Neural Network Based Resource Allocation using Run Time Instrumentation with Virtual Machine Migration in Cloud Computing

Anitha N, Anirban Basu

Abstract— The enterprise level and the market level both are seeing a huge growth in the cloud computing. The resource is accessed in a large with better way and also globally. An individual or organization can lease the computational or storage resources, in return reducing the cost of the infrastructure. The resources optimization is one of the major issue faced in the cloud computing for the cloud service providers. Most of the optimization of resources allocation is done after the calculation of the resources needed and on the go. In this paper, a mathematical system model for the resource allocation using neural network with run time instrumentation has been proposed. The proposed model shows the better resource utilization.

Index Terms— Cloud Computing, Deep Inspection, Instrumentation, Machine Learning, Neural Network,

I. INTRODUCTION

The RAS (Resource Allocation Strategy) is a strategy used in-order to allocate and utilize the resources in cloud computing so that the application resources requirements are met [1]. It usually encompasses types of resources, amount of resources that is needed by the application that is using the cloud. The efficient strategy must satisfy certain criteria such as less resource conflict, possibly zero scarceness, less resource shattering, over provisioning and under provisioning. Today most of the service level agreement (SLA)[1] for the application based on cloud only consider the execution times and number of transactions. The SLA is used for allocating the resources. It will be loss to the cloud provider if the resources provided are more than the specified in SLA and SLA is affected if less resource is provided. Here in this paper we are monitoring the application of the user and based on the reading we will further assign the resources, while making sure that there is no under provision or over provision. Procedure for Paper Submission.

II. RELATED WORK

Topology Aware Resource Allocation [2] is a resource allocation technique that is used in the Infrastructure-as-a-service. The resources are allocated independently to the hosted application in the current IaaS system as they are not aware of their resources requirements. This can cause a huge performance deficit for the application which may be distributed data intensive. To overcome they proposed a system “whatif” which will guide the IaaS in resource allocation. A lightweight simulator along with a prediction engine for estimating the performance of a resource has been used and also the genetic algorithm for identifying the best solution. When compared to other resources allocation system the TARA has reduced the time of completion of job to 50%. The Linear Scheduling for tasks and Resources (LSTR) is a algorithm [3] that is used for scheduling which will first perform the task and then resource is allocated. Since the LSTR is used along with a server node, the system throughput and resource utilization is maximized. Both the resource allocation and resource consumption is combined to improve the utilization. The scheduling algorithm main focus is on the distribution of the resources, so that the QoS parameters are maximized. The QoS parameters are the cost function. In order to maximize the resource utilization the scheduling algorithm is designed based on carried out tasks, strategy of scheduling and available virtual machines. Adaptive Resonance Theory-2(ART2) neural network [4] is used for prefetching and clustering the resources. The request of resources is a three tuple (n, rt, d) where rt is a ready time for execution , n is the number count of virtual machines, d is completion deadline. The weights of the neural network is calculated as :

$$Z_i = \frac{u_i}{1-d}$$

Where u_i is the sub-layer output.

In Job Scheduling using fuzzy neural network algorithm [5], from the federal database, the training data is considered as an input. Then for the neural network this converted input is passed .Later the system resources mapping with user task are decided by neural network. In neural network, if the user tasks are not mapped with the System resources then using back propagation algorithm the tasks are reclassified and continue with the alike operations. The federal database is used for maintaining and updating the mapping of training data. For implementing dynamic resource allocation algorithm Hopfield Neural Network has been used [6].The inputs and output relationship of the different neurons is given by the function:

$$V_i=f(U_i)= 1/1+e^{-\alpha U_i}$$

where α is the gain of the neurons.

U_i and V_i are the input and output of the i-th neuron.

In a multi-tier distributed Systems the resource allocation is done by creating different virtual machines [7]. The problem is modeled such that each virtual machine is associated with SLA with profit P_{ij} .Using the concept of knapsack problem, a objective function is formed and based on this resource allocation is done.

Manuscript Received on January 2015.

Anitha N, Research Scholar VTU, Department of ISE, EPCET, Karnataka, Bangalore, India.

Dr. Anirban Basu, Research Head, Department of CSE, EPCET, Karnataka, Bangalore, India.

