Revenue Maximization for Cloud Computing Environment through Resource Sharing by SLA

Asish Mishra
M tech Student, School of Computer Engineering, KIIT University Bhubaneswar, India

Muheet Ahmed Butt
Scientist, PG Department of Computer Science, Unvierstiy of Kashmir, J&K India:

Majid Zaman
Scientist, Directorate of IT&SS, University of Kashmir, J&K,India

ABSTRACT
Cloud computing is receiving a great demand among the budding and old enterprises depending on various Information Technology Services in the present IT Scenario. Cloud provides various services which mostly depend upon a good decision making process to handle various requests from service consumers. The Cloud users are mostly divided into two premium or priority and basic users. As far as priority users are concerned the resources are reserved well in advance and it shows a strong negotiation between the service providers and end users. If there is lack of resources the provider has to pay some reimbursement to the user as per the agreement called as Service Level Agreement (SLA), whereas the elementary user gets the service but has to make more than the premium user. In this case providers always attempt is made to reduce the repayment and maximize the revenue. Here the goal of our paper is to achieve the service pooling mechanism with an energy-aware allocation method in PAAS model to save the overall expenses. Experimental outcomes in the proposed research also demonstrate how the projected outline is able to handle cost, revenue, penalty and efficient server utilization for cloud services**.**

Keywords
Cloud Computing, Server Utilization, Power Saving, Revenue
.

1. INTRODUCTION
Cloud computing is a design for empowering ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources. (e.g., networks, servers, storage, applications, and services) which are very flexible for utilization as compared to other technologies present in the market. It has different type of model such as service model and deployment model. During the excessive demand of resource in cloud there should be a method which will be able to provide resources as per the consumer demand. That's why we have proposed a shared based resource model which will help us to take a decision at the time of heavy demand of resources from the consumers.

The total resource is the combination of the all resources of each individual service provider. It is very difficult for an individual cloud service provider to provide a resource during high demand period. This model mainly focuses on the provider perspective to provide the user the demanded resource with maximizing its revenue.

There are numerous challenges at cloud service level architecture. The proposed research is mostly focused on revenue maximization for Cloud computing environment through resource sharing by SLA. The unavailability of resources in the Cloud Environment creates an obstacle for the cloud provider in revenue maximization. The failure in providing the resource to the consumer results in providing various penalties in terms of capital and service.

The proposed research aims to develop a cloud billing system for revenue maximization for the provider keeping in view both premium and basic type of consumers where it can extend the local resource to global resource so as to provide a better service during overflow situation. This solution will help Cloud Service Providers to maximize profits through resource sharing by SLA applications among various users.

2. RELATED WORK
A Performance management and revenue maximization by application of scheduling is a hot topic in both cloud and grid environment. The function of job scheduling system is responsible for best resource selection and allocation. The research work on revenue optimization of cloud can be broadly divided two types. For optimal server utilization and energy efficient allocation queuing theory is applicable to handle both type of systems [1]. The queuing model which is used to manage the power consumption of Server is given by service level agreement (SLA).in [2] the customer has to wait for a service which has an unknown waiting time due to unavailable of resource. [3] Is focused to minimizing the Energy-response time but is does not consider the cost factor of lost job.[4] is focused on optimize the performance by application of hierarchical job scheduling. In recent years, more and more researches are going to study the QoS for various scheduling mechanisms, but in [5][6][7][8] research is focused on cloud computing scheduling in a dynamic approach. However most of these researches rarely maintain the QoS with respect to better service scheduling.

Apart from this, very less research has been done about the maximization of profit with respect to better service. The main condition for the existence of the cloud system is to get maximum profit. In this paper, the proposed model will handle various factors pertaining to Cloud Computing Environment like better pricing, resource availability and power saving so that while analyzing these environments better dynamic decisions could be taken.

3. CLOUD PLAYERS
Cloud computing has become a common platform for various service providers who can offer different kinds of services to the clients who expect to find most of the services they need at a single place. There are mainly two types of players in cloud computing environment-one is customer, another is cloud service provider. Accordingly these two perspectives can be considered namely customer perspective and provider perspective. Customer perspective deals with the customer who can get many benefits from cloud computing

environment with low cost for service, less maintenance cost, ubiquitous access in a global level with wide range of choices. But it has restricted resource allocation during high demand phases in the system. Provider perspective deals with large market scale, stable revenue inflow, lower entry barrier and better strategic positioning.

