A Survey on Resource Provisioning Using Machine Learning in Cloud Computing

[1]Manali Trivedi,[2]Prof.Hinal Somani

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

***Abstract*— Cloud Computing is widely used for sharing of data, Information and Resources. Resource provisioning is used for providing resources in Cloud Computing. Cloud Computing uses different techniques for resource provisioning. Resource Provisioning has many problems like auto scaling and workload for CPU Utilization . Machine Learning is used for resource Provisioning so Provisioning can be done Efficiently. In this review paper we have reviewed some techniques which are used for CPU Utilization and also for scaling and some technique uses machine learning techniques. Proposed model Accepts Request from the user than Request Manager manages the requests than ML Processor check which request wants how much resources on basis of previous requests and than Resource Provisioning is execute and scaling decision is performed.**

***Keywords*— Cloud Computing , Machine Learning , Resource Provisioning, CPU Utilization**

I. INTRODUCTION

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Essential Characteristics of Cloud are as follow.

1. Rapid Elasticity : It is defined as the ability to scale resources both up and down as needed.
2. 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.
3. On-Demand Self-Service: It means that a consumer can use cloud services as needed without any human interaction with the cloud provider.
4. 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.
5. 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.

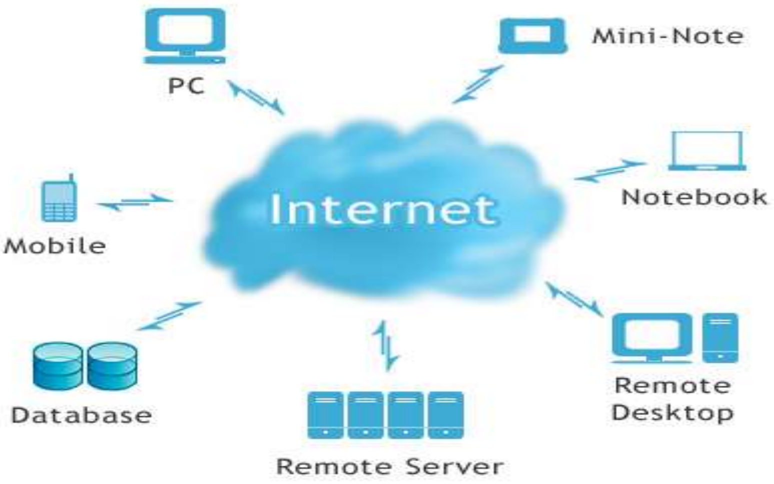

Fig. 1: Cloud Computing

Machine learning provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.
Face book news feed is the example of machine learning.

Resource provisioning is the allocation of a cloud provider's resources to a customer. When a cloud provider accepts a request from a customer, it must create the appropriate number of virtual machines (VMs) and allocate resources to support them. Resource Provisioning might affect the performance, scaling and also CPU Utilization. In the Proposed model CPU Utilization is checked and then Scaling decision is performed and machine learning processor predict the data and scaling decision is done.

II. METHODS USED

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]

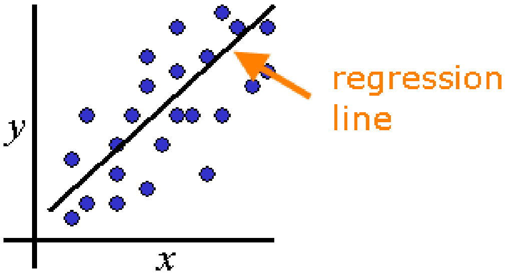

Fig. 2: Regression

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]

