Efficient Auto-Scaling Mechanism in the Cloud Environment Using Proactive Approach

Aneri Parekh[1], Prof. Gayatri S. Pandi (Jain)[2].

[1]*Student (Master of Engineering), Computer Engineering, L.J. Institute of Engineering and Technology , Gujarat, India.*

[2] *H.O.D.PG Departments, Computer Engineering, L.J. Institute of Engineering and Technology , Gujarat, India.*

ABSTRACT

Cloud computing is latest emerging technology for large scale distributed computing and parallel computing. Cloud computing gives large pool of shared resources, information, packets at any instances of time. Auto-scaling is the strategy that has ability to adjust the available resources to meet the user demands. To facilitate users with availability of resources seamlessly. Auto scaling is cloud computing feature that allows users to automatically scale cloud services, like virtual machine and server capacities, up or down, depending on user on-demand. Proactive auto-scaling mechanisms predict the workload ahead such that the auto-scaler can make decision based on the expected workload instead of waiting for trigger. Proactive auto-scaling mechanism is efficient then reactive auto-scaling because in reactive auto-scaling approach, the auto-scaling decision would be triggered by a predefined set of events. In current cloud computing environment, management of data reliability has become a challenge. For data-intensive scientific applications, storing data in the cloud with the typical 3-replica replication strategy for managing the data reliability would incur huge storage cost. In this paper we work on machine learning base effective approach for auto-scaling mechanism in these systems we work on design effective approach for automatically scale the capacity of cloud. We also consider machine learning concept for selecting appropriate node and scale according. Work on parameters like QOS, Accuracy, efficiency etc.

Keyword:- *Cloud computing, Proactive Auto-scaling, Auto-scaling..*

1. INTRODUCTION

Cloud computing offers small and big organizations, the opportunity to scale their computing resources. It is done by either increasing or decreasing the required resources. Cloud computing provides delivery of resources on demand over the internet. It also provides the users to store and access the data stored by them on cloud. It provides metered service, so that users are asked to pay only for what they use. Cloud provides elasticity by scaling up as computing needs increase and then scaling down again as demands decrease. Auto-scaling is the strategy that has ability to adjust the available resources to meet the user demands. To facilitate users with availability of resources seamlessly. Auto scaling is cloud computing feature that allows users to automatically scale cloud services, like virtual machine and server capacities, up or down, depending on user on-demand. Proactive auto-scaling mechanisms predict the workload ahead such that the auto-scaler can make decision based on the expected workload instead of waiting for trigger.

2. WORKING TECHNOLOGY

2.1 BASIC CONCEPT

Cloud computing services fall into three broad categories:

1. Infrastructure as a service (IaaS)
2. Platform as a service (PaaS)
3. Software as a service (SaaS).

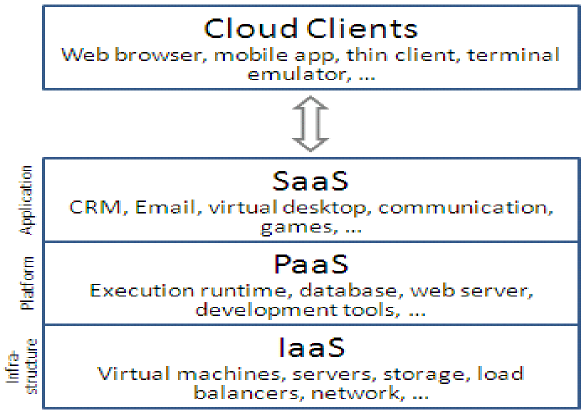
Cloud Clients
Web browser, mobile app, thin client, terminal emulator, ...
SaaS
CRM, Email, virtual desktop, communication, games, ...
PaaS
Execution runtime, database, web server, development tools, ...
IaaS
Virtual machines, servers, storage, load balancers, network, ...
Application
Platform
Infra-structure

Fig. 1 – Cloud Services[13]

2.2.1 Infrastructure-as-a-service (IaaS)

IaaS allows us to rent IT infrastructure—servers and virtual machines (VMs), storage, networks, operating systems—from a cloud provider on a pay-as-you-go basis. It enables companies to deliver applications more efficiently by removing the complexities involved with managing their own infrastructure. IaaS enables fast deployment of applications, and improves the agility of IT services by instantly adding computing processing power and storage capacity when needed.

