Load Balancing in Cloud Computing using Observer's Algorithm with Dynamic Weight Table

Anjali Singh
M. Tech Scholar (CSE)
SKIT Jaipur,
27.anjali01@gmail.com

Mahender Kumar Beniwal
Reader (CSE & IT),
SKIT Jaipur
mbeniwal@gmail.com

Abstract-- Cloud computing is emerging technology which is a new standard of large scale distributed computing and parallel computing. It provides shared resources, information, software packages and other resources as per client requirements at specific time. As cloud computing is growing rapidly and more users are attracted towards utility computing, better and fast service needs to be provided. For better management of available good load balancing techniques are required. So load balancing in cloud becoming more interested area of research. And through better load balancing in cloud, performance is increased and user gets better services. Here in this paper we have discussed many different load balancing techniques used to solve the issue in cloud computing environment**.**

Keywords: *Cloud Computing, Cloud Service Model, Load balancing, virtualizations load balancing algorithm*.

*****

I. INTRODUCTION

Cloud computing can be defined as "Utilizing the internet to provide technology enabling services to people and organizations. Cloud computing enables consumers to access resources online through the internet, from anywhere, at any time without worrying about technical/physical management and maintenance issues of the original resources". Resources of cloud computing are dynamic and scalable. Cloud computing is an on demand service in which shared resources, information, software and other devices are provided according to the clients requirement at specific time. Capital and operational costs can be cut using cloud computing.

1.1. Cloud computing

According to the US National Institute of Standards and Technologies (NIST)[1], Cloud computing is "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" Cloud computing can be modeled broadly in two ways: Deployment model and Service model. Deployment model includes management of cloud infrastructure while Service model includes services that can be accessed on a cloud computing platform. Figure 1.1 illustrating the three basic service layers that constitute the cloud computing. It provides three basic services that are Software as a Service, Platform as a Service and Infrastructure as a Service.

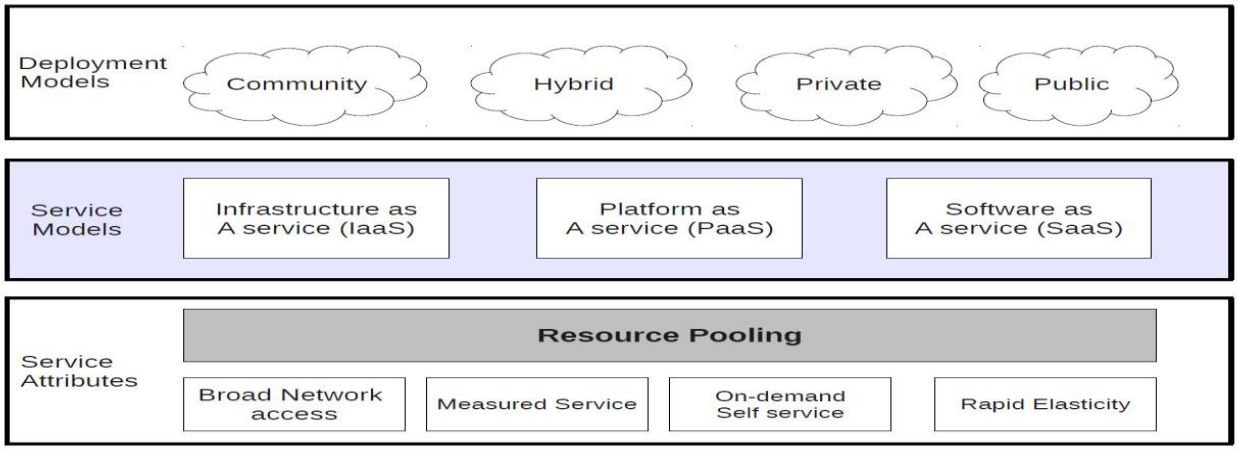

Fig.1.1 The NIST model of cloud computing

Characteristics

The essential characteristics of cloud computing design are:

1. **Broad network access**: implies widespread, heterogeneous network accessibility for thin, thick, mobile and other commonly used compute mediums.

