Esha Sarkar Int. Journal of Engineering Research and Applications www.ijera.com
ISSN : 2248-9622, Vol. 4, Issue 9(Version 4), September 2014, pp.114-116

RESEARCH ARTICLE OPEN ACCESS

LOAD MANAGEMENT IN CLOUD ENVIRONMENT

Esha Sarkar*, Ch. Sekhar**
*(Department of Computer Science, JNTUK University, Visakhapatnam-46)
** (Department of Computer Science, JNTUK University, India)

ABSTRACT:
Cloud computing is an on demand service in which shared resources, information, software and other devices are provided to the end user as per their requirement at a specific time. A cloud consists of several elements such as clients, datacenters and distributed servers. There are n number of clients and end users involved in cloud environment. These clients may make requests to the cloud system simultaneously, making it difficult for the cloud to manage the entire load at a time. The load can be CPU load, memory load, delay or network load. This might cause inconvenience to the clients as there may be delay in the response time or it might affect the performance and efficiency of the cloud environment. So, the concept of load balancing is very important in cloud computing to improve the efficiency of the cloud. Good load balancing makes cloud computing more efficient and improves user satisfaction. This paper gives an approach to balance the incoming load in cloud environment by making partitions of the public cloud.
Keywords- Load Balancing, Cloud Partition, Main Controller, Load Balancers

I. INTRODUCTION

Cloud computing is the latest technology that delivers computing resources as a service such as infrastructure, storage, application development platforms, software etc. the advantages of cloud computing over traditional computing include agility, lower entry cost, device independency, location independency and scalability. The essential characteristics of cloud computing are: on demand self-service, network access, resource pooling, rapid elasticity, measured service [1]. Cloud computing is replacing the existing technologies as allows the end users to use its services without worrying about the infrastructure, installation, setup etc and offers them to pay for only for what they use. The on-demand service may be the software resources (software as a service, SAAS), physical resources (platform as a service, PAAS), or the Infrastructure (Infrastructure as a service, IAAS). There are three categories of cloud [2]:

Public cloud: A public cloud is available to the clients by the third party service provider via the internet. A service provider makes resources such as applications and storage available to the general public over the internet. Public services may be free or offered on a pay per usage model.

Private cloud: in private cloud the computing resources is fully dedicated to a particular firm or organization. No other organization can use the infrastructure.

Hybrid Cloud: Hybrid cloud is a composition of two or more clouds that remains distinct but bound together, offering the benefits of multiple deployment models.

Cloud is made up of huge resources. Management of these resources requires proper planning and proper layout. Cloud computing is efficient and scalable but maintaining the stability of processing so many requests in the cloud computing is a very complex problem [3]. cloud computing is very significant. It improves the efficiency and performance of cloud computing. As there are tremendous increase in traditional use of internet due to which uneven distribution of workload can occur, and it may cause some server overloaded and other under loaded which in turn may cause server crash. Workload distribution problem in cloud computing is very crucial and complex task till today, because the request arrival pattern on the cloud is not predictable and the capability of different servers in the cloud differ. In this paper we give an approach to manage the load in cloud environment by using the concept of load balancing and cloud partitioning, which simplifies the load balancing concept.

II. LOAD BALANCING

Load balancing is the process of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time [4]. Load balancing ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. Load balancing in cloud computing systems is really a challenge now. Load Balancing is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on

predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server [5]. Load balancing algorithms are two types based on the cloud computing environment, whether it is static or dynamic:

1. Static load balancing algorithm
2. Dynamic load balancing algorithm

In static cloud environment, the cloud service provider loads all the homogeneous resources. In this case, the cloud needs to have prior information about the nodes capacity, memory, processing speed, performance of the nodes, etc. These requirements cannot be changed at the run time. All the information about the system is known in advance, and the load balancing strategy is made by load balancing algorithm at compile time.
In dynamic cloud environment, the cloud service provider loads the heterogeneous resources. Here the cloud system cannot depend on the prior information. The user requirement can change during the run-time. Load balancing does not consider the previous state or behavior of the system; it depends on the present behavior of the system. The important things to consider while developing such algorithm are: estimation of load, comparison of load, stability of different system, performance of system, interaction between the nodes, nature of work to be transferred, selecting of nodes, etc. Dynamic algorithm is implemented at run time.
There are many load balancing algorithms, for example Equally Spread Current Execution Algorithm, Round Robin, Ant Colony algorithm, Random algorithm, the Weight Round Robin and the Dynamic Round Robin. Some of the classical load balancing methods is similar to the allocation method in the operating system, for example, the Round Robin algorithm and the First Come First Served (FCFS) rules.

III. PROPOSED SYSTEM

Our proposed system implements the load balancing concept in the public cloud environment. The public cloud has many nodes in different geographical locations. In our system, we divide the public cloud into several partitions to manage and simplify the load balancing in cloud environment, which is huge and complex. The partitioning is based on the geographical locations. Thus a cloud partition is subarea of public cloud and each cloud partition has many nodes of particular geographical region. Our proposed model has a Main Controller and Load Balancers. Main Controller chooses the suitable cloud partition for processing the incoming requests based on the status of the cloud partition. The Load Balancers are present in each cloud partition, which chooses the load balancing strategy. When the request arrives in the cloud; the Main Controller decides which cloud partition should receive the request. Then in the cloud partition the Load Balancer decides which nodes in the cloud partition should process the requests. This decision is done by the load balancing algorithm. The cloud partition can have three status based on which the Main Controller chooses a particular partition.

IDLE: In this status, most of the nodes are in idle state.
NORMAL: In this status, some of the nodes are in idle status while some others are overloaded.
HEAVY: In this status, most of the nodes are overloaded.
The main controller will choose the partition that is idle or normal in status. If the cloud partition status is heavy, the Main Controller will forward the request to another cloud partition that has normal status.

