An enhanced QoS Architecture based Framework for Ranking of Cloud Services

Mr.K.Saravanan [#1], M.Lakshmi Kantham [#2]
[1] Assistant Professor, [2] PG Scholar
Department of Computer Science and Engineering
Anna University Tirunelveli Region,
Tirunelveli – 627007
Tamilnadu

Abstract: **With the rapid growth of cloud computing, many organizations such as Amazon, IBM and HP started to offer cloud services to various consumers. From the customer's point of view, it is very difficult to choose which service is best one to use and what the criteria for their selection are. Determining the best cloud computing service for a specific application is a challenge and often determines the success of the underlying business of the service consumers. In some situations, due to the vast number of requests, the providers are not able to deliver the requested services within requested time. To avoid this scenario, advanced reservation scheme is proposed which provides the guaranteed delivery of resources. Currently there is no standard framework for ranking service for the customers to select the appropriate provider to fit their application and the advanced reservation mechanism which provides the customers to access their services at a right time. A novel framework for ranking and advanced reservation of cloud services is proposed which is based on a set of cloud computing specific performance and a Quality of Service (QoS) attributes. It provides an automatic best fit and a guaranteed delivery.**

Keywords: Inter cloud, Cloud Coordinator, Cloud Exchange, QoS, SLA, AHP.

1. INTRODUCTION

"A cloud is a type of parallel and distributed systems consisting of a collection of inter connected and virtualised computers that are dynamically provisioned and presented as one are more unified computing resources based on service level agreements established through negotiation between the service provider and customer." Cloud services can be acquired in different abstraction levels namely IaaS, SaaS and PaaS. This paper concerned with IaaS, Where the customer can lease the hardware resources. The hardware resources may typically but not restricted to virtualised resources.

The resources available in a single data centre are limited even though data centres may contain thousands of physical machines able to host tens of thousands of virtual machines when there is a large demand is observed.

A single data centre is unable to provide all the requested resources. In order to satisfy the customers need some of the SLA's may be altered. To avoid this scenario, we allow the applications by enabling it to scale across multiple independent cloud data centres following market based trading and negotiation of resources between providers and brokers.

Thus we concentrate on federated cloud, which consist of multiple clouds and a cloud exchange unit. Due to the economical advantage of cloud computing, many organisations are moving their existing applications and started building the new application on the cloud infrastructure. This switching to the cloud infrastructure provides the higher flexibility and higher scalability. The customers can access their services from anywhere in the world with the internet connection. And the cloud infrastructure is scalable, i.e., it can serve any number of requests at a time.

With the growth of these public cloud offerings, a strong confusion raised among the cloud customers to choose which provider can satisfy their QoS requirements. There are many providers' offers similar services at different cost with different features. For example, one provider may offer cheaper storage whereas renting powerful virtual machines from them are so expensive.

Enabling the ranking mechanism may help the customers to choose the provider whose service resulting best performance to their application. Choosing the appropriate provider for specific application is a challenge and often determines the success of the underlying business.

The rest of this paper is organised as related work, system overview, hierarchical structure of QoS attributes for ranking mechanism, formation of RSRM and RSRV, performance evaluation and comparative analysis, conclusions and future enhancements and references.

2. RELATED WORK

Federated cloud concepts were addressed in [5], [6], where scalability is increased by inter cloud negotiation i.e., any number of customers requests can be satisfied at a time. And the mapping function [6], is implemented by continuous double action, sensor unit was used to predict the geographic distribution of users. Advanced reservation strategies were encountered in [14], [15] which allow the customers to reserve the resources in a prior manner. But no co-reservation is allowed in [14]. Overlapped Advance Reservation Strategy (OARS) gives mitigation of negative effects brought about by advance reservation but the lower rejection rate at the price slightly increasing the violations of reservation [15]. Alternate offers protocol and broker's negotiation strategy were used for advanced reservation in [7]. But it supports only the negotiation for timeslots and number of resources. An algorithm to minimize total cost of resource provisioning and to avoid over-provisioning and under-provisioning is proposed using reserved instances which guarantee the resources to reserved users. But there is an underlying assumption is that there are always providers willing to sell call options [10]. SLA-oriented Dynamic Provisioning Algorithm supports integration of market based provisioning policies and virtualization technologies for flexible allocation of resources to applications. However it does not include customer-driven service management, computational risk management and autonomic management of Clouds which improve the system efficiency, minimization of SLA violation and the profitability of service providers [11]. Comparison of different cloud services can be obtained through Service Measurement Index (SMI) and Analytic Hierarchy Process (AHP). But the ranking algorithm proposed here cannot cope with variation in QoS attributes such as performance by Adopting fuzzy sets [8]. Singular Value Decomposition Technique (SVD) determines the best service provider for a user application with a specific set of requirements. It provides an automatic best-fit procedure which does not require a formal knowledge model. However there is no standard way to allow a universal description format and semantics [12].