2. **Rapid Elasticity**: simply means additional capacity remains available and accessible on an 'as needed' basis, and is recovered back to the pool when no longer needed for alternative allocation.

3. **Measured service**: as noted above means that all variables of resource consumption are tracked in capacity that users can be automatically billed for their consumption.

4. **On demand self service**: implies a customer can order service via the web or some other method at any point in time, 24×7, which becomes immediately available for his or her use.

1.2. Deployment Models - Type of Clouds [2][3][4][5]

A deployment model defines the purpose of the cloud and the nature of how the cloud is located. Based on the domain in which the clouds are used, clouds can be divided into four categories

1.2.1. Public cloud

A cloud is called a "public cloud" when the services are rendered over a network that is open for public use. Public cloud services may be free or offered on a pay-per-usage model. Public clouds are owned and operated by companies that use them to offer rapid access to affordable computing resources to other organizations or individuals. With public cloud services, users don't need to purchase hardware, software or supporting infrastructure, which is owned and managed by providers.

1.2.2. Private cloud

Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party, and hosted either internally or externally. A private cloud is owned and operated by a single company that controls the way virtualized resources and automated services are customized and used by various lines of business and constituent groups. Private clouds exist to take advantage of many of cloud's efficiencies, while providing more control of resources and steering clear of multi-tenancy.

1.2.3. Hybrid cloud

A hybrid cloud uses a private cloud foundation combined with the strategic use of public cloud services. The reality is a private cloud can't exist in isolation from the rest of a company's IT resources and the public cloud. Most companies with private clouds will evolve to manage workloads across data centers, private clouds and public clouds—thereby creating hybrid clouds.

A hybrid storage cloud uses a combination of public and private storage clouds. Hybrid storage clouds are often useful for archiving and backup functions, allowing local data to be replicated to a public cloud.

1.2.4. Community cloud

A community cloud may be established where several organizations have similar requirements and seek to share infrastructure so as to realize some of the benefits of cloud computing. With the costs spread over fewer users than a public cloud (but more than a single tenant) this option is more expensive but may offer a higher level of privacy, security and/or policy compliance. Examples of community cloud include Google's "Gov Cloud".

1.3. Service Models - Services provided through cloud computing [2],[3],[4],[5],[6]

In the service model different cloud types are an expression of the manner in which infrastructure is deployed. You can think of the cloud as the boundary between where a client's network, management, and responsibilities ends and the cloud service provider's begins. As cloud computing has developed, different vendors offer clouds that have different services associated with them. The portfolio of services offered adds another set of definitions called the service model.

Clouds provide three types of services to the customers.

These are : SaaS, PaaS and IaaS

1.3.1 Software-as- a-Service (SaaS)

This model provides users with business specific capabilities such as email or customer management. This model provides environment with pre-installed software and applications with infrastructure for specific business capabilities developed by third parties in the cloud . Client will access these applications by using devices, which can be connected with Internet and Internet browser installed to access cloud applications. Most popular example is software on demand.

1.3.2 Platform as Service (PaaS)

In this model clients create the software using tools and libraries from the provider. Clients also control software deployment and configuration settings. The provider provides the network, servers and storage. It provides a framework and platform to developer to create and deploy applications on host or service provider infrastructure. The customer does not require to be worried or be much concerned about the underlying hardware; they solely have to be follow and manage the in-operational environment with the interface provided to them. They can use any language supported by the cloud service provider to create their application in that environment e.g., *Google App Engine* that provides clients to run their applications on Google's infrastructure.

1.3.3 Infrastructure as a Service (IaaS)

In Infrastructure as a Service computing model, client will borrow the necessary hardware resources for building their own framework. They'll customize their entire framework with the available virtual machines, virtual network, memory, etc. the virtual resources used by client is managed programmatically so he is not much concerned about managing physical resources. IaaS provide the computing infrastructure as a fully outsourced service.

1.4 Virtualization [7] [8] [15]

Virtualization is used to install a software that allow multiple instances of virtual server applications. It is a very useful concept in context of cloud systems. Virtualizations means "something which is not real , but gives all the facilities of a real . It is the software implementation of a computer which will execute different programs like a real machine.

