Efficient Auto Scaling Mechanism in the Cloud Environment Using Proactive Approach

Aneri parekh[1], Prof. Gayatri S. Pandi (Jain)[2]
[1]M.E Student, Dept. of Computer Engineering, L.J College of Eng. & Tech., Ahmedabad, Gujarat, India [2]Head of the department, Computer Engineering Department, L.J College of Eng. & Tech. Ahmedabad, 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 papers we work on machine learning base effective approach for auto-scaling mechanism in these system we work on design effective approach for automatically virtual machine scale in cloud. And also consider machine learning concept using machine learning select appropriate node and according to that vm start scaling. Also work on parameters like latency, load balance, overhead etc.

Keywords: - *Proactive Auto-scaling, Auto-scaling , machine learning.*

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.

Cloud computing is internet based computing where virtually shared servers provide software, infrastructure, platform, devices and other resources to customers on a pay-as-you- use basis. Users can access these services available on the "Internet cloud" without having any previous know-how on managing the resources involved.

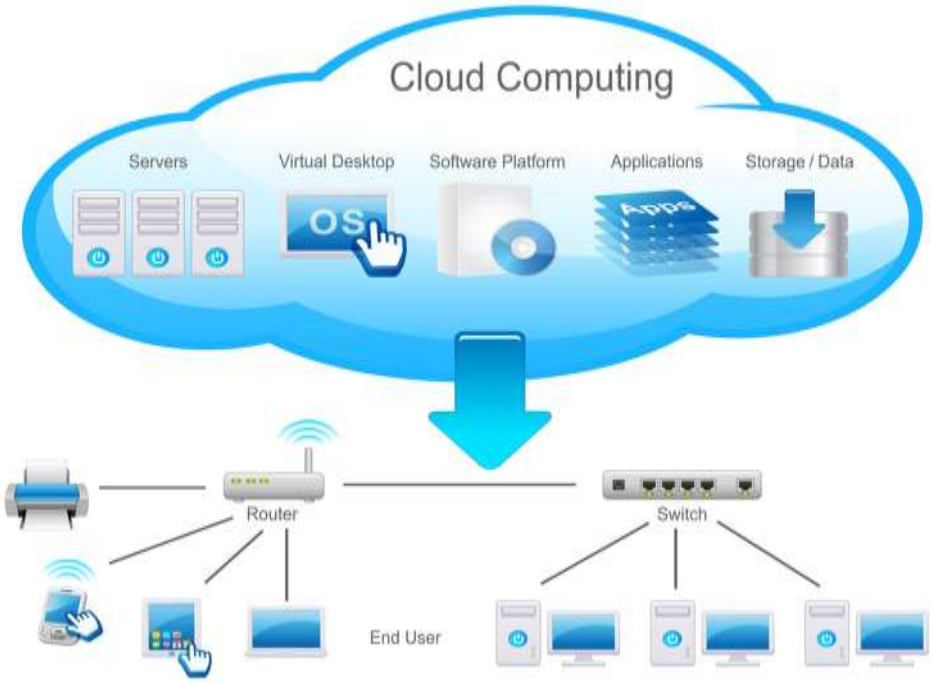

Fig-1: Cloud Computing

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 defined situation. It is closely related to and builds upon, the idea of load balancing.

Proactive mechanisms predict the workload ahead such that the auto-scaler can make decision based on the expected workload instead of waiting for trigger. Proactive 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. Proactive Auto Scaling is one of the benefit of cloud.

Predictive or proactive auto-scaling techniques try to anticipate to future needs and consequently acquire or release resources in advance, to have them ready when they are needed. Proactive scaling system enables providers to schedule capacity changes that match the expected changes in application demand.

2. Related Work

[1] Dimiter R. Avresky, Pierangelo Di Sanzo∗ , Alessandro Pellegrini , 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.

[2] Simon Spinner, Nikolas Herbst and Samuel Kounev,Xiaoyun Zhu ,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.