3. SYSTEM OVERVIEW

The overall system relies on federated cloud, which consists of multiple clouds and a cloud exchange unit. Each cloud has a coordinator element, which collects the request from the users and checks it as well as then provide the requested resources. If the resource is not available within that cloud, the coordinator will submit the request to the cloud exchange unit. The providers who have an alternative resource will also submit their offers to the cloud exchange unit. Thus the cloud exchange unit have the collections of requests and offers from the various coordinators. This cloud exchange unit will publish all the requests and offers. Then the coordinators will choose the required offer/request and start negotiating with the other coordinators. The cloud exchange unit is also responsible for advance reservation when the user ask for advance reservation for particular resources for a particular time; it will check whether the resource will be available for the requested timeslot. If it is available, it will lock the reserved resources for the requested the time slot. The cloud exchange unit also provide ranking for resources. When there are multiple service providers, there will be confusion that which service they can use and what is the basis for their selection. To avoid this scenario ranking mechanism is included. The overall architectural diagram (Fig.1) and the major components description are depicted below:

3.1 Cloud coordinator element:

Each multi cloud contain a coordinator element which collects all the offers within that cloud and submit it to the central cloud exchange unit. Every coordinator can negotiate with all the other coordinators. Thus the main role of this coordinator element is to represent all the available service providers within each cloud to the market place.

3.2 Cloud exchange unit:

Cloud exchange unit acts as a central market place between the cloud customers and the service providers by collecting all the requests and offers. It generates a unique id for each and every requests and offers. Cloud exchange unit also responsible for dynamic provisioning of resources, advanced reservation and ranking of cloud services.

3.3 Dynamic provisioning:

The cloud exchange unit will provide mapping or all the submitted offers and request after issuing the unique id. Then the corresponding mapped coordinators can start their negotiation directly.

3.4 Advanced reservation:

Advanced reservation unit generate a unique reservation id for every reservation request. After that it will get the users choice of resources, reservation time and period. With those details the reservation unit will check the availability. If the requested resources are available for the requested time the unit will lock the reserved resources for the reserved period.

3.5 Ranking unit:

The ranking unit provide ranking of cloud services by forming hierarchical structure of the QoS attributes. The QoS attributes are computed and classified as top level, first level and second level. Then the relative weights for each attributes are assigned randomly. After that the RSRM and RSRV are calculated for each attributes. Finally all the second level RSRVs are aggregated to compute the RSRM of first level attributes and all the first level RSRVs aggregated to find the finalRSRM