2.2.2 Platform as a service (PaaS)

Platform-as-a-service (PaaS) refers to cloud computing services that supply an on- demand environment for developing, testing, delivering and managing software applications. PaaS is designed to make it easier for developers to quickly create web or mobile apps, without worrying about setting up or managing the underlying infrastructure of servers, storage, network and databases needed for development.

2.2.3 Software as a Service (SaaS):

Software-as-a-service (SaaS) is a method for delivering software applications over the Internet, on demand and typically on a subscription basis. With SaaS, cloud providers host and manage the software application and underlying infrastructure and handle any maintenance, like software upgrades and security patching. With SaaS, vendor makes the required software available to a business on subscription basis.

IJARIIE

3 Characteristics of Cloud

1. On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

2. Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

3. Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

4. Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

5. Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

4 Benefits:

The following are some of the possible benefit's for those who offer cloud computing based services and applications:

1. **Unlimited storage capacity** — storing information in the cloud gives you almost unlimited storage capacity. Hence, you no more need to worry about running out of storage space or increasing your current storage space availability.
2. **Backup and Disaster Recovery** — since all your data is stored in the cloud, backing it up and restoring the same is relatively much easier than storing the same on a physical device. Furthermore, most cloud service providers are usually competent enough to handle recovery of information.
3. **Scalability/Flexibility** — Companies can start with a small deployment and grow to a large deployment fairly rapidly, and then scale back if necessary. Also, the flexibility of cloud computing allows companies to use extra resources at peak times, enabling them to satisfy consumer demands.
4. **Cost Savings** — Companies can reduce their capital expenditures and use operational expenditures for increasing their computing capabilities. This is a lower barrier to entry and also requires fewer in-house IT resources to provide system support.
5. **Reliability** — Services using multiple redundant sites can support business continuity and disaster recovery.
6. **Maintenance** — Cloud service providers do the system maintenance, and access is through APIs that do not require application installations onto PCs, thus further reducing maintenance requirements.
7. **Mobile Accessible** — Mobile workers have increased productivity due to systems accessible in an infrastructure available from anywhere.

5 Auto-scaling

Auto-Scaling is cloud computing feature that allow user to automatic scale cloud services, like Virtual Machine (VM) and server capacities, Up or Down, depending on defining situation. It is closely related to and builds upon, the idea of load balancing.

Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. We can create collections of EC2 instances, called *Auto Scaling groups*. We can specify the minimum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes below this size.

If you specify the desired capacity, either when you create the group or at any time thereafter, Auto Scaling ensures that your group has this many instances. If you specify scaling policies, then Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

6. RELATED WORKS

6.1 LITERATURE REVIEW

6.1.1 Proactive Scalability and Management of Resources in Hybrid Clouds via Machine Learning

In Proactive Scalability and Management of Resources in Hybrid Clouds via Machine Learning[1], Goal is to present a novel framework for supporting the management and optimization of application subject to software anomalies and deployed on large scale cloud architectures, composed of different geographically distributed cloud regions. The framework uses machine learning models for predicting failures caused by accumulation of anomalies. It introduces a novel workload balancing approach and a proactive system scale up/scale down technique.

