Enhancing Resource Provisioning Based on CPU Utilization Using Machine Learning in Cloud Computing

Manali J Trivedi[1], Prof.Saket Swarndeep [2]

[1] *M.E Student, Dept. of Computer Engineering, L.J College of Eng. & Tech., Gujarat, India*
[2] *Assistant Professor, Dept. of Computer Engineering, L.J College of Eng. & Tech., Gujarat, India*

ABSTRACT

Cloud Computing is a recent buzzword in IT industry. Cloud computing is used for sharing of data ,information and resources. Machine learning used for prediction. Resource Provisioning provide resources to the end user. It has problems like auto scaling and resource utilization .A perfect Resource provisioning aims at avoid under provisioning and over provisioning of resources. Machine learning technique predict which request needs resources in which amount from the past data and Cloud computing checks the resource utilization. This paper aims to accurate resource provisioning and also scaling decision is accurate and it uses SVM algorithm for the better prediction of resources.

Keywords:- *Cloud Computing* **,** *Machine Learning , Resource Provisioning, Prediction*

1. INTRODUCTION

Cloud computing is a type of Internet-based computing that provides computer, shared processing resources and data to computers and other devices on demand. It is a model for providing ubiquitous, on-demand access to a shared pool of configurable computing resources (e.g., computer networks, servers, storage, applications and services),which can be rapidly provisioned and released with minimal management effort. Cloud computing gives users and companies with the ability to save and process their data in mediator data centers and that are stores at distant location from the user–rang across a city to across the world.

Essential Characteristics of Cloud are as follow.

- Rapid Elasticity : It is defined as the ability to scale resources both up and down as needed.
- Measured Service: Cloud services are controlled and monitored by the cloud provider. This is crucial for billing, access control, resource optimization, capacity planning and other tasks.
- On-Demand Self-Service: It means that a consumer can use cloud services as needed without any human interaction with the cloud provider.
- Ubiquitous Network Access: It means that the cloud provider's capabilities are available over the network and can be accessed through standard mechanisms by both thick and thin clients.

- Resource Pooling: It allows a cloud provider to serve its consumers via a multi-tenant model. Physical and virtual resources are assigned and reassigned according to consumer demand.

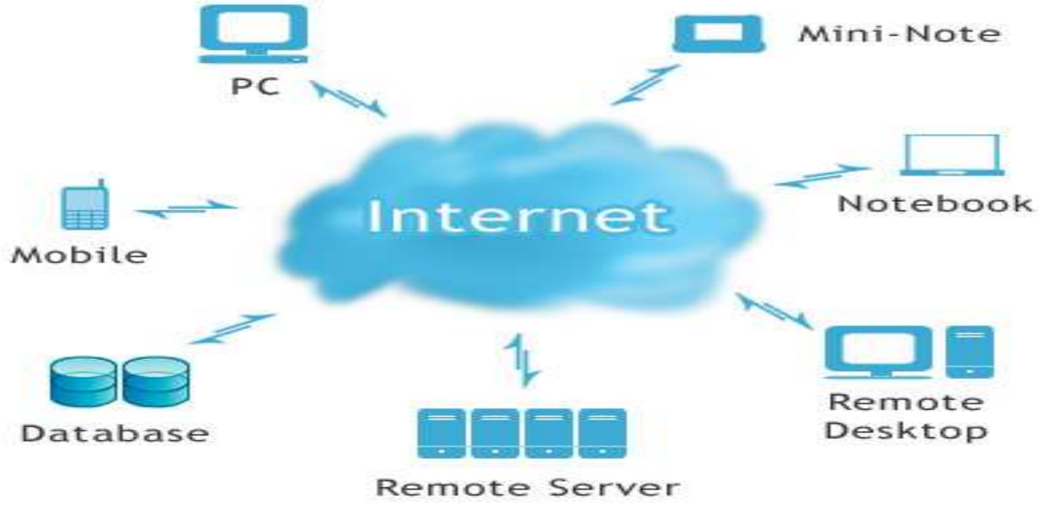

Fig.- 1: Cloud Computing

Machine learning is the part of computer science and it provide computers the ability to learn without being explicitly programmed. Machine learning consider the study and build an algorithms that can learn from previous data and make predictions on given current data. This algorithms defeat static program instructions and it also gives the result for given or input data.