III. PROPOSED SOLUTION

The proposed solution for resource allocation dynamically [1] consists of 3 parts

1. Run time Instrumentation
2. Deep inspection using neural network
3. Resource management with VM migration

The instrumentation will contain information such as the applications current usage and the present running status. Each VM will have an instrumentation process running as agent which will collect information such as the Memory, CPU, Disk Usage and Bandwidth usage, IO read/write time waiting and bandwidth currently allocated. The data is gathered often and subject to deep inspection. The installing probes in the VM will send this information to the instrumentation process running. During the deep inspection stage, a comparison is made between the transaction per second(TPS) for running application and parameters of the resources usage and then assessment will be made to recognize if there is any need to increase or decrease the resources needed. In order to arrive at a standard resources needed, the resources baseline values to TPS mapping is kept. Here only the SaaS will provide the opportunity to calculate the baseline value, the software is then provided by the cloud service provider, the rigorous performance testing is done to arrive at a benchmark. But in situation when the user application of resource utilization vary depending on which order the processes are run, in this case we cannot get a proper benchmark for the utilization of resource. Soft computing algorithm such as the fuzzy logic or the Neural Network can be used to overcome these issues and understand the resources demands properly. To obtain quick response time feed-forward neural network is used. For different output resource, input TPS the network is trained and the best output value is provided based on that. The different parameters obtained from instrumentation step are provided as input to first layer. Then in inspection step the feed-forward neural network is assigned to decide whether to increase or decrease the resources needed. Let the allocated resource currently to the application be Q_R , *a* be the incremented step size and *b* be the decremented step size. For the next period time, the resource allocated for the application is calculated based on

$$Q_R = Q_R + at_i - b(t-t_i)$$

Where t_i is the time that resource was idle without usage.

IV. SYSTEM MODEL

The system model will mathematically analyze the proposed resource management algorithm. Let $V_1,V_2\ldots V_N$ be the VM running on Host $H_1,H_2\ldots H_N$ initially. For each VM there is expected TPS which depends on the application mapped to the VM as $E_{TP1},E_{TP2},\ldots E_{TPN}$. The objective of the resource management algorithm is such that each VM is able to meet its expected TPS. In order to meet the TPS, the resource has to be allocated from the host pool. The resource required for the VM will be dynamically allocated from the resource pool of corresponding host and when it is not possible to meet the resource requirement locally the VM will be migrated to other host which can meet it resource requirement. Let the current TPS of each VM be C_{TP1}, $C_{TP2},\ldots.C_{TPN}$. For any VM if the $E_{TPx} - C_{TPx} > \beta$, then resource management is necessary for that VM_x. β is the threshold for TPS tolerance. The neural network will be invoked with the current TPS and the expected TPS, to get the resource step size IR for each of the resource. If the $\sum$ (IR +CR) < TR, then no VM migration is needed.

Where

IR is the step size for the resource
CR is the current resource used at VM
TR is the total resource available at Host

Neural network will give the step size with in a time bound of Ω to give the step size IR. After the step size is given, the VM will be allocated resource according to step size. The time required for resource capture from pool is bound by λ. The total time needed to re-adjust the resource to meet the TPS is bound by $\Omega+\lambda$. As in any non linear systems there may be oscillation before the VM TPS value is stabilized around the expected TPS. The stabilization time ρ is dependent on the β as given below:

$$\rho = F(\beta) \qquad (1)$$

F is non linear function dependent on the resource variables.

In case the resource pool of the host in which VM is currently running is not able to meet the VM requirements, the decision to shift must be taken. But the decision must not be made based on one observation alone; it should be made on M consequent interval. In this case, the stability time is given as

$$\rho = M^*t_O + t_M + F(\beta) \qquad (2)$$

where

t_O is the observation time
t_M is the migration time.

The stabilization time for meeting the desired TPS is always guaranteed in our solution and also resource wastage is controlled thereby paving the way for better resource utilization by meeting the QOS.

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

The proposed algorithm based on dynamic instrumentation for resource allocation is implemented in cloud-sim simulator. In the cloud-sim the cloudlets have varied TPS based on poison distribution. The new scheduling mechanism is compared with existing resource allocation in cloud-sim such as round robin and throttled mechanism. With the listed below conditions, the simulation is conducted as tabulated in the table 1.

