ISSN 2278-5612

Utility Based Resource Allocation Model for Cloud Services

J.Antony Judi[1], F.Ezhil Mary Arasi[2], Dr.S.Govindarajan[3]

[1] Senior Systems Analysts, Accenture Private Limited, Chennai
antony.judi@accenture.com
[2] Asst.Prof, Department of MCA,SRM University, Chennai
ezhilmary.f@ktr.srmuniv.ac.in
[3] Professor, Department of MCA,SRM University, Chennai
govindaraja.n@ktr.srmuniv.ac.in

ABSTRACT

Minimizing Resource allocation problems under the demand and price uncertainty in cloud computing environments is the motivation to explore a resource provisioning strategy for cloud consumers. In this paper a utilization-based optimal cloud (UBOC) algorithm is proposed to minimize the total cost for provisioning resources in a certain time period. To make an optimal decision, the demand uncertainty from cloud consumer side and price uncertainty from cloud providers are taken into account to adjust the tradeoff between on-demand and oversubscribed costs. Using this UBOC user can share cloud resources and pay based on the usage and the results show that this algorithm can minimize the total cost under uncertainty. It also provisions the resources to remove the demand uncertainty.

Keyword: Clouds Services, Resource Allocation, Cost

Council for Innovative Research
Peer Review Research Publishing System
Journal: International Journal of Management & Information Technology

Vol. 6, No. 3
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 2278-5612

1. INTRODUCTION

Cloud computing is a popular trend in current computing which attempts to provide cheap and easy access to computational resources. Cloud computing is the use of computing resources (hardware and software), that are delivered as a service over a network. End users access cloud-based applications through a web browser or a light-weight desktop or mobile app while the business software and user's data are stored on servers at a remote location.

Cloud Computing adopts concepts from Service-oriented Architecture (SOA) that can help the user to break these problems into services that can be integrated to provide a solution. Cloud Computing provides all of its resources as services, and makes use of the well-established standards and best practices gained in the domain of SOA to allow global and easy access to cloud services in a standardized way.

In computing clouds, it is desirable to avoid wasting resources as a result of under-utilization and to avoid lengthy response times as a result of over-utilization. This research ensures fair resource allocation between sites and users, dynamic load changes adaptability, minimize the total cost for provisioning resources in a certain time period, overcome the demand uncertainty from cloud consumer side and price uncertainty from cloud service providers.

2. LITERATURE SURVEY

Dynamic resource management is one of the most challenging problems in cloud environment. This problem has attracted a lot of attention from the research community in the last few years. In the following we provide a review of most relevant prior work.

Multi-dimensional resource allocation for single tier applications in the cloud computing system is presented in [1].SLA model based on the response time of the applications is considered to model the profit optimization problem. This problem is solved with generating an initial solution and using local optimization techniques. Tang et al. [2] presents a dynamic resource provisioning technique for the case of very large number of servers and application sizes. The proposed heuristic solution for this NP-hard problem is focused on the scalability aspects of the solution. Virtualization management policies are presented in [3] to handle the performance, efficiency and stability of a server system. The results show that effective dynamic resource management can greatly reduce the operation cost of the system and improve the stability of the applications.

In [4], Zhang and Ardagna extend the early work of [5] and present a problem statement with clients that have discrete utility functions. The authors propose a heuristic to solve the problem of assigning different client classes to different servers to maximize the total profit. Ardagna et al. [6] extend this work to profit (revenue) optimization for continuous utility functions

in a multi-tier virtualized environment. The authors use a complex model for energy calculation to increase the accuracy and solve the problem by generating a feasible solution and improving it by local search. The availability of servers is considered as a new constraint to the problem in [7]. A heuristic to maximize the profit via decreasing the energy consumption in cloud systems is presented in [8], where an adaptive search based on turning servers on or off is proposed. Resource allocation for tasks with fixed memory, disc and processing requirements is presented in [9].

References [10]-[12] use mathematical or economics based models to formulate the profit optimization problem.The problem of gossip resource management address in the existing is related to two lines of research, which are application placement and load balancing in processor networks. Application placement in data centers is often done through mapping a set of applications onto a set of machines (nodes) such that some utility function is maximized under resource constraints. Each node has a specific CPU capacity and memory capacity. When the node is allocated and user requires smaller memory than available memory, no provision is available. The solutions from these works have been incorporated in middleware products .While these product solutions, in a similar way as this scheme does, allow for computing an allocation that maximizes the utility but does not scale to system sizes. This process also requires global synchronization.

