International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Exploratory Learning of Resource Management in Private Cloud Environment

Vipul Chudasama, Madhuri Bhavsar

Abstract: ***Paper: Scientific and Web applications are major sources of Internet traffic that requires resources such as Memory ,CPU and Network are on demand. Cloud computing and virtualization are the boons for such resource demand applications from various users. Service models of cloud computing provide a platform for many applications to use resources as pay per use model. In Cloud, Auto-scaling with manage Service Level Agreement (SLA) of resources is one of the main challenges to meet the current demand for resources. To maintain the performance of the cloud, which provision resources based on a heuristic for workload prediction is prime importance. In this paper, we address auto-scaling as a problem to forecast near-future demand of resource using a KNN machine learning methods suggest the optimized model for the dynamic variation of CPU utilization.***

Keywords **: Virtualization, Service Level Agreement, Resource provisioning, Resource prediction.**

I. INTRODUCTION

Cloud computing is now everywhere to offer most user-oriented computing services. Pay as you go model becomes the well-known model to assist users who want services from the known service providers. Service providers of Cloud take help of technology like virtualization to offer elastic compute service to a large number of users. Users and Cloud service providers follow a contract known as SLA that helps to monitor and manage the budgetary requirements.

Cloud vendors impose generalized rules to offer to scale decision of user's resource requirements. These rules are in the form of a threshold based on the current usage of the resources. The rules suffer from low response which may not maintain performance in-demand resources. In many cases, the exact scaling trigger is difficult to judge by the assigned rule. For example, a sudden demand for computing resources will trigger two compute resources while one of them is not utilized that would lead to over-provisioning of resources. Over provisioning of resources also increases the costs and wastage of data centre energy. While in the same case if a trigger is not activated will result in under-provisioning in resources. Due to the rule triggered the desired Quality of Service(QoS) could not be achieved in a given time. It is difficult to manage resources in the cloud with some static rules and hence efficient resource management technique [1] is required. SLA (Service Level Agreement) is closely related to QoS of the service. An efficient resource management technique will reduce SLA violation, increase customer satisfaction and QoS. Different types of applications fall under the different classes of QoS requirements [2] for the computing environment. The demand for resources in the near future may be decided by the VM parameters link CPU , memory utilization, Response time and availability. The role of auto-scaling is also considered to maintain a fair balance between QoS and SLA. IBM model [3] of autonomous auto-scaling with MAPE (Monitor-Analyze-Plan-Execute) flow provides current resource monitoring and action associated with the cloud environments. In MAPE cycle process where Discovery of the knowledge is achieved in Analysis phase. This process can be simple or complex in nature. Various approaches are designed based on cost[4] for application resources. A review on a web application based auto-scaling [5] also provides good direction. Service-based business model [6] gives the complex process of an auto-scaling in a cloud environment.

In given context our proposed work described in this paper has following contributions: 1) Design of auto-scaling model for prediction, 2) A regression time series based approach using machine learning method to predict future resource demand of CPU utilization of VM instances, 3) Evaluate the model with the standard regression-based measure.

The rest of the paper is organized as follows. In section 2, we discuss related literature review to highlight our work with existing work. In section 3, we propose a model of predictive time series of auto-scaling system and experiment using dataset [7]. Finally, section 4 include conclusions.

II. RELATED WORK

This section presents a literature review of related work in the field of resource provisioning and auto-scaling in the Cloud environment. Resource provisioning in such a dynamic environment is also one of the challenging tasks where research work is grouped into two categories.

The first category focuses on Resources provisioning techniques which can be further classified that proactive and reactive systems. Reactive systems [6]-[12] manage the resource provisioning with fine-tune the application performance parameters.

Manuscript published on 30 September 2019
* Correspondence Author
Vipul Chudasama*, Department of Computer Science and Engineering, Nirma University,Ahmedabad ,India.
Email:vipul.chudasama@nirmauni.ac.in
Dr. Madhuri Bhavsar, Department of Computer Science and Engineering, Nirma University,Ahmedabad ,India.