Virtualization is related to cloud, because using virtualization an end user can use different services of a cloud. The remote datacenter will provide different services in a fully or partially virtualized manner.

There are two types of virtualization found in case of clouds as given below:

1.4.1 Full virtualization

In case of full virtualization a complete installation of one machine is done on another machine. It will result in a virtual machine which will have all the software's that are present in the actual server. Here the remote datacenter delivers the services in a fully virtualized manner.

1.4.2 Para virtualization

Para virtualization is basically partial virtualization. In para virtualization, the hardware allows multiple operating systems to run on single machine by efficient use of system resources such as memory and processor. E.g. VMware software. Here all the services are not fully available, rather the services are provided partially.

1.5 Load Balancing [7][8][10][11][12]

Load balancing is defined as distributing processing and communications activity evenly, across a computer network so that no single device is overwhelmed. Load Balancing is a process in which the total load to the individual node is reassigned to make resource utilization effective with the improved response time of job, and also remove the conditions in which some nodes are overloaded and some are under loaded.

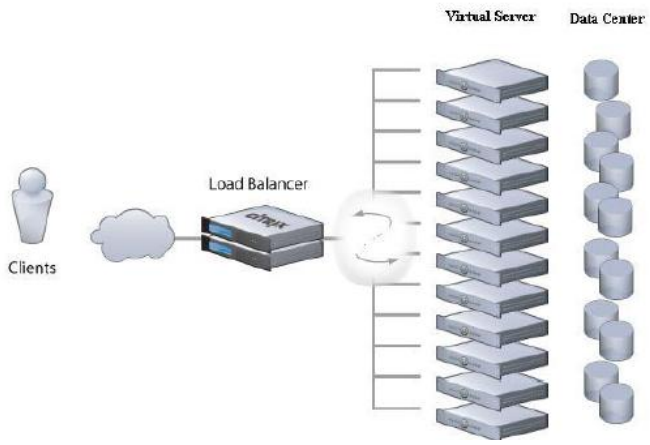

Fig 1.2 There are mainly two types of load balancing algorithms:[8]

1.5.1 Static Algorithm

In static algorithm the traffic is divided evenly among the servers. This algorithm requires a prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of system. Static algorithm is proper in the system which has low variation in load

Types of static algorithm: There are various types of static algorithm. Some of these are given below:

1.5.1.1. Random Algorithm [9][10]

In this, the request from the client is assigned to any random server from the list, it will keep on selecting the nodes in a different order each time. There is no way of sharing the load because of which server gets overloaded. Disadvantage of this algorithm is that it leads to underutilization of some servers and overutilization of some servers

Pros: Random Scheduling load balancing algorithm is simple to implement.

Cons: It can lead to overloading of one server while under-utilization of others.

1.5.1.2 Round Robin [15]

This algorithm does not take into account the previous load state of a node at the time of allocating jobs. It uses the round robin scheduling algorithm for allocating jobs. It selects the first node randomly and then, allocates jobs to all other nodes in a round robin manner. In Round Robin the requests are instantly distributed among all the servers evenly so load is shared evenly among all. This algorithm is better than the random algorithm.

Though the work load distributions between processors are equal but the job processing time for different processes are not same. And further the running time of any process is not known prior to execution, so at any point of time some nodes may be heavily loaded and others remain idle. This algorithm is mostly used in web servers where Http requests are of similar nature and distributed equally.

There are several limitations to the working of this algorithm. Nodes must be identical in capacity else

 ISSN: 2321-8169
3383 - 3388

performance will be based on speed of the slowest node in the cluster .works better only if the entire server has same configuration , but this is hard to achieve in cloud environment.