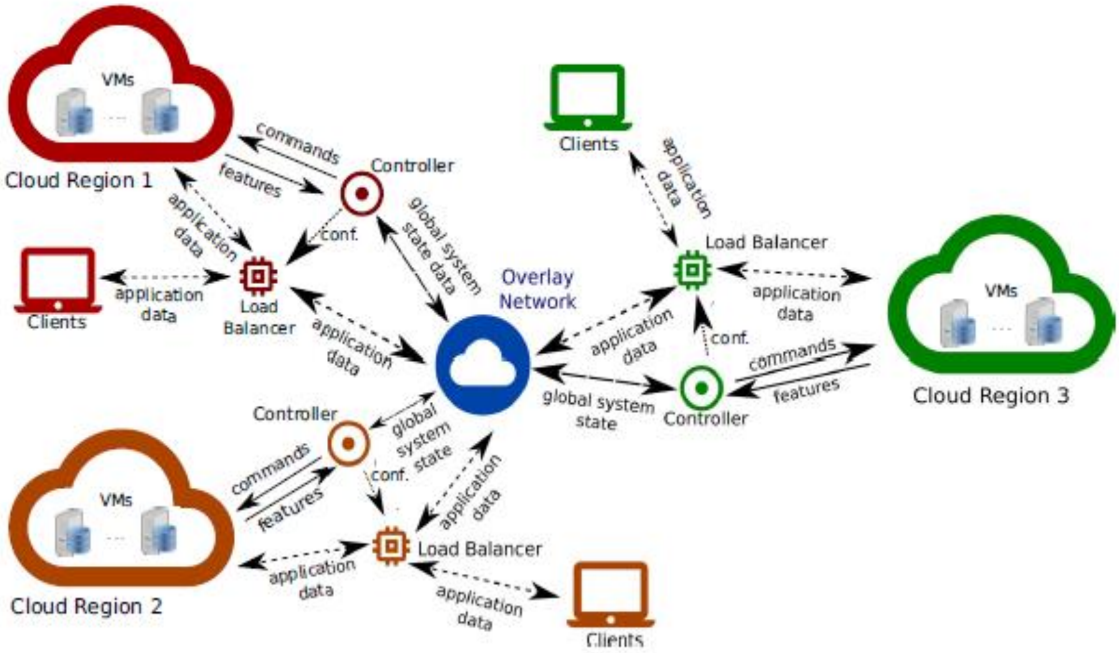

Fig. 2 - Global System Architecture with 3 Cloud Regions [1]

6.1.2 Proactive Memory Scaling of Virtualized Application

In Proactive Memory Scaling of Virtualized Application [2], Goal is to Proactive Approach is based on a control loop which proactively adds or removes memory resources to VMs match its future workload demand and to improve application availability and performance. It enables to plan reconfiguration in advance and schedule it to execute during phase of low application load (e.g., at night).This has following benefits: a) reconfiguration failures at OS level are avoided b) if application restart is required, the impact on performance and availability can be significantly reduced.

TABLE 1 COMPARISON OF CONTROLLERS

	No control	Reactive	Proactive
Mean response time	7,567 ms	1,211 ms	52 ms
Timeouts	84	285	0
Errors	8493	1485	337
Time of reduced availability	176 min	33 min	4 min

6.1.3 Instance Type Selection in Proactive Horizontal Auto-Scaling

In Instance Type Selection in Proactive Horizontal Auto-Scaling, The goal is to predict the future scaling actions for providing application resources in advance using proactive auto-scaling system. Exact quantity of VMs that one needs to run is determined according to predicted future demand across all resource dimensions, instance type used to execute application. Instance type characterizes VM in terms of its resource capacities (CPU, memory, disk, etc.). The user is charged previously agreed usage fee for each interval of time of length smaller or equal to minimal accountable usage interval over which instance is allocated, that consists in billing cycle practiced by IaaS provider. Auto-scaling service aims at periodically triggering infrastructure capacity planning (number and type of VMs) and provisioning actions needed to acceptable workload fluctuations experienced by application.

