International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 474 - 477

Neural Network Prediction based Dynamic Resource Scheduling for Cloud System

M Uma
M. Tech. (CSE), Dept. of CSE, SRM University
NCR Campus Modinagar Ghaziabad
Utter Pradesh, India
uma.m279@gmail.com

Partha Sarathi Chakraborty
Assistant Professor, Dept. of CSE, SRM University
NCR Campus Modinagar Ghaziabad
Utter Pradesh, India
parthasarathi@live.com

***Abstract*—**Cloud computing is known as a internet based model for providing shared and on demand accessing of the resources (CPU, memory, processor, etc.). It is known as a dynamic service provider using very large scalable and virtualized resources over the Internet. With the help of cloud computing and virtualization technology, large number of online services can run over virtual machines (VMs), which in turn will reduce the number of physical servers. However, maintaining and managing the resources demand dynamically for these virtual machines with changing demand of resources while maintaining the service level agreement (SLA) is a challenging task for the cloud provider. Dynamic resource scheduling is a way to help manage the resource demand for virtual machines to handle variable workload without SLA violation. In this paper, we introduce Neural based prediction strategy to enable elastic scaling of resources for cloud systems. Unlike traditional static approach which do not consider the VM workload variability in account and dynamic approaches which sometimes predict under estimate of resources or over estimate of the resource, here we consider both workload fluctuations of VMs and prediction estimation problem into account. Neural based prediction strategy will first predict the VM resource demand based on Artificial Neural Network (ANN) model, to achieve resource allocation for cloud applications on each VM. Once the prediction is done, we than apply dynamic resource scheduling to consolidate the virtual machines with adaptive resource allocation, to reduce the number of active physical server while satisfying the SLA.

Keywords- *Cloud System, Virtualization, Dynamic Resource Scheduling, Artificial Neural Network*

*****

I. INTRODUCTION

Cloud computing deliver three types of services which are infrastructure as a service, platform as a service, and software as a service which are made available to its customers through pay as you-go-basis. Cloud system offers a platform for hosting large, distributed Internet services, such as e-commerce, social network and so on[1] and [4]. Cloud client use virtual machines running over physical server to deploy their applications with all the required resources for guaranteed performance provided in service level agreement (SLA) [3].

Over cloud the workload of each VM is not fixed all the time, rather it fluctuates all the time, in order to guarantee performance at peak demand cloud provider over provision the VM capacity which sometimes leads to waste of cloud resources. Thus one of the major goals of cloud provider is to provide proper handling of resources to the virtual machine while managing the variable workload without SLA violation.

With Virtualization and cloud computing it possible to consolidate multiple online services on VMs in a small cluster of physical server [6], [7] and [8]. These virtual machines can be consolidated statically and dynamically. Static allocation method, assume that VM resource demand is known well before in advance and it will not consider the VM workload variability into account, which means that the VM will be allocated fixed number of resources well in advance and will not change the number of resource allocated to VM if required. The problem with static approach is if there is less resource required not required at all, there will be wastage of VM resources, and if the resource allocated to the VM is less than the resource required, than the VM will be ideal and not provide the service. In cloud computing VMs will get service from physical machines (PMs), now PM should have the sufficient resources to meet the demand of all the VMs running on it, otherwise PM will be overloaded and it will not provide the resources to some VMs. Those VMs will be idle for certain time and this situation can degrade the performance of its VMs such situations can be avoided by dynamic provisioning and consolidation approach. In dynamic allocation takes VM workload variability into account. But with dynamic approach that predict the VM resource demand sometimes give over estimations of the resources which lead to wastage of VM resources, and sometimes give under estimation of resources which may lead to resource conflict, both the estimations leads to SLA violation.

In this paper, we introduce Neural based prediction strategy to automate elastic resource scaling for cloud systems. In this paper we consider both the dynamic workload fluctuations of VMs and prediction estimation into account. Neural based prediction strategy will first predict the VM resource demand based on Artificial Neural Network (ANN) model, to achieve resource allocation for cloud applications on each VM. Then we apply dynamic resource scheduling algorithm to dynamically consolidate the VMs with adaptive resource allocation to reduce the number of physical machines (PMs) while satisfying the SLA.