In this paper an optimal cloud Service Level Agreement (OCSLA) algorithm is proposed to minimize the total cost for provisioning resources in a certain time period

3. UTILIZATION-BASED CLOUD RESOURCE ALLOCATION

3.1. Utilization based Optimal Cloud Algorithm

Step-0: In this step the under provisioning and over provisioning threshold point measured by the past day performance of specific cloud service. Let denote under provisioning threshold point as $\tilde{U}P_{th}$ and over provisioning point as $\check{O}P_{th}$.

$$\tilde{U}P_{th} = \sum_{i}^{i<=0} RCi$$

//RC-> Resource Consumer

$$\check{O}P_{th} = \sum_{i=11}^{\infty} RCi$$

Step-1: Sub problem solution: In this Step, allocation of cloud service cost according to the level of $\tilde{U}P_{th}$ and $\check{O}P_{th}$ is carried out. For example,

If n = ŨƤ$_{th}$

// n-number of registered cloud resource consumers

Set ŨƤ$_{cost}$ = ŨƤ$_{cost1}$

// ŨƤ$_{cost}$ –Under Provisioning Cost

Else if n> ŎƤ$_{th}$

Set ŎƤ$_{cost}$= ŎƤ$_{cost2}$

// ŎƤ$_{cost2}$ –Over Provisioning Cost

Step-2: Convergence checking: In Step-2, the convergence such as time limit, size limit of ŨƤ$_{customer}$ and ŎƤ$_{customer}$ are carried out .

ŨƤ$_{customer}$ -> Customer at Under provisioning Region.

ŎƤ$_{customer}$ -> Customer at Over provisioning Region.

Step-3: Auto Repayment. In Step-3, Auto repayment process carried out for each and every success outcome of convergence checking (from Step-2).To initiate Repayment process, following calculations are carried out automatically,

ƮƯ=$\sum_{i=1}^{n} CUi$

ƮƯ-> Total Usage,

CUi->Consumed Units

ḂƯ=$\sum_{i=1}^{n} UCUi$

UCU-> Unconsumed Units

ḂƯ->Bending Usage

ƮɃ= ƮƯ̇ x C/U

C/U->Cost per Units,

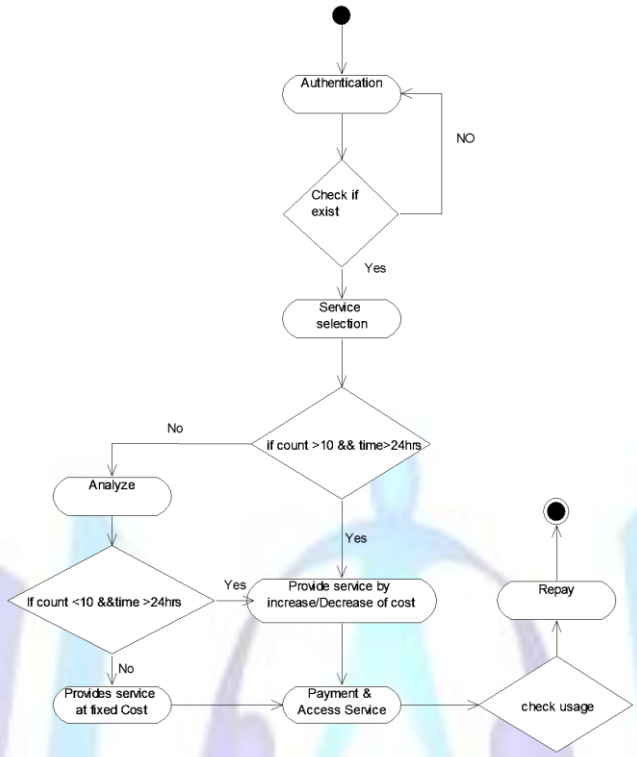
ƮɃ=Total Bill

ṚĂ= ḂƯ x C/U

ṚĂ-> Repayable Amount.

3.2. Utility-Based Cloud Resource Allocation Architecture

The components of the utilty-based architecture that supports resource allocation in multi-Cloud environments (Fig. 1) are as follows:



4. IMPLEMENTATION

The implementation of the work is divided into four phases: Cloud services, Cloud setting, Service Usage and Consumption.