Machine learning algorithms are of 4 types which are as follows:

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

Resource provisioning means to provide resources to a customer when user demand it. User wants resources so it send request to the Cloud provider as a Cloud provider accepts a request from a user, it must create the number of virtual machines (VMs) and allocate resources to the particular request to support them. There are three types of resource provisioning Static provisioning, dynamic provisioning and User self provisioning . Some applications that have predict resources in advance and generally they are not changing at later stage, that time “static provisioning" will be used. Resources are dynamically changing when user need it so provider allocate resources to the user on the fly and removes when not needed it is called Dynamic provisioning, the user can purchase resources from the cloud provider via a web from it is called user self provisioning.

2.RELATED WORK

In[1] authors used to set CPU Thresholds that triggering of auto scaling policies. They use impact of both utilization threshold and scaling size factor on performance of Cloud Computing Services during Provisioning process. Upper Threshold of CPU Utilization is used so efficiently deal with load spikes. They also use the metrics such as Response time and cost. On the basis of some Results they found that Resource utilization and Response time have impact on the Performance and cost of cloud services. They solved some optimization problems for Upper CPU Utilization thresholds and scaling size based on input loads, cost and Response time. Some optimization setting minimize the cost for number of allocated instances and provide acceptable SLO for Cloud Computing Services.

In[2] authors use Small Cloud for Resource Provisioning. They use Machine learning technique that is regression and based on Regression selecting and provisioning of virtual machine is done with minimal configuration to meet Web applications workload in small cloud provider. Non –linear regression method not only obtains CPU Demands but also check RAM Required to handle workload in different scales. First setup Workload system architecture that represent small cloud provider than implement main part of architecture to achieve virtualization on hardware. From Http request workload tests are performed on platform. The proposed model assist resource provisioning and capacity planning for virtual computing resources. They analyze proposed model as case study they found model improve small cloud provider’s Resource Utilization. It also eliminate OverProvisioning another benefit is identifying underutilized resources in order to provide large amount of VMs which allow small cloud provider to serve more small and medium Enterprise tenants.

In[3] authors developed cloud client prediction model on TPC-W benchmark web application and use Machine learning techniques for prediction. Three machine learning techniques used and that are : Linear Regression, Neural Network, Support Vector machine. They use two SLA metrics – response Time and Throughput. So Client can take more robust scaling decision. They extend the experiment time by 200%.Then Random workload pattern is employed and find the results using three Machine learning techniques and Support vector regression method give best prediction accuracy over both Neural Network and Linear Regression. SLA Metrics Response Time and Throughput degraded before an application reaches its set CPU Thresholds. Model also checks workload pattern impact on the database server and bottleneck occur than they use High Memory/CPU Infrastructure and they also include database server.

In[4] author proposed model that concerns dynamic provisioning of cloud resources performed by an intermediary enterprise that provides private cloud for single client enterprise and acquired resources from public cloud. Proactive technique introduced for auto scaling of resources that changes the number of resources for private cloud dynamically based on system load. The machine learning engine is used for predicting future workload pattern from the past workload pattern. From the experiments they found the proposed system reduce the user cost and broker cost. and the number of resource in pool used by user request need not be predict priori and controlled dynamically.

In[5] authors introduce unsupervised machine learning methods to dynamically provision multitier Web applications, while observing user-defined performance goals. The proposed technique operates in real time and uses learning techniques to identify workload patterns from access logs, reactively identifies bottlenecks for specific workload patterns, and dynamically builds resource allocation policies for each particular workload. Proposed model work in two parts : Workload pattern identification and Resource provisioning policy learning. From the workload pattern first partitioning of application’s URI space into request with similar resource utilization characteristics. and workload pattern uses probabilistic distribution model after uses policy learning algorithm for adaptive resource allocation to multitier web application. The proposed model not require prior knowledge of application’s resource utilization and minimize the overhead needed to monitor, detect and resolve bottlenecks. Proposed model meet service level agreement at minimal cost.

3.MACHINE LEARNING METHODS

1)Regression

Regression is a measure of the relation between the mean value of one variable and corresponding values of other variable. One wishes to find some simple pattern in the data – a functional relationship between the X and Y components of the data. For example, one wishes to find a linear function that best predicts a baby’s birth weight on the basis of ultrasound measures of his head circumference, abdominal circumference, and femur length.[6]

2)Neural Network

Networks of non-linear elements, interconnected through adjustable weights, play a prominent role in machine learning. They are called neural networks because the non-linear elements have as their inputs a weighted sum of the outputs of other elements— much like networks of biological neurons do. These networks commonly use the threshold element which we encountered in study of linearly separable Boolean functions.[7]

3)Support Vector machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering.

