A Load Balancing Approach for Increasing the Resource Utilization by Minimizing the Number of Active Servers

Govind Singhi[*], Damodar Tiwari[#]
govindsinghi1223@gmail.com[*], damodarptiwari21@gmail.com[#]

Abstract- **Cloud computing is one of the fastest growing technologies in the field of computer science. It became so popular due to the online, cheap and pay as use scheme. It is business model, so provider wants to achieve more and more profit. Resource utilization is one of the effective way to increase the provider because if we increase the resource utilization, it will lead to minimize the number of server result in reduce the energy consumption. This approach proposed a load balancing approach for the cloud environment that minimizes the number of active server and increase the profit of the provider. In order to increase the profit this approach first place the VM to the PM which gives to higher profit and during the time of migration we select VM which gives lower profit Our proposed approach uses two static lower and upper thresholds. Experiment result shows that proposed approach minimize the number of active server.**

Key Terms- Distributed computing, On demand resources, Cloud computing, Virtualization, Upper and lower threshold, VM consolidation.

1. INTRODUCTION

In the cloud computing, the computing resources are provided to the client through virtualization, via internet. The large scale computing infrastructure is established by cloud providers to make availability of online computing services in flexible manner so the user find easiness to use the computing services [1].

According to NIST cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources. The computing resources include networks, servers, storage, applications, and services. In cloud computing, the shared pool of computing resources can be rapidly provisioned and released [2].

In cloud, several physical machines (PM) are connected to each other in the form of cluster. Virtualization [3-6] is the enabling technology in the cloud, which divide the physical resources to the multiple parts via virtual machine (VM). When any user needs resources scheduler assign the resources of these PM to the user through the VM. Each user has its own VM and the resource requirement of the VM can be change dynamically at run time. Due to this load balancing in the cloud is the challenging task. To balance the PM, VM migration approach is used which transfer the VM from one PM to another PM.

This paper, we discuss the overview of cloud computing with their components basic model. The goal of the paper is provide a complete study of cloud computing with different types. In section 2, we discuss the background knowledge of the cloud computing with their framework. Section 3 gives brief descriptions about types of cloud as public, private, community and hybrid cloud. In section 4, we present the types of cloud services as software as a service, platform as a service and infrastructure or hardware as a service. Section 5 discuss about different advantages and challenges of cloud computing. Section 6 concludes the paper with the focus on the future work.

2. DELIVERY AND DEPLOYMENT MODEL

In the cloud computing there are three types of service delivery model [5] as software as a service (SAAS), platform as a service (PAAS), and infrastructure as a service (also known as hardware as a service). It can be design by three different model i.e., private cloud, public, hybrid and community cloud as shows in Figure 1.

In Software-as-a-Service (SaaS) delivery model only software is provide on demand to the client. There is no need to installed software on the client side. In Platform-as-a-Service (PaaS) delivery model complete platform which required designing new application is provide to the client. It is mainly use by the developer. In infrastructure-as-a-Service (IaaS) delivery model complete computing environment i.e., hardware, software, network etc., are provide to the user,

There are four types of cloud deployment model [7] in the cloud computing known as public, private, community and hybrid cloud.

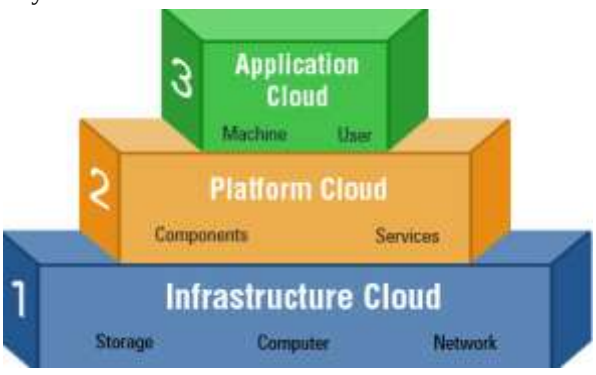

Figure 1: Cloud Computing Models

In private cloud all computer are connected locally. Service running in private cloud cannot be access from outside the network. It is more secure and less scalable as compare to the other cloud.

Public cloud is model of cloud where all users are allowed to access the services using internet. The user need only internet connection and web browser to access with pay per use scheme. All the services with infrastructure of cloud provider are available on the internet. User need to subscribe the application and make enable to use it. Community cloud includes number of organization to share their services to increase resource utilization of cloud infrastructure.

The cloud infrastructure is not limited to only one organization. Hybrid cloud combines both public and private cloud with their advantages. Hybrid cloud offers the benefits of both the public and private cloud. The hybrid cloud is the good solution for purely business oriented concept because many modern businesses have a wide range of concerns to support users demand.

3. CLOUD ATTRIBUTE

The cloud computing have various types of characteristics and attributes [8, 9]. The first most famous attributes is pay per use concept. The basic characteristics are given below.