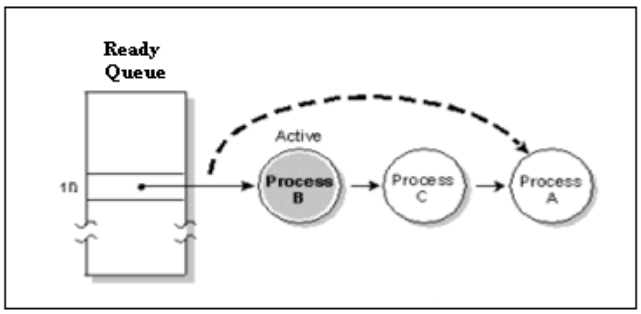

Fig: 1.3. Round Robin Algorithm

1.5.1.3 Weighted Round Robin

Weighted round robin was proposed to solve the problem of round robin. It is improved version of Round Robin. In this algorithm, each node is assigned a specific weight. Depending on the weight assigned to the node, it would receive appropriate number of requests. Each instance of server gets the load assigned depending on its processing capability, which depends on how that instance is behaving. One can assign a weight to each server in the group so that if one server is capable of handling twice as much load as the other, the powerful server gets a weight of 2. In such cases, the IP sprayer will assign two requests to the powerful server for each request assigned to the weaker one.

If the weights assigned to all the nodes are equal, then each node will receive same traffic. In cloud computing system, precise prediction of execution time is not possible therefore, this algorithm is not preferred.

Since, it does not consider the processing time each server is taking in responding to the request it may lead to the degraded service to the client

1.5.2 Dynamic Algorithm. [9][11][12][16]

In dynamic algorithm the lightest server in the whole network or system is searched and preferred for balancing a load. For this real time communication with network is needed which can increase the traffic in the system. Here current state of the system is used to make decisions to manage the load.

Following are some of the dynamic load balancing methods / techniques

1.5.2.1 Fastest response time Load Balancing

It depends on response time of server, in this method the request is routed from client to server which has fastest response time. The main disadvantage of this algorithm is that it is not possible that every server responds to request in seconds, which in turn leads to congestion in the network.

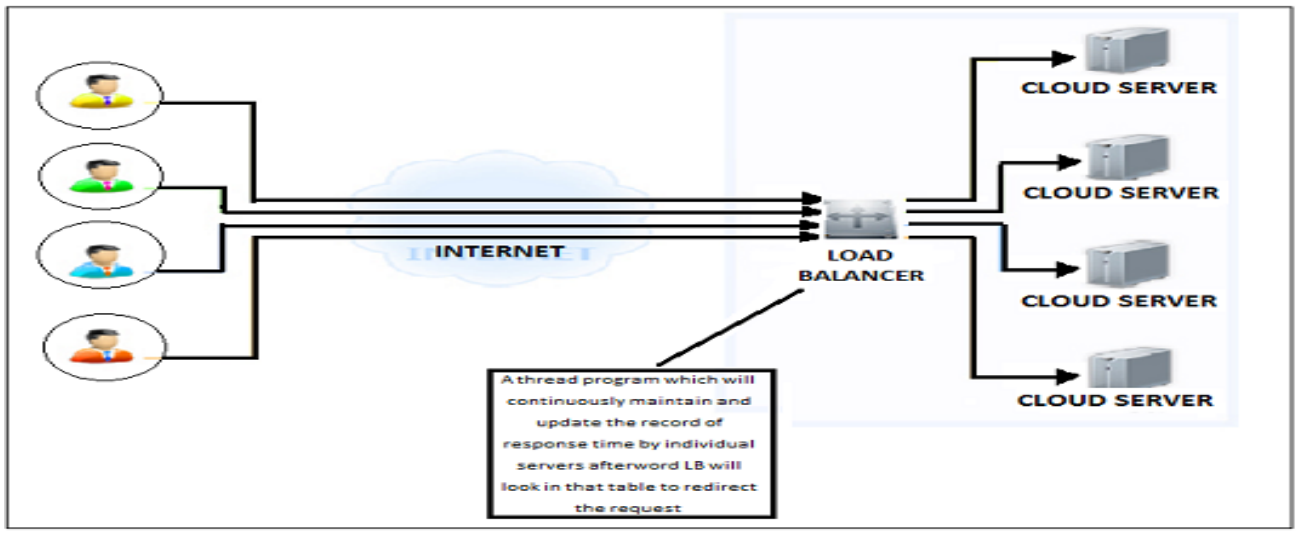

Fig.1.4 fastest response

1.5.2.2 Least Connection Load Balancing

This method depends on the least connection. It consists of middle server which passes new request from client to that server which has least connection at that point of time. It is best technique where all application running web server have same infrastructure. Its advantage is its main disadvantage also, as If there are two applications having different infrastructure i.e. HTML application and other uses J2EE or xml, it will lead to bottleneck of connections, as all connections will require to have different round trip time as its dependent on the server from where the request is originated