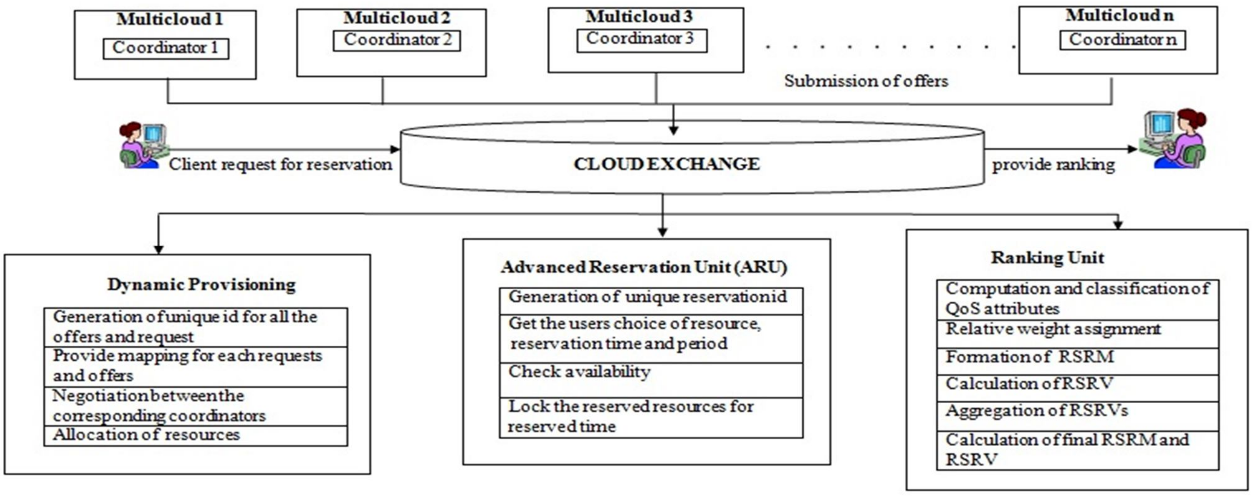

Fig. 1

4. HIERARCHICAL STRUCTURE OF QOS ATTRIBUTES FOR RANKING MECHANISM:

The hierarchical structure of QoS attributes provides the classification of QoS attributes needed by the customers for selecting the appropriate service providers based on: cost, performance, assurance, security, usability, agility, accountability.

4.1 Cost:

The main objective of the cloud computing is to minimise the cost spent on the hardware and infrastructure. Thus the cost plays the key role while deciding the service provider. Cost can be of two modes likely on demand cost and reservation cost.

4.1.1 On-Demand cost:

On-Demand cost can be collected only for the usage of resources. It does not include any minimum charges. And in this ranking scheme on demand cost can be calculated by VM cost, data cost and storage cost.

$$VM\ cost = Vm\ costs\ per\ hour \times no.of.Vms\ uptime\ in\ hours$$

$$Data\ cost = (data\ in\ (GB) \times data\ in\ rate) + (data\ out\ (GB) \times data\ out\ rate)$$

$$storage\ cost = storage\ rate \times storage\ amount\ (GB)$$

4.1.2 Reservation cost:

Reserving the resources may reduce the cost compared to on demand cost. But in includes onetime registration cost. Therefore reservation cost can be calculated by

$$reservation\ cost = usage\ fee + onetime\ fee$$

4.2 Performance:

Performance is the important factor for ranking of cloud services. The service's performance can be evaluated by service response time, accuracy, stability, interoperability and suitability.

4.2.1 Service response time:

To get the better performance, the service response time should be minimum that is how fast the service can be made available for usage. Service measurement can be measured by

$$service\ response\ time = \sum_i \frac{T_i}{n}$$

Where $T_i$the time between user's request and the response

4.2.3 Stability:

Stability is defined as the variability in the performance of a service. For storage, it is the variance in average read and write time and for computational resource it as defined as

TOP LEVEL ATTRIBUTES	WEIGHTS	1st LEVEL ATTRIBUTES	WEIGHTS	2nd LEVEL ATTRIBUTES	WEIGHTS
COST	0.3	On Going Cost	1	VM cost	0.6
				Data	0.2
				Storage	0.2
PERFORMANCE	0.25	Service Response Time	0.3	Range	0.5
				average value	0.5
		Accuracy	0.2	Degree of proximity	1
		Stability	0.2	Storage	0.6
				computational resource	0.4
		Interoperability	0.2	user experience	1
		Suitability	0.1	essential requirements	0.6
				non-essential requirements	0.4
ASSURANCE	0.1	Availability	0.5	% of access time	1
		Service Stability	0.2	CPU	0.4
				Memory	0.3
				upload time	0.3
		Serviceability	0.2	free support	0.7
				type of support	0.3
		Reliability	0.1	chance of failure	1
SECURITY	0.1	Confidentiality	0.4	--------	--------
		Data Integrity	0.3	--------	--------
		Privacy	0.3	--------	--------
USABILITY	0.1	Transparency	0.4	Frequency of changes in the services	1
		Access ability	0.3	Time	1
		Install ability	0.2	Time	1
		Learn ability	0.1	Time	1
AGILITY	0.1	Capacity	0.5	CPU	0.5
				Memory	0.3
				Disk	0.2
		Elasticity	0.2	Time	1
		Adaptability	0.1	Time	1
		Flexibility	0.1	--------	--------
		Portability	0.1	--------	--------
ACCOUNTABILITY	0.05	Sustainability	0.4	DCiE and PUE	0.5
				DPPE	0.5
		Audit ability	0.3	% of accuracy	1
		Compliance	0.3	------------	--------

$$stability\ of\ a\ computational\ resource = \frac{\frac{\alpha_{avg,i} - \alpha_{SLA,i}}{t}}{n}$$

