Group Based Load Balancing Algorithm in Cloud Computing Virtualization

[1]Rishi Bhardwaj, [2]Sangeeta Mittal,
[1]Student, [2]Assistant Professor,
[1] Department of Computer Science,
[1] Jaypee Institute of Information Technology
Noida, India

***Abstract* - Efficient allocation of tasks to available resources is a critical problem in cloud computing. Load balancing algorithms are basic approach to optimized resource allocation. This paper presents a static load balancing algorithm to allocate tasks to virtual machines in optimized way. The algorithm is based on a group based allocation strategy. The strategy determines the allocation of virtual machines in a group based on type of task and the configuration of the virtual machine. Using this strategy running cost of virtual machines is optimized. Compared to other algorithms like round robin and weighted round robin, proposed algorithm achieves better CPU utilization.**

***Index Terms* - Cloud Computing; Virtualization; XEN; load balancing.**

I. INTRODUCTION

With the speedy development in the cloud computing area as more and more users are using the internet. It has become important to use optimal solution for every task. The growth of user request to the server is becoming high, which require the higher computing at the server end. The host server requires to response the user with the minimum response time in order to cut the cost and improve the user experience.

System virtual machine are widely used from personal computer to large organization. System virtualization act as powerful means of abstraction of upcoming applications upon which shared resources can be flexibly allocated to VM-hosted applications. The emergence of virtual machine has brought tremendous effect on computing field by migrating the task to virtual machine that have memory according to the task requirement. Cloud systems are unique in their unprecedented ability of elasticity, i.e. the ability to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible.

With the help of virtualization, we create the virtual system that can be called as virtual cluster. The virtual cluster means binding homogenous and heterogeneous system together and then virtualizing multi-computing resources out of one computing resource via virtualization, thus constructing a huge cluster system that is similar to the physical cluster.

There are thousands of computing tasks on cloud computing environment. It will be a tough task of validating these tasks and allocating these tasks to a virtual machine. The use of tradition load balancing algorithm is not sufficient as no. Of requests increase. Load balancing is the way to produce resources and maintaining parallelism. Cutting down the response time and increasing the throughput is the goal of the load balancing. Load balancing has three rules that are Location rule, Distribution rule, Selection Rule. Cloud infrastructure has various types of load on a system that includes CPU load, Memory Load, Network Load.

The contribution of this paper is as follows:

- Minimizing the response time: It allows user requests to be responded as soon as possible. This will decrease the total time for requests.
- Maximum system throughput: It achieves the maximum performance and maximum resource utilization to avoid the cost as much as possible.
- Ensuring the quality of the service based on the different-2 user request and allocating the request to the relevant virtual machine.
- Minimizing the system overhead caused by the network file system and load balancing algorithms.

The rest of this paper is organized as follows. Section 2 describes related work in cloud computing and load balance. The architecture and the design of our system are introduced in Section 3. Section 4 is the experimental result and evaluation. Section 5 is the conclusion.

II. RELATED WORK

A. Cloud Computing

Currently cloud computing is the technology that is widely going to use in every area of information technology. Cloud computing is the development of parallel computing, distributed computing and grid computing and is the combination of the other services like virtualization, software-as-a-service (SaaS), Platform-as-a-Service (Pass) and Infrastructure-as-a Service (IaaS). The service of cloud computing pretty much depends on user requirement. A user has to pay for the shared IT resources

like network, server, and storage. The main benefit of the cloud computing is its scalability and quick processing for rapid development.

1) *Scalable according to requirement:* The cloud infrastructure is scalable according to the requirement. All the hardware resources that need are scalable for the user and service providers also. Cloud computing is a very large infrastructure and its size increasing day by day. Google cloud computing have millions of servers for its users, IBM, Amazon and other service providers are also have a very high number of servers.

2) *Virtualization:* Virtualization is the main element of cloud computing arena. All the resources provided by cloud computing are provided through virtualization technology. Virtualization reduces the cost and provide a basic element for coupling between the hardware and software.

