LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING PARADIGM- A SURVEY

[1]Rajat Sharma, [2]Prof. Sachin Upadhyay

***Abstract*: Cloud computing is a conversational expression used to describe a variety of different types of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. Cloud computing is a term without a commonly accepted demonstrable scientific or technical definition. In science, cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. As cloud computing is based on the concept of virtual machines. So the most important aspect is to equally share the load on these virtual machines to get the best out of them. This paper elaborates the notion of cloud load balancing. It provides a complete study of existing cloud load balancing techniques.**

1 INTRODUCTION

Cloud computing [1,2] is a recent technological development in the computing field in which mainly focused on designing of services which can be provided to the users in same way as the basic utilities like food, water, gas, electricity and telephony. In this technology services are developed and hosted on the cloud (a network designed for storing data called datacenter)and then these services are offered to users always whenever they want to use.

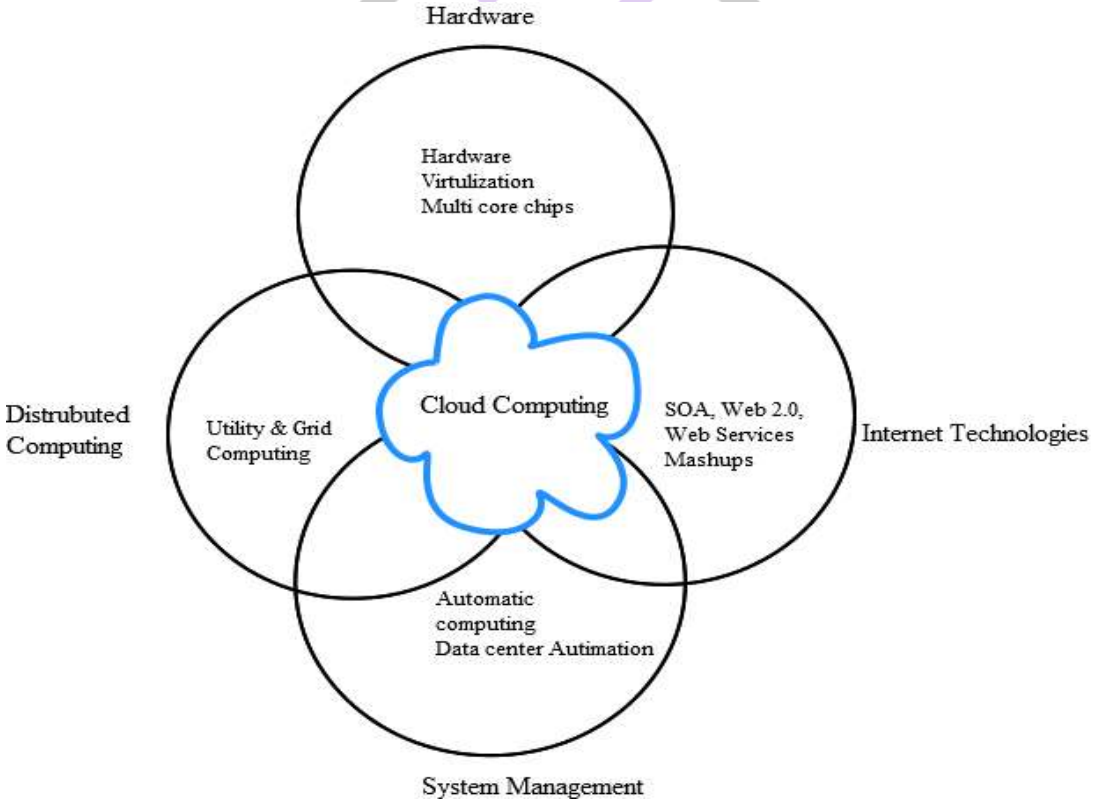

Figure 1: BasicRoots of CloudComputing

The cloud hosted services are delivered to users in pay-per-use, multi-tenancy, scalability, self-operability, on-demand and cost effective manner. Cloud computing is become popular because of above mention services offered to users. All the services offered by servers to users are provided by cloud service provider (CSP) which is working same as the ISP (Internet service provider) in the internet computing. In the internet technology some innovative development in virtualization and distributed computing and accessing of high speed network with low cost attract focus of users toward this technology. This technology is designed with the new concept of services provisioning to users without purchasing of these services and stored on their local memory

2. Cloud Computing Benefits

Cloud computing provided so many services to their users in which some of the very popular services are listed below [3,6,7] –

- **On-demand self service**

A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

- **Broad network access**

Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

- **Resource pooling**

The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

- **Rapid elasticity (Scalability)**

Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time.

- **Measured service**

Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service. Improves the capability of the resource provisioning to the users in such a way so that they just feel like to work with a separate resource using the concept of virtualization.

- **Pay-Per-Use Cost**

This purportedly lowers barriers [6] to entry, as infrastructure is typically provided by a third party and does not need to be purchased for one- time or infrequent intensive computing tasks. Pricing on a utility computing basis is fine-grained, with usage-based options and fewer IT skills are required for implementation (in-house). The e- FISCAL project's state-of-the-art repository contains several articles looking into cost aspects in more detail, most of them concluding that costs savings depend on the type of activities supported and the type of infrastructure available in-house.