[3] F´abio Morais, Raquel Lopes, Francisco Brasileiro , In Instance Type Selection in Proactive Horizontal Auto-Scaling[3] , 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.

[4] Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy, Ali El-Haraki, 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.

[5] Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy ,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).

3. Proposed Work

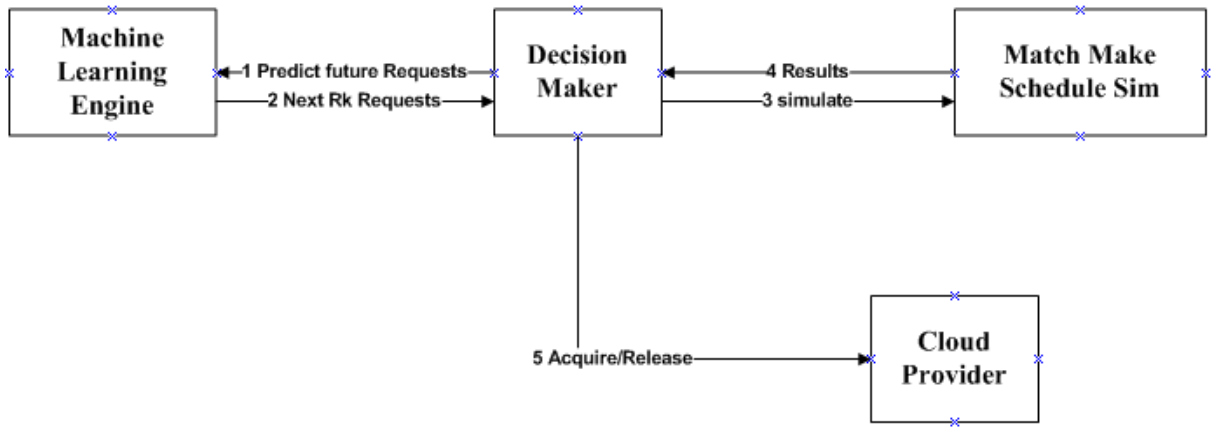

Fig -2 : Auto-scaler Flow Diagram

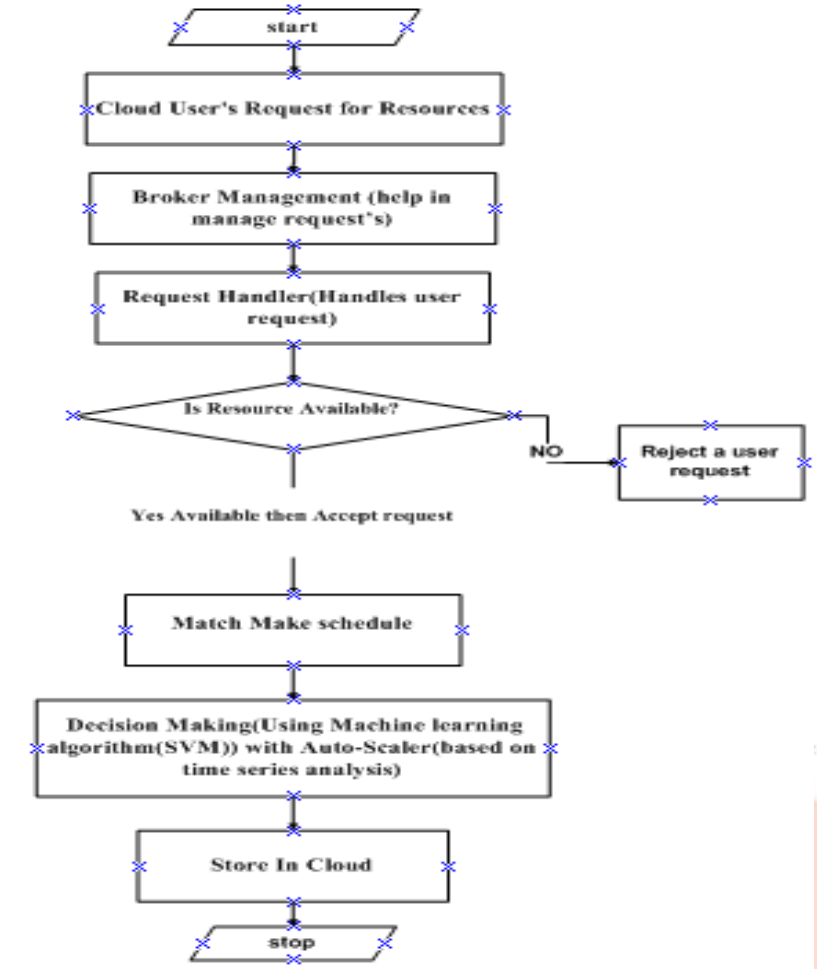

Fig-3 : Flow Chart for the proposed system

3.1 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 requests for the next k requests to arrive from MLE, demoted by Rk. DM simulates the resource management operations for these predicted future requests using MMS. 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.

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

4. Result Analysis

We are predicting future workload using proactive auto-scaling and machine learning and based on the Predicted workload the virtual machine will be allocated to the incoming request and if required additional resource then VM added to it for allocate to the request