Where α can be computational unit, network unit or storage unit of the resource; $\alpha_{avg,i}$ is the observed average performance of the user i who leased the cloud service; $\alpha_{SLA,i}$ is the promised value in the SLA; t is the service time; and n is the total number of users.

4.2.2 Accuracy:

Accuracy of the service can be defined as the degree of proximity to the user's actual values to the expected values.

$$accuracy = \sum i\ \frac{fi}{n}$$

Where f_i = No.of.times provider fails to satisfy the promised value for user I over time t.

4.2.3 Stability:

Stability is defined as the variability in the performance of a service. For storage, it is the variance in average read and write time and for computational resource it as defined as

$$stability\ of\ a\ computational\ resource = \frac{\frac{\alpha_{avg,i} - \alpha_{SLA,i}}{t}}{n}$$

Where α can be computational unit, network unit or storage unit of the resource; $\alpha_{avg,i}$ is the observed average performance of the user i who leased the cloud service; $\alpha_{SLA,i}$ is the promised value in the SLA; t is the service time; and n is the total number of users.

4.2.4 Interoperability:

Interoperability is the ability of a service to interact with other services offered by same provider or other providers.

$$interoperabilty = \frac{No.of.platforms\ offered\ by\ the\ provider}{No.of.platforms\ required\ by\ users\ for\ interoperability}$$

4.2.5 Suitability:

It is defined as the degree to which the customer's requirements are met by a cloud provider

$$suitability = \frac{No.of.non-essential\ features\ provided\ by\ service}{no.of.non-essential\ features\ required\ by\ the\ customer}$$

4.3 Assurance:

Assurance indicates the likelihood of the service performance as expected or promised in the SLA. The cloud customers will choose the services only If the providers guarantees the assurance. Mostly the customers will expect better availability, service stability, service ability and reliability.

4.3.1 Availability:

It is defined as the percentage of time a customer can access the service

$$availability = \frac{(total\ service\ time)-(total\ time\ for\ which\ service\ was\ not\ available)}{total\ service\ time}$$

4.3.2 Reliability:

Reliability can be defined as how a service operates without failure during a given time and condition.

$$reliability = probability\ of\ violation * p_{mttf}$$

$$= \frac{1-numfailure}{n} * p_{mttf}$$

Where p_{mttf} is the promised mean time to failure.

4.4 Security:

Security is the major criteria for every customer. Hosting data under a third party is always a critical issue which require a stringent security policies employed by the cloud service providers. Security includes confidentiality, data integrity and privacy.

4.4.1 Confidentiality:

Confidentiality is an integral component of security. It ensures that the information stored on the cloud is protected against the unintended or unauthorized access. It is also defined as the percentage of authorized access.

$$confidentiality = \frac{total\ no.of.access\ to\ the\ service - no.of.unauthorized\ access\ to\ the\ service}{total\ no.of.access\ to\ the\ service}$$

4.4.2 Data integrity:

Data integrity is the assurance that the information kept on the cloud can only be accessed and modified by the authorized one. It is also defined as the accuracy and consistency of the data

$$data\ integrity = \frac{percentage\ of\ accuracy\ after\ modification}{percentage\ of\ acuuracy\ before\ modification}$$

4.4.3 Privacy:

Privacy is defined as the state or condition of being free from being observed or disturbed by others. That is the information kept on the cloud should not be disturbed by others.

$$privacy = \frac{no.of.third\ party\ access}{toal\ no.of.access}$$

4.5 Usability:

The ease of using a cloud service is defined by the attributes of usability. Usability of the service can be measured by the average time taken by the customers to learn, install and understand it.