Various threshold levels generate alerts which tune the scaling parameter in the system [13]. Proactive systems using machine learning techniques [14]-[18] to react with future load variation and achieve the good system performance in the cloud. Service Level Agreement (SLA) based model [19] with workload clustering captures the future state of the system. Majority of the proactive systems use workload trace to predict future demand of cloud resource using a machine learning approach. A Datacenter workload represents the time series based behaviour [20]. Instance-based learning is evaluated on real workload which is used to extract features [21]. A comprehensive study in a different aspect of effective resource provisioning for applications was presented [22]. In the present approach, we used machine learning algorithms to model such behaviour of cloud for resource provisioning.

The second category focuses on the auto-scaling of the demand resources in vertical and horizontal directions. Auto-scaling based model is implemented with MAPE loop. The objective of auto-scaling is to advance the performance of a cloud system with a managed environment. Proactive systems used different phases to provide a solution for managed resources in the cloud. In MAPE, an analysis process with load prediction [14],[23],[24] using proactive and reactive approaches discussed where a CPU as a predicted resource was considered. In MAPE, a planning process with various architectures are studied and scaling rules are presented [25]. Another work was discussed on planning focuses on using learning automata with reactive and proactive methods [26].

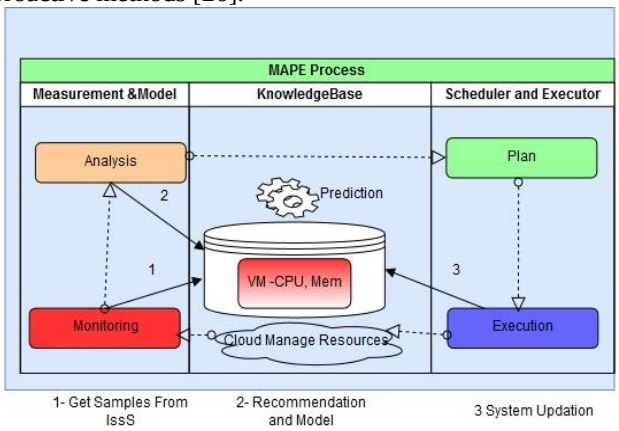

Figure 1: Life cycle of MAPE process for Cloud resource provisioning

Proposed Figure 1, MAPE process for cloud resource provisioning has a life cycle which starts with the monitoring process. The monitoring process is simple to get the status of resource utilization from cloud system [7][27]. Various attributes of physical resources like CPU, memory, hard disk, network bandwidth are considered in their work [21][28]. SLA parameters like MIPS and Memory are also taken into account in the monitoring process. Next, the analysis process which is most important for resource provisioning which explores the simple or complex discovery of resource utilization knowledge using derived rule sets or machine learning respectively. Next, the planning process will schedule the decision of scaling of resources. Execution process updates such information taken by the planning process and execute it. The motivation for such a cycle process is to improve the analysis process [29] so that a better plan can be provided for scaled resources. Utilization of resources is model as a time series problem, where a window of length m points [t-m .. t-1] are used for training for the current time t. The future state at time t is predicted to forecast near future resource utilization.VM sizing [30] for the web server is considered for resource provisioning. Linear Regression (LR), Auto Regressive Moving Average (ARMA) and Moving Average (MA) are considered to predict a number of request for cloud user for resource provisioning. QoS parameter like response time and rejection rate [16] for VM provisioning using ARIMA model gives high accuracy with simulation results. Major algorithms in the family of machine learning algorithms which explores knowledge from the historical data to train model for new problem instances. Here we relate the problem of resource provisioning for host application using a KNN model and validate model with a metric. When designing resource provisioning techniques, CPU utilization is considered to measure the policy of a cloud.