2. Literature Review

2.1 Load Balancing in Cloud Computing

In a scenario of limited servers available at data center, if the request submitted are high than the capacity of the data center, its overall performance degrades. In such cases load balancer is used to improve the performance of data center. Load balancing is a technique to distribute load among multiple entities such as CPUs, disk drives, server or any other type of device. The goal of load balancing is primarily to obtain much greater utilization of resources. Load balancing can be provided either through hardware or software.

- **Hardware Load Balancing**

Load balancing can be provided through the specialized devices such as a multilayer switch that can route the packets to the destination or the cluster. Hardware based load balancing is complex in configuration & maintenance, and not suitable for hosted environment.

- **Software Based Load Balancing**

Load balancing can also be achieved through the software either using operating system or as an add-on application. Software based load balancing is simple to deploy and have the performance similar to that of hardware based load balancing. Some software based load balancing includes those bundles with Microsoft azure or Linux and add on such as PM proxy.

- **Load Balancer**

Load balancer manages the traffic flow between various servers. Load Balancer is placed between the server and the client and distributes the load among the available servers depending upon the algorithm of the Load balancer. Load balancer is not only improves the response time of cloud applications but also ensures the optimum utilization of the resources.

2.2 Major Load Balancer Algorithm [4,5]

To improve the performance in different types of cloud a number of Load balancing techniques are used. These Load balancing techniques are selected as per the load received. Depending on the load and the distance from the user, applications are directed to the data center to optimize the performance.

Considering the importance of load balancing, major operating system such as Microsoft windows and Linux are providing the load balancer as build-in capability or it can be implemented as add-on software which is having the comprehensive option. These load balancing techniques can be implemented at the user end or at data center end. When implemented at user end at that time it is known as service broker policy.

- **Service Broker Policy**

In cloud computing load management is required so that the request submitted should take minimum time and to be routed to the appropriate type of cloudlet. In case of one data center is overloaded at that point of time, provision should exist to divert the traffic to other data center.

Traffic routing between user base and DC is performed by a service broker that decides which data center is to be used for a particular user base. Three types of service broker policies which are supporting three different routing policies are [5]:

- **Closest Data Center Based Routing**

CDC is based on the quickest path available from the user base to the data center with minimum latency. Service broker will transmit the traffic to DC with minimum transmission delay.

- **Optimum Response Time**

ORT Service broker policy actively monitor the performance of all the data center and directs the traffic to a datacenter which estimates to give the best response time to the end user at the time it queried.

- **Reconfigure Dynamically with Load**

This service policy is an extension to CDC and deals with the scaling of resources as per the load it receives. In case of overload it increases the number of VMs in the data center and reduces the VMs in case of less load.

- **Data Center Controller**

DC controller is one of the most important entities in cloud. It manages the creation and destruction of VMs and performs the request routing from user base to the concerned VM. It consists of VM Load balancer to determine which VM should be assigned to the cloudlet for processing. Currently the three VM Load balancing techniques existing in cloud analyst are :

- **Round Robin**

Round robin performs the basic type of load balancing and functions simply by providing the list of IP address of cloudlet. It allocates first IP address to the first requester then second IP address to the second requestor for a fixed interval of time known as time slice. If the request is unable to finish within the given slice time, it will have to wait for the next cycle to get it turn for execution. This will continue till submitted tasks are not completed.

- **Active Monitoring Load Balancer**

This load balancer find outs the active VM and also to event out the active task at any point of time.

- **Throttled Load balancer**

This load balancing technique ensures that only a per-defined number of internet cloudlets are allocated to a single VM at any point of time. If more groups are presents in the data center than the number of available VM than some of the requests have to be queued until the next VM is available.

- **Other Load Balancing Algorithm**

A number of load balancing algorithms existing which are distributing the load among the data center. Each of them has their own functionality. Some of the major load balancing algorithms has been discussed as follows:

3. Conclusion:

In this paper, we have proposed a survey of load balancing methods. In cloud computing load balancing is one of the main issue. When client is requesting for service it should be available to the client. When any node is overloaded with job at that time load balancer has to set that load on another free node. Therefore load balancing is necessary in cloud computing. So in this paper we have discussed all the existing techniques for Load balancing.

References:

[1] National Institute of Standards and Technology- Computer Security Resource Center -www.csrc.nist.gov
[2] http://en.wikipedia.org/wiki/Cloud_computing.
[3] YashpalsinhJadeja and KiritModi, "Cloud Computing - Concepts, Architecture and Challenges", International Conference on Computing, Electronics and Electrical Technologies [ICCEET], IEEE-2012.
[4] Mr. Nitin S. More, Mrs. Swapnaja R. Hiray and Mrs. SmitaShukla Patel," Load Balancing and Resource Monitoring in Cloud", International Journal of Advances in Computing and Information Researches ISSN: 22774068, Volume 1– No.2, April 2012.
[5] R. X. T. and X. F. Z,"A Load Balancing Strategy Based on the Combination of Static and Dynamic, in Database Technology and Applications (DBTA)",2nd International Workshop,2010.
[6] Mohammed Fazil Ali, Ahmed Muhammad Barnawi, Abul Bashar, "Modeling and Simulation Strategies for Performance Evaluation of Cloud Computing Systems" , International Journal of Information Studies Volume 4 PP 148-160 Number 3 July 2012.
[7] Adabi, D. J. (2009), "Data management in the cloud: Limitations and opportunities", IEEE Data Engineering Bulletin, Vol. 32, No.1, pp. 4 12.