4.5.1 Learnability:

It is the capability of the service to enable the customers to learn how to use the service

$$learnability = \sum i \frac{no.of.users\ using\ the\ service\ i}{total\ no.of.users}$$

4.5.2 Installability:

Installability can be defined as the capability of the service to be installed in a specific environment.

$$installability = \frac{no.of.platforms\ supported\ by\ the\ service}{total\ no.of.platforms\ available}$$

4.5.3 Easiness:

It is the capability of the customers of being understood. That is how much the customers are aware about the service functionality. That is how easy to understand

$$easiness = \frac{total\ no.of.services\ used\ by\ the\ customers}{total\ no.of\ services\ offered\ by\ the\ providers}$$

4.6 Agility:

Agility is the most important advantage of cloud computing. That is agility is the ability of a service to expand and change quickly without much expenditure. Agility is calculated by the various sub- factors like elasticity, adaptability, flexibility and portability.

4.6.1 Elasticity:

Elasticity is defined as scalability that is how much a cloud service can be scaled.

$$elasticity = \frac{mean\ time\ taken\ to\ expand\ or\ contract\ the\ service\ capacity}{maximum\ capacity\ of\ the\ service}$$

4.6.2 Adaptability:

Adaptability is the ability of the service provider to adjust changes in the service based on the customer request.

$$adaptability = time\ taken\ to\ adapt\ the\ changes\ or\ upgrading\ the\ service\ to\ the\ higher\ level$$

4.6.3 Flexibility:

Flexibility can be defined as the capability of being changed or adapted repeatedly without much expenditure.

$$flexibility = \frac{adaptability\ of\ the\ service}{maximium\ capacity\ of\ the\ service}$$

4.6.4 Portability:

Portability can be defined as the capability of running over two or more services $portability = \frac{no.of.services\ used}{total\ no.of.services}$

Portability can be defined as the capability of running over two or more services $portability = \frac{no.of.services\ used}{total\ no.of.services}$

4.7 Accountability:

This characteristic is used to create a trust of a customer on a cloud service provider. Accountability includes sustainability, audit ability and compliance.

4.7.1 Sustainability:

Sustainability is defined in terms of the environmental impact of the cloud service used. It is measured by the energy efficiency of the cloud service

4.7.2 Auditability:

It is the ability to achieve the accurate results. Thus audit ability can be defined as the percentage of getting expected results.

$$auditability = \frac{obtained\ resultant\ value}{expected\ resultant\ valuue} \times 100$$

4.7.3 Compliance:

It is the metric that ensures the rules regulations of data management.

$$compliance = \frac{no.of.rules\ and\ regulations\ met}{total\ no.of.rules\ and\ regulations}$$

5. FORMATION OF (RSRM) AND (RSRV):

The above attributes are classified as top level, first level and second level to form the hierarchical structure. For each level define the corresponding first level and second level attributes and Compute the values of all the attributes. Assign the relative weights for each attributes randomly. For each second level attributes compute the RSRM by comparing the values of one provider with others. Then calculate the RSRV by multiplying the RSRM with their corresponding weights. Then these RSRVs are combined to get the RSRM of first level attributes. Finally all the first level RSRVs are aggregated to find the final RSRM and RSRV.

6. EXPERIMENTAL STUDY:

Here the multi cloud environment is created with four separate clouds. First step is to get the resource availability from each cloud which is showed in figure 3(a). Then the availability status of each cloud is sent to their cloud coordinator which is represented by figure 3(b). And every coordinator will send the resource availability to the central cloud exchange unit and this has been depicted in figure 3(c). The central cloud exchange unit also collects the customer's requests for dynamic provisioning, advanced reservation which is illustrated in figure 3(d). Every reservation requests will be stored in the database with the reservation id , reservation period etc.

Fig. 3(a)

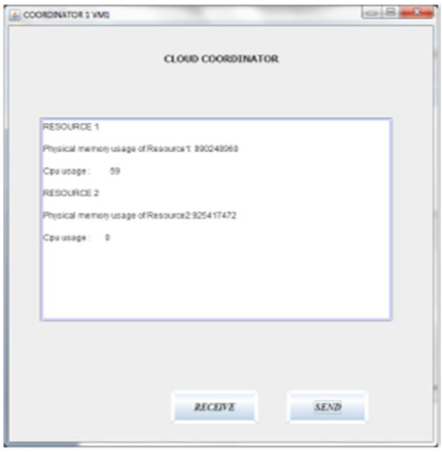

Fig. 3(b)