A. KNN (K -Nearest Neighbour) Regression

A regression function which will learn the relationship between the dependent variable and independent variable. In order to achieve the goal to establish such function here, we have used KNN regression model. For such model, the training dataset having sufficient sample are used. The training dataset has *s* samples which are described by xi where $xi=\{xi1,xi2,\ldots,xi_m,y_{m+1}\}$. The function will be model like $f:x\rightarrow y$. By opting a local average of the training dataset function will learn the relationship. By using the value of *k* which defines the performance of KNN algorithm and finding the best values of k is needed. In order to achieve good performance of the model to minimizes, prediction losses smaller value of k is preferred. Cross-validation is a process by which we can estimate the accuracy and validity of a model. There are two methods of cross-validation m-fold validation and leave one out cross-validation. Here we have used 10 fold cross-validation to train the model.

III. EXPERIMENT

Real workload data Materna [7] which is a full cloud service provider for ITC based projects is considered for current work. This time-series data represents one month of data of business-critical applications. The dataset contains performance parameters of 520 VMs in such a distributed environment. Each VM files contains 12 performance features such as CPU core, Memory capacity provisioned, Memory usage and write throughput, Disk size, Disk read and Network Throughput, CPU capacity provisioned, CPU usage,. VM execution data are collected for every five minutes interval from which CPU utilization data is considered for analysis Historical time series data of CPU utilization are sampled and provided to Analysis process. Next, the Analysis process will run at Δt time. First, for every five minutes, it monitors the cloud resources like CPU, memory, and Disk and stores this information as part of the monitoring process. Next, it sees the difference of current time and analysis time to run the operations. The first operation in that is to get the stored dataset of VM utilization with CPU utilization. To perform learning with input and output variables to model relationship, a regression approach is used. Training dataset samples consist of CPU utilization of the past 30 minutes. Best value of tuning parameters will be selected as per the cross validation method for given test samples and training samples.

The output label for VM utilization which is based on CPU load will be considered for the next process to plan VM.As shown in Table 1 , CPU utilization of selected VMs are considered for analysis.

Table 1: CPU load variation in VMs

CPU Utilization %	VM1	VM2	VM3	VM4
Minimum	0.37	1.88	0.17	2.01
Maximum	46.49	100	80.16	100
standard deviation	3.2	9.14	11.39	16.05

Figure 2 provides information about samples of a VM CPU utilization for 1 Month.

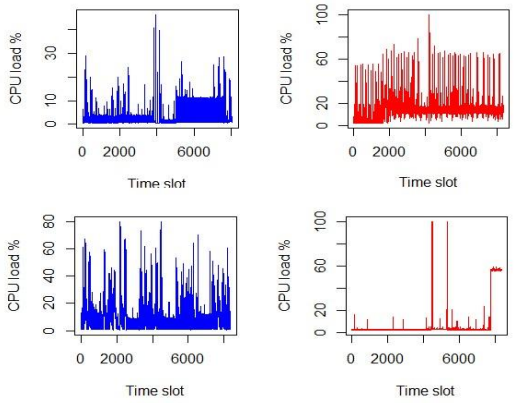

Figure 2 One month CPU Utilization of VM1,VM2,VM3,VM4

A. KNN-Auto-scale Algorithm

Each example of training consist of five input variables and an output variable. The total utilization values of each VM (VMU_t,VMU_{t+1} ,VMU_{t+2},VMU_{t+3}, .. VMU_{t+5}) and output will be VM utilization after five minutes VMU_{t+6}.

To evaluate accuracy of models, Mean absolute error (MAE) metric is used. In MAE, e_t is the difference of actual output and predicted output and n is the number of observations in dataset for which prediction is done. A lower value of MAE indicates good model. Table 2 shows comparison of Mean Absolute Error(MAE) of KNN machine learning algorithm with sliding window(SW).