 ISSN: 2321-8169
3383 - 3388

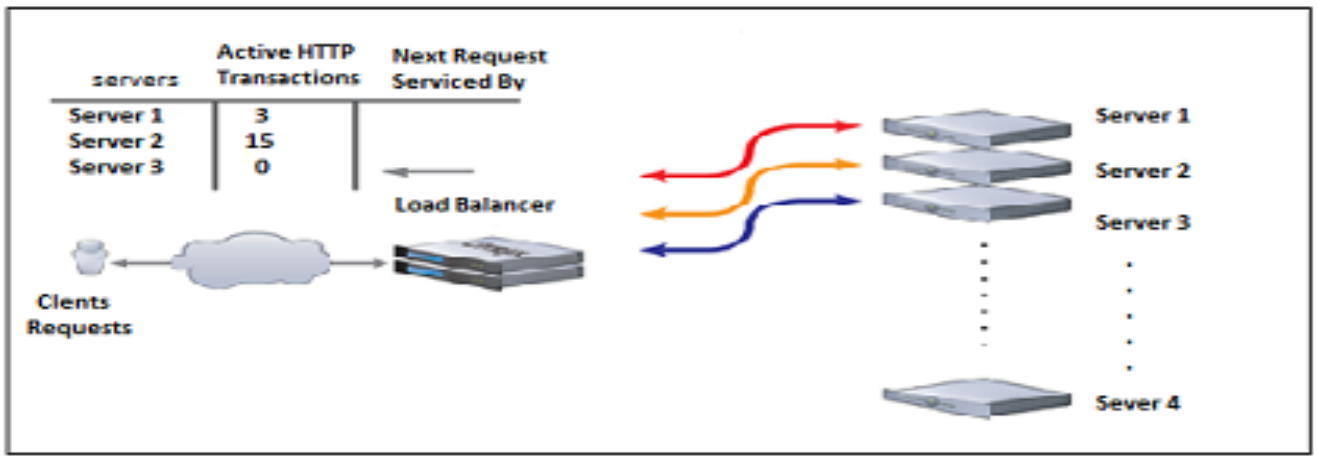

Fig.1.5 Least Connection Scheduling Algorithm

1.6 Problem statement

Analyze the effect of increasing the load on the server in cloud environment with and without applying the proposed load balancing observer algorithm (ii) designing an efficient approach of decreasing the response time of responses generated by server by modifying the observer load balancing algorithm.

1.7 Objective of the study

(i) To analyze the effect of increasing the load on the server in cloud environment with and without applying the proposed load balancing observer algorithm

(ii) Designing an efficient approach of decreasing the response time of responses generated by server by modifying the observer load balancing algorithm.

1.8 Tools and Technology Used

The load balancer application consist of five sections,

(i) A client which can get the response time delay,

(ii) A Server which will respond back for each incoming request,

(iii) A client application which will produce request flood to increase load on server,

(iv) A load balancer with NAT translation capability and implementing observer,

(v) A java thread which will automatically update the dynamic weight database which contain info about servers.

So to develop this whole simulator of load balancer we used java language, socket programming, multithreading, and java networking packages.