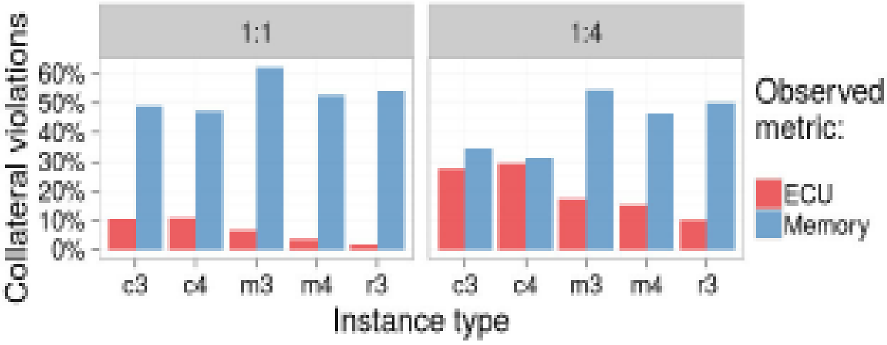

Fig. 3 SLO violations when a single resource is considered [3]

6.1.4 Automatic Resource Provisioning: A Machine Learning Based Proactive Approach

In Automatic Resource Provisioning: A Machine Learning Based Proactive Approach[4] , Paper concerns dynamic provisioning of cloud resources performed by an intermediary enterprise that provides a private cloud (also referred to as a virtual private cloud) for a single client enterprise using resources acquired on demand from a public cloud. A new proactive technique for auto-scaling of resources that changes the number of resources for the private cloud dynamically based on system load is proposed. The technique that supports both on-demand and advance reservation requests uses machine learning to predict future workload based on past workload.

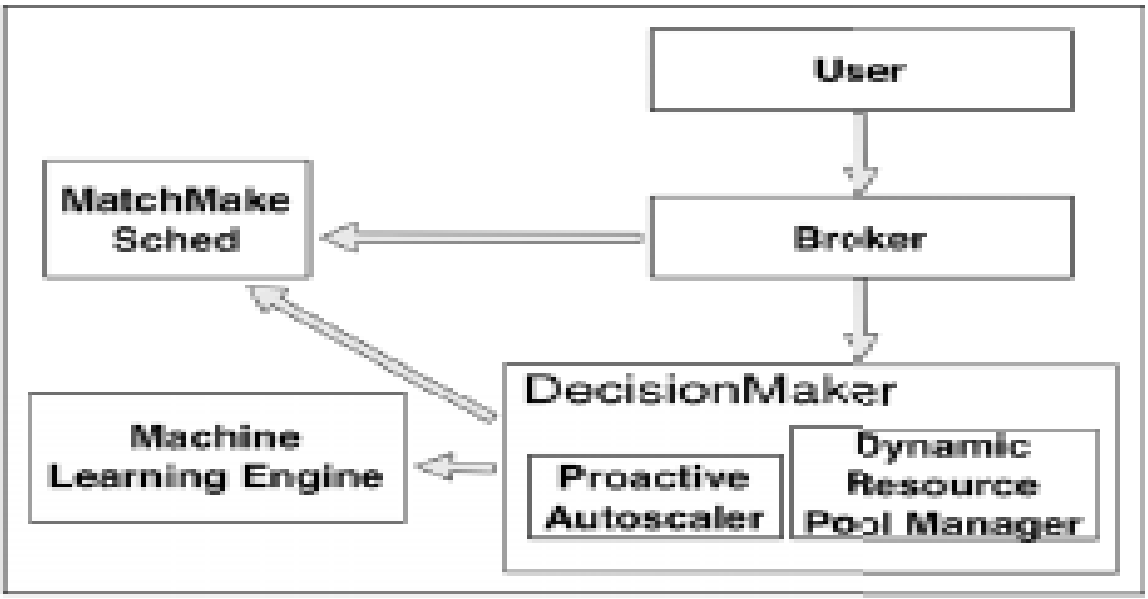

Fig. 4 System Architecture [4]

6.1.5 Predictive Auto-Scaling Techniques For Clouds Subjected To Requests With Service Level Agreements

In Predictive Auto-Scaling Techniques For Clouds Subjected To Requests With Service Level Agreements, Paper focuses on automatic provisioning of cloud resources performed by an intermediary enterprise that provides a virtual private cloud for a single client enterprise by using resources from a public cloud. And auto-scaling techniques for dynamically controlling the number of resources used by the client enterprise. To focus on proactive auto-scaling that is based on predictions of future workload based on the past workload. The primary goal of the auto-scaling techniques is to achieve a profit for the intermediary enterprise while maintaining a desired grade of service for the client enterprise. The technique supports both on demand requests and requests with service level agreements (SLAs).

