

Resource Based (Adaptive) load balancing

Introduction:

For network services to be highly available and reliable, load balancing is essential. It is essential to evenly divide network traffic over several servers in order to accommodate the rising demand and maintain a consistent user experience as a result of the increase in internet usage. One technique for balancing network traffic among several servers is the Source IP Hash load balancing approach. It is also used to prevent any single resource from becoming a bottleneck, which can result in decreased performance and increased latency.

Types of load balancers:

There are two main types of load balancers: static and adaptive.

Regardless of the resources' availability or usage, a static load balancer uniformly distributes requests across all of them. On the other hand, an adaptive load balancer decides how to distribute incoming requests after considering the availability and usage of each resource. An example of an adaptive load balancer is resource-based (or adaptive) load balancing.

Resource based load balancing:(Adaptive load balancing)

It is a method for evenly distributing workloads in distributed systems. It operates by evenly dividing the workload among all system nodes according to the resources available. This method assures that no node is overcrowded, which is advantageous for applications that demand a lot of resources. Maximizing resource use and reducing request response times are the two objectives of resource-based load balancing.

Methodology:

By creating a hash value from the source IP address of incoming network packets, the Source IP Hash load balancing technique operates. The server to which the packets should be sent is then determined using the hash value. The client will have a consistent experience thanks to this technique, which makes sure that all packets from the same source IP address are routed to the same server. Without requiring complicated algorithms or configurations, the Source IP Hash technique offers a quick and efficient solution to guarantee that network traffic is split equitably among servers.

Implementation:

1. Monitoring: The first step is to monitor the performance and resource usage of each server in the network. This can be achieved using various monitoring tools such as Nagios, Zabbix, or New Relic.
2. Resource allocation: Once the performance and resource usage of each server are known, the next step is to allocate resources to each server based on their performance and resource utilization. This involves dynamically adjusting the resources allocated to each server based on the workload and resource utilization at any given time.
3. Load balancing: Once the resources have been allocated to each server, the next step is to distribute the workload across the servers in a way that maximizes resource utilization and minimizes the risk of server overloading. This involves dynamically adjusting the workload distribution based on the performance and resource utilization of each server.

Advantages:

- Consistent user experience: The Source IP Hash load balancing approach ensures a consistent user experience by sending all packets from the same source IP address to the same server. This helps to provide a great user experience by preventing sessions from being interrupted when traffic is switched between servers.
- Scalability: The Source IP Hash load balancing technique is scalable, permitting an increase in the number of servers without degrading the load balancing's efficiency. This is crucial in settings where the user population is quickly growing since it enables the addition of new user servers to accommodate the growing demand.
- Fair distribution of traffic: By ensuring that traffic is evenly dispersed among servers, the Source IP Hash load balancing mechanism avoids any one server from becoming overloaded. This is crucial for reducing downtime because an overloaded server can deny users access to services.

Disadvantage:

- Inefficient use of resources: When many source IP addresses point to the same server, the server may get overloaded, which can result in an inefficient use of resources. Users may experience decreased speed and more latency as a result of this.

- Lack of control: It is challenging to limit traffic spikes when using the Source IP Hash load balancing method because it offers no control over the distribution of traffic. Due to this, a single server may get overloaded during moments of high traffic, denying users of their services.

Conclusion:

Resource-based (adaptive) load balancing is a useful method for dividing workload among several servers in a way that maximizes resource utilization and raises service availability. To guarantee that the strategy is successful in maximizing resource consumption and enhancing performance, it must be implemented with rigorous monitoring, resource allocation, and load balancing.