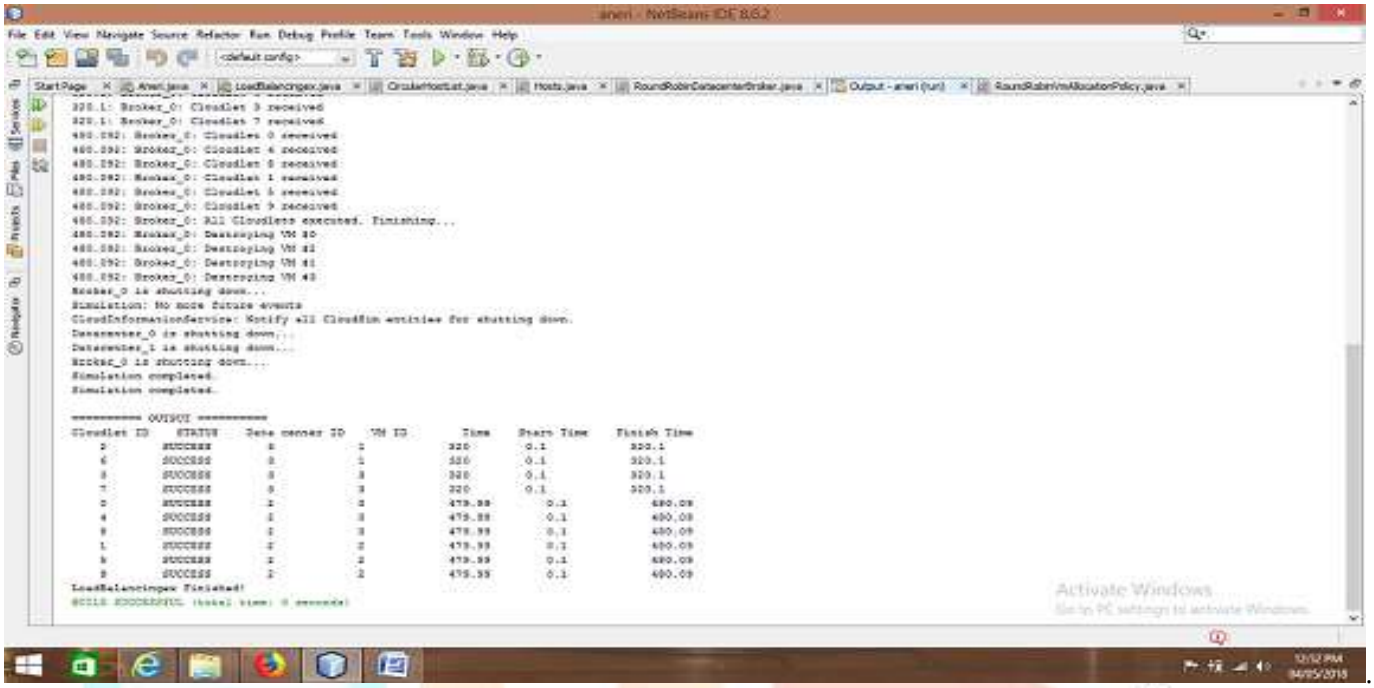

Fig-4: scale down on virtual machine

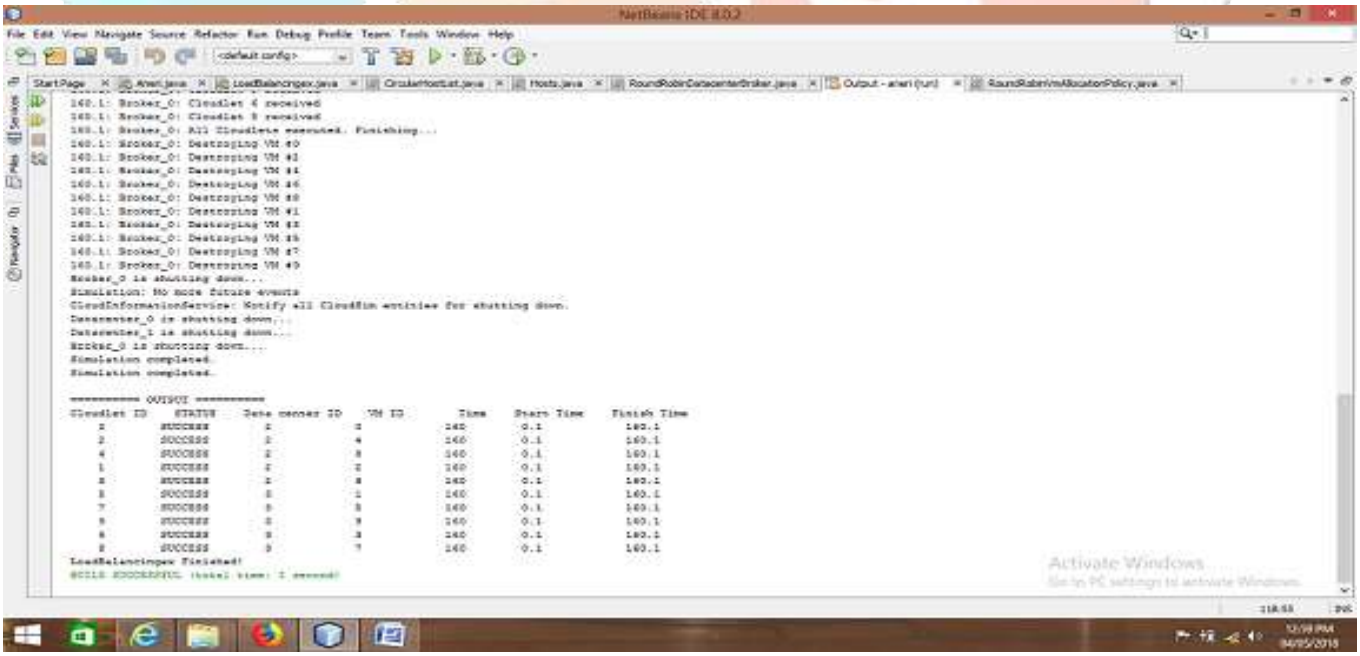

Fig-5: Scale up on virtual machine

5. Conclusion

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, details of various Proactive Auto-Scaling Usages have been presented. Proposed work aims at proactive auto-scaling that is based on predictions of future workload based on past workload . Goal of the auto-scaling techniques is to Improve forecasting Accuracy. In proposed work investigated the sensitivity of auto-scaling mechanisms to prediction results by evaluating influence of performance predictions accuracy on auto-scaling action.

6. 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.

[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.

[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.

[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-base learning%2F&psig=AOvVaw2zRwu6dvmFxMipLTnhtbVR&ust=1510816283114589

[Access on 26/9/2017, 11:10:55]

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