International Journal of Engineering & Technology, 7 (4.10) (2018) 413-416

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Efficient Cloud Resource Scaling based on Prediction Approaches

K Dinesh Kumar[1*] and E Umamaheswari[2]

[1,2]School of Computing Science and Engineering, VIT University
**Corresponding author E-mail: kdinesh.kumar2015@vit.ac.in*

Abstract

Resource Scaling is one of the important job in cloud environment while adapting resource configurations due to elasticity mechanism. In the view of cloud computing, resource scaling mechanism hold the assurance of QoS (Quality of Service), So, one of the key challenging task in cloud environment is, resource scaling. Effective scaling mechanism gives an optimal solutions for computational problems while achieving QoS and avoiding SLA (Service Level Agreement) violations. To enhance resource scaling mechanism in cloud environment, predicting future workload to the each application in different manners like number of physical machines, number of virtual machines, number of requests and resource utilization etc., is an essential step. According to the prediction results, resource scaling can be done in the right time, while preventing QoS dropping and SLA violations. To achieve efficient resource scaling, proposed approach lease advantages of fuzzy time series and machine learning algorithms. The proposed approach is able to reach effective resource scaling mechanism with better results.

Keywords: *Cloud Computing; Fuzzy Time Series; Prediction Approaches; Resource Scaling; Workload.*

1. Introduction

Cloud computing is a trending technology in computing area where providing resources like computation power, storage, network and applications etc., are delivered as a service through internet. The main services are Software-as-a-service (SaaS), Platform-as-a-service (PaaS) and Infrastructure-as-a-service (IaaS). Apart from these services, delivering services in different categories, such as network, security, API and test environments etc. In the current scenario, information technology organizations requires huge amount of running servers to increase their economy which can leads to raised purchase cost of hardware and environmental pollution. The maintenance cost of underutilized resources (datacenters, servers etc.) effects on company profits. Additionally, extra resources requiring during peak situations to avoid SLA violations.

As per the SLA's, cloud service provider should be handle bulk requests. This is the main challenge to scale up the cloud infrastructure based on the workload. In cloud environment, high resource configuration (vCPUs and vRAMs) applications have fluctuation resources demands due to maintain massive requests. For example, the offer sales days many e-commerce applications such as Flipkart, Amazon, e-bay and Snapdeal etc., gets huge requests from customers. Sometimes servers will hang due to inefficient resources and e-commerce provider's loss their goodwill. Hence, resource scaling is the important task in cloud computing environments. Auto-scaling basically depends on the virtualization technique and sometimes delaying to launch new virtual machine which can leads reputation loss of cloud providers. Therefore, the efficient resource scaling can be achieved by using prediction algorithms, to scaling the resources as per the workload behavior.

The Challenge of estimating future workload of the application, is the prediction of correct amount of required resources, has been undertook through proactive approaches. Proactive methods estimating the future required resources in a way the resource scaling manager has efficient time to scale the cloud resources before facing the bottleneck situation. If sudden increase of application workload, the resource scaling manager scale up the virtual resources structure according to the predicted future workload of the application. On the other hand, based on the workload reduction, resource scaling manager will change the allocated resources configuration. So, by using prediction methods, can avoid SLA violation and can reach QoS's.

The rest of the paper is organized as follows: Section 2 describes prediction approaches in different manner such as metrics of the applications. Section 3 presents the related work of prediction methods. Section 4 describes proposed approach of this paper. Finally, the paper is concluded in Section 5.

2. Prediction Approaches

The main challenge of the prediction is, to estimate the future behavior of the application with respective metrics based on the historical information which is collected before. Prediction can be done with different metrics of the application [1]. Resource prediction approach must be proactive approach. Normally, in reactive methods, either increasing or decreasing the resources depends on predefined threshold values, which can lead more time taking to migrating or launching new virtual machines. Final result is provider might lose their customers. In contrast, proactive approaches combines with learning algorithms to predict future workload of the application. With prediction results, resource scaling manager make the decisions like either scale-in or scale-out.

The main branches of the prediction are workload and performance. Prediction in two manner, is important step for the efficient resource scaling management in cloud environment. Based on the future workload, the efficient resource scaling manager must be detect the correct amount of resources to satisfy QoS parameters like response time, CPU utilization, reliability and availability. Figure 1 represents characteristics and challenges of the prediction methods.

The efficiency of prediction model evaluated by the accurate predicted results. Prediction model should be give accurate outputs which are close to the actual values. To improve the accuracy, prediction approach analyse the historical data of the application and develop the new trained model by using learning algorithms. Generally, cloud computing environment able to change their resource configuration continuously. So, resources adaptation helps to change configuration of resources by using prediction results with better results. The different dimensions of the applications for prediction are as follows.

