ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Efficient Dynamic Resource Management in Cloud

Sowmiya S M[1], Dr. R. Udayakumar*[2]

Assistant Professor, Dept. of Information Technology, Jerusalem College of Engineering, Chennai, Tamil Nadu, India

Associate Professor, Dept. of Information Technology, Bharath University, Chennai, Tamil Nadu, India

* Corresponding Author

ABSTRACT: Cloud computing is service sharing technology. One of the important characteristics of cloud is, it provides scalable resources needed for the application hosted on it.

As cloud-based services become more dynamic, resource provisioning becomes more challenging. In the proposed system, a power efficient resource allocation scheme using the bio-inspired ant colony framework is introduced; in which customers are willing to host their applications on the provider's cloud with a given SLA requirements for performance such as throughput and response time. Since, the data centers hosting the applications consume huge amounts of energy and cause huge operational costs, solutions that reduce energy consumption as well as operational costs are gaining importance. Thus, load balancing is done so as to transfer the load fairly among the available servers and also to hibernate the servers that are idle for a particular time. To optimize this approach, Load forecasting is done to allocate minimum number of active servers to meet the current demand with a less power consumption.

KEYWORDS: Bio-inspired, Cloud computing, Load balancing, Load forecasting, Resource allocation.

I. INTRODUCTION

Cloud computing is an emerging technology that is thought to realize the vision of utility computing where customers pay for computing services. Cloud customers no longer need to worry about the costs associated with under-provisioning and over provisioning. Many researchers focused mainly on hosting high performance applications in clouds without considering the energy efficiency. Since the energy costs are increasing, the need for optimizing cost of data center resources is also increasing. This will not only reduce the energy consumption but also the operational cost. So, cloud resources need to be allocated not only to satisfy QoS requirements specified by users through SLAs, but also to reduce energy usage.

The cloud servers load changes dynamically. So, an efficient mechanism is needed to take this dynamism into account and allocate the resources to the services so that minimum number of servers will be used for hosting the services. Thus, less operational cost and energy consumption are achieved. As the Ant colony mechanisms are helpful for adapting to dynamic behavior of the loads in the system, the above objective can be achieved using intelligent ant agents for monitoring.

This paper also attempts to impact the resource allocation problem by looking at the specific characteristics of the ability to predict future loads. The relationship between the ability to predict the future loads and accuracy of the load prediction must be analyzed in order to optimally allocate the cloud resources. The neural network prediction method is examined to forecast the future load demand profiles. So, the goal is to ensure that incoming requests are being properly serviced with a minimum amount of required power.[2] The rest of the paper is organized as follows. In the next section, we present some of the related work in this direction. Section III describes the system architecture used for

10.15680/ijircce.2015.0303269

the resource allocation. In section IV, the details about the system methodology is discussed. Section V outlines the future work and section VI concludes the paper.

.

II. RELATED WORK

Cloud computing has its root deep into ground and in the market. The evolution of cloud computing is one of the major advances in the computing area as well as in economics [7] of using computing. There are three major technologies which represent cloud computing: Platform-as-a-service (PaaS), Software-as-a-Service (SaaS) and Infrastructure-as-a-Service (IaaS). The Infrastructure-as-a-Service (IaaS) as a model of cloud computing service delivery represents hardware platform services. This kind of service enables scaling of bandwidth, memory, computing power and storage. The pricing is done according to leased computing power, bandwidth and storage.

Users of this platform can sign a contract with service provider for the certain amount of time or can rely on pay-as-you go model. The best side of this model is pricing compared to traditional way of computing. User pays for the resources that they use [3]. Virtualization lets a single resource (such as server, OS, application or storage device) appear as multiple logical resources; or making multiple physical resources (such as storage devices or servers) appears as a single logical resource. Adoption of virtualization technologies to data centre [5] environment helps to increase server utilization thus ensuring doing more with less servers. Specifically, the ability to dynamically distribute server workloads in a virtualized server environment and intelligent consolidation of virtualized machines (VM) on the physical servers helps turning off physical machines during periods of low activity, and bringing them back up when the demand increases.

Various ongoing research efforts are trying to reduce the power consumption of the data centre through Virtual Machines (VMs) migration with the goal of server consolidation [6]. The main idea is to execute the VMs on as few physical machines as possible to concentrate the workload and to efficiently use the physical servers. For instance, we can have two physical servers that run one VM each and both are not using their maximum computational capacity; hence, we decide to allocate both VMs in the same physical server, so that we can profitably switch one server off. It comes without saying that the power saving will be more considerable in large-scale data centers where several physical servers can be turned off to save more energy.