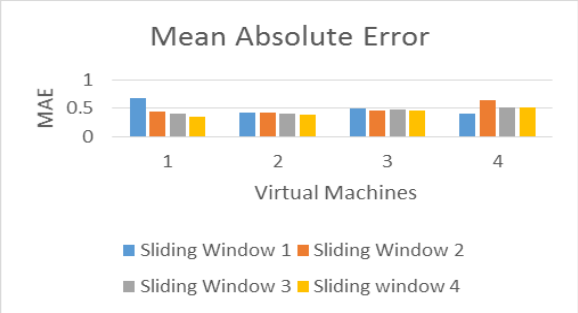

Figure 3 MAE of VM Utilization using KNN

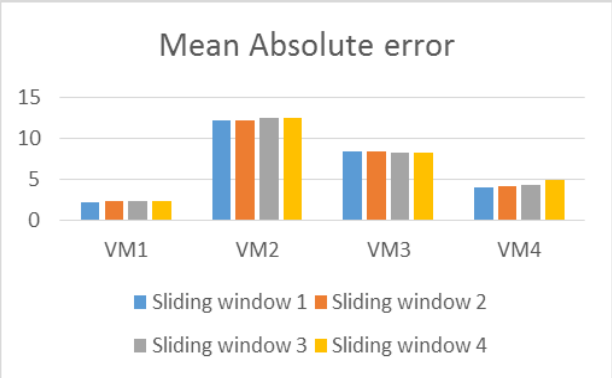

Figure 4 MAE of VM Utilization using multiple linear regression

IV. CONCLUSION

In this paper, we proposed an exploratory path to build an efficient prediction model for resource provisioning in the cloud. In the recent scenario, optimization of resources for scientific or business applications are very important. To understand and predict the dynamic nature of the cloud environment, we explored machine learning algorithms having time as a factor to derive decision. We also evaluated our experiment with the standard metric for validating the proposed methods. The accuracy of KNN as the proposed learning method is encouraging and demonstrating optimized resource prediction. Auto-scaling with such method will provide effective support to the private cloud environment in future.

REFERENCES

1. S. Sigh, I. Chaa, Cloud resource provisioning: survey, status and future research directions, Knowledge and Information Systems 49 (3) (2016) 1005–1069.
2. S. Sinh, I. Chna, Resource provisioning and scheduling in clouds: Qos perspective, The Journal of Supercomputing 72 (3) (2016) 926–960.
3. A. Computing, et al., An architectural blueprint for autonomic computing, IBM White Paper 31 (2006) 1–6.
4. T.Lorido-Botran, J. Miguel-Alonso, J. A. Lozano, A review of autoscaling techniques for elastic applications in cloud environments, Journal of grid computing 12 (4) (2014) 559–592.
5. C.Qu, R. N. Calheiros, R. Buyya, Auto-scaling web applications in clouds: A taxonomy and survey, ACM Computing Surveys (CSUR) 51 (4) (2018) 73.
6. M. Mohamed, M. Amziani, D. Bela¨ıd, S. Tata, T. Melliti, An autonomic approach to manage elasticity of business processes in the cloud, Future Generation Computer Systems 50 (2015) 49–61.
7. T. G. W. Archive, The grid workloads archive URL http://gwa.ewi.tudelft.nl/
8. H. R. Qavami, S. Jamali, M. K. Akbari, B. Javadi, Dynamic resource provisioning in cloud computing: a heuristic markovian approach, in: International Conference on Cloud Computing, Springer, 2013, pp. 102–111.
9. G. Molto, M. Caballer, C. de Alfonso, Automatic memory-based verti-´ cal elasticity and oversubscription on cloud platforms, Future Generation Computer Systems 56 (2016) 1–10.
10. P. Padala, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, A. Merchant, K. Salem, Adaptive control of virtualized resources in utility computing environments, in: ACM SIGOPS Operating Systems Review, Vol. 41, ACM, 2007, pp. 289–302.
11. Q. Zhu, G. Agrawal, Resource provisioning with budget constraints for adaptive applications in cloud environments, in: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, ACM, 2010, pp. 304–307.