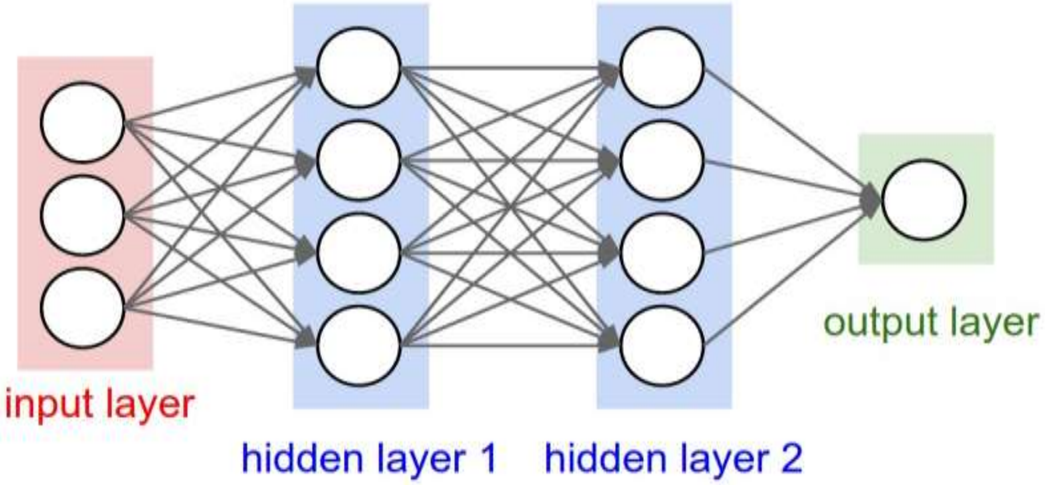

Fig. 3 Neural Network

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.

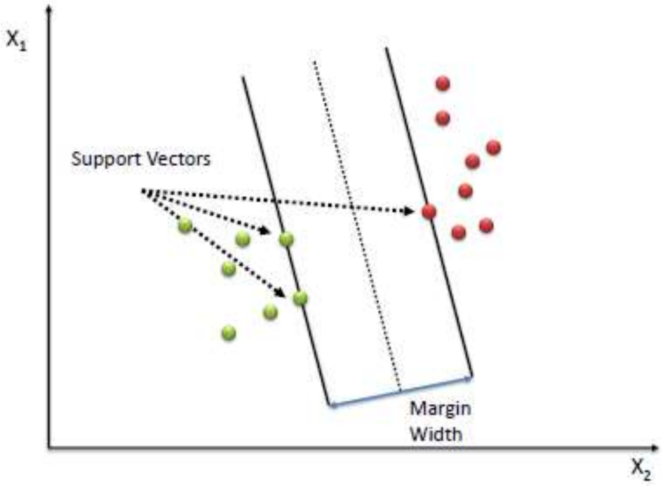

Fig. 4 Support Vector Machine

III. 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.

IV. COMPARISON TABLE

Sr. No.	Paper Title	Publication	Advantages	Research Gap
1	Impact of CPU Utilization Thresholds and Scaling Size on Auto scaling Cloud Resources.	IEEE 2013	Cost is minimize in terms of allocated resources.	It uses upper CPU Utilization threshold
2	Workload Regression-based Resource Provisioning for Small Cloud Providers	IEEE 2016	Resource utilization is improved and identify under utilized resources	Coefficient of model are related to hardware
3	Cloud Client Prediction Models Using Machine Learning Techniques.	IEEE 2013	SVR has superior prediction accuracy and response time	Work load pattern creates bottleneck on Database server
4	Automatic Resource Provisioning: a Machine Learning based Proactive approach	IEEE 2014	Model predicts the characteristics of future requests and generate profit for intermediary cloud provider	Auto scaling is not used Storage and network resources.
5	Unsupervised Learning of Dynamic Resource Provisioning Policies for Cloud-Hosted Multitier Web Applications	IEEE 2015	It minimize the overhead needed to monitor , detect and resolve bottlenecks and also meet SLA with minimal cost.	Problem with real time varying workloads and also dynamically changing workload and for scale down actions.

V. CONCLUSION

Resource Provisioning in cloud computing used for Resource prediction and also for Service Level Agreement and it uses machine learning algorithm for better and efficient prediction. Resource provisioning sometimes violet the service level agreement and also it is allocate virtual machine but when not needed it is not scale down the virtual machines.

VI. 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 23/11/2016

[7] http://ai.stanford.edu/~nilsson/MLBOOK.pdf accessed on 5:11:00 23/11/2016