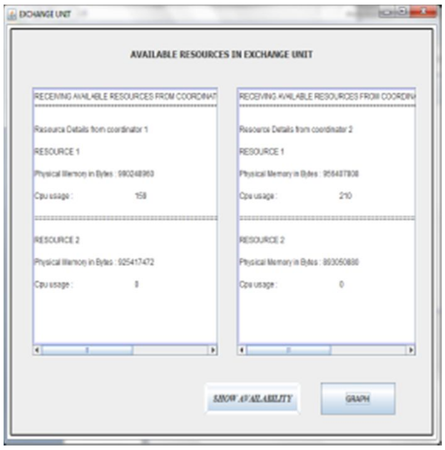

Fig. 3(c)

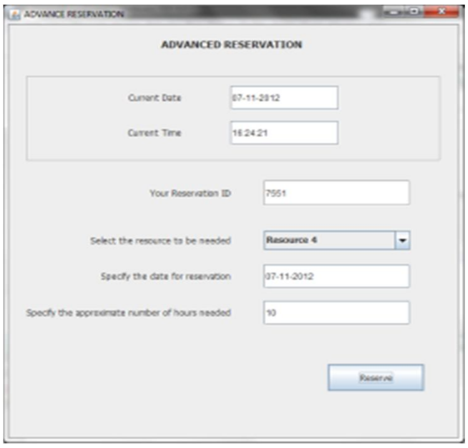

Fig. 3(d)

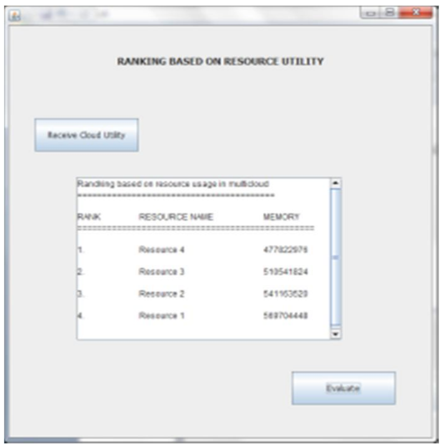

Fig. 3(e)

7. COMPARATIVE ANALYSIS:

The providers can expand their resource availability when the number of customers increases in case of federated cloud where as in ordinary cloud the resource unavailability may occur, which is clearly compared in figure 4.1 When the QoS is increasing, the number of SLA violations is decreased, which is clearly illustrated in figure 4.2. And while using the advanced reservation scheme, the provider can assure the guaranteed delivery of resources, where as in on demand access the provider cannot assure. This scenario is evidently produced in figure 4.3. Then finally the cost is compared with on-demand access and advanced reservation mechanism. The on-demand access does not involve any one time registration fee it includes only usage cost. But in case of advanced reservation scheme the customer have to pay some amount as one time registration fee as well as the usage cost. Which are presented in the figure 4.4 and 4.5. But in the advanced reservation scheme the usage cost is lower than the usage cost in on-demand access. Thus the advanced reservation scheme reduces the cost as well as provides guaranteed delivery.