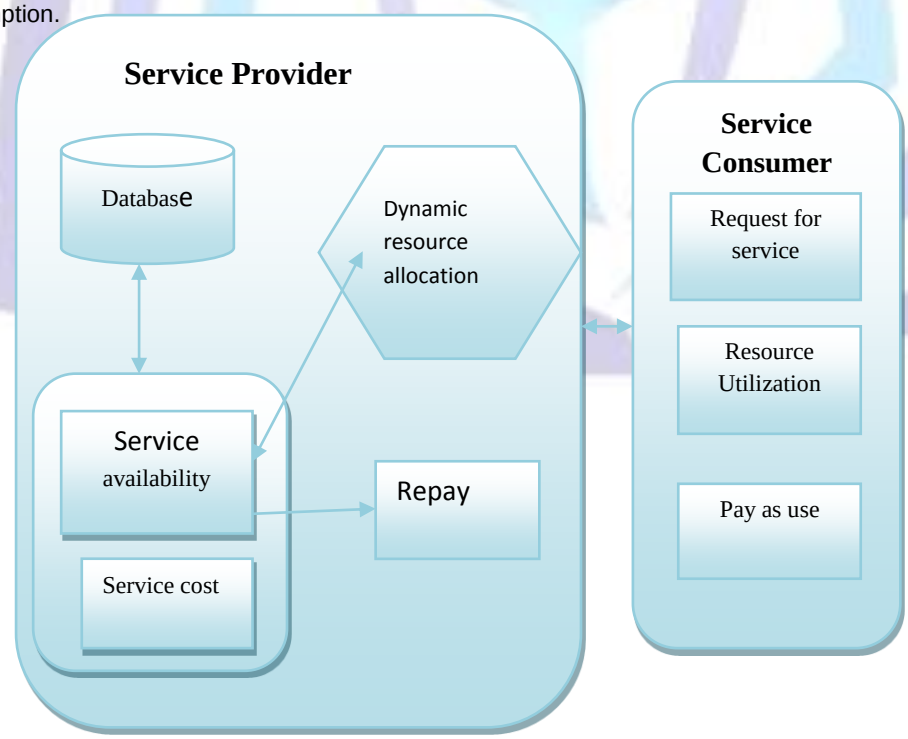


Fig 1: Cost-Based Cloud Resource allocation Architecture

Cloud Services

In cloud services phases all the available cloud storage services and web services cost and its time limit are available to customers. Customers can see all the services which are provided by the cloud service provider and their cost value and time to consume the services. Also the customer can see the services which are owned by them previously and use those services.

Cloud setting

The user can select the services based on their requirements. Registered user is allowed to create new settings such as maximum time limit, maximum budget, and maximum storage consumed and allocation of resources. User can buy services for the limited period of the time.

Service usage

Authenticated user is allowed to use the available cloud storage services and web services based on cost and time period of the services. The cost of the services can be automatically increased and decreased based on the past day performance. If the number of customer is more than 10 then the cost for the service can be increased by 10 percent of the actual cost and if the resource cannot be consumed by no one then the cost of the resource can be decreased by 10 percent of the actual cost. This will help to remove the under provisioning and over provision of the resources. Customer can upload the files in their storage area download the files from their also able to see the available memory and see the files uploaded and download. Using web services customer can provide security to their data, they can decrypt and encrypt their images and files and store them safely.

Consumption

This phase shows the current consumption details and the total cost of services for that particular user. It also provides the repayment for the services. If the owned size cannot be consumed by the customer with in time limit of 24 hours, after 24 hours automatically repayment process can be completed and the amount is repaid to user and the files from the storage will be deleted.The cost-based Cloud resource allocation scenario (Fig. 2) is as follows.