We first predict the future resource demand of each VM, and then the future load of a PM is predicted by aggregating all the predicted resource demand of each VM running on PM. If for a particular resource if it is predicted that the demand exceeds than that of its respective PM, than that PM will be considered overloaded in the next time interval.

In this situation some of the VMs should be migrated from that PM until the predicted load does not lead to overload the cloud system. Basic idea of this project is to effectively predict

the resources and reserve sufficient resource for VMs for guaranteed performance.

Thus the design goal of the predictor is to ensure that the predicted resources are no less than the actual demand in next time interval. Once we have predicted the resources we than decide the physical servers needed and placement of VMs. Here we consider both live migration and resource conflict into account.

Rest of the paper is organized into: section II Related Work, section III system architecture overview, and section IV ANN based resource prediction and dynamic resource scheduling and section V conclusion.

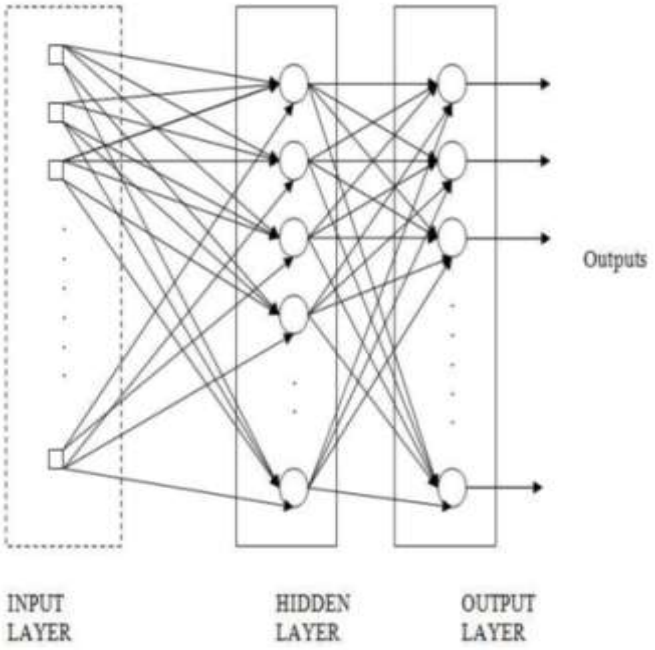

Figure1 Basic Neural Network Structure.

II. RELATED WORK

Scaling Techniques are used to determine the number of servers required, Cloud system scaling techniques such as Amazon Auto Scaling help cloud customers to manage cloud resources [5]. These scaling techniques are used to determine how many servers are needed and placing VMs over the physical servers, there are various techniques used such as machine learning, queuing theory, and control theory and so on to dynamically manage the workload among the servers. There are many other schemes used to dynamically determine the physical servers, required and correct placement of VMs, such as VM consolidation schemes.

But VM consolidation method does not consider VM workload variability into account. With virtualization it is possible to live virtual machine migration. Adya and Collopy worked to identify the effectiveness of Neural Networks (NN) for forecasting and prediction. They came up to the conclusion that neural network are well suited for the use of prediction, but need to be validated against a simple and well-accepted alternative method to show the direct value of this approach. Live migration is the ability to check if the physical machine is overloaded or have very less resources or free, and if the PM is overloaded its overloaded VM will be migrated to the PM where there is less load, and if the PM is free it will turn off that PM in order to save power and minimize the number of PMs. Sandpiper approach trigger the VM migration when it detects a overload condition and it is predicted to continue overload in the next few time interval. Sandpiper provides two ways for VM migration: gray-box and black-box approach. But their approach use short-term prediction, and most of the time it may launch migration with a delay and resource conflicts have occurred. In our approach we will consider both this live migration and resource into account.

