Load Balancer

Load balancing is a technique used in system design to distribute workloads across multiple resources (servers) to improve performance and scalability.

What is a load balancer?

Load balancer is a device or software that sits between clients and a group of servers and distributes workloads evenly to prevent any one server from becoming overloaded. By balancing the workload, load balancer can improve throughput and performance of the system. It also increases its reliability by preventing servers from becoming overloaded and potentially failing.

Why do we need a load balancer?

When thousands of users request a service simultaneously, it can be a challenge to allocate these requests across multiple servers to ensure that system can handle the load.

If the load on servers increases too much, it can slow down the website and make it harder for users to get a fast and reliable response. One way to address this issue is to increase number of servers, but this brings its own challenges: How to distribute requests evenly across these servers? In this situation, load balancer can help to solve this problem by distributing requests across multiple servers in a way that ensures balanced workload. This will empower system to handle a large volume of requests.

Let's understand from another perspective!

Suppose we have several clients sending requests to a single server. When number of requests increases, there will be two critical issues

Server overloading:

There is a limit to how many requests a single server can handle. If number of requests exceeds this limit, server may become overloaded and unable to function properly.

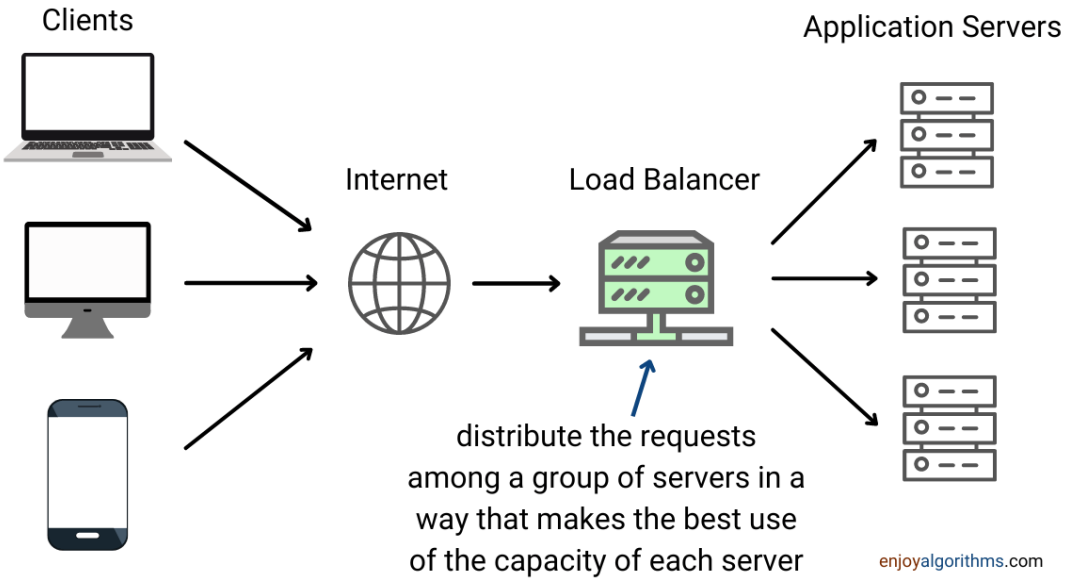
Single point of failure:

If the single server goes down for any reason, the entire application will become unavailable to users for a period of time. This can result in a poor user experience and impact overall reliability of the system.

We can solve this scalability problem in two ways:

Vertical scaling
We can increase the power of our current server. However, there are limits to how much we can increase the capabilities of a single machine.

Horizontal scaling
We can add more servers to our system. In this situation, we can use a load balancer to distribute requests across multiple servers and increase our ability to handle a large number of requests by adding more servers. In addition to this, load balancer can also ensure that the service remains available, even if one of the servers goes offline. It continuously checks the server health and prevents traffic from being sent to servers that are unable to fulfil requests.



Where do we add a load balancer?

Load balancers can be placed at different points in a system to distribute workload. Some common places to use load balancers are

Between clients and frontend web servers
Between frontend web servers and backend application servers
Between backend application servers and cache servers
Between cache servers and database servers

Types of load balancers
Software load balancers -Are more flexible and offer more options for customization.

HAProxy: A TCP load balancer.
NGINX: An HTTP load balancer with SSL termination support.
mod_athena: Apache-based HTTP load balancer.
Varnish: A reverse proxy-based load balancer.

Hardware load balancers-on the other hand, are physical devices that are installed in a network. They are generally less flexible and offer fewer options for customization.

F5 BIG-IP load balancer
CISCO system catalyst
Coytepoint load balancer
Citrix NetScaler

Advantages of load balancing
Availability and scalability
Prevent server overload and single points of failure
Additional functionality:encryption, authentication.

Critical concepts to explore further

What is the difference between Load Balancer and Reverse Proxy?
Different Categories of Load Balancing: 1) Layer 4 (L4) load balancer 2) Layer 7 (L7) load balancer 3) Global server load balancing (GSLB)
Health check feature of the load balancer.
DNS load balancing vs Hardware load balancing
The application load balancer in designing several systems
Cloud load balancing