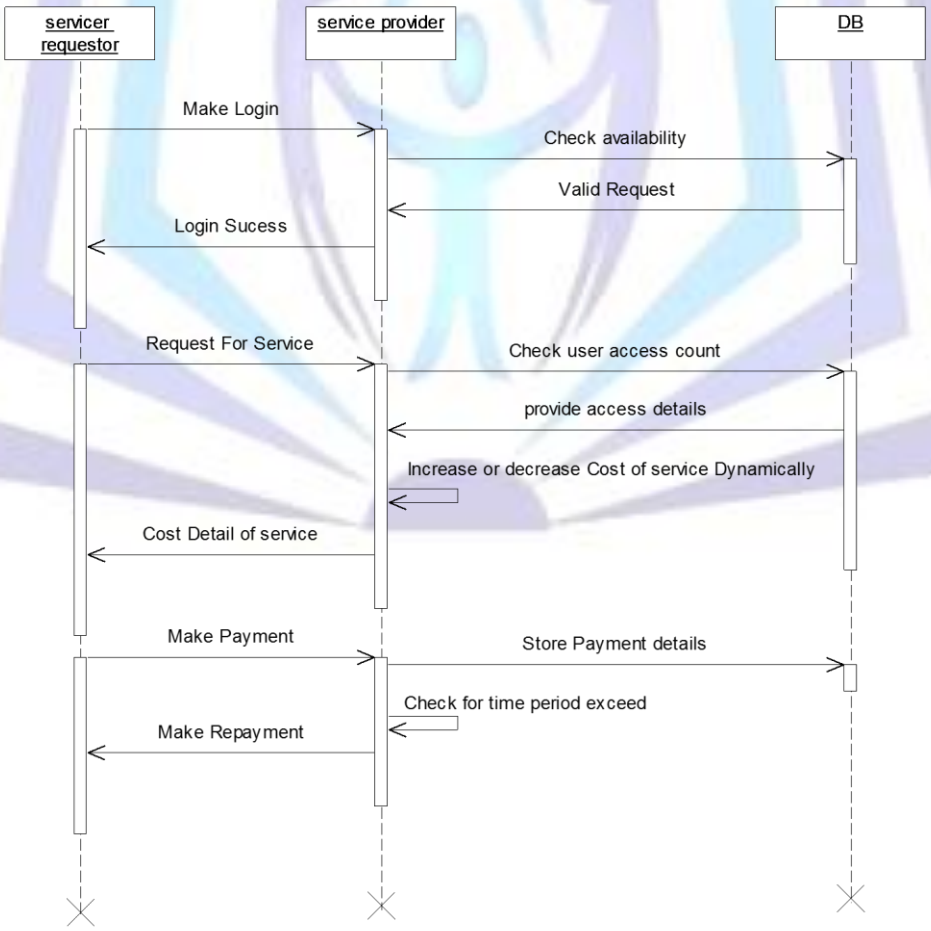


Fig 2: Cloud resource allocation interaction protocol

CONCLUSION

This paper presents an algorithm to meet design goals for dynamic resource allocation in large cloud. The main tasks which have been accomplished are: fairness allocation of resources with respect to sites, scalability of middleware layer in terms of number of machines as well as host in cloud and provisioning of the resources**.** So that it can be utilized by a large number of customers within cloud. This algorithm is also very beneficial for the customers since they have to pay as much as they use, the extra amount can be paid back to the customers after a specified period of time. The performance of this algorithm was evaluated through simulation. Hence it provides a more efficient and flexible way to dynamically allocate resources in cloud environment.

REFERENCES

[1] H. Goudarzi and M. Pedram, "Maximizing profit in the cloud computing system via resource allocation," *Int'l workshop on Data Center Performance,* Minneapolis, MN, Jun. 2011.

[2] C. Tang, M. Steinder, M. Spreitzer, and G. Pacifci. A scalable application placement controller for enterprise data centers. *Int'l Conf. on WWW*, 2007.

[3] S. Kumar, V. Talwar, V. Kumar, P. Ranganathan, K. Schwan.vManage: loosely coupled platform and virtualization management in data centers. *Int'l Conf. on Autonomic Computing*, 2009.

[4] L. Zhang and D. Ardagna. SLA based profit optimization in autonomic computing systems. *The Second Int. Conf. on Service Oriented Computing*, November 2004.

[5] Z. Liu, M. S. Squillante and J. L. Wolf. On maximizing servicelevel- agreement profits. *The Third ACM Conference on Electronic Commerce*, 2001.

[6] D. Ardagna, B. Panicucci, M. Trubian and L. Zhang. Energy- Aware Autonomic Resource Allocation in Multi-Tier Virtualized Environments. *IEEE Transactions on Services Computing*, 2010.

[7] B. Addis, D. Ardagna, B. Panicucci, Z. Li. Autonomic Management of Cloud Service Centers with Availability Guarantees. *IEEE 3rd International Conference on Cloud Computing*, July 2010.

[8] M. Mazzucco, D. Dyachuk, R. Deters. Maximizing Cloud Providers' Revenues via Energy Aware Allocation Policies. *IEEE 3rd International Conference on Cloud Computing*, July 2010.

[9] F. Chang, J. Ren, R. Viswanathan. Optimal Resource Allocation in Clouds. IEEE 3rd International Conference on Cloud Computing, July 2010.

[10] A. Chandra, W. Gongt and P. Shenoy. Dynamic resource allocation for shared clusters using online measurements. *ACM SIGMETRICS*, 2003.

[11] C. Santos, X. Zhu, and H. Crowder. A mathematical optimization approach for resource allocation in large scale clusters. *TechnicalReport HPL-2002-64, HP Labs*, March 2002.

[12] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat and R. P.Doyle. Managing energy and server resources in hosting centers.*ACM SOSP*, 2001.

Authors Biography

J. Antony Judi, Sr Systems Analyst. Accenture Private Limited, Over Ten years of experience in client facing roles carrying out effective test management, test planning, execution, reporting and software test automation management skilled in Data warehouse testing, Web services testing, testing in MIS and Banking domain.

F. Ezhil Mary Arasi, Asst. Professor in SRM University has been serving the Education Profession for the past 9+ years. Earlier she was working as lecturer for St. Joseph's College, Trichy for about 3 years. Currently, she carries out Research in Service Oriented Architecture. She has guided more than 35 students in completing their MCA Graduation Projects.

Dr. S. Govindarajan, a PhD in Information Technology (IT), earlier had served major software MNC as a General Manager. Currently he works for SRM University as a Professor for Computer Applications Department since 2009. He has published more than 10 journals, which include 3 International Publications.