

Load Balancing: Frequently Asked Questions

- 🔍 **What is a load balancer?**
 - 🔍 **Why load balance?**
 - 🔍 **How does load balancing work?**
 - 🔍 **What are load balancing methods/strategies/ schedules?**
 - 🔍 **What is server health checking?**
 - 🔍 **What is persistence?**
 - 🔍 **What is SSL acceleration and off load?**
 - 🔍 **Why Offload SSL?**
 - 🔍 **What is a next generation load balancer or ADC (Application Delivery Controller)?**
 - 🔍 **What is web acceleration?**
 - 🔍 **Layer 4 vs Layer 7**
-

What is a load balancer?

A network load balancer is an appliance, software or virtual appliance used to split network load across multiple servers.

Why load balance?

The main reasons for using a load balancer are:

- ➔To increase capacity of the system,
 - ➔To improve performance (normally related to increasing concurrency),
 - ➔To provide resilience
-

How does load balancing work?

The basic principle is that network traffic is sent to a shared IP in many cases called a virtual IP (VIP), or listening IP. This VIP is an address that is attached to the load balancer. Once the load balancer receives a request on this VIP it will need to make a decision on where to send it. This decision is normally controlled by a "load balancing method/ strategy", a "Server health check" or, in the case of a next generation device, a rule set.

The request is then sent to the appropriate server and the server will produce a response (hopefully). Depending on the type of device, the response will be sent either back to the load balancer, in the case of a Layer 7 device, or more typically with a layer 4 device directly back to the end user (normally via its default gateway).

In the case of a proxy based load balancer, the request from the web server can be returned to the load balancer and manipulated before being sent back to the user. This manipulation could involve content substitution, compression. Indeed some top end devices offer full scripting capability. Depending on the capability of the load balancer, in many cases it is desirable for the same user to be sent

back to the same web server. This is generally referred to as persistence.

What are load balancing methods/strategies/ schedules?

Typically a load balancing method or strategy is used to decide how the load balancer chooses where to send the request. There are many strategies available depending on the vendor, however a few common ones are listed below:

Round robin: The most simple method, each server takes a turn.

Least number of connections: The load balancer will keep track of the number of connections a server has and send the next request to the server with the least connections.

Weighted: Typically servers are allocated a percentage capability as one server could be twice as powerful as another. Weighted methods are useful if the load balancer does not know the real and actual performance of the server.

Fastest response time: This method is normally only available on more advanced products. The request will be sent to the fastest responding server.

Server agent: A client is installed on the server that communicates with the load balancer. This is sometimes required when you are using a basic load balancer that has direct server return. I.e. it does not know how many actual connections the server has or how well it is responding as it does not get the responses from the servers.

Methods such as server agent and weighted try to guess what the performance should be like for the next request whilst methods such as fastest response time actually know what the server is doing in real time.

What is server health checking?

Server health checking is the ability of the load balancer to run a test against the servers to determine if they are providing service.

Ping: This is the most simple method, however it is not very reliable as the server can be up whilst the web service could be down.

TCP connect: This is a more sophisticated method which can check if a service is up and running like a service on port 80 for web.

Simple HTTP GET: This will make a HTTP GET request to the web server and typically check for a header response such as 200 OK.

Full HTTP GET: This will make a HTTP GET and check the actual content body for a correct response. This feature is only available on some of the more advanced products but is the superior method for web apps as its will check that the actual application is available.

What is persistence?

Persistence is a feature that is required by many web applications. Once a user has interacted with a particular server all subsequent requests are sent to the same server thus persisting to that particular server. It is normally required when the session state is stored locally to the web server as opposed to a database.

IP based persistence: This is a simple method but is not very effective due to the “super proxies” i.e. a user could connect to the site via a range of external IPs thus breaking the persistence.

Cookie based: Layer 7 devices can take advantage of setting a load balancer cookie with the persistence information. This is a more reliable method and does not suffer from the super proxy problem.

There are also a number of vendor specific methods that use header information and other parameters found in the content such as SSL session ID to track uniqueness.

What is SSL acceleration and off load?

SSL acceleration or SSL offload is the ability for a load balancer to establish a secure tunnel with the client thus in most cases replacing the requirement for the web server to perform SSL.

In order for the load balancer to perform this function it must be configured with an SSL certificate either self generated or signed by a certificate authority. By default SSL is initiated to the server on port 443, as such, the load balancer should be configured to either terminate this or simply pass it straight through to the web server.

Some load balancers offer the facility to import and export certificates from common web servers such as MS IIS and Apache.

Why Offload SSL?

SSL (Secure Sockets Layer) can be a very CPU intensive operation thus reducing the speed and capacity of the web server. In addition there can be cost saving and management simplifications of having all the certificates stored in one place. And yes one certificate can be used for many web servers.

What is a next generation load balancer or ADC (Application Delivery Controller)?

An ADC can be described as a next generation load balancer. Typically they provide features to improve capacity whilst speeding up system performance.

They tend to offer a richer feature set including advanced load balancing, content caching, content compression, connection management, connection Pooling ,SSL, advanced routing, highly configurable server health monitoring and content manipulation.

These devices tend to run at the Layer 7 level.

What is web acceleration?

Web acceleration typically consists of a number of methods including HTTP compression and connection management/ pooling.

Compression can dramatically reduce the size of a typical web page by 70-90%. This means that over a real internet connection i.e. with some latency, the page will be compressed, delivered and uncompressed much faster than a page without compression. Over a modem or contention ADSI this can be 3-4x faster. Many commercial web sites use compression to speed up the delivery of the site to end users. Other effects of compression include a reduction in the amount of data sent over the network that could lead to reduced bandwidth bills! As the pages are served faster, session times are reduced and in most cases this will lead to an increase in server capacity.

Layer 4 vs Layer 7

Layer 4 devices are still commonly available although their market share has been reducing significantly as layer 7 devices and ADC's are becoming more powerful and cost effective. Some Layer 7 devices can run at over 3.2gig! This together with n=n scalability, a feature which is often found with layer 7 devices, means that it is very unlikely there are any web sites in the world that can't take advantage of this technology.

Although layer 4 devices offer little functionality compared to layer 7 devices and ADC's, they have traditionally been cheaper. This is no longer the case with many of the new ADC/l7 load balancing