On-demand self-service: in the cloud computing there I no human interaction. Ever thing is in the form of online services.

Broad network access: The cloud users have many of the option to access the services. There is many of the service providers who offer the services with effective service cost model.

Resource pooling: The cloud computing resources are pooled to serve large number of cloud user with virtual resources. Resources include storage media; processing capacity, primary memory, and network bandwidth etc are available to their client.

Rapid elasticity: As the cloud is a distributed collection of server and datacenter so capabilities are increasable with minimum effort. When consumer demand increases then cloud provide can increase the resource capacity by adding a new datacenter or server at any time.

Measured service: Cloud systems control and optimize all the installation and configuration related issues with automatically. Resource utilization, process scheduling and other infrastructure management work can be monitored, controlled, and reported only automated system and providing transparency for both the provider and consumer.

4. LITERATURE SURVEY

Y. Song [10] et al. describes about the allocation policy. The paper has introduced a RAINBOW prototype through which multi tier resource scheduling is done. The allocation of resource based of priority. The evaluation result says that paper is capable of improving resource allocation for both critical and less priority jobs. Problem with method is that it does not apply virtual machine migration policy for the optimization.

D. Gmach [11] et al. describes about the threshold based approach a threshold based reactive approach to dynamic workload handling. The paper has tired detecting the underutilized and over utilized work load and initiates migration as required. This approach is not much suited for IaaS environment.

A. Beloglazov et al. [12], proposed load balancing approach for the cloud. The proposed energy-aware allocation heuristics provision data center resources to client applications in a way that improves energy efficiency of the data center, while delivering the negotiated Quality of Service (QoS). In this approach they are using double threshold value for the load balance. When the utilization is beyond the upper level threshold system will be consider as an overloaded system. Similarly when the utilization is beyond the lower threshold system is called to be underloaded.

T. Wood et al. [13], present Sandpiper, a system that automates the task of monitoring and detecting hotspots, determining a new mapping of physical to virtual resources and initiating the necessary migrations. Sandpiper implements a black-box approach that is fully OS- and application-agnostic and a gray-box approach that exploits OS- and application-level statistics.

5. PROPOSED WORK

In cloud each data center can have number of PM and each PM has several VM. Load on the PM is depends on the VM load which can be change dynamic. Load balancing in the cloud is a challenging task because load on PM is change very frequently.

To balance the PM we are using lower and upper threshold where lower threshold shows the underloaded situation and upper threshold represent the overloaded situation. Value of these thresholds can be static or dynamic. In our proposed approach we are using static value of the lower and upper threshold which is calculated on the basis of number of migration.

Since each PM have several PM, so load on the PM is depends on the number of the VM in the PM. CPU, RAM and bandwidth are the main resources in the cloud, but PM performance is mainly depends on the CPU utilization. Hence to calculate the load on the PM we considered only CPU utilization as a decision parameter. CPU utilization of the VM is given by the following equation:

$$VL_{load} = \frac{\text{TotalRequestedMips}}{\text{Total MIPS of the host}}$$

Based on the above equation we calculate the load of each VM. Now total load on the PM is given by the following formula:

$$HL_n = \frac{\sum_{i=1}^{m} VL_i}{m}$$

Where-

m is the number of VM into the PM
VL_i is the load of the i^{th} VM.
HL_n is the load of the n^{th} PM

After calculating the load on the PM we compare it with lower and upper threshold of the PM. If load on the PM is larger than it upper threshold then PM is overloaded and some VM has to be migrated. Similarly when load on the PM is lower than its lower threshold the PM is underloaded and all VM running on it will be migrated known as server consolidation. Hence in both cases CM migration technique is used for balance the PM.

VM migration is one of the important features of the cloud; because it allows the provider to transfer the running VM from one PM to another PM. VM migration involved five steps:

i. Calculate order in which VM is scheduled
ii. Load estimation on the PM and VM.
iii. Estimation of lower and upper limits
iv. Calculate future load by using AR prediction model
v. Candidate VM selection for the migration by using optimization technique.
vi. Selection of candidate PM for hosting the selected VM.

5.1. Calculate order in which VM is scheduled

In order to calculate the order of the VM in which they are placed, our approach uses the cost which they will gives to the provider for using their resources. Following equation is used to calculate the cost

$$T_i = \frac{x}{\text{MIPS of the VM} * 0.7}$$

5.2 Load Estimation on the PM and VM

In cloud load on the PM is propositional to the number of VM running on the PM. Following equation is use d to find load.

$$H_{load} = \sum_{i=1}^{n} \frac{VLi}{n}$$

Where-

n = number of VM
VL_i = load of i VM.

5.3 Estimation of lower and upper limits

Proposed approach used 20 and 80 as a lower and upper approach.