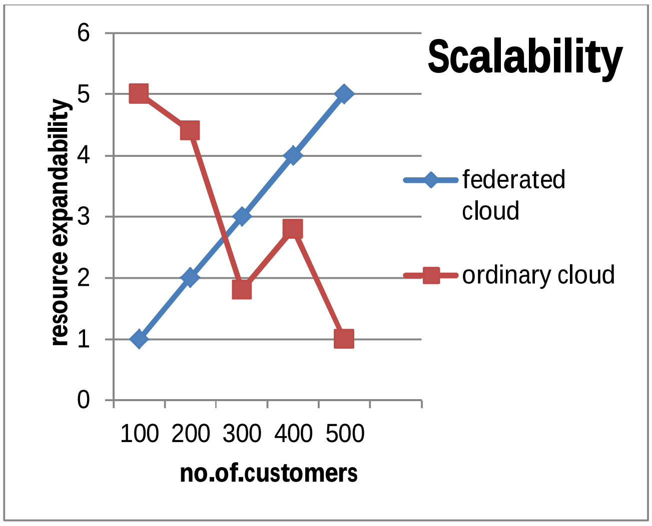

Fig. 4.1

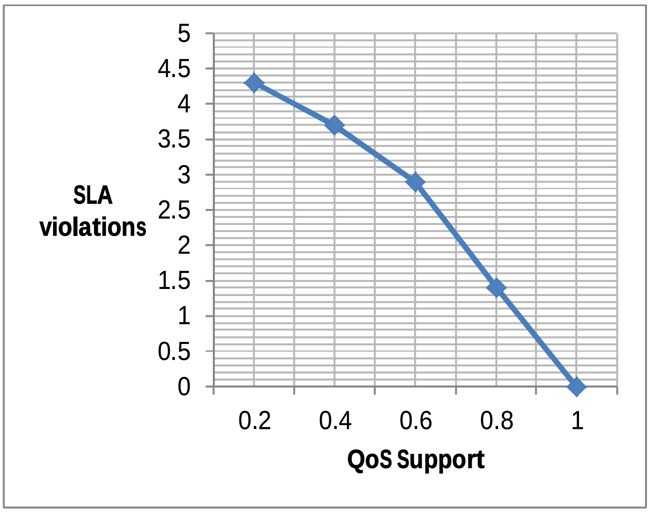

Fig. 4.2

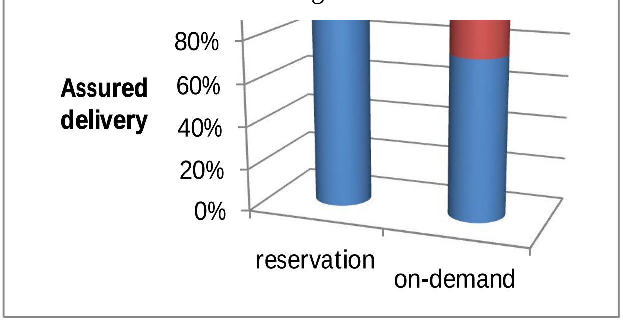

Fig. 4.3

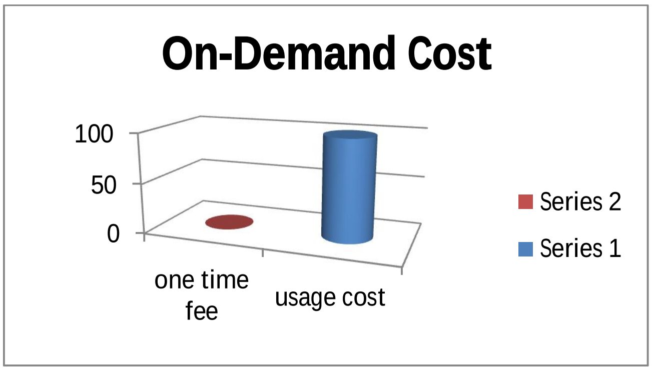

Fig. 4.4

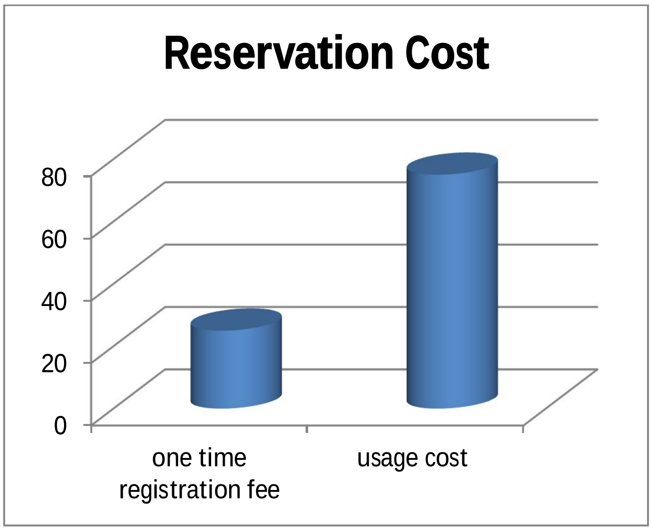

Fig. 4.5

8. CONCLUSION AND FUTURE WORK:

This novel framework acts as an improvement over existing methods already being employed. The main advantage of this model is, it increases the scalability and provides ranking, advanced reservation schemes by which the customer can access the right resources at right time without fail. It helps the cloud customers to choose the best service provider who satisfy their QoS requirements. The user who wants to choose the providers based on the specific requirements can also use this framework by neglecting all the unnecessary attributes or by increasing the weight of their required attributes.