II. DESIGN AND IMPLEMENTATION DETAIL

A more efficient load balancer has been designed which is a combination of Least connection and Minimum response time Load Balancer algorithm It has been named as "Observer algorithm with dynamic weight table concept and CPU efficiency(Instruction per second)"

By this algorithm we compare and found that proposed algorithm takes response time less than the original algorithm with more CPU utilization

III. EXPERIMENT AND RESULTS TABULAR FORM

Number of clients	Original algorithm response time (ms)	Proposed algorithm response time
600	150-200	100 – 174
1500	300-342	200-253

No of clients	Single server with no optimization
1(with one request at a time)	Less than a millisecond
1(with multiple request at a time)	Less than a millisecond
200 clients	20 to 55 ms
600 clients	49 to 123 ms
1500 clients	Above 300 ms
>2000	500 ms to 1 min

To get the above results we have simulate the client server architecture on a single pc to develop the code we have used the concept of least connection algorithm, least response time dynamic weight table which will get update by a java thread and consisting the set of best possible server to redirect a request from client after getting all the possible set of best servers we compare the CPU efficiency of all of them and then select the best one out of them.

IV. CONCLUSION

We have studied various load balancing algorithm and modified the original observer algorithm by adding the dynamic weighted table with processing capability of servers to manage the load on cloud servers.

REFERENCES

[1] Peter Mell Timothy Grance, Special Publication 800-145 the NIST Definition of Cloud Computing Special Publication 800-145, National Institute of Standards and Technology September 2011

[2] (http://www.ibm.com/cloud-computing/in/en/what-is-cloud-computing.html- visited site in April 2014.

[3] Cloud Computing, A Practical Approach By: Toby Velte ISBN:9780070683518 Publisher: Tata McGraw-Hill Education India Year of publishing: 2009 Format: Paperback

[4] Cloud Computing: Theory and Practice By: Dan C. Marinescu ISBN:9780124046276 Publisher: Elsevier Science & Technology Year of publishing: 2013 Format: Paperback No of Pages: 416

[5] Cloud computing bible ISBN:9788126529803 Publisher: Wiley india Pvt. Ltd Year of publishing : 2011 Format: Paperback , by Barrie Sosinsky

[6] http://www.cloud-competence-center.com/understanding/cloud-computing-servicemodels - assessed in April 2014.

[7] Cloud Computing and Virtualization By: Abhay Bhadani ISBN: 9783639347777 Publisher: VDM Verlag Year of publishing: 2011 Format: Paperback /soft back No of Pages: 116 pp

[8] Tushar Desai, Jignesh Prajapati , A Survey of Various Load Balancing Techniques and Challenges in Cloud Computing, International Journal of Scientific & Technology Research , Volume 2, Issue 11, November 2013 , ISSN 2277-8616

[9] Soumya Ray and Ajanta De Sarkar Execution analysis of load balancing algorithms in cloud computing environment , International Journal on Cloud Computing: Services and Architecture (IJCCSA),Vol.2, No.5, October 2012

[10] Abhijit A. Rajguru, S.S. Apte ―A Comparative Performance Analysis of Load Balancing Algorithms in Distributed System using Qualitative Parameters‖ International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-3, August 2012.

[11] Rajesh George Rajan1, V. Jeyakrishnan, ―A Survey on Load Balancing in Cloud Computing Environments‖, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

[12] Zubair Khan, Ravendra Singh, Jahangir Alam, Shailesh Saxena ‗Classification of load balancing conditions for parallel and distributed systems' . IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011 ISSN (Online): 1694-0814 www.IJCSI.org

[13] Altman E., Kameda H., and Hosokawa Y. Nash equilibria in load balancing in distributed computer systems. International Game Theory Review, 4(2):91–100, 2002.

[14] El-Zogdhy S. F., Kameda H., and Li J. Numerical studies on braess-like paradoxes for non-cooperative load balancing in distributed computer systems. International 76workshop (Networking Games and Resource Allocation), July 2002.

[15] Sotomayor B RS Montero IM Llorente and I.Foster , ‗Virtual infrastructure management In private and hybrid clouds ,' in IEEE internet computing , Vol 13, No. 5 , pp 14-22, 2009

[16] Daniel Grosu. M.S. ," Load Balancing in Distributed Systems: A Game Theoretic Approach " Dissertation for Doctor of Philosophy in Computer Science at University of Texas, May 2003.