5.4 Calculate the future load by using AR prediction model

In order to avoid the migration due to the temporary peak load our approach uses AR prediction model.

$X_t = C + \emptyset_1 X_{t-1} + + \emptyset_2 X_{t-2} + + \emptyset_3 X_{t-3}$ ----------

--$+ \emptyset_p X_{t-p} + \in_t$

Where-

$C = (1 - \sum_{i=1}^{p} \emptyset_i)\ \mu$
P= Order of the AR model
μ = mean of the history data
$\emptyset_i$ = weighting coefficient
$\in_t$ = white noise

5.5 Candidate VM selection for the migration by using optimization technique

Step-1: Calculate profit for each VM to place the VM
Step-2: for each host in the hostList do
Step-3: PM_{load}←host.util()
Step-4: if PM_{load}< lower_threshold
Step-5: Calculate the load after 10 second
Step-6: If load is still less than lower threshold then
Step-7: Move or migrate all VM in order to save power.
Step-8: else
Step-9: Not migrate the VM
Step-10: end if
Step-11: end if
Step-12: if PM_{load}> upper_threshold
Step-13: Calculate the load after 10 second
Step-14: If load is still greater than upper threshold then
Step-15: Arrange all VM into decreasing order of priority
Step-16: Choose the last VM (VM with lower value) form the VMList and migrate them.
Step- 17: end if
Step-18: end for

5.6 Selection of the candidate PM for hosting the selected VM.

Step-1: Add all new and migrated VM in the VMList
Step-2: Calculate profit for each VM
Step-1: For all VM in VMList do
Seep-2: Select VM with higher profit
Step-3: for all PM in PMList
Step-4: PMload←host.util()
Step-5: if PMload > lower_threshold && PMload <upper_threshold
Step-6: Add PM into the newPMList
Step-7: end if
Step- 8: for all PM in the newPMList do
Step-9: Select largest PM and placed VM with higher profit
Step-10:end for
Step-11: end for

6. RESULT EVALUATION

To evaluate the performance of the propose approach CLoudSim simulator [14] is used. . We create one center and the number of PM in this data center is vary i.e., 20 PM. Each PM in the data center has 1000, 2000 and 3000 MIPS, 10000MB of RAM and bandwidth of 100000 bit/s. For each virtual machine on host ram size is 128, 512 and 1024 MB and bandwidth size is 2500 bit/sec.

MIPS for each VM generated randomly between 250, 500, 750 and 1000. First and second experiment is performed to find the value of lower and upper thresholds respectively.

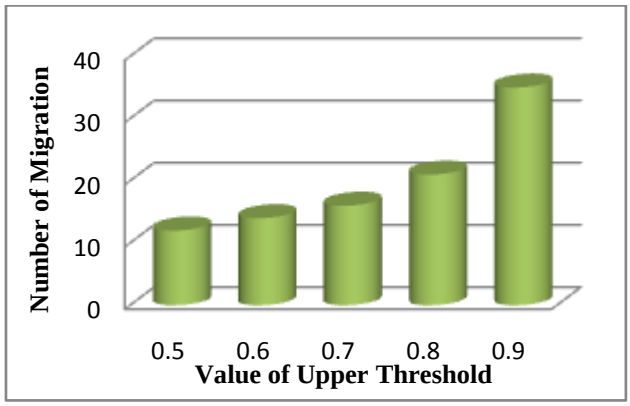

Figure 2: Number of Migration v/s Different Values of Upper Threshold

Figure 2 shows the number of migration for the different value of lower and upper threshold. As shown in graph, number of migration increases with thresholds values but when the value of upper CPU threshold is change to 0.8 to 0.9, number of migration is increasing dramatically. So we set the value of upper threshold is 0.8 or 80%.

$$T_{upper} = 80$$

After deciding the value of CPU lower and upper threshold, we will check the number of active server for placing 40, 45 and 50 VM in our propose approach and competitive approach.

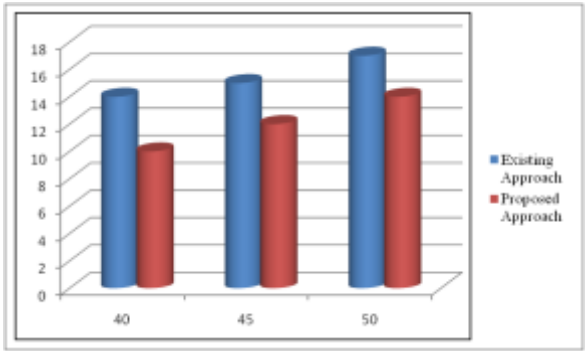

Figure 3: Number of Active PM for 20 VM

Figure 3 shows that minimum numbers of active servers are required in our proposed approach. Hence it increases the resource utilization.