Table 1: configuration setup

No of data center	1
No of host in data center	50
TPS value for cloudlets	Poisson distribution from 1 MPS to 30 MPS
Initial No of VM	10
No of cloudlet	120
No of Resource Profile for Hosts	3
Resource Profile 1	60 MPS CPU, 1GB RAM, 60 GB disk
Resource Profile 2	80 MPS CPU, 2GB RAM, 120 GB disk
Resource Profile 3	100 MPS CPU, 4 GB RAM , 140 GB disk

The conformance of SLA is measured through the cloudlets which are meeting the response time. The simulated results as shown in fig 1 shows that the proposed systems has a better response time as compared with existing methods. It is observed from fig 1 that the response time is better when compared to the conditions as tabulated in table 1 of [1] and we could achieve this due to the virtual machine migration.

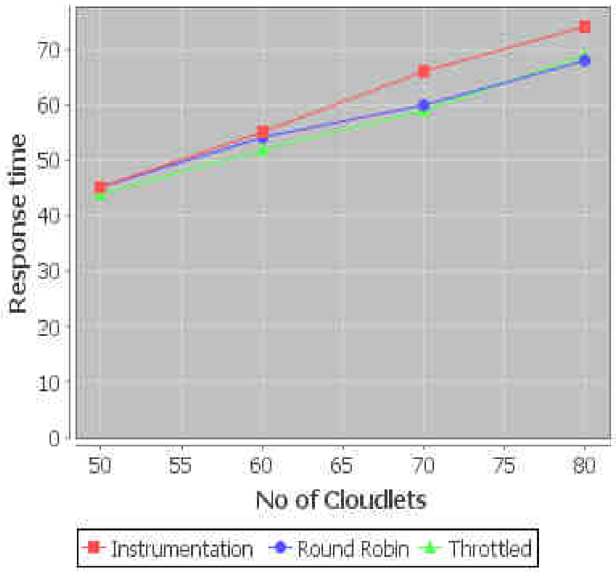

Fig. 1: Graph of Response Time

VI. CONCLUSION AND ENHANCEMENTS

The proposed mathematical system model proved to have better response time using neural network by instrumentation and also better resource utilization by controlling the resource wastage. Further advancements in cloud computing are leading to a promising future for collaborative cloud computing (CCC), where globally-scattered distributed cloud resources belonging to different organizations or individuals (i.e., entities) are collectively used in a cooperative manner to provide services. Due to the autonomous features of entities in CCC , the resource allocation is a complex. Hence in future we plan to propose a dynamic resource allocation mechanism to allocate resources to CCC cloud consumers.

REFERENCES

[1] Anitha N and Anirban Basu, “Dynamic Resource Allocation in Cloud using Runtime Instrumentation”, International conference on Communication and Computing ICC 2014 and Elsevier Science and Technology Publications June 2014, PP 482-490 **(Self referenced paper)**

[2] Gunho Lee, Niraj Tolia, Parthasarathy Ranganathan, and Randy H. Katz, “Topology aware resource allocation for data-intensive workloads”, ACMSIG COMM Computer communication Review, 41(1):120--124, 2011.

[3] Abhirami S.P and shalini Ramanathan, ”Linear Scheduling Strategy for Resource Allocation in Cloud Environment”, International Journal Of Cloud Computing: services and architecture ,Volume 2:N01 Feb 2012.

[4] Dr.T.R. Gopalakrishnan Nair, P Jayarekha,” Pre-allocation Strategies of Computational Resources in Cloud Computing using Adaptive Resonance Theory-2”,International Journal on Cloud Computing: Services and Architecture(IJCCSA),Vol.1, No.2, August 2011,PP 31-41.

[5] V. Venkatesa Kumar and K. Dinesh”, Job Scheduling Using Fuzzy Neural Network Algorithm in Cloud Environment”, Bonfring International Journal of Man Machine Interface, Vol. 2, No. 1, March 2012, PP 1-6

[6] Daniel Calabuig, José Monserrat, David Gómez-Barquero, and Narcís Cardona ,“Hopfield Neural Network Algorithm for Dynamic Resource Allocation in WCDMA Systems”, IEEE 2006 PP 40-44.

[7] Paolo Campegiani ,Universit`a di Roma Tor Vergata” A Genetic Algorithm to Solve the Virtual Machines Resources Allocation Problem in Multi-tier Distributed Systems”.