Our work uses neural based prediction strategy. Neural network can be implemented using MATLAB, which is a strong programming tool for neural based prediction [11]. Neural Networks are widely used for forecasting problems [12].Artificial Neural Networks are proven universal approximators [13] and are able to forecast both linear [15] and nonlinear time series [16]. Zhang, Patuwo, and Hu [16] show multiple other fields where prediction by ANN was successfully implemented.

III. SYSTEM ARCHITECTURE

System architecture of our prediction based scheduling is shown in figure 2. Below Figure consists of following main components: Control plane and the cloud cluster. Control plane consist of ANN predictor tool, system statistic collector, resource demand predictor and resource scheduler. Cluster consists of number of physical machines, virtual machines and cloud broker. Overview of this architecture is it collects the resource usage of the virtual machines based on the statistic collected it predicts the resource demand in next time interval, which is used to perform the dynamic resource scheduling. Main aim of this paper is to predict sufficient resources with minimum number of physical server while satisfying the service level agreement.

Cloud Broker this module is connected to the data centre which collects the resource usage statistics of the VMs running over cluster of physical machine in the data centre, and performs migration action from the Resource scheduler, if the Virtual machine doesn't have enough resources to handle the workload.

System statistic collector this module collects the resource usage statistic from cloud broker placed inside the cluster and define the measurement interval after which the resource statistic to be collected.

Resource Demand Predictor this module is responsible for predicting the resource demand for the next time interval, based on resource usage collected from the previous time intervals, using Artificial Neural Network (ANN). ANN predictor tool predicts the resource demand and fed into Resource Demand Predictor.

Resource Scheduler this module is responsible for virtualizes resource scheduling. Predicted resource data from Resource demand predictor is fed into the scheduler which than determines the number of active PMs, placement of VMs and necessary VM migration operations. The goal of Resource scheduler is to minimize the number of active physical machines while satisfying the service level agreement (SLA).

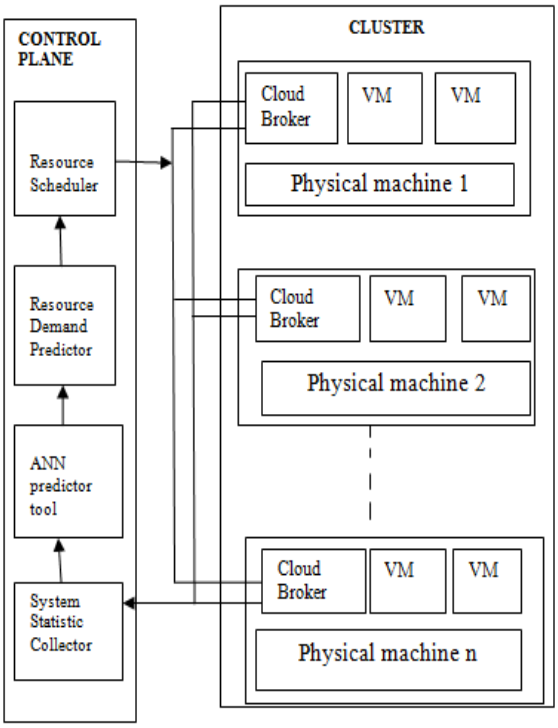

Figure2 System Architecture

IV. NEURAL NETWORK BASED RESOURCE DEMAND PREDICTION AND DYNAMIC RESOURCE SCHEDULING

This section is divided into two parts. First part describes how ANN is used to predict the resource demand based on the previous resource usage statistics obtained from the system statistic collector and presents the steps of calculation of prediction of time series.

Second part describes how ANN prediction results for dynamic resource scheduling to manage the resource for virtual machines in order to handle variable workload while satisfying the SLA.

Here Neural Network Toolbox (nntool) of MATLAB environment is used to implement ANN [11]. The program MATLAB with Neural Network Toolbox is a strong tool for prediction. Neural network toolbox creates, train and simulate the neural network, it takes previous outcome of cloud data centre as input through the Cloud broker connected to the data centre and predict the output, shown in figure 3. Steps for neural based prediction are as follow, first download the input output time series data, next is the choice of the type of neural network to be used, next we have to set up the number of layers of the network, followed by selecting the number of neurons, transfer functions necessary for prediction calculations. Once the neural network is built, learning and testing process are necessary to be run.