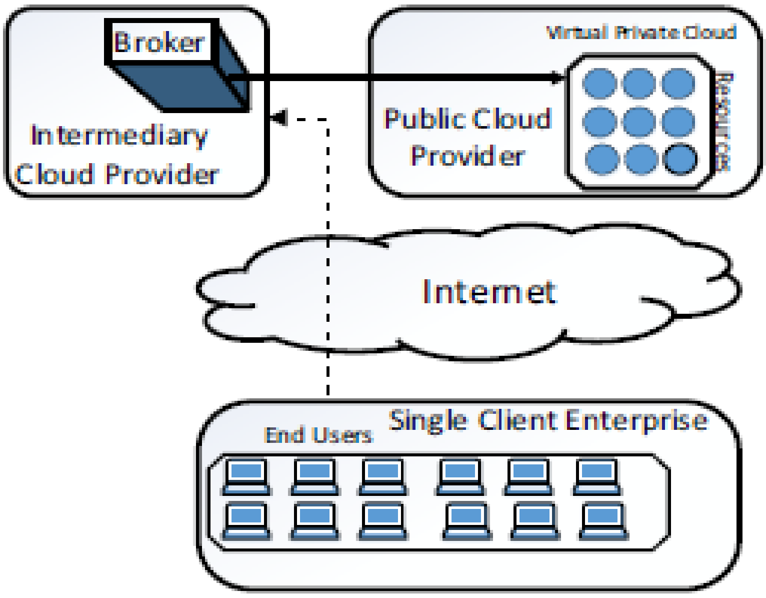

Fig. 5 System Overview [5]

6.2 COMPARATIVE TABLE

Table -2: Comparative Table

Sr. No.	Paper Title	Publish Year	Method	Advantages	Future Scope
1	Proactive Scalability and Management Of Resources in Hybrid Clouds Via Machine Learning[1]	2015	I-ACMF(Inter Autonomic Cloud Manager) Architecture, MTTF Prediction Model , Election algorithm, load balancing approach.	I-ACMF has been able to cope with different incoming client request rates in different cloud regions, with variations due to scale up/ scale down of system.	Plan to address issues related to communication over geographical scale in I-ACMF. & plan to use an overlay network to support reliable and efficient communication between cloud regions.
2	Proactive Memory Scaling Of Virtualized Application[2]	2015	splitTs method, Proactive approach.	- splitTs significantly improves the forecasting accuracy. -proactive approach can reduce impact of reconfigurations on application availability and performance of the application by more than 80% compared to reactive controller.	-extend the descriptive modeling capabilities for capturing relevant meta-information to the forecasting accuracy. -extend the set of experiments as soon as a standardized benchmark methodology for evaluating resource management mechanisms.
3	Instance Type Selection In Proactive Horizontal Auto-Scaling[3]	2016	Optimal instance type selection	- Cost savings at the expenses of a small number of SLO (service level objective) violations. -Reduce provisioning cost. -Multiple dimensions is essential to keep SLO violations low and reduce the occurrence of SLO violations.	- proactive solution to improve predictions accuracy by exploring the use of different predictor models. Evaluate our proactive auto-scaling solution by carrying out measurement experiments in a real IaaS environment.