3) Multi node architecture: Cloud computing provides the multi node architecture for the storage and data processing. So if there is a single node failure, then it will not influence the whole process. Cloud computing also provides higher reliability for data.

4) *Universality:* The services provided by the cloud computing are not customized for a specific application. The users can choose the different application according to their needs. A system can run different types of application.

5) *Scalability:* Cloud platform provide the scalability according to user requirement. We can change resources dynamically according to the number of users and type of applications.

B. Load Balancing

In cloud computing, Load balancing distributes the workload among the multiple computing nodes, processes, disk, or other resources in order to get optimal resource utilization, maximizing throughput, minimizing response time and for reducing the computation time for a task. The main aim to load balancing is to achieve the effective resource utilization. Load balancing has been the hot topic in cloud computing, grid computing and distributed computing.

Load balancing has different meanings: First, it puts a large number of data processing and concurrent access to multiple nodes to reduce the time users waiting for the response. Second, it divides the workload from heavy loaded node to multiple nodes to improve the resource utilization of each node and to reduce the time in processing of task. Load balancing have simple meaning like assigning tasks to each person rather than assigning tasks to a single person. All these are done by monitoring the system in real time for CPU utilization, memory of VMs, network speed and the weight factor of virtual machine. After evaluating the VMs with the help of all these data, load balancer allocates the task to the specific node. Allocation strategy is varies from application to application, like for opening Firefox in a virtual machine we do not require very high configuration virtual machine, but for application like some high graphic game or Adobe Photoshop type of heavy application we require virtual machine that have high configuration. Load balancing can be achieved at platform layer, server layer and protocol layer. Load balancing can be achieved with the help of some load balancing tools. The main objective of load balancer is to improve the utilization of the system resources, reducing the cost and increasing the system response time for users.

There are two types of load balancing algorithms.

1) *Static load balancing:* In static load balancing algorithm we know the size of the task to be allocated to the virtual machine so our load balancer can calculate that which virtual machine is most suitable for that task. That is done by evaluating the load balancing information of the each virtual machine and checking that how much memory utilization required by the task. So basically static balancing is the distribution of the task prior to the execution of the task to a virtual machine. The static load balancing algorithm is easy to implement on systems and provide better result in case we know the size of the task. Some basic static load balancing algorithm includes the Round Robin, randomized algorithm etc.

2) *Dynamic load balancing:* There are some problems in which we cannot statically partition the work among the processors. In these types of problems static partition is not possible or can lead to serious imbalance of load in the system. To solve that issue we use to work with dynamic load balancing in virtual machine. So during the execution of process work is dynamically transferred among the virtual machines form a virtual machine that have heavy workload to the virtual machine that have light workload. The all the process is done with the help of virtual machine migration in the virtual machines. Dynamic load balancing have centralized and decentralized type of algorithm where in centralized case task is handed out from centralized location.

A load balancing algorithm is generally divided in to four steps.

- Load measurement of a node
- Load information gathering
- Initiation rule
- Operational rule

III. ALGORITHM AND LOAD CALCULATION

A. Architecture of the System

According to the load of the each system, we are going to allocate the task to the virtual machine that have load value less than the threshold value of the load. We have set our threshold value to 70% CPU utilization of the system. If VM load value is less than this value than we will allocate the task to the virtual machine.

We have a load monitor in our system that keep monitoring the load of all the virtual machine in a periodic manner and keep updating the load value in the database. The host operating system is calculating the load of all the virtual machine and host operating system generates the request to the guest operating system. A request can be like opening the Firefox browser, a text file, word file, or any other work that a Linux system can perform.