Resource allocation in a large-scale cloud environment can be configured for green computing objectives under CPU and memory constraints. A generic gossip protocol (GRMP-Q), [1] is proposed for resource, which can be instantiated to aim at minimizing power consumption through server consolidation, while satisfying a changing load pattern.[5]

The increasing demand for cloud computing resources has led to a commensurate increase in the operating power consumption of the systems that comprise the cloud. A novel framework [2] was introduced in which it combines both load demand prediction and stochastic state transition models. This model lead to optimal cloud resource allocation by minimizing energy consumed while maintaining required performance levels.

Many researchers implemented IAAS in their applications with Eucalyptus for its user friendly GUI and compatibility. EUCALYPTUS [3] – an open source software framework for cloud computing that implements Infrastructure as a Service (IaaS); systems that give users the ability to run and control entire virtual machine instances deployed across a variety physical resources.

Node Controller controls the execution, inspection, and terminating of VM instances on the host where it runs.
Cluster Controller gathers information about and schedules VM execution on specific node controllers, as well as manages virtual instance network.

Storage Controller (Walrus) is a put/get storage service that implements Amazon's S3 interface, providing a mechanism for storing and accessing virtual machine images and user data.

Cloud Controller is the entry-point into the cloud for users and administrators. It queries node managers for information about resources, makes high level scheduling decisions, and implements them by making requests to cluster controllers.

III. SYSTEM ARCHITECTURE

The main aim of our resource allocation is to allocate the online service requests for applications which are CPU and memory intensive. To achieve the objective of adaptive resource allocation for satisfying the service requests of customers, we use the following architecture. [4]

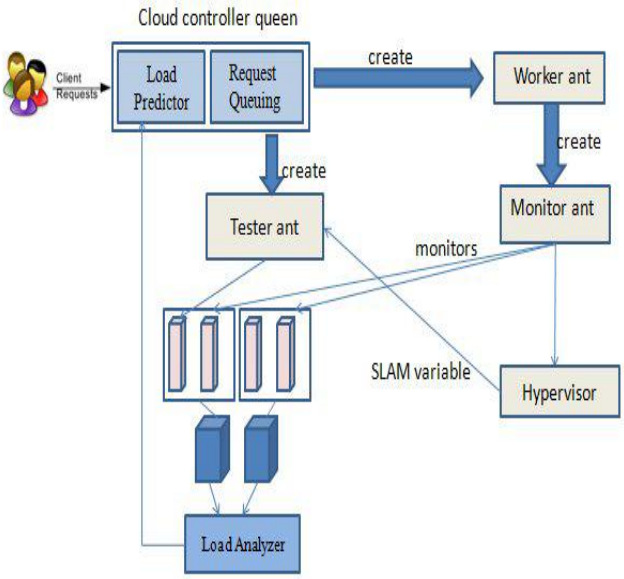

Fig 1 Cloud Architecture with different ant agents

The three ant agents used in this architecture are worker ant, monitor ant and tester ant. The architecture describes how the cloud controller acts as an interface between the cloud service provider and external users. Based on the current load, the load analyzer analyses the load and feeds it as an input to the load predictor for future load prediction.

Users/Brokers: Users or brokers acting on their behalf submit service requests to the cloud via cloud controller for processing.
Cloud Controller: It acts as the interface between the cloud service provider and external users/brokers. It acts similar to the Queen in the ant colony.
Virtual Machines (VMs): This is where the applications of customers will be deployed. We can dynamically create, start,
stop and migrate these VMs depending on our requirement, from one physical machine to another.
Physical Machines: These are the physical computing servers that will provide hardware infrastructure for creating virtual machines.

IV. SYSTEM METHODOLOGY

The power consumption of each server in the data center along with the resource capabilities such as CPU processing power and primary memory are gathered before admitting them into cloud. This information consisting of Node Id, Processing Power, Memory and Power Consumption are stored in a table called Cloud Resource Table.

The request from the user is queued in the queen cloud controller. To launch a virtual machine for the user request an appropriate image for the user request is chosen and based on the key-pair, an instance is created for the user request with a particular virtual machine type.

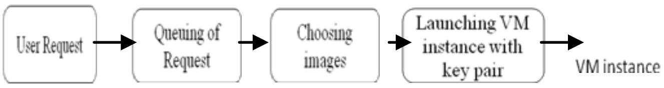

Fig 2 Virtual machine instance creation

A. Cloud Controller & Queen Ant:

The requests from the customers consisting of the following,
are given to the controller.
(i) Throughput (THPUT) (In %)
(ii) Avg. Response Time(RTIME)
(iii) Application Code
(iv) Operating System

Cloud controller maintains a queue(Q) for storing the service requests for hosting the applications. It enqueues each of the service request received, in this queue.[7]

It generates the tester and worker ants periodically. The movement of these ant agents is modeled in the following way. The tester ant maintains a Visited Node list which is initially empty. Each node in the cloud maintains a list of neighboring node's information.[8] Whenever the tester ant reaches a node, it updates the controller about the current utilization and randomly chooses an unvisited neighboring node. When all the nodes are covered, it makes the Visited Node list empty and continues again in the same way.

