IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Load Balancing in cloud computing using Partitioning Method

Patel Kajal Miteshbhai
Assistant Professor
VBTMCA

***Abstract*- Recently Cloud computing has become one of the popular techniques adopted by both industry and academia providing a flexible and efficient way to store and retrieve the data files. The Service Provider plays an important role in transmitting information across the cloud. Load balancing is one of the critical components for efficient operations in the cloud computing environment. Load balancing in the cloud computing environment has an important impact on the performance. Good load balancing makes cloud computing more efficient and improves user satisfaction computing in order to have a better performance and less response time in the cloud computing. This paper gives a better load balancing model for the public cloud using cloud partitioning concept with a switch mechanism to choose different strategies for different situations.**

Keyword- load balancing model; public cloud; cloud partition; game theory;

I. Introduction

Cloud computing is the latest version of computing technology. Cloud means that the applications and services that are offered from data centre to all over the world. It involves a many number of nodes that are connected by the communication network like the Internet. Simply cloud computing use the internet to access someone else's software that is running on someone else's hardware in someone else's data centre. An Environment that are created in client's computer from an on line application stored on the cloud and run through a web browser.

Cloud computing has the advantage of giving a high-performance, pay-as-you-go, flexible, on-demand service. A model that giving information technology services in which resources are received from the internet by web based tools and applications have a direct connection to a server data and software packages that are stored in the servers. Cloud computing model provides the information access as far as an electronic device has access to the web and that type of model provides users to access remotely. Cloud data centres and end users are distributed geographically across the cloud computing environment.

Fig 1: Cloud Computing Architecture

Internet technology is quick growing and being used more expensively with the Cloud Computing that became a famous topic of academic and industry that became latest computing mechanism. It provides computing ability to meet the requirements of the community. On demand application services access from anywhere in the world is used by businesses and users.

Cloud computing system has several storage devices, servers, data centres, virtual machines etc which are communicated in an efficient way. Nowadays, computing models heavily depends on the virtualization technology that makes the server feasible for independent. Cloud computing is the concept based on the virtualization. Virtualization is one method that creates what are called virtual servers that run on a cluster of a number of real servers. Virtualization provides smaller number of high-powered servers that create a higher number of low powered servers while decrementing the overall cost in power, space, and other infrastructure.

There are various cloud computing services that are different from our traditional web service because of five basic principles behind cloud computing. These basic principles are: virtualization, resource pooling, automatic/easy resource deployment, elasticity, metered billing. These principles make cloud computing more automation, cost-savings and flexibility to the end users.

In cloud computing system there are mainly three major components like clients, distributed servers, and data centre. Each component has particular benefits that plays definite role in specific area.

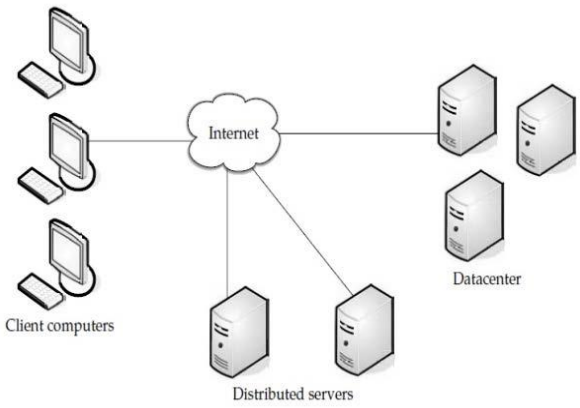

Fig 2: Three components make up a cloud computing [5]

- **Types of cloud computing**

1. **Public cloud**
   - **a.** Computing infrastructure is hosted by cloud vendor premises.
   - **b.** It can be shared by various organizations.
   - **c.** For Example: Amazon, Google, Microsoft, Sales force

2. **Private cloud**
   - **a.** The computing infrastructure dedicated to a particular organization and not shared with other organization.
   - **b.** High expensive and high secure than the public cloud.
   - **c.** For Example: HP data center, IBM, Sun, Oracle, 3tera

3. **Hybrid cloud**
   - **a.** Use of both public and private together is called hybrid cloud.
   - **b.** Organization may host critical application on private cloud.
   - **c.** Less security concerns on public cloud.
   - **d.** Hybrid cloud approach allows a business to take advantage of the flexibility, scalability, cost-effectiveness, and efficiency.

- **Services of Cloud Computing**

The word Service means that there are different types of applications provided by different servers across the cloud which is generally known as “as a service”[5]. Cloud computing provide applications and services over the internet. There are three type of services provided by cloud computing.