After the calculation it is necessary to evaluate the results of prediction. The results of prediction could be exported for graphical visualisation; this prediction result is then passed to resource scheduler for sufficiently allocating the resources for VMs. Sample result is shown in figure 4.

This section describe Resource Scheduler module of our system architecture. Here first we have to learn migration overhead analysis, and problem formulation. Migration overhead is to test migration overhead of different kinds of VMs. Problem formulation will take the cluster of homogenous PMs as shown in figure 2 over which the VMs operate in order to collect the resource consumption of these VMs. Based on the statistics prediction decision and scheduling decision are made.

Based on the prediction result obtained from ANN prediction tool, we determine the number of physical servers required and correct placement of virtual machines over it and allocate them required resources, once we have allocated resources like cpu, memory consumption, processor and so on to virtual machines, next step is to divide the physical server into available set, free set and busy set, available set consist of physical servers which are lightly loaded or those PMs whose capacity doesn't exceed than that of the virtual machines, free set consist of those PMs whose resources are not utilized by the VMs placed over it, and busy set are consist of those PMs for which VM resource demand exceed than the capacity of PM.

This step is done to determine the virtual machines that are overloaded, in order to perform the migration, once migrate list is formed from the overloaded VMs which are marked by the busy set, this step is done to avoid the resource conflict, next we perform the migration of all the VMs present in the busy set by moving them into the available set and free set, until predicted load does not lead to overload the system. Next step is determine which systems are free and are lightly loaded, in order to transfer the lightly loaded VMs to those that PMs that can accommodate these VMs and those systems that are free are power off in order to consolidate the VMs into less number of physical servers, The main goal of consolidation phase is to reduce the number of physical servers and avoid resource conflict while satisfying the service level agreement (SLA). Relative Error (RE) to evaluate the prediction accuracy on each prediction step, shown in figure 5. RE is the difference between prediction value and real resource demand value at different time interval.