- SLA violation: Allocating inefficient resources to the application leads SLA violation. The result is cloud provider might lose their revenues. Identifying this violations, in manner of applications could help to scaling the allocated resources.
- Number of Users: With this metric information, can decide which application have many customers.
- Resources Utilization: This dimension gives the information about amount of resource utilization. This results helps to make decisions like whether applications are either in over-utilization or in under-utilization stage.
- Number of Requests: Based on incoming requests, prediction approach can make decisions to scale the resources.
- Number of PMs and VMs: The number of allocated resources such as virtual machines (VMs) and physical machines (PMs) to the applications, gives idea about future needed resources.

Prediction, in manner of different dimensions gives the better results rather than considering single dimension of the application.

3. Related Work

In this section, discusses about prediction models [1] and these models are categorized into three types: (i) computational intelligence, (ii) analytical, and (iii) simulation models. The computational intelligence models are mainly depends on neural networks, fuzzy logic and genetic algorithms etc. The Analytical are mainly depends on regression models, hidden markov model and queuing theory model etc. Few simulation tools used for prediction of resources such as Memory buddies and Overdriver etc. So far, authors proposed many prediction approaches for different aims and aspects.

Dinda et al. [3] developed a simulation toolkit to predict future CPU workload of resources according to their CPU usage information. The authors used some linear prediction approaches such as autoregressive (AR), moving average (MA), autoregressive integrated moving average (ARIMA) and autoregressive moving average (ARMA) for predicting CPU workload. Based on the prediction results, scaling can be done. Nahrstedt et al. [4] proposed a multi resource prediction model to predict the future workload of server by finding correlation among other resources. This correlation concludes the same resources behavior. Sometimes CPU load increase suddenly due to massive requests, in these situations linear prediction approaches cannot gives the accurate results. So, finding correlation among resources helps for better results. Guitart et al. [5] proposed a prediction algorithm for efficient energy scheduling in virtual machines, infrastructure, resource and SLA's of a cloud environment. The proposed model takes all advantages of exponential smoothing, moving average, double exponential smoothing and linear regression. The prediction approaches predicts the results individually, after that, based on lowest absolute error, is chosen as the prediction results.

Estrella et al. [7] proposed prediction model for resource adaptation by integrating different linear prediction models. Proposed algorithm used a genetic algorithm to integrating all linear prediction model. The main advantage of their model is, it does not required training model before scaling. Mainly, their prediction approach predicts by integrating the prediction value of dissimilar time series prediction models and combining different time series by using genetic algorithm. Kumar et al. [6] proposed prediction approach to predict the utilization of virtual machines by using ant colony approach. Chen et al. [8] addressed a dynamic prediction approach to forecast CPU workload of cloud resources. Proposed