Sr. No.	Paper Title	Publish Year	Method	Advantages	Future Scope
4	Automatic Resource Provisioning: A Machine Learning Based Proactive Approach[4]	2014	Proactive approach, Deadlineline driven algorithm, Support vector machine algorithm, Linear Regression algorithm, System architecture.	- Reduction of cost for the client. - Effectively lead to profit for intermediary enterprise. -Reduce cost for capacity planning	- Performance Analysis Of Proposed System using real workload traces. - Application of online training to a system subject to such workload traces. - Extending the autoscaling technique to storage and network resources warrants
5	Predictive Auto-Scaling Techniques For Clouds Subjected To Requests With Service Level Agreements[5]	2015	Proactive algorithm, support vector machine algorithm, Linear Regression algorithm, system architecture, deadlne driven algorithm.	- Reduction of cost for the client. - Effectively lead to profit for intermediary enterprise.	- Adaptation of these techniques to auto-scaling on clouds. - Performance analysis of proposed system using real workload traces & hybrid system that functions as System II during training period (accepting all the client request) and switches to the proactive system (System I) at end of training period.

7. Proposed Work:

In proposed algorithm we use machine learning algorithm to predict more accurately future workload and Proactive Approach to predict the workload ahead such that the auto-scaler can make decision based on the expected workload instead of waiting for trigger.To increase the Accuracy to Predict the future workload and reduction of cost for the client, effectively lead to Profit for intermediary enterprise.

4.2 Proposed Solution

Step-1 Cloud user will request for resources.

Step-2 Broker Management will handle the request of user and request will pass out to request handler Request handler will firstly inquire to matchMakeScheduler that if the resources are available or not. If resources are available, it places the request.

Step-3 Once a request enters MMS, a matchmaking algorithm determines a resource on which the request can be executed. A scheduling algorithm determines the order in which the requests allocated on a given resource are executed.

Step-4 Based on the information collected by MMS, Decision Maker (DM) which manages resource provisioning. DM, which responsible auto-scaling, DM implements a proactive approach for auto-scaling resources by using a helper module. The helper module, known as MLE, uses a machine learning algorithm to predict the future workload.

Step-5 Auto-Scaler work that, DM(decision maker) requests for the next k requests to arrive from MLE(Machine Learning Engine), demoted by Rk. DM simulates the resource management operations for these predicted future requests using MMS(Matchmakeschedule). Based on the output of MMS, DM decides whether to acquire new resources or change the stop time for existing resources After receiving the characteristics of the k predicted requests from MLE, DM invokes a simulation of the MMS(Matchmakeschedule).

Step-6 Then this decision is passed to Cloud. And Store in Cloud.

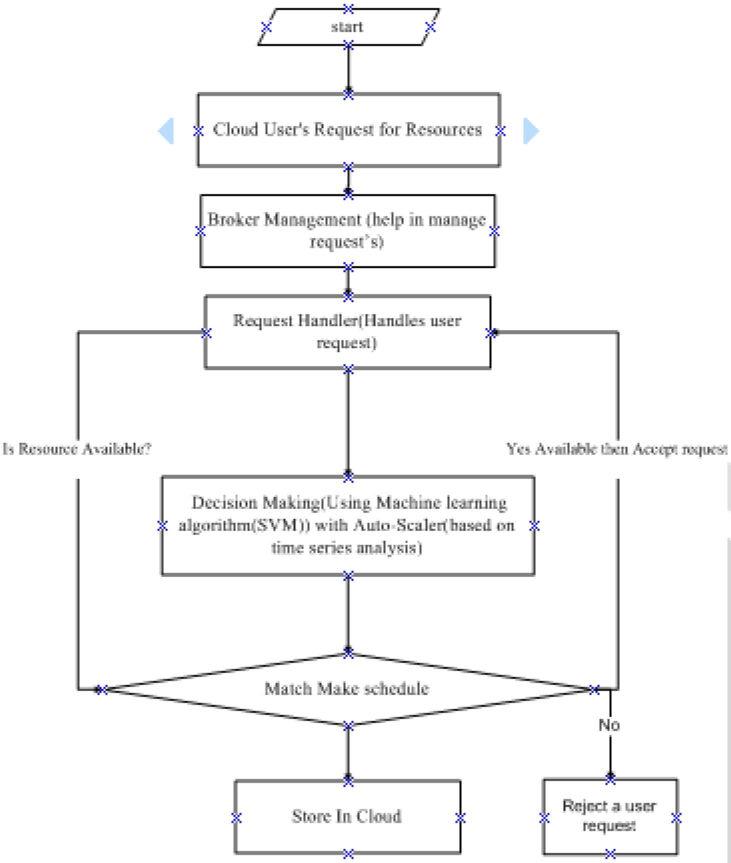
start
Cloud User's Request for Resources
Broker Management (help in manage request's)
Request Handler(Handles user request)
Is Resource Available?
Yes Available then Accept request
Decision Making(Using Machine learning algorithm(SVM)) with Auto-Scaler(based on time series analysis)
Match Make schedule
No
Store In Cloud
Reject a user request