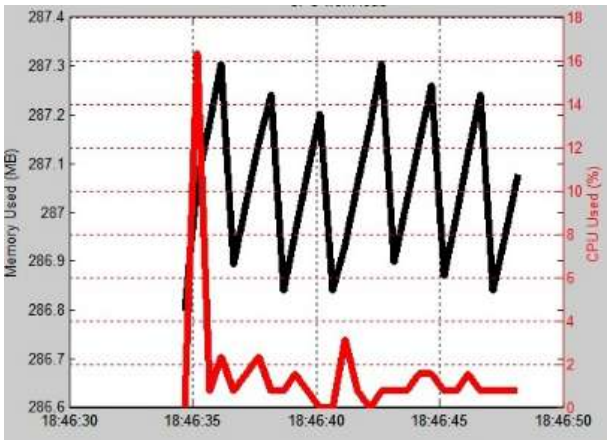

Figure3 Resource Utilization

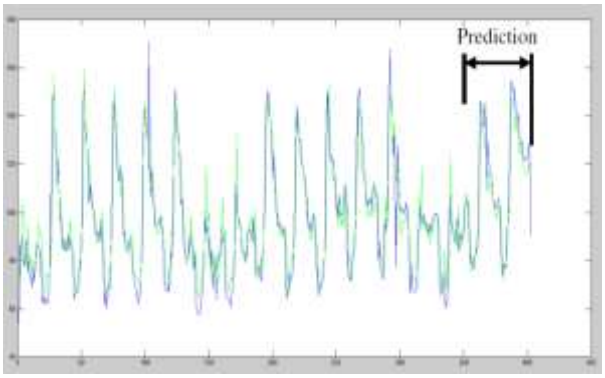

Figure4 ANN predicted data

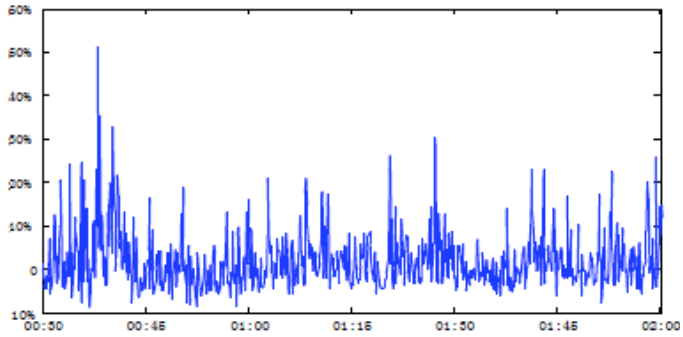

Figure5 RE values of Neural based prediction

V. CONCLUSION

This paper has presented that dynamic resource scheduling is managed more effectively and efficiently, using neural network based prediction strategy.

This paper sought to improve the resource utilization and manage to handle the variable workload of virtual machines with minimum physical server while satisfying SLA.

It is found that with neural network based prediction over-provisioning of the allocated resource could be improved efficiently.

REFERENCES

[1] Abdulaziz Aljabre,“ Cloud Computing for Increased Business Value”, International Journal of Business and Social Science, Vol. 3 No. 1; January 2012

[2] P. Patel, A. Ranabahu, and A. Sheth, “Service level agreement in cloud computing,” UKPEW, 2009.

[3] P. Patel, A. Ranabahu, and A. Sheth, “Service level agreement in cloud computing,” UKPEW, 2009.

[4] Nikolaus Huber, Marcel von Quast, Micahal Hauck and Samuel Konev, “Evaluating and Modeling Virtualization Performance Overhead for Cloud Environments”

[5] “Amazon elastic compute cloud.” [Online]. Available: http://aws. amazon. com/ec2/

[6] J. Sahoo, S. Mohapatra, and R. Lath, “Virtualization: A survey on concepts, taxonomy and associated security issues,” *in Computer and Network Technology (ICCNT), 2010 Second International Conference on. IEEE*, 2010, pp. 222–226.

[7] A. Kivity, Y. Kamay, D. Laor, U. Lublin, and A. Liguori, “kvm: the linux virtual machine monitor,” *in Proceedings of the Linux Symposium*, vol. 1, 2007, pp. 225–230.

[8] P. Barham, B. Dragovic, K. Fraser, S. Hand,T. Harris, A.Ho, R. Neugebauer, I. Pratt, and A. Warfield, “Xen and the art of virtualization,” *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, pp. 164–177, 2003.

[9] Y.Hashimotox and K. Aida, “Evaluation of performance degradation in hpc applications with vm consolidation,” *in Networking and Computing (ICNC), 2012 Third International Conference on. IEEE*, 2012, pp. 273–277.

[10] J. G. F. Hermenier, X. Lorca and J. Lawall, “Entropy: a consolidation manager for clusters,” *In Proc. VEE*, 2009.

[11] http://in.mathworks.com/help/nnet/gs/neural-network-time-series-prediction-and-modeling.html?refresh=true.

[12] A. Lapedes and R. Farber, “Nonlinear Signal Processing Using Neural Networks: Prediction and System Modelling,” Los Alamos National Laboratory, Los Alamos, NM, Tech. Rep. LA-UR-87-2662, 1987.

[13] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feed forward networks are universal approximators,” Neural Networks, vol. 2, no. 5, 1989, pp. 359–366.

[14] K. Hornik, “Approximation capabilities of multilayer feed forward networks,” Neural Networks, vol. 4, no. 2, 1991, pp. 251–257.

[15] G. P. Zhang, “An investigation of neural networks for linear time-series forecasting,” Computers and Operations Research, vol. 28, no. 12, 2001,pp. 1183–1202.

[16] G. Zhang, B. E. Patuwo, and M. Y. Hu, “Forecasting with artificial neural networks: The state of the art,” International Journal of Forecasting, vol. 14, no. 1, 1998, pp. 35–62.

[17] “Rubis online auction system,” http://rubis. ow2. org/.