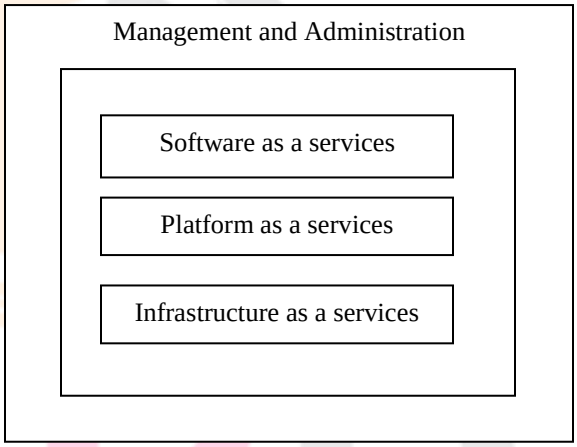

Fig 3: Model of cloud services

- **Load balancing**

Load balancing is the process of giving effective resource utilization by reassigning the total load to the individual nodes of the collective system and minimizing the response time of the job. Load balancing technique used to make sure that none of the node is in idle state while other nodes are being utilized”. Balance the lode among many nodes that you can distribute the load to another node which has lightly loaded. Generally load balancing algorithms can be divided into two categories as static and dynamic algorithms.

In static load balancing algorithm, all the information about the system is known in prior, and the load balancing strategy has been made by load balancing algorithm at compile time. Static load

balancing algorithm assign load to machines by their processing capability but do not consider dynamic changes of these attributes at run-time. Static algorithms are suitable for homogeneous and give stable environment Thus, static algorithm give poor results whenever attributes are dynamically changing. Generally used static algorithms are Round Robin (RR) & Weighted Round Robin (WRR), Improved Round Robin(IRR).

In dynamic load balancing algorithm, all the information about the system is not known in prior, and the load balancing strategy has been made by load balancing algorithm at run time. So, dividing the load during runtime is known as Dynamic Load Balancing technique. Dynamic load balancing algorithm gather information and run times conditions of machines and according to gathered characteristics assign and dynamically reassign the load among machines. Dynamic Load balancing strategies change according to their real statement of the system. Thus, Dynamic algorithms are higher flexible than static algorithm and take different types of system attributes during the run-time. Generally Least connection (LC) and weighted least connection (WLC) are used as dynamic load balancing algorithms.

II. Proposed Methodology

Proposed Methodology introduces the load balancing strategy based on the cloud partitioning concept. Cloud partitioning is one method to make partitions of large public cloud in some partitions of cloud. A cloud partition has various nodes that belongs to a particular area, these subarea of the public cloud based on the different geographic locations. For choosing node in partition calculate load degree and based on that load degree node status taken.After creating the cloud partitions, the load balancing starts. When a job comes at the system then the main controller deciding that which cloud partition should retrieve the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, then partitioning can be done locally otherwise the job should be transferred to another partition [6].

- **Main controller and balancers**

The load balance solution is done by the main controller and the balancers. The main controller first gives jobs to the suitable cloud partition and then communicates with the balancers in each partition that refresh the status information. When main controller deals the information for each partition then the smaller data sets will lead to the higher processing rates. The balancers in each partition collects the status information from every node and then select the right strategy to distribute the jobs. The relationship between the main controller and the balancers is shown in Fig 4.

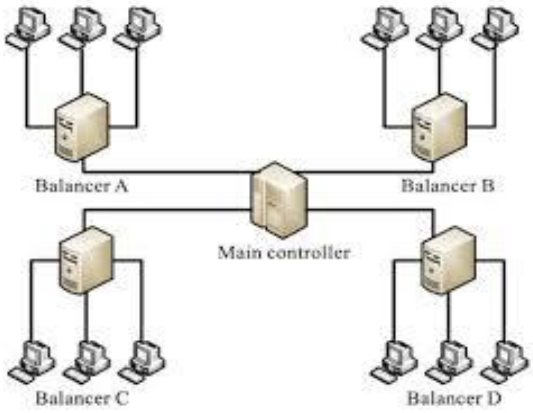

Fig 4: Relationships between the main controllers, the balancers and the nodes

- **Assigning jobs to the cloud partition**

When a job arrives at the public cloud, then the first step is to select the right partition. The cloud partition status can be divided into three categories:

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 others are overloaded.

HEAVY: In this status most of the nodes are overloaded.

- **Assigning jobs to the nodes in the cloud**

The cloud partition balancer collects load information from every node to evaluate cloud partition status. This evaluation of each node's load status plays very important role. Thus, the first task is to define the load degree of each node.

The node's load degree is related to several parameters like static parameters and dynamic parameters. There are various static parameters like the number of CPU's, the CPU processing speeds, the memory size, etc and Dynamic parameters like the memory utilization ratio, the CPU utilization ratio, the network bandwidth, etc. The load degree is evaluated from these parameters and Based on the evaluated load_degree (N) there are three nodes load status defined as:

Step 1: Define the load parameter set: F = {F1, F2,…..,fm}with each Fi(1<=i<=m, Fi € [0,1])) where parameter being either static or dynamic and M represents the total number of the parameters.

Step 2: evaluate the load degree as:

Load degree(N) = $\sum_{i=1}^{n} \alpha i$ Fi

Where αi represents weights that may differ for various kinds of jobs and N represents the current node.

Step 3: Three node load status level are defined as :

Idle

When Load degree(N) = 0

there is no job being processed by this node thus, the node status is charged to Idle.

Normal

When 0 <Load_degree(N)≤ Load_degree high

Then the node is charged to normal status and it can process the other jobs.