Fig. 6 Flow Chart of the proposed system

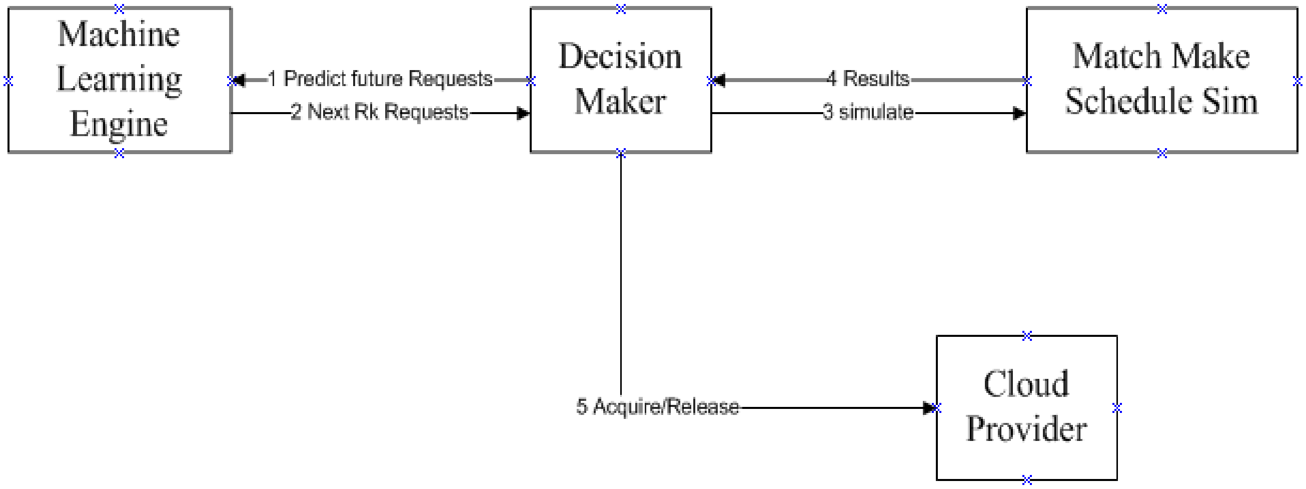

Fig. 7 Auto-scaler Flow Diagram

8. CONCLUSIONS

Cloud computing is an emerging trend and as the resource demands of users are increasing there is need to provide efficient resources to them with ease. In our research, the details of various Proactive Auto-Scaling Usages have been presented. The proposed work aims at proactive auto-scaling that is based on predictions of future workload based on the past workload and Efficiently Instance type selection based on demand of multidimensional resources. The goal of the auto-scaling techniques is to achieve a profit for the intermediary enterprise, Reduce the cost of client enterprise, Reduce Provisioning and Improve forecasting Accuracy. In proposed work we investigated the sensitivity of auto-scaling mechanisms to the prediction results by evaluating the influence of performance predictions accuracy on the auto-scaling actions.

