DEPARTMENT OF STATISTICS MCA, PU

CS-44 Unit 5

Load Balancing in Cloud Computing

Cloud load balancing is defined as the method of splitting workloads and computing properties in a cloud computing. It enables enterprise to manage workload demands or application demands by distributing resources among numerous computers, networks or servers. Cloud load balancing includes holding the circulation of workload traffic and demands that exist over the Internet.

As the traffic on the internet growing rapidly, which is about 100% annually of the present traffic. Hence, the workload on the server growing so fast which leads to the overloading of servers mainly for popular web server. There are two elementary solutions to overcome the problem of overloading on the servers-

- First is a single-server solution in which the server is upgraded to a higher performance server. However, the new server may also be overloaded soon, demanding another upgrade. Moreover, the upgrading process is arduous and expensive.
- Second is a multiple-server solution in which a scalable service system on a cluster of servers is built. That's why it is more cost effective as well as more scalable to build a server cluster system for network services.

Load balancing is beneficial with almost any type of service, like HTTP, SMTP, DNS, FTP, and POP/IMAP. It also rises reliability through redundancy. The balancing service is provided by a dedicated hardware device or program. Cloud-based servers farms can attain more precise scalability and availability using server load balancing.

Load balancing solutions can be categorized into two types –

1. **Software-based load balancers:** Software-based load balancers run on standard hardware (desktop, PCs) and standard operating systems.
2. **Hardware-based load balancer:** Hardware-based load balancers are dedicated boxes which include Application Specific Integrated Circuits

(ASICs) adapted for a particular use. ASICs allows high speed promoting of network traffic and are frequently used for transport-level load balancing because hardware-based load balancing is faster in comparison to software solution.

Examples of Load Balancers

1. **Direct Routing Requesting Dispatching Technique:** This approach of request dispatching is like to the one implemented in IBM's Net Dispatcher. A real server and load balancer share the virtual IP address. In this, load balancer takes an interface constructed with the virtual IP address that accepts request packets and it directly routes the packet to the selected servers.
2. **Dispatcher-Based Load Balancing Cluster:** A dispatcher does smart load balancing by utilizing server availability, workload, capability and other user-defined criteria to regulate where to send a TCP/IP request. The dispatcher module of a load balancer can split HTTP requests among various nodes in a cluster. The dispatcher splits the load among many servers in a cluster so the services of various nodes seem like a virtual service on an only IP address; consumers interrelate as if it were a solo server, without having an information about the back-end infrastructure.
3. **Linux Virtual Load Balancer:** It is an opensource enhanced load balancing solution used to build extremely scalable and extremely available network services such as HTTP, POP3, FTP, SMTP, media and caching and Voice Over Internet Protocol (VoIP). It is simple and powerful product made for load balancing and fail-over. The load balancer itself is the primary entry point of server cluster systems and can execute Internet Protocol Virtual Server (IPVS), which implements transport-layer load balancing in the Linux kernel also known as Layer-4 switching.