4. PROPOSED WORK

In this section we have described proposed model for accurate resource provisioning. This model will improve resource provisioning so resources neither be overprovision nor under provision and scaling decision can be better. First User send a request. Request is going to request manger. Request Manager manages the request than forward to the Queue manager

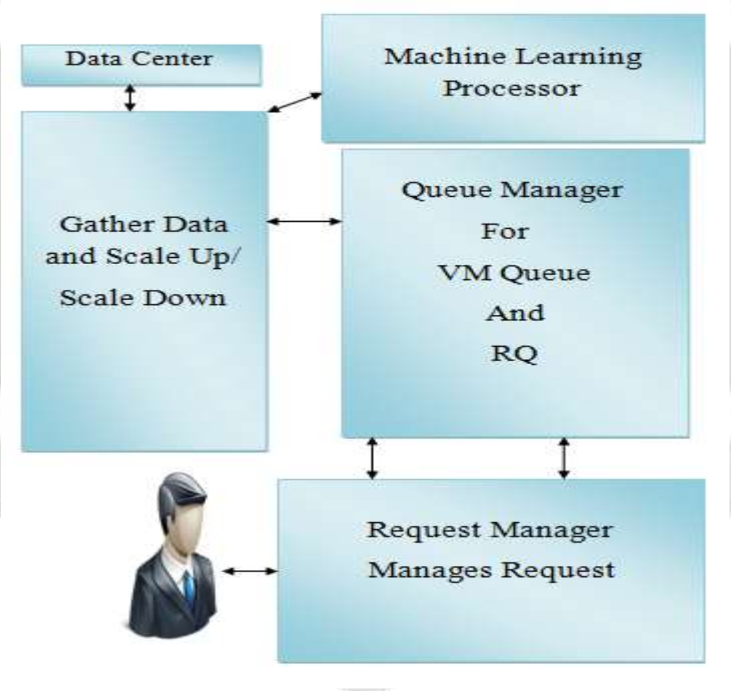
Data Center
Machine Learning Processor
Gather Data and Scale Up/ Scale Down
Queue Manager For VM Queue And RQ
Request Manager Manages Request

Fig.- 2: Proposed System

For Request Data is Gathered from data center than calculate how much resource request wants than on the basis of the data it will check either to scale the resources up or down also at that time machine learning processor parallel execute with it .Machine learning gives better prediction and on the basis of previous data find the resource requirement for request. Than request is map with resource and than it will execute.

5. RESULT ANALYSIS

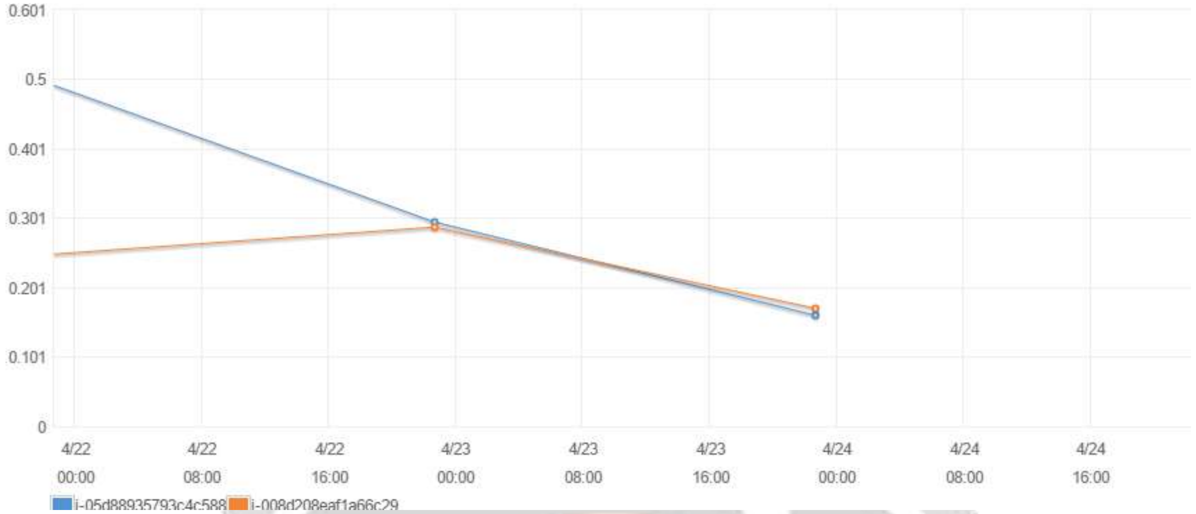

Fig.- 3: CPU Usage of Two Instances

Two Virtual Machines are running and calculate CPU Usage for the incoming request it firstly gives high utilization of CPU than as resource provisioning is done in proper way so CPU usage will be less for the particular request.