9. REFERENCES

[1]Saurabh Kumar Garg, Steve Versteeg and Rajkumar Buyya, "A framework for ranking of cloud computing services", (2012) S.K. Garg et al. / Future Generation Computer Systems.

[2]Zibin zheng, Yelei zhang and Michael R.lyu, "Cloud Rank: A QoS driven component framework ranking for cloud computing", 2010 29th IEEE International Symposium on Reliable Distributed Systems.

[3] Zibin Zheng, Xinmaio Wu, Yilei Zhang, Michael R. Lyu and Jianmin Wang, "QoS Ranking Prediction for Cloud Services", (2012) IEEE Transactions On Parallel And Distributed Systems.

[4] Ryan K L Ko, Bu Sung Lee and T. Muthu Nesa Beula et al., "Towards Achieving Accountability, Auditability and Trust in Cloud Computing", Siani Pearson (2011) / International Journal of Engineering Science and Technology (IJEST).

[5]Rodrigo N. Calheiros, Adel NadjaranToosi, Christian Vecchiola and Rajkumar Buyya, "A coordinator for scaling elastic applications across multiple clouds", Future Generation Computer Systems, Volume 28, No. 8, Pages: 1350-1362, ISSN: 0167-739X, Elsevier Science, Amsterdam, The Netherlands, October 2012.

[6]R.Buyya, R.Ranjan and R.N.Calheiros, "InterCloud: Utility-Oriented Federation of Cloud Computing Environments for Scaling of Application Services", in: Proceedings of the 10th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP'10, Springer, Busan, South Korea, 2010, pp. 13–31.

[7]S. Venugopal, X. Chu, and R. Buyya, "A Negotiation Mechanism for Advance Resource Reservation using the Alternate Offers Protocol", in: Proceedings of the 16th International Workshop on Quality of Service, IWQoS'08,
IEEE Computer Society, Enschede, Netherlands, 2008, pp. 40–49.

[8]Saurabh Kumar Garg, Steve Versteeg and Rajkumar Buyya, "SMICloud: A Framework for Comparing and Ranking Cloud Services", Proceedings of the 4th IEEE/ACM International Conference on Utility and Cloud Computing (UCC 2011, IEEE CS Press, USA), Melbourne, Australia, December 5-7, 2011.

[9]S.K. Garg, C. Vecchiola and R. Buyya, "Mandi: a market exchange for trading utility and cloud computing services", The Journal of Supercomputing (2011) http://dx.doi.org/10.1007/s11227-011-0568-6.

[10]Adel NadjaranToosi, Ruppa K. Thulasiram and Rajkumar Buyya, "Financial Option Market Model for Federated Cloud Environments", Technical Report CLOUDS-TR-2012-1, Cloud Computing and Distributed Systems Laboratory, The University of Melbourne, March 10, 2012.

[11]Rajkumar Buyya, Saurabh Kumar Garg, and Rodrigo N. Calheiros, "SLA-Oriented Resource Provisioning for Cloud Computing: Challenges, Architecture, and Solutions", Proceedings of the 2011 IEEE International Conference on Cloud and Service Computing (CSC 2011, IEEE Press, USA), Hong Kong, China, December 12-14, 2011.

[12]HoiChan and Thomas J.Watson, "Ranking and Mapping of Applications to Cloud Computing Services by SVD", Res.Center, IBM, Hawthorne, NY, USA Trieu Chieu Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP **Page(s):** 362 -369 Conference Publications.

[13]Bahman Javadi, Parimala Thulasiraman and Rajkumar Buyya, "Cloud Resource Provisioning to Extend the Capacity of Local Resources in the Presence of Failures", HPCC-183The 14th IEEE International Conference on High Performance Computing and Communications (HPCC-2012).

[14]Rui min and muthucumaru maheswaran, "Scheduling advance reservations with priorities In grid computing systems", PDCS 2001 - Parallel and Distributed Computing and Systems ISSN: 1027-2658; ISBN: 0-88986-307-5.

[15]Zhe Huang, Peng Xiao and Dongbo Liu, "An Overlapped Advance Reservation Strategy for Grid Resources", Journal of Information & Computational Science 9: 8 (2012) 2211–2220.