Overloaded

When Load_degree high ≤ Load_degree(N)

Then the node is not available and can't receive the jobs until it returns to the normal status.

The cloud partition balancers creates the load status tables which store the load degree results. Each balancer has a Load Status Table and refreshes it in each fixed time period T. The balancers use the table to calculate the partition status. Each partition status has a various load balancing solution. When the job comes at the cloud partition then the job assigns to the nodes by balancer based on its current load strategy. This strategy is updated by the balancers as the cloud partition status changes.

III. Cloud Partition Load Balancing Strategy

Good load balance will improve the performance of the whole cloud. There is no one common method that can provide all possible strategies. Different methods have been developed in improving existing solutions to resolve problems.

A relatively simple method can be used for the partition idle status compare to normal partition status. The load balancers update methods according to the status updates.

Here, the idle status uses an improved Round Robin algorithm while the normal status uses a game theory based load balancing strategy.

- **Load balance strategy for the idle status**

When the cloud partition is in idle state, then there are many computing resources available and relatively less jobs are coming. Thus, in this situation cloud partition has the ability to process jobs as fast as possible and simple load balancing method can be used.

The one of the simplest load balancing algorithms is the Round Robin algorithm, which passes each new request to the next server in the queue. This algorithm does not record the status of each connection thus, it does not store status information. Whereas in the regular Round Robin algorithm, each node has a same opportunity to be chosen. In public cloud, configuration and performance of each node will not equal so, this method may overload some nodes. So, an improved Round Robin(IRR) algorithm is used, which is called "Round Robin algorithm based on the load degree evaluation".

When the load status table is refreshed by the balancer, at this moment, if any job comes at the cloud partition, then it creates the inconsistent problem. The system status will have updated but the information will still be old so, this may create an erroneous load strategy choice and an erroneous nodes order. To resolve this problem, there are two Load Status Tables created as: Load Status Table 1 and Load Status Table 2. A flag is assigned to each table to indicate Read or Write.

When the flag is "Read", then the Round Robin based on the load degree evaluation algorithm is using this table.

When the flag is "Write", then the table is being refreshed and new information is written into this table.

- **Load balancing strategy for the normal status**

When the cloud partition is normal, jobs are coming faster than in the idle state and the situation is more complex, so there are various strategy used for the load balancing. Each user wants that his jobs completed in the shortest time, so the public cloud requires a method that can complete the jobs of all users with reasonable response time and provide high through. Compare this algorithm with other methods to show that their algorithm was less complexity that give better performance. Based on game theory Aote and Kharat gave a dynamic load balancing model. This model is related on the dynamic load status of the system with the users being the decision makers in a non-cooperative game. Previous studies have shown that the load balancing strategy for a cloud partition in the normal load status can be viewed as a non-cooperative game.

IV. Conclusion

In a large-scale cloud computing environment the cloud data centers and end users are geographically distributed across the internet, so it remains strong and great potential for the future. Load balancing using cloud partitioning helps to achieve a high user satisfaction with shorter response time and higher resource utilization ratio by ensuring an efficient and fair allocation of every computing resource.

V. Future Work

This work is a conceptual framework, more work is needed to implement the framework and resolve new problems. Some important points are Cloud division rules, How to set the refresh period, better load status evaluation and Find other load balance strategy are needed. Other load balance strategies may provide better results, so tests are needed to compare different strategies. Many tests are needed to guarantee system availability and efficiency.

References

[1] Gaochao Xu, Junjie Pang, and Xiaodong Fu , "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud TSINGHUA SCIENCE AND TECHNOLOGY ISSN l1007-0214l l04/12l lpp34-39 Volume 18, Number 1, February 2013

[2] Mangal Nath Tiwari, Kamalendra Kumar Gautam, Dr Rakesh Kumar Katare, "Analysis of Public Cloud Load Balancing using Partitioning Method and Game Theory" ISSN: 2277 128X Volume 4, Issue 2, February 2014 ISSN: 2277 128X

[3] S.Hemachander @ Harikrishna, R.Backiyalakshmi, "A Game Theory Modal Based On Cloud Computing For Public Cloud" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. XII (Mar-Apr. 2014), PP 48-53

[4] Shrikant M. Lanjewar, Susmit S. Surwade,Sachin P. Patil,Pratik S. Ghumatkar, Prof Y.B.GURAV, "Load Balancing In Public Cloud" IOSR Journal of Compute Engineering (IOSR- JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume16,Issue 1,VI (Feb. 2014),PP 82-87

[5] Nidhi Bedi, Shakti Arora, "A Secure Load Balancing Technique based on CloudPartitioning for Public Cloud Infrastructure" ISSN 2348 – 7968 IJISET – International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, July 2014

[6] Azizkhan F Pathan1, S. B. Mallikarjuna, "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud" International Journal of Information & Computation Technology ISSN 0974-2239 Volume 4, Number 16 (2014), pp. 1605-1610

[7] Shilpa S, Prof. Shubhada Kulkarni, "An area Based Cloud Partitioning for Balancing the Public Cloud using Game Theory Approach" AEIJST – July 2014 Vol 2 Issue 7 ISSN – 2348- 6732