12. Q. Zhang, L. Cherkasova, E. Smirni, A regression-based analytic model for dynamic resource provisioning of multi-tier applications, in: Autonomic Computing, 2007. ICAC'07. Fourth International Conference on, IEEE, 2007, pp. 27–27.
13. M. Z. Hasan, E. Magana, A. Clemm, L. Tucker, S. L. D. Gudreddi, Integrated and autonomic cloud resource scaling, in: Network Operations and Management Symposium (NOMS), 2012 IEEE, IEEE, 2012, pp. 1327– 1334.
14. S.Islam, J. Keung, K. Lee, A. Liu, Empirical prediction models for adaptive resource provisioning in the cloud, Future Generation Computer Systems 28 (1) (2012) 155–162.
15. A.A.Bankole, S. A. Ajila, Predicting cloud resource provisioning using machine learning techniques (2013) 1–
16. 4doi:10.1109/CCECE.2013.6567848.
17. R.N.Calheiros, E. Masoumi, R. Ranjan, R. Buyya, Workload prediction using Arima model and its impact on cloud applications' QoS, IEEE Transactions on Cloud Computing 3 (4) (2015) 449–458.
18. M. Ghobaei-Arani, S. Jabbehdari, M. A. Pourmina, An autonomic approach for resource provisioning of cloud services, Cluster Computing 19 (3) (2016) 1017–1036.
19. H. Zhang, G. Jiang, K. Yoshihira, H. Chen, A. Saxena, Intelligent workload factoring for a hybrid cloud computing model, in: Services-I, 2009 World Conference on, IEEE, 2009, pp. 701–708.
20. S. Singh, I. Chana, Q-aware: Quality of service-based cloud resource provisioning, Computers & Electrical Engineering 47 (2015) 138–160.
21. N. Roy, A. Dubey, A. Gokhale, Efficient autoscaling in the cloud using predictive models for workload forecasting, in: Cloud Computing (CLOUD), 2011 IEEE International Conference on, IEEE, 2011, pp. 500– 507.
22. H. Li, D. Groep, L. Wolters, An evaluation of learning and heuristic techniques for the application run time predictions, in: Proceedings of 11th Annual Conference of the Advance School for Computing and Imaging (ASCI), Netherlands, Citeseer, 2005.
23. M. Amiri, L. Mohammad-Khanli, Survey on prediction models of applications for resources provisioning in cloud, Journal of Network and Computer Applications 82 (2017) 93–113.
24. N. R. Herbst, N. Huber, S. Kounev, E. Amrehn, Self-adaptive workload classification and forecasting for proactive resource provisioning, Concurrency and computation: practice and experience 26 (12) (2014) 2053– 2078.
25. M. D. de Assunc¸ao, C. H. Cardonha, M. A. Netto, R. L. Cunha, Impact of˜ user patience on auto-scaling resource capacity for cloud services, Future Generation Computer Systems 55 (2016) 41–50.
26. E. Casalicchio, L. Silvestri, Mechanisms for sla provisioning in cloudbased service providers, Computer Networks 57 (3) (2013) 795–810.
27. M. Fallah, M. G. Arani, M. Maeen, Nasla: Novel auto scaling approach based on learning automata for web application in cloud computing environment, International Journal of Computer Applications 113 (2).
28. C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, M. A. Kozuch, Heterogeneity and dynamicity of clouds at scale: Google trace analysis, in: Proceedings of the Third ACM Symposium on Cloud Computing, ACM, 2012, p. 7.
29. J.-J. Jheng, F.-H. Tseng, H.-C. Chao, L.-D. Chou, A novel vm workload prediction using grey forecasting model in cloud data center, in: Information Networking (ICOIN), 2014 International Conference on, IEEE, 2014, pp. 40–45.
30. S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, K. Dutta, Modeling virtualized applications using machine learning techniques, in: ACM Sigplan Notices, Vol. 47, ACM, 2012, pp. 3–14.
31. M. Mishra, A. Das, P. Kulkarni, A. Sahoo, Dynamic resource management using virtual machine migrations, IEEE Communications Magazine 50 (9) (2012) 34–40. doi:10.1109/MCOM.2012.6295709.

AUTHORS PROFILE

Vipul Chudasama, Assistant Professor, Department of Computer Science and Engineering, Nirma University ,ISTE life member.

Dr. Madhuri Bhavsar, Head , Department of Computer Science and Engineering, Nirma University ,ISTE life member.