithms. In BMHA Jobs are combined together when they are arriving in the system. The BMHA scheduling algorithm will start after a fixed period of time.
- Min Min algorithm, First Come First Served Scheduling algorithm (FCFS), Max Min algorithm, Round Robin scheduling algorithm (RR), and are the examples of BMHA based algorithms.[8]
- In our daily life example websites are the example of load balancing. Users could face many problems without Load balancing like delays, timeouts and long system responses.[8]
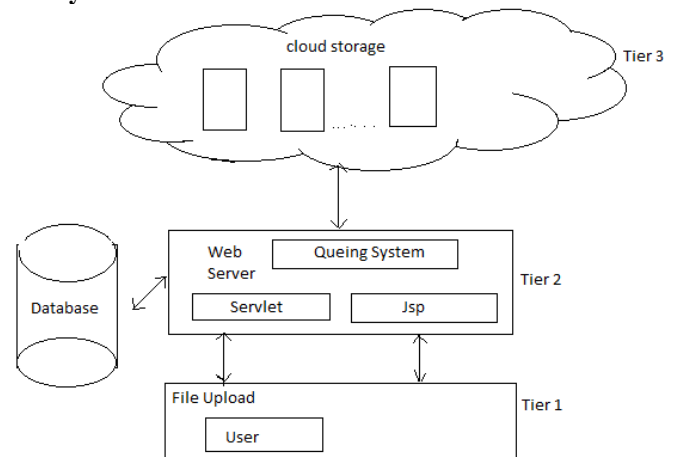
3.4 System Architecture

Figure 2: System Architecture

A novel renting scheme for service providers is proposed, which can satisfy quality-of-service requirements as well as obtain more profit. A double resource renting RR scheme which is also called as double quality guaranteed RR Scheme (DQG-RR) is proposed for service providers. It combines long-term renting with short-term renting, which not only satisfy quality-of-service requirements under the varying system workload, but also reduce wastage of resources. we are going to implement and improve QOS of cloud service provider by balancing workload of VMs. Whenever workload of one VM increased then the request get transfer and processed by another VM. We are going to develop load balancing algorithm to maximize the profit of our cloud. A multi-server system adopted is modeled as an M/M/m+D queuing model and the performance indicators such as the average service charge, the ratio of requests that need short term servers are analyzed. The optimal configuration problem of service providers for profit maximization is formulated and two kinds of optimal solutions, i.e., the ideal solutions and the actual solutions, are obtained respectively.

A few major goals of load balancing algorithms:

- **Cost effectiveness:** primary aim is to achieve an overall improvement in system performance at a reasonable cost.
- **Scalability and flexibility:** the distributed system in which the algorithm is implemented may change in size or topology. So the algorithm must be scalable and flexible enough to allow such changes to be handled easily.
- **Priority:** prioritization of the resources or jobs need to be done on beforehand through the algorithm itself for better service to the important or high prioritized jobs in spite of equal service provision for all the jobs regardless of their origin.

4. Conclusion

A novel double quality guaranteed renting scheme is proposed which guarantee the Qos to the customers and maximize the profit of the service providers. M/m/m+D queuing model is used for multi-server system which takes many factors into consideration request workload, SLA, rental cost of servers etc . This paper presents a concept of

180

Cloud Computing along with load balancing which consist many factors like Cost effectiveness , scalability, flexibility and priority.

References

[1] J. Cao, K. Hwang, K. Li, and A. Y. Zomaya, "Optimal multiserver configuration for profit maximization in cloud computing," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1087–1096, 2013.

[2] May Al-Roomi, Shaikha Al-Ebrahim, Sabika Buqrais and Imtiaz Ahmad, " Cloud Computing Pricing Models: A Survey", International Journal of Grid and Distributed Computing, Vol.6, No.5, pp.93-106, 2013

[3] Richa, Hari Singh, "A Review of SaaS Profit Maximization in Cloud Computing", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015

[4] Dr. P. Balakumar, Deepa.V, "Qos Supported SLA for Profit maximization of multi-server configuration in cloud computing", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 3, March – 2014

[5] Jing Mei, Kenli Li, Member, IEEE, Aijia Ouyang and Keqin Li, Fellow, IEEE, "A Profit Maximization Scheme with Guaranteed Quality of Service in Cloud Computing", I EEE TRANSACTIONS ON COMPUTERS, VOL. 1, NO 2015

[6] H. Xu and B. Li, "Dynamic cloud pricing for revenue maximization," IEEE Trans. Cloud Computing, vol. 1, no. 2, pp. 158–171, July 2013.

[7] Y.-J. Chiang and Y.-C. Ouyang, "Profit optimization in sla-aware cloud services with a finite capacity queuing model," Math. Probl. Eng., 2014.

[8] Rajwinder Kaur and Pawan Luthra, "Load Balancing in Cloud Computing", Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC

Author Profile

Kshitija. S. Sontakke Research Scholar RMD Sinhgad School of Engineering, Savitribai Phule Pune University. She has received B.E. in Computer Science and Engineering from Brahmdevdada Mane Institute of Technology, Solapur University, Solapur (2012). Currently she is pursuing M.E. in Computer Engineering from RMD Sinhgad School of Engineering, Warje, Pune, Savitribai Phule Pune University

Prof. Karan S. Mashal Prof. Karan Mashal received M.tech in Information Technology. He is currently working as an Assistant Professor in Computer Engineering, RMD Sinhgad School of Engineering, Savitribai Phule Pune University. His area of interest is web development, operating system and smart system. He has 5 years of teaching experience.