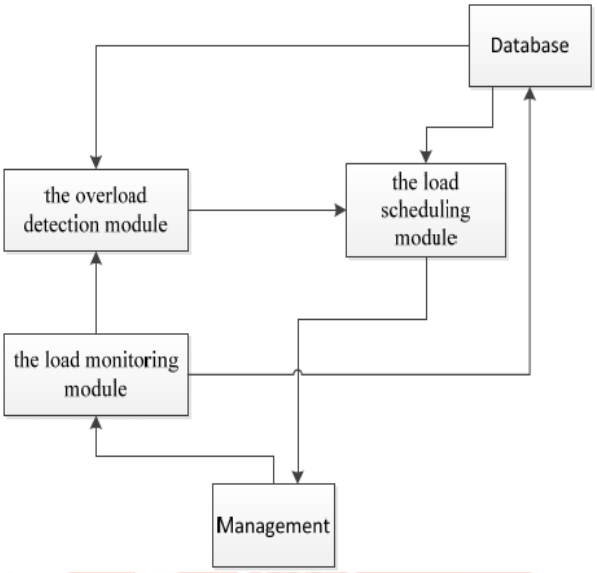

Figure 1. Design of the system

The database will keep the current load value of the system. The overload detection module will detect whether the system is overloaded or not. The load monitoring module will monitor the load of the systems. The load scheduling module will schedule the task in virtual machines.

B. Load calculation of the system

Load calculation of the system is done by the data provided by the XEN hypervisor. Xen hypervisor provide the data of the each virtual machine. Every virtual machine has a virtual CPU associate with it so we can calculate the load of the virtual machine.

1) CPU Utilization

Every virtual machine V_k have CPU utilization C_k associated with it and also have the number of virtual CPU VC_k. Load of the virtual Machine V_k is calculated as L_k dividing the total load by the number of virtual CPU. We have used the L_k to check the system status.

$$L_k = (c)_k \div (vc)_k$$

So we will have the load of every virtual machine in our database now. We have connected all the virtual machines locally so we will not have any network delay. XEN provide all the detail of the load according to RAM of virtual machines.

C. Load Balancing Algorithm

The load balancing algorithm is based on the group based strategy. We have created the virtual group of virtual machine. A group can have one or more than one virtual machine. So there will be total n (R_1,R_2….R_n) number of the virtual resource group in the system.

First, we will find the average load of each virtual resource group. This can be find by the total addition of the load on the virtual machines in the group and then divided by the total number of virtual machines in the virtual resource group.

We define the number of nodes is n, the utilization of each CPU is the A_k CPU load indicators is:

$$CPU = \frac{\sum_{k=1}^{n} A_k}{n}$$

After calculating the load of each virtual resource we will select the virtual resource group. If the load value of the virtual resource group is less than the threshold value, then we will select that virtual resource group else we will calculate the load value of next virtual resource group.
Now we will check the each virtual machine of the group in the selected virtual resource group and allocate the task to the first virtual machine that have load value less than the threshold value.
The algorithm is described as follows.

```
// make the virtual; resource group
Group= R1, R2, ....Rn
//calculate the load of virtual resource group
For i =0 to n
        Li = (V1+V2+…..+Vn)/n
        If Li Threshold value
                Select virtual resource group
                Break;
//Run the algorithm for selected virtual resource group
For j=0 to K
        If  Lj<Threshold value
                Allocate task to virtual machine
                Break;
```

Figure 2 Pseudo code for the algorithm

The process of calculating the load and then selecting the virtual resource in shown in the pseudo code

IV. EXPIREMETAL SETUP AND RESULTS

For experimental setup we have installed the Xen hypervisor in the Linux System.
Configuration of Linux System
Linux Version 12.04, System RAM = 4GB
Memory = 500GB, Dell Inspiron N5010

For communication between virtual machine we have set up the Network File System on the Virtual machine. The host system is made NFS Server and for Guest operating System we have set up the NFS Client. We have generated the request from the host operating system to and our server will allocated the request to the guest operating system. When a request is allocated to the guest operating system by the server then guest operating system will do the assigned task automatically. Number of time an algorithm run shows that how many times our algorithm needs to be checked for that purpose. The data is real time, so that it can vary according to the time.