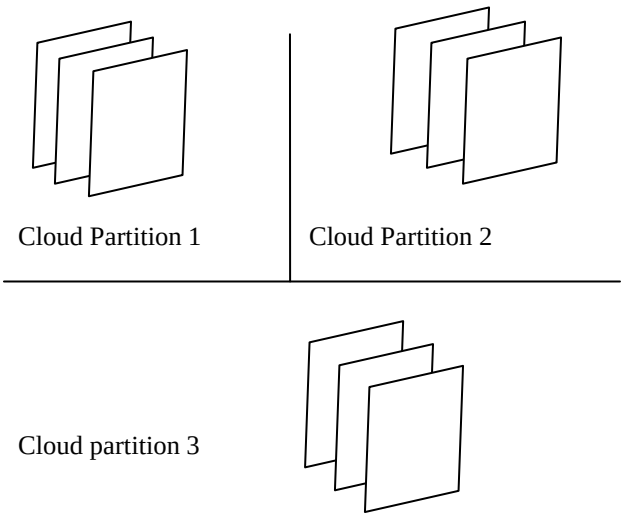

Fig 1: Cloud Partition Model

Algorithm 1 Best Partition Algorithm [6]

```
begin
while job do
SearchBestPartition (request);
If partitionState==idle||partitionState== normal then
 Send Job to Partition;
 else
 search for another Partition;
 end if
 end while
 end
```

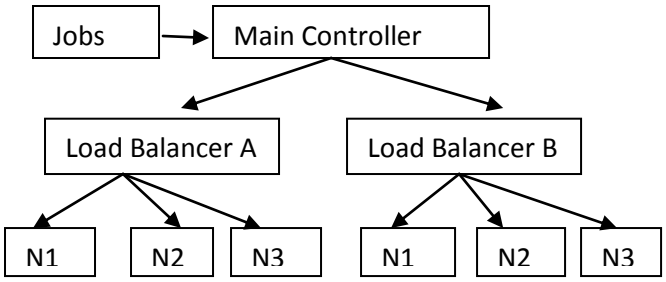

Fig 2: System Architecture

The Load Balancer gathers the load information from every node to evaluate the cloud partition status. The first step is to find the load degree of each node. The load degree of node is related to various static and dynamic parameters. The load status of a node can be:
IDLE: when LD (N) =0
NORMAL: 0<LD (N) <= High_LD
OVERLOADED: High_LD<=LD (N)
Where, LD (N) is the load degree of node N.
The Load Balancer will choose the node that is either idle or normal in status. Then, it will assign the request to that node for further processing. The load degree results are input to the load status table created by the Load Balancer. Each Balancer has a load status table, which is used to calculate the cloud partition status.

Algorithm 2: Request Allocation to Node

```
set counter=0;
Select node sequentially
If counter< max load capacity of node
        Assign request to node && counter++;
else select next node
Check for node
        If counter< max capacity of node
        Assign request to node && counter++;
        If counter == max capacity of node
        Request assignment not possible;
        wait for some nodes to become free;
        end if
end else
end if
```

This algorithm describes that each node in the cloud partition can handle up to certain number of requests. The Load Balancer selects the nodes sequentially. Each node is associated with a counter. The Load Balancer will check the node whether the counter is less than the maximum number of requests it can handle. If it is so, then it will assign the requests to that particular node. If the counter is equal to the maximum number of requests, then the Load Balancer will not assign the requests to it and will check the next node and so on.

IV. CONCLUSION AND FUTURE WORK

Load Balancing is an essential task in Cloud Computing environment to achieve maximum utilization of resources. Various works has been done in this research area to enhance the performance and efficiency of the cloud system. Our proposed system gives a simple approach of load balancing in the cloud using the concept of cloud partitioning. However, more work needs to be done in this area. In future study, other load balancing algorithms can be found out, because other algorithms may give better load balancing results. Various load balancing algorithms must be compared. More efficient algorithms should be designed so that the distribution of the load among the nodes in the cloud partition will be simpler and reduce the time complexity. Cloud division is not a simple task. Many things need to be considered while making cloud partition, such as the nodes in a cluster may be still apart.

V. ACKNOWLEDGEMENT

I would like to thank and acknowledge my guide Mr. Ch. Sekhar, for his support, his supervision and his precious time. He supervised the work over the past few months and advised many innovative ideas, helpful suggestion and valuable advice.

REFERENCES

[1.] Ruxandra Stefania PETRE, "*Data Mining in Cloud Computing*" published in "Database Systems Journal vol.III, no.3/2012"
[2.] kashish Ara Shakil, Mansaf Aslam,"*Data Management in cloud based environment using k-median clustering technique*" published in "IJCA"
[3.] Mohammad Manzoor Hussain, Anandkumar Biyani, Bhavana Bidarkar, "*Cloud Partitioning for Public Clouds using Load Balancing Model*", published in "International Journal of Engineering Development and Research"
[4.] Mangal Nath Tiwari, Kamalendra Kumar Gautam, Dr Rakesh Kumar Katare, "*Analysis of public cloud Load Balancing using Partitioning Method and Game Theory*", published in "International journal of Advanced Research in Computer Science and Software Engineering"
[5.] Mayanka Katyal, Atul Mishra, "*A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment*", published in," International Journal of Distributed and Cloud Computing".
[6.] Gaochao Xu, Junjie Pang, and Xiaodong Fu, " *A Load Balancing Model Based on Cloud Partitioning for the Public Cloud*" published in "IEEE TRANSACTIONS ON CLOUD COMPUTING YEAR 2013"