B. Worker Ant:

Whenever a service request received in the queue, one of the worker ants creates a VM with a specific CPU processing power and memory etc, if accepted. So, worker ants are always looking in the queue to check if there are some pending requests to be processed. The worker ant is only responsible for deploying the request on a VM.

C. Monitor Agent:

The Monitor ant monitors the hosted application. It calculates the average response time and throughput of the hosted application. Based on the response time and throughput Monitor ant calculates the MO variable. The MO variable is used for load balancing among the nodes in the cluster. Based on the MO variable the tester ant decides whether to balance the load or not. The MO variable can have three values:

□ 0(No Balancing)
□ 1(Recommended for Balancing)
□ 2(Need for Balancing)

The MO variable values are calculated based on the SLA Response time, throughput and the response time and throughput of the node as follows:

MO variable 0
Response time 10% less and throughput 10% more than SLA
MO variable 1
Response time 5-10% less and throughput 5-10% more than SLA

MO variable 2
Response time and throughput reaching SLA .

D. *Tester Ant*

The main job of the tester ants is to get the utilization and power consumption information from each of the node and to update the available node's list. It also takes the load balancing decisions.[6]

It is assumed that if the CPU and memory utilization is below 80% it is assumed that the node does not need any balancing. If the CPU utilization is 80% and memory utilization is 80% in a node it is considered to be a desirable utilization and if these are above 90% then it is considered to be peak. Based on these criteria MO variable is calculated.

If the *MO variable is 2* the following algorithm is called

```
1. Search for nodes in cloud resource table that have remaining resources as 50%
2. If nodes are available having enough resources then
      a. Select the first node
      b. If the selected node is standby node then
            i. Wake-up the selected node
3. Migrate to the new created VM
4. Configure the router to share the request with the new VM created
```

If the *MO variable is 1*, the following algorithm is called

```
1. Search for nodes in cloud resource table that have remaining resources as 30%
2. If nodes are available having enough resources then
      a. Select the first node
3. Migrate the VM
4. Configure the router to share the request with the new VM created
```

To sort the nodes in descending order

$$\frac{\text{Processing power of the node (in Ghz)}}{\text{Power consumption of CPU (in Watts)}} = \text{PPW}$$

$$\frac{\text{Memory capacity of node (in Gb)}}{\text{Power consumption of Memory (in Watts)}} = \text{MPW}$$

PPW = Processing Power per Watt,
MPW = Memory Consumption per Watt

E. *Load Forecasting*

The load predicting device is responsible for predicting the incoming load and the load analyzing device examines the current performance of all the available cloud resource nodes. The cloud load predictor architecture is given below:[9]

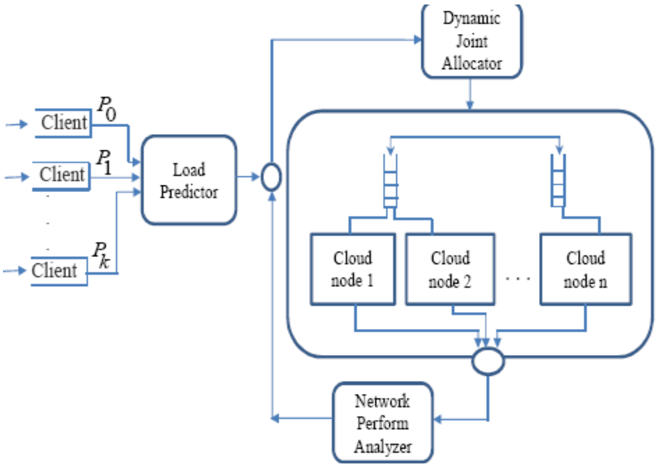

Fig 3 Cloud Load Predictor Architecture

Load prediction is done using neural network predictor. The neural network perceptrons are applied to predict the future load of applications running in the cloud.[10] The perceptrons are trained in a supervised manner with back propagation algorithm. Perceptrons uses the values: Number of external inputs, one internal input, Threshold and produces a single output. All the inputs have weights attached to the input patterns that modify the input values to the neural network.

The main feature of perceptrons is that they can be trained to behave in a certain way. [11]So the perceptrons learns like this: it produces an output, compares the output to what the output should be, and then adjusts itself a little bit. After repeating this cycle enough times, the perceptrons will have converged to the correct behavior.[12]

V. Future Directions

The future work includes developing a version of the protocol for a heterogeneous cloud environment in which CPU and memory capacities vary across machines, develop a distributed mechanism that efficiently places new sites and make the protocol robust to machine failures.[13]

VI. Conclusion

With this proposed system, a significant contribution towards a resource management using eucalyptus middleware for cloud environments is engineered. The components of the middleware and load forecasting can be used to meet the design goals for resource management. Bio-inspired computing produces a solution to the resource allocation problem for a dynamically changing resource demand.