9. ACKNOWLEDGEMENTS

The authors are very much grateful to Department of Computer Engineering, L.J Institute of Engineering & Technology, Ahmadabad, from Gujarat Technological University; for giving opportunity to do research work on Proactive Auto-scaling .Two authors Parekh Aneri, and Prof. Gayatri S. Pandi (Jain) , are also grateful to management team of L.J Institute of Engineering & Technology, Ahmadabad, from Gujarat Technological University; for giving constant encouragement to do research work in the Department.

10. REFERENCES

[1]Dimiter R. Avresky, Pierangelo Di Sanzo∗, Alessandro Pellegrini† , Bruno Ciciani‡ , Luca Forte§ , “Proactive Scalability and Management of Resources in Hybrid Clouds via Machine Learning”, 2015 IEEE 14th International Symposium on Network Computing and Applications pages 114-119.

[2] Simon Spinner, Nikolas Herbst and Samuel Kounev,Xiaoyun Zhu, Lei Lu, Mustafa Uysal and Rean Griffith, “Proactive Memory Scaling of Virtualized Applications”,2015 IEEE 8th International Conference on Cloud Computing pages 277-284.

[3] F´abio Morais, Raquel Lopes, Francisco Brasileiro “Instance Type Selection in Proactive Horizontal Auto-Scaling”,2016 IEEE 8th International Conference on Cloud Computing Technology and Science pages 102-109.

[4] Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy, Ali El-Haraki, “Automatic Resource Provisioning: a Machine Learning based Proactive approach” ,2014 IEEE 6th International Conference on Cloud Computing Technology and Science pages 168-173.

[5] Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy, Ali El-Haraki,“Predictive Auto-scaling Techniques for Clouds Subjected to Requests with Service Level Agreements”, 2015 IEEE World Congress on Services pages 311-318.

[6] Pranali Gajjar1, Brona Shah2, “Survey on Different Auto Scaling Techniques in Cloud Computing Environment”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015, ISSN (Online) 2278-1021.

[7] What is Auto Scaling?, https://docs.rightscale.com/fag/What-is_Auto-Scaling.html.

[8] R.S. Shariffdeen, D.T.S.P. Munasinghe, H.S. Bhathiya, U.K.J.U. Bandara, and H.M.N. Dilum Bandara, “Workload and Resource Aware Proactive Auto-Scaler for PaaS Cloud” ,2016 IEEE 9th International Conference on Cloud Computing, pages 11-18.

[9] M.Kriushanth, L. Arockiam and G. Justy Mirobi, ”Auto Scaling in Cloud Computing: An Overview” , International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013, ISSN (Online) : 2278-1021.

[10] Tania Lorido-Botran, Jose Miguel-Alonso, Jose A Lozano, “Auto-scaling Techniques for Elastic Applications in Cloud Environments”,Journal of Grid Computing, pages 1-34, 2014. ISSN 1570-7873.

[11]https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjUgd9g8DXAhUIpI8KHWmHBKIQjRwIBw&url=https%3A%2F%2Fwww.pinterest.com%2Fperrmeg%2Fcloud-baselearning%2F&psig=AOvVaw2zRwu6dvmFxMipLTnhtbVR&ust=1510816283114589, [access on 26/9/2017, 11:10:55].

[12]https://www.google.co.in/search?q=cloud+computing+services&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjQ3d2c_MLJAhWQCY4KHb82COMQ_AUIBygB&biw=1366&bih=623#imgrc=1c58z6WIAi, [access on 26/9/2017, 10:10:55].

[13]http://www.cse.unsw.edu.au/~cs9321/16s1/lectures/lec11/introductiontocloudcomputing.pdf, [access on 26/10/2017, 11:11:54]

BIOGRAPHIES

	Parekh Aneri received the B.E. degree in Computer Sciences and Engineering from Gujarat Technological University in 2016 and student of M.E. in Computer Engineering, L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; Currently he is doing research work in Proactive Auto-Scaling.
	Gayatri S. Pandi (Jain) is Associate Professor and Head of the Department of PG Department., L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; more than 15 years' experience in teaching.