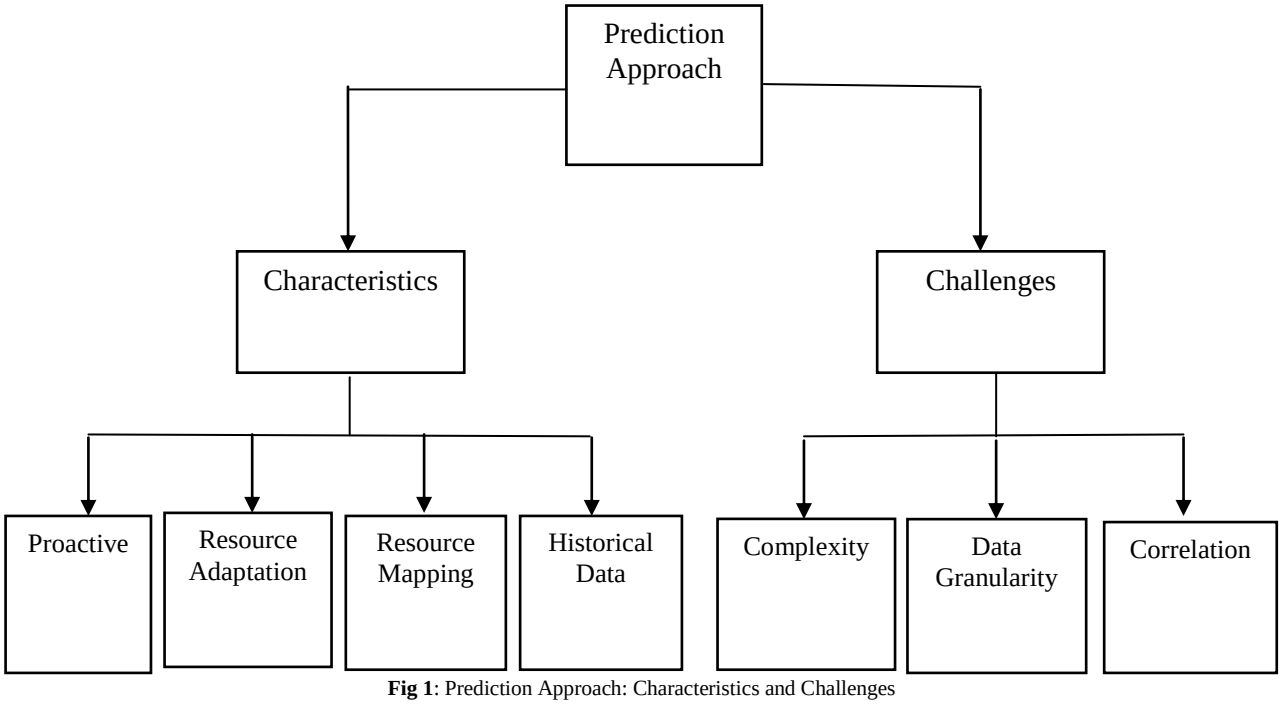

Fig 1: Prediction Approach: Characteristics and Challenges

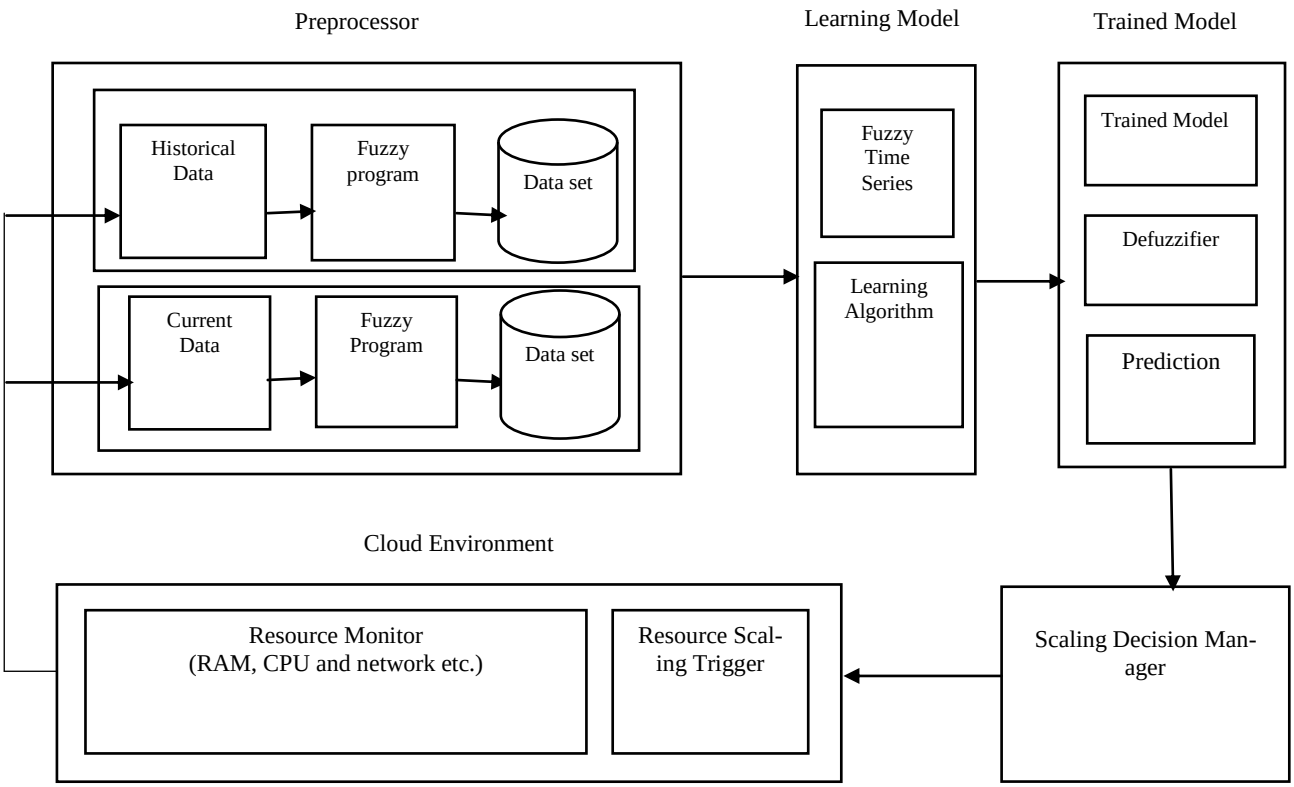

Fig 2: Proposed Proactive Resource Scaling Model

Model integrated linear and non-linear forecasting models. It consists of two layers, the old values with prediction error rates are replaced with new prediction approach values in first layer. To replace error rate values, proposed model using exponential smoothing, autoregression, weighted nearest models. Zhu et al. [9] proposed adaptive prediction approach by using fuzzy clustering based fuzzy neural networks. For effective results they used fuzzy c-means integrating with subtractive cluster algorithm.