Table 1 shows the CPU utilization of the virtual machine in XEN hypervisor.

Table 1: CPU Utilization of the Virtual Machine

VM NAME	VM1(CPU)	VM2(CPU)
Idle State	1%	1%
Gedit text editor	10%	9%
Firefox(client is using)	47%	67%
Libreoffice	86%	97%
Firefox+Liberoffice	99%	140%
Gedit +Firefox	87%	123%

Table 2 shows the number of time algorithm required to run for allocation of the task.

Table 2: Analysis of the Algorithm

No. of virtual machine	No. of Virtual resource group	No. of Application	No. of times algorithm runs	Average Response time (MS)
1	1	1	1G+1I	1
1	1	3	1G+2I	1.33
2	2	4	2G+5I	1.75
3	2	7	2G+8I	1.45
3	1	10	1G+12	1.3

Here G means running of the group based of algorithm and I mean running of the intergroup based algorithm. The worst case algorithm complexity is the O(M+N). Where M is the total number of virtual resource group and N is the virtual machine in that resource group. This algorithm is beneficial when we have a large number of virtual machine in our system. So that we can make the group of the virtual machine to allocate the request more precisely.

V. CONCLUSIONS

This paper proposes a static load balancing algorithm in cloud computing environment by considering the real time CPU utilization of the virtual machine. A Group Based strategy has been developed to allocate the request to virtual machine. Load calculation is performed on the real time data provided by the XEN hypervisor. The request can be made to the machine if and only if the load value of machine is less than the threshold value. This made the machine more reliable. The algorithm is found to be very beneficial in the case if one have numerous virtual machines, as group formed will be large then.
Further this algorithm can be improved by the dividing the virtual machine in the grades. So that we can assign a special type of task to a special type of virtual resource group. This algorithm can be implemented before the round robin approach in datacenters.

REFERENCES

[1] Joel Gibson, Darren Eveleigh, Robin Rondeau, Qing Tan, "Benefits and Challenges of Three Cloud Computing Service Models", 2012 Fourth International Conference on Computational Aspects of Social Networks.
[2] Mohiuddin Ahmed1, Abu Sina Md. Raju Chowdhury, Mustaq Ahmed, Md. Mahmudul Hasan Rafee, "An Advanced Survey on Cloud Computing and State-of-the-art Research Issues", International Journal of Computer Science Issues.
[3] R.Ramya, M.Krishsanth, L.Arockiam, "A State-of-Art Load Balancing Algorithms in Cloud Computing", International Journal of Computer Applications (0975 – 8887) Volume 95– No.19, June 2014.
[4] Sharrukh Zaman, Daniel Grosu, "A Combinatorial Auction-Based Mechanism for Dynamic VM Provisioning and Allocation in Clouds", IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 1, NO. 2, JULY-DECEMBER 2013.
[5] Haozheng Ren, Yihua Lan, Chao Yin, "The Load Balancing Algorithm in Cloud Computing Environment", 2012 2nd International Conference on Computer Science and Network Technology.
[6] Zhiyun Zheng, Ying Zhang, Zhenfei Wang, Xingjin Zhang, "The Multi-processor Load Balance Scheduler Based on XEN", 2012 International Conference on Systems and Informatics (ICSAI 2012).
[7] Xu Chaoqun, Zhuang Yi and Zhu Wei, "A load balancing aglorithm with key resource relavance for virtual cluster", International Journal of Grid and Distributed Computing Vol.6, No.5 (2013), pp.1-16.
[8] Xiaona Ren, Rongheng Lin,Hua Zou, "A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast", "IEEE CCIS2011".
[9] Priyesh Kanungo, "Load Measurement Issues in Dynamic Load Balancing in Distributed Computing Environment", "International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 10, October 2013".