4. PROPOSED GUIDELINES FOR SLA'S

The Cloud Service Provider has to follow certain rules for optimizing various services provided by it during the high demand period. So, To make this proposed model a novel one certain rules have to followed that can be helpful to the Cloud Service Provider to maintain the global sharing of its resources. When there is an optimal demand from the users which not dealt properly can result in a system overflow which in turn can malfunction certain vital services provided by the provider.

In the proposed model the CSP has to reserve 10% of its total resources for the global sharing but which can be used only after enabling the global accesses permission. To enable a provider to use an extra resource beyond 10% of already allocated ones the provider will have to go thought a reservation process for its premium users which will be having a different price structure. The validity of this external reservation facility will be time bound which will be assigned during the reservation process only.

The CLP will have to setup a special high speed bandwidth network with the sharing center so as to minimize the processing time and enhance the efficiency. The CLP can utilize the unused resources present in the sharing center which means that provider can sell more than 10% resource to the sharing center to get better revenue. These are the proposed golden rules which are the back bone of our model which play a very important role in maximizing the revenue of the CLP.

5. PROPOSED SYSTEM MODEL

At certain point of time the Cloud Service Provider has two types of resources available, one is small "s" (local resource) and another is Capital "S"(shared resource) .The provider provides these two types of resources according to increase in service rate as per the users demand. The user has to pay a suitable amount according to the service package at point of service request.

Resources are the backbone for creating a Cloud Based Service Environment for running various kinds of jobs. These jobs are controlled by two types of conditions i.e. "overflow" and "underflow". The size of job and duration of its use is decided by the user. The jobs are provided with the local or underflow resource i.e. "s". The processing of job mainly follows the M/M/C/C queuing model to serve the customer request for any service. If the numbers of request (Rq_{Total}) is more than the numbers of local resources("s")then request will be forward to the newly shared queue, where the jobs are provided with "S" (shared or overflow resource). The shared system follows the M/M/C/N model that provides the service to various users under Cloud. The service level extends from "C" to "N" so rate of lost jobs is minimized and the revenue can be maximized up to creation limit. The details of the proposed model are shown in the Figure 1.

Depending on the use the users are classified as the "premium user "and the "instant user". Than main difference between two users is the premium user reserve their resource by paying some advance money for service allocation, whereas the instant user has to pay a high amount of money so as to use the resource instantly. Both the users are capable of processing request "R_{q1}"and "R_{q2}" respectively.

"R_{q1}"+ "R_{q2}"= "R_{qTotal}"……(eq. 1)

The level one system is capable of handling the only when $s>R_{qtotal}$ (underflow state).The pricing structure of underflow is set by the local service charges of the local user provider only. Similarly, the level two system only handles the request which satisfy two conditions $s< R_{qtotal}$ and $S> R_{qtotal}$ (overflow).The pricing of overflow is high as compare to underflow because it follows the shared charging system

5.1 Service request processing system

The service requests are served in two ways constituting of Underflow Service Request processing and Overflow Service Request Processing. One is based on availability of local resource and second is based on type of customer.

The local system is classified into two subsystems according to the type of user-premium and instant. Both the subsystems treat their customer separately. The processing of the requests are served by M/GI/C/C (Markovian Arrival follows general distribution).Here,in this type of model the number of input is equal to number of servers. If all the servers are busy then the job will be simply lost as there is no waiting space. Hence, the rate of loss is high. Before entering into the system, jobs will first check he service capability.

5.2 Revenue calculation system

The revenue calculation process is mainly done by provider's perspective only.

