Techniques of Load Balancing:

Round Robin

A simple method of load balancing servers. Multiple identical servers are configured to provide the same services. All are configured to use the same Internet domain name, but each has a unique IP address. A DNS server has a list of all the unique IP addresses that are associated with the Internet domain name. When requests for the IP Address associated with the Internet domain name are received, the addresses are returned in a rotating sequential manner.

#### Weighted Round Robin

Weighted Round Robin builds on the simple Round Robin load balancing method. In the weighted version, each server in the pool is given a static numerical weighting. Servers with higher ratings get more requests sent to them.

#### Least Connection

Neither Round Robin or Weighted Round Robin take the current server load into consideration when distributing requests. The Least Connection method does take the current server load into consideration. The current request goes to the server that is servicing the least number of active sessions at the current time.

#### Weighted Least Connection

Builds on the Least Connection method. Like in the Weighted Round Robin method each server is given a numerical value. The load balancer uses this when allocating requests to servers. If two servers have the same number of active connections, then the server with the higher weighting will be allocated the new request.

Agent-Based Adaptive Load Balancing

Each server in the pool has an agent that reports on its current load to the load balancer. This real-time information is used when deciding which server is best placed to handle a request. This is used in conjunction with other techniques such as Weighted Round Robin and Weighted Least Connection.

Chained Failover (Fixed Weighted)

In this method, a predetermined order of servers is configured in a chain. All requests are sent to the first server in the chain. If it can’t accept any more requests the next server in the chain is sent all requests, then the third server. And so on.

Weighted Response Time

This method uses the response information from a server health check to determine the server that is responding fastest at a particular time. The next server access request is then sent to that server. This ensures that any servers that are under heavy load, and which will respond more slowly, are not sent new requests. This allows the load to even out on the available server pool over time.