4. Proposed Approach

The proposed model, resource scaling can done in the way of proactive method. Proposed approach mainly have four components: (i) preprocessor, (ii) learning algorithm, (iii) trained model, and (iv) scaling decision manager. The main role of the first component is to collect the all metrics information (historical data and current usage data) from resource monitor and converted into fuzzy time series, which are input datasets for learning algorithms. In the second phase, all fuzzy time series transformed into group multivariate fuzzy time series and transformed as input for learning algorithm. In the third phase, trained model gives the output values for defuzzifier to separate the values according to the respective metrics and finally it predicts future resource usage. In the final phase, based on the prediction results scaling manager make the decisions to scale the resources.

Pseudocode: Proactive Scaling Model

1. Begin
2. Normalize all historical time series and current usage time series at each point 't'.
3. Fuzzy time series as inputs for learning algorithm: F_1 (t), F_2 (t), F_3 (t) … F_n (t).
4. Develops a learning algorithm for next observation values.
5. Pass the values to trained model.
6. At each point of time 't', gathered new resources data.
7. Compare the results with new resources data.
8. Make the decision such as either scale-in or scale-out

The goal of the model is to enhance prediction approaches to scale resources in advance. Accuracy is the essential for the proposed approach.

Many prediction methods such as autoregression, moving average, back-propagation neural network, autoregression integrated moving average and exponential smoothing etc., are employed to forecast cloud resource workloads. Integration of prediction approaches improve the prediction accuracy. Accuracy is one of main requirements while developing efficient resource manager is to avoid SLA violation.

5. Conclusion

The proposed work mainly concentrate on developing a proactive resource scaling model for cloud environment. Proposed model addressed comprehensive approach, which consists of four components to get the accuracy results. This model take all advantages of fuzzy time series approach to transfers information into group multivariate time series and learning algorithms to predict the forecasting values. Based on forecasting values scaling decision manager, change the resource configuration.

References

[1] Amiri, Maryam, and Leyli Mohammad Khanli. "Survey on prediction models of applications for resources provisioning in cloud." Journal of Network and Computer Applications (2017).

[2] Liang, Q., Zhang, J., Zhang, Y.H., Liang, J.M., 2014. "The placement method of resources and applications based on request prediction in cloud data center." Inf. Sci. 279,735745.

[3] P.A. Dinda, "Design, implementation, and performance of an extensible toolkit for resource prediction in distributed systems", Parallel Distributed Systems. IEEE Trans. 17 (2) (2006) 160–173.

[4] J. Liang, K. Nahrstedt, Y. Zhou, "Adaptive multi-resource prediction in distributed resource sharing environment", in: Cluster Computing and the Grid, 2004. CCGrid 2004. IEEE International Symposium on, 2004, pp. 293–300.

[5] J. Subirats, J. Guitart, "Assessing and forecasting energy efficiency on cloud computing platforms", Future Generation Computer Systems 45 (2015) 70–94.
[6] Kumar, K Dinesh; Umamaheswari, E. "An Authenticated, Secure Virtualization Management System in Cloud Computing." Asian Journal of Pharmaceutical and Clinical Research**,** [S.I.], p. 45-48, Apr. 2017. ISSN 2455-3891.
[7] V.R. Messias, J.C. Estrella, R. Ehlers, M.J. Santana, R.C. Santana, S. Reiff Marganiec, "Combining time series prediction models using genetic algorithm to autoscaling web applications hosted in the cloud infrastructure", Neural Computer Applications (2016) 1–24.
[8] J. Cao, J. Fu, M. Li, J. Chen, "CPU load prediction for cloud environment based on a dynamic ensemble model", Software Practice Exp. 44 (7) (2014) 793–804.
[9] Z. Chen, Y. Zhu, Y. Di, S. Feng, "Self-adaptive prediction of cloud resource demands using ensemble model and subtractive-fuzzy clustering based fuzzy neural network", Computer Intelligence Neuroscience 2015 (2015) 17.
[10] Tran, Dang et al. "A Proactive Cloud Scaling Model Based on Fuzzy Time Series and SLA Awareness." Procedia Computer Science 108 (2017): 365-374.
[11] Zhang, H., Jiang, G., Yoshihira, K., Chen, H., 2014. "Proactive workload management in hybrid cloud computing." IEEE Trans. Network and Service Management 11 (1), 99100.
[12] Amiri, M., Feizi-Derakhshi, M.R., Mohammad-Khanli, L. "IDS fitted Q improvement using fuzzy approach for resource provisioning in cloud." Journal of Intelligent & Fuzzy Systems 32 (1) (2017): 229-240.