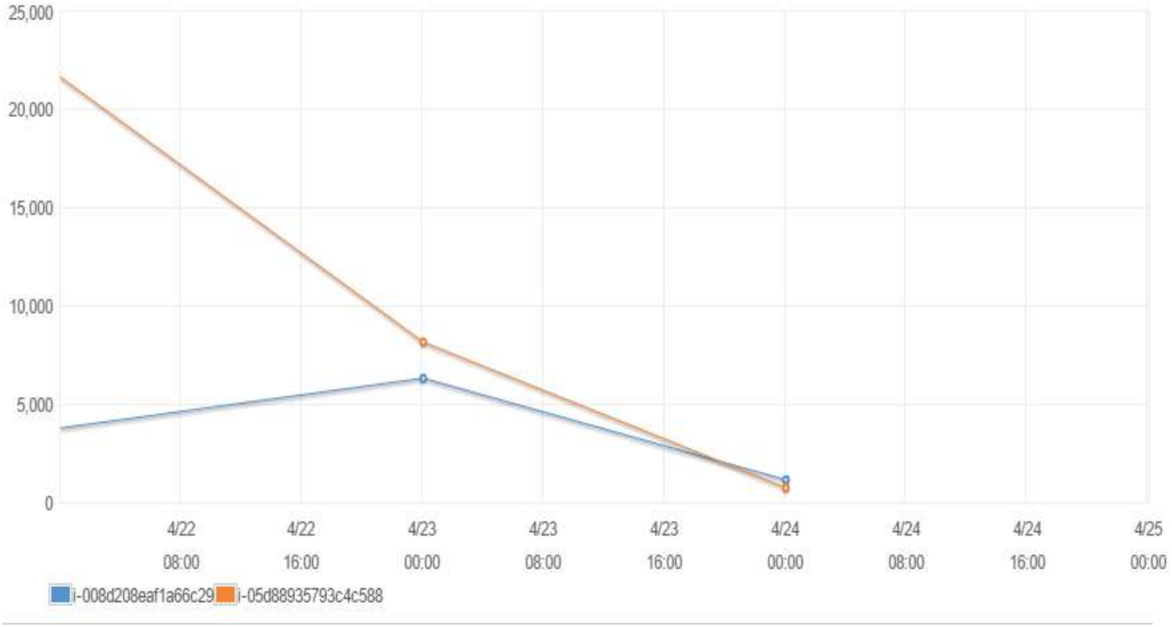

Fig.- 4: Response Time of Two Instances

Two Virtual Machines are running and checks response time for the each incoming request it firstly gives high Response time for one virtual machine so it will stop that instance which gives the higher response time and run other virtual machine which gives less response time.

6. CONCLUSIONS

Nowadays trend of cloud computing is increased among people and due to this their demands for resources is also increased. So there is possibility of resources to be either over provisioned or under provisioned, so it is necessary to manage these resources. So in this work, we are overcoming this problem by using SVM, which will predict the future requirement of resources based on the previous data.

7. REFERENCES

[1] F. Al-Haidari, M. Sqalli,K. Salah "Impact of CPU Utilization Thresholds and Scaling Size on Auto scaling Cloud Resources.", in IEEE International Conference on Cloud Computing Technology and Science ,DOI: 10.1109/CloudCom.2013.142 , pp. 256 -261 , Dec. 2013

[2] Bruno Yuji Lino Kimura,Roberto Sadao Yokoyama, Thiago Oliveira Miranda "Workload Regression-based Resource Provisioning for Small Cloud Providers." in IEEE Symposium on Computers and Communication, DOI**:** 10.1109/ISCC.2016.7543757 ,Aug. 2016.

[3] Samuel A. Ajila Akindele A. Bankole "Cloud Client Prediction Models Using Machine Learning Techniques " in IEEE 37th Annual Computer Software and Applications Conference , DOI 10.1109/COMPSAC.2013.21,pp. 134-142 , July 2013.

[4] Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy ,Ali El-Haraki "Automatic Resource Provisioning: a Machine Learning based Proactive approach" IEEE 6th International Conference on Cloud Computing Technology and Science , DOI 10.1109/CloudCom.2014.147 ,pp. 168-173 , Dec. 2014.

[5] Waheed Iqbal, Mathew N. Dailey, and David Carrera "Unsupervised Learning of Dynamic Resource Provisioning Policies for Cloud-Hosted Multitier Web Applications" DOI: 10.1109/JSYST.2015.2424998,pp. 1-12 ,May 2015.

[6]http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf Accessed on 5:09:00 22/03/2017

[7] http://ai.stanford.edu/~nilsson/MLBOOK.pdf accessed on 5:11:00 25/03/2017