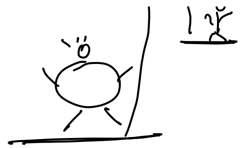
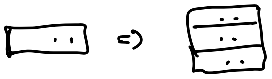
Load Balancer



Load Balancer: I will take care that your servers are not overloaded

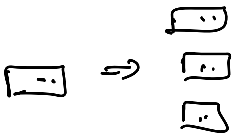
Vertical Scaling:

Increasing the capacity of already existing server.



Horizontal Scaling:

Increasing the number of servers



Load Balancer:

Load Balancer comes into play when the servers are horizontally scaled.

A load balancer is a type of reverse proxy that distributes traffic across servers

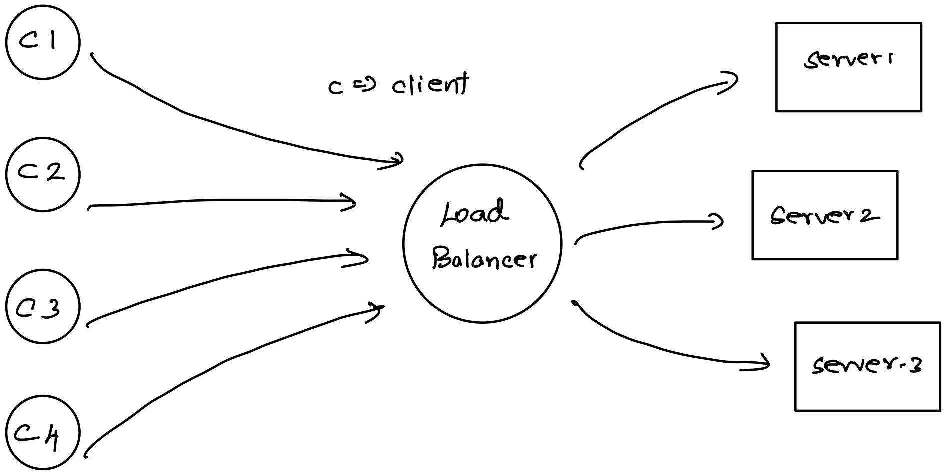
This distribution is based on strategies such as round-robin, weighted Round Robin, performance based selection

top to bottom again top to bottom ...

(choosing the server with the best performance metrics - like fastest response time or the least amount of traffic), based on IP address, based on

paths (Eg: praveen.com/payments => server1
praveen.com/notes => server2)

→ if an IP from India comes use server 1 else use server 2



A load balancer need not only be placed between a client and a server, it can be placed between servers & databases, at the DNS layer etc

Load Balancer at the DNS layer -> DNS Round Robin.

Sometimes a domain name can have multiple IP address

Eg

```
Terminal

>> dig hotstar.com

IP Addresses :

      108.158.251.103
      108.158.251.113
      108.158.251.98
```

This is where a DNS load balancer comes into play.

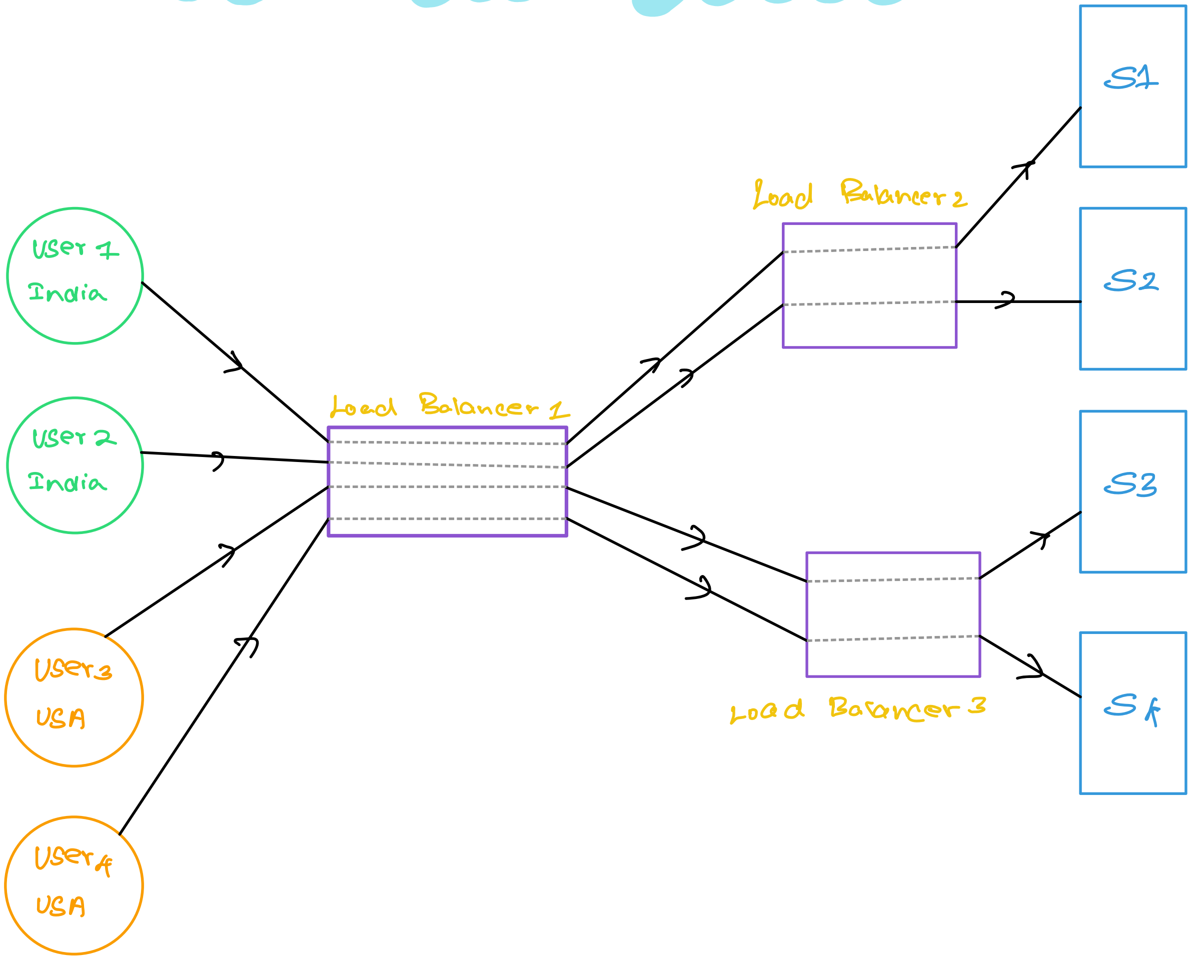
A DNS load balancer distributes the traffic by returning different IP address for the same domain name.

Based on Round-Robin strategy

Hotspots :

When distributing workload across a set of servers, sometimes the workload might be spread unevenly. This happens because the sharding key or the hashing function we use might be sub optimal. Thus some servers will receive more traffic than the others creating a hotspot.

MULTIPLE LOAD BALANCERS



Load Balancer 1 — distributes based on the IP Addr
Eg: users from India use the Top server Else use bottom server

Load Balancer 2 & 3 — Distributes based on Round Robin