REFERENCES

[1] Mike Spreitzer, Fetahi Wuhib and Rolf Stadler(2012) A Gossip Protocol for Dynamic Resource Management in Large Cloud Environments, IEEE transactions on Network and Service Management.
[2] Sree Latha R., Vijayaraj R., Azhagiya Singam E.R., Chitra K., Subramanian V., "3D-QSAR and Docking Studies on the HEPT Derivatives of HIV-1 Reverse Transcriptase", Chemical Biology and Drug Design, ISSN : 1747-0285, 78(3) (2011) pp.418-426.
[3] John J.Prevost, KranthiManoj Nagothu, Brian Kelley and Mo Jamshidi, (2011). Prediction of Cloud Data Center Networks Loads Using Stochastic and Neural networks, System of Systems Engineering (SoSE), 6th International Conference .
[4] Masthan K.M.K., Aravindha Babu N., Dash K.C., Elumalai M., "Advanced diagnostic aids in oral cancer", Asian Pacific Journal of Cancer Prevention, ISSN: 1513-7368, 13(8) (2012) pp.3573-3576.
[5] Nurmi, D.; Wolski, R.; Grzegorczyk, C.; Obertelli, G.; Soman, S.; Youseff, L.; Zagorodnov, D(2009) The Eucalyptus Open Source Cloud Computing System in the proceedings of Cluster Computing and the Grid, 9th IEEE/ACM International Symposium.
[6] S.Bhardwaj, L. Jain, S. Jain, Cloud Computing(2010): A Study of Infrastructure as a Service (IAAS), International Journal of Engineering and Information Technology .
[7] Tamilselvi N., Dhamotharan R., Krishnamoorthy P., Shivakumar, "Anatomical studies of Indigofera aspalathoides Vahl (Fabaceae)", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384 , 3(2) (2011) pp.738-746.

[8] V.K. Mohan Raj, R.Shriram (2012) A Study on Server Sleep State Transition to Reduce Power Consumption in a Virtualized Server Cluster Environment, Communication Systems and Networks, 4th International Conference.

[9] Devi M., Jeyanthi Rebecca L., Sumathy S., "Bactericidal activity of the lactic acid bacteria Lactobacillus delbreukii", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384 , 5(2) (2013) pp.176-180.

[10] Antonio Corradi, Mario Fanelli, Luca Foschini (2011) Increasing Cloud Power Efficiency through Consolidation Techniques, Computers and Communications, IEEE Symposium.

[11] Lizhe Wang, Jie Tao, Marcel Kunze, Alvaro Canales Castellanos, David Kramer, Wolfgang Karl(2008), Scientific Cloud Computing: Early Definition and Experience in the proceedings of High Performance Computing and Communications, 10th IEEE International Conference.

[12] Reddy Seshadri V., Suchitra M.M., Reddy Y.M., Reddy Prabhakar E., "Beneficial and detrimental actions of free radicals: A review", Journal of Global Pharma Technology, ISSN : 0975-8542, 2(5) (2010) pp.3-11.

[13]B Karthik, TVUK Kumar, A Selvaraj, Test Data Compression Architecture for Lowpower VLSI Testing, World Applied Sciences Journal 29 (8), PP 1035-1038, 2014.

[14].M.Sundararajan .Lakshmi,"Biometric Security system using Face Recognition", Publication of International Journal of Pattern Recognition and Research. July 2009 pp. 125-134.

[15].M.Sundararajan," Optical Sensor Based Instrumentation for correlative analysis of Human ECG and Breathing Signal", Publication of International Journal of Electronics Engineering Research, Research India Publication, Volume 1 Number 4(2009). Pp 287-298.

[16]C.Lakshmi & Dr.M.Sundararajan**, "**The Chernoff Criterion Based Common Vector Method: A Novel Quadratic Subspace Classifier for Face Recognition**"** Indian Research Review, Vol.1, No.1, Dec, 2009.

[17]M.Sundararajan & P.Manikandan," Discrete wavelet features extractions for Iris recognition based biometric Security", Publication of International Journal of Electronics Engineering Research, Research India Publication, Volume 2 Number 2(2010).pp. 237-241.

[18]M.Sundararajan, C.Lakshmi & .M.Ponnavaikko, "Improved kernel common vector method for face recognition varying in background conditions", proceeding of Springer – LNCS 6026- pp.175-186 (2010).ISSN 0302-9743**.(Ref. Jor – Anne-II)**