Revtotal = Revunderflow + Revoverflow

Revunderflow = Revprimium + Revinstant

Revoverflow = Revhybrid

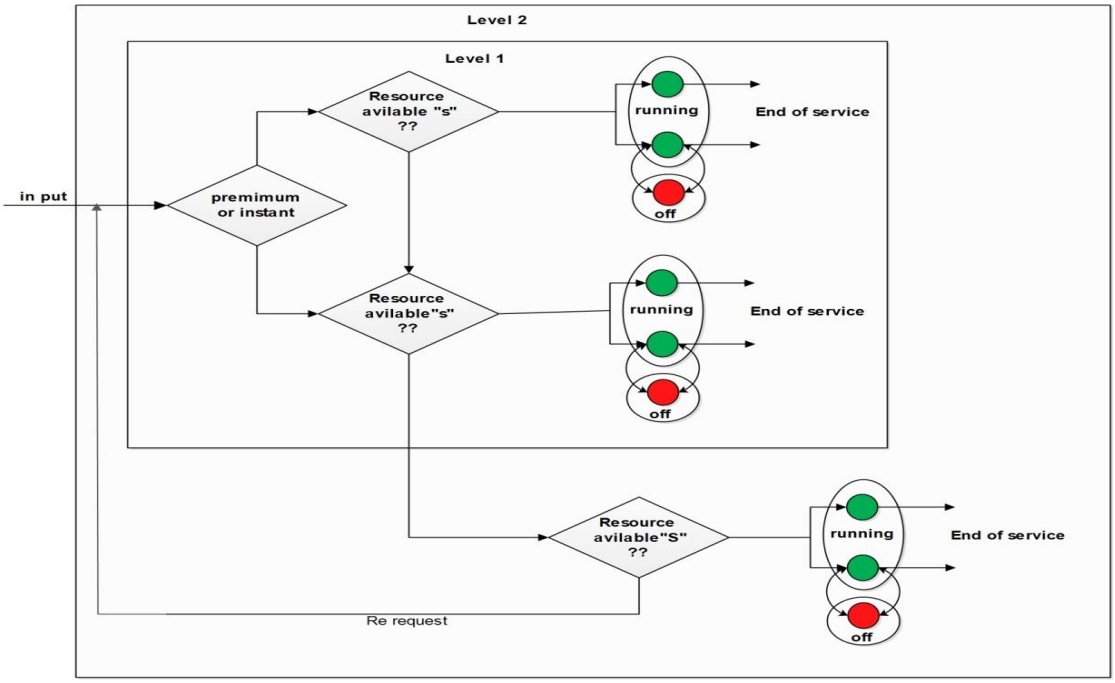

Fig 1: Cloud request processing model for revenue maximizing

5.2.1 Under flow state Revenue collection:

The total revenue of underflow case is controlled by the local provider. Here, revenue is classified into two types according to the user, one is revenue for a premium user and another is revenue for instant user.

Rev (premium) = (Cp /μ) * T1 -r * P1
Rev(instant) =(Ci /μ) * T2 -r * P2

Where,
Cp =Charge paid by premium type customer
Ci=Charge paid by instant type customer
1/ μ=Average service time
Pi = Power Consumption of 'i'(The power is calculated from[3])
Ds =Local penalty paid by the premium customer due to unavailability of resources
Ti =Throughput of i th local subsystem
r=Fixed electric cost per unit power consumption

5.2.2 Over flow state Revenue collection:

It is the shared system created by multiple service providers to handle high resource demand during lack of local resource.

Both the type of customer has to pay through the shared billing system. The system is called Hybrid system because it is capable of serving both type of customer in a similar manner.

$Revhybrid = (C_H/\mu)*T_H - r*P_H$

Where C_H =Charge paid by the Hybrid customer 'H'

6. CONCLUSION AND FEATURE WORK

This paper proposes a model to determine the optimized revenue of cloud provider by application of power saving and also capable to handle the maximum number of service request . Where both the instant and premium customers get well service in overflow and underflow state conditions (i.e. overflow to extension of hybrid state). The services are served on best effort basics.

The optimal allocation and revenue maximization is possible due to application of shared system and power saving .The penalty of the provider is also reduced because hybrid system is cable to handle the service request during the overflow state.

For future work, the distributed algorithm can be implemented with an SLA policy for quick resource sharing among the providers.

7. REFERENCES

[1] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing server energy and operational costs in hosting centers," SIGMETRICS Perform. Eval. Rev., vol. 33, no. 1, pp. 303–314, 2005 .

[2] D. Dyachuk and M. Mazzucco, "On Allocation Policies for Power and Performance," in Proceedings of the 11th ACM/IEEE Grid, October 2010

[3] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. Kozuch, "Optimality Analysis of Energy-Performance Trade-off for Server Farm Management," in Proceedings of the 28th Performance 2010, November 2010

[4] Zhiguang, S. and L. Chuang. *Modeling and Performance Evaluation of Hierarchical Job Scheduling on the Grids*. in *Grid and Cooperative Computing, 2007. GCC 2007. Sixth International Conference on*. 2007.

[5] Weidong, H., Y. Yang, and L. Chuang. *Qos Performance Analysis for Grid Services Dynamic Scheduling System*. in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*. 2007.

[6] Afzal, A., A.S. McGough, and J. Darlington, *Capacity planning and scheduling in Grid computing environments.* Future Generation Computer Systems 2008. **2008**(24): p. 404-414.

[7] Kiran, M., et al. *A prediction module to optimize scheduling in a grid computing environment*. in *Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on*. 2008.

[8] *Yuan-Shun, D., X. Min, and P. Kim-Leng, Availability Modeling and Cost Optimization for the Grid Resource Management System. Systems*, Man and Cybernetics, Part A, IEEE Transactions on, 2008. **38**(1): p. 170-179.