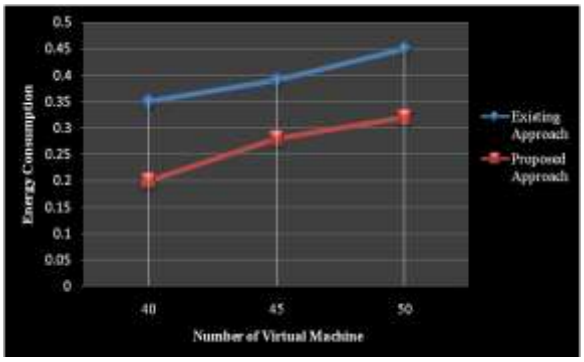

Figure 4: Energy Consumed by the Datacenter

Figure 4 shows the energy consumed by the PM when 40, 45 and 50 VM are placed on the PM.

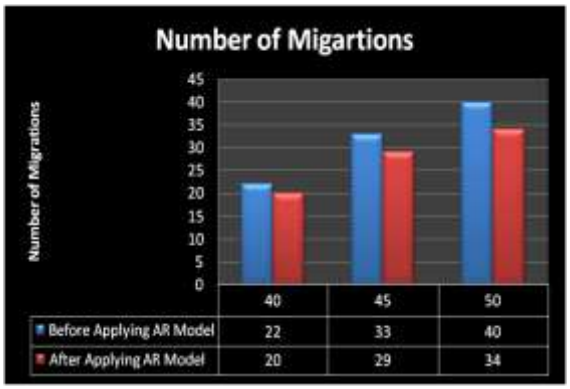

Figure 5: Number of Migrations in Proposed Approach

Figure 5 shows comparison in the number of migration before applying the AR model and the number of migration after applying the AR model. After applying the AR model number of migration are reduced.

7. CONCLUSION

This paper proposed load balancing approach for the cloud. VM are migrated when the load on the PM is less than the lower or greater than its upper threshold. This approach uses static value of lower and upper threshold i.e., 20 and 80 respectively. These threshold values are calculated based on the number of migration. Proposed approach minimum numbers of active servers are required in our proposed approach. Hence it increases the resource utilization.

Experimental result shows that proposed approach gives better results and minimizes the number of active servers and number of migrations. It also used the AR prediction model to minimize the number of migrations.

REFERENCES

[1]. "Cloud computing," [Online] available: http://en.wikipedia.org/wiki/ Cloud_computing, July 2014.

[2]. R. K. Gupta and R. K. Pateriya, "A Complete Theoretical Review on Virtual Machine Migration in

Cloud Environment", International Journal of Cloud Computing and Services Science (IJ-CLOSER), Vol.3, No.3, June 2014, pp. 172-178.

[3]. R. K. Gupta and R.K. Pateriya," Survey on Virtual Machine Placement Techniques in Cloud Computing Environment", International Journal on Cloud Computing: Services and Architecture (IJCCSA) ,Vol. 4, No. 4, August 2014, pp. 1 -7.

[4]. "Hypervisor," [Online] available: http://en.wikipedia.org/wiki/ hypervisor, June 2014

[5]. R. Santhosh and T. Ravichandran, "A Survey on Cloud-Based Scheduling Algorithms", International Journal of Communications and Engineering (IJCE), Volume 01– No.1, Issue: 01 May2013.

[6]. D. R. Sahu and D. S. Tomar, "Analysis of Web Application Code Vulnerabilities using Secure Coding Standards", Springer, Arabian Journal for Science and Engineering, Vol. 42, Issue 2, pp. 885–895, 2016.

[7]. S. K. Mandal and P. M. Khilar, "Efficient Virtual Machine Placement for On-Demand Access to Infrastructure Resources in Cloud Computing", International Journal of Computer Applications (IJCA), Vol. 68, No.12, April 2013.

[8]. "Xen, virtual machine manager in Cloud computing," [online] available: http://www.xen.org, April 2013.

[9]. A. Jain et Al., "A Threshold Band Based Model For Automatic Load Balancing in Cloud Environment", in proc. of IEEE International Conference on Cloud Computing in Emerging Markets, pp 1-7, 2013.

[10]. Y. Song, "Multi-Tiered On-Demand resource scheduling for VM-Based data center" In Proc. of the 2009 9th IEEE/ACM Intl. Symp. on Cluster Computing,155, 2009.

[11]. D. Gmach, "Resource pool management: Reactive versus proactive or let Ss be friends". Computer Networks, 2009

[12]. A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers". 10th IEEE/ACM Intl. Symp. on Cluster, Cloud and Grid Computing ,2010.

[13]. T. Wood, et Al., "Black-box and gray-box strategies for virtual machine migration" Proceedings of the 4th USENIX conference on Networked systems design & implementation, 2005. PP. 17-28.

[14]. R. Calheiros, et Al. "CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services", 2011.

