Availability

⁂ while designing a system always find which part of the system is critical thus should be highly available

Eg: In Stripe (Payment Service) the payment page should be highly available whereas the dashboard (used by business owners) can be less available.

Availability :

Availability is the measure of how resistant a system is to failures

Nines

Availability is measured using "nines"

If a system has 90% availability (ie it has a downtime of (36.5 days) per year) then the system has "one nine" of availability

percentage	nines	downtime per year
90%	one nine	36.53 days
95%	one and half nines	18.26 days
99%	two nines	3.65 days
99.9%	three nines	8.77 hours
99.999%	five nines	5.26 minutes

→ Gold Standard

High Availability (HA):

Used to describe systems that have particularly high levels of availability, typically 5 nines or more.

SLO

SLO stands for "Service level Objective". This is an agreement given by the service providers to the customers on the system's availability and other agreements.

SLA

SLA stands for "service level agreements". A group of SLOs form a SLA

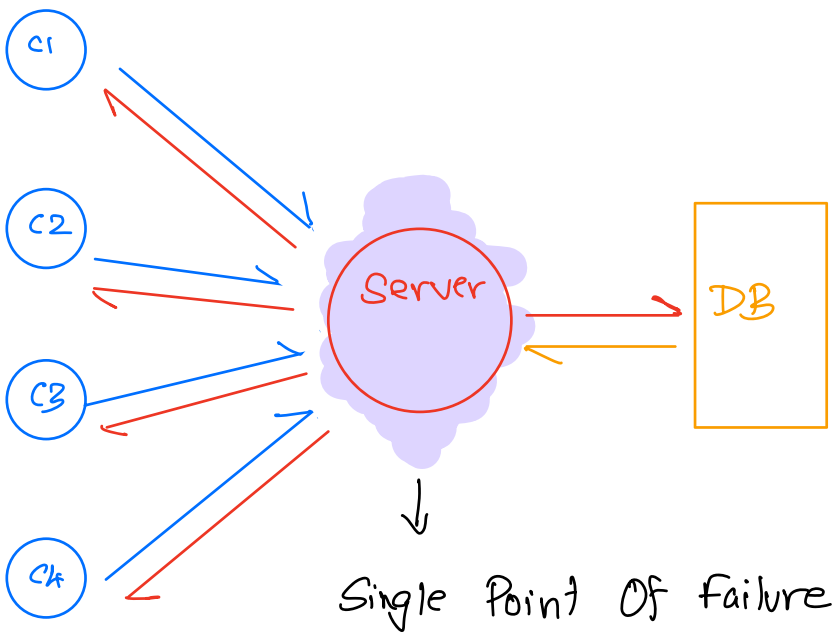
High Availability

To ensure high availability make sure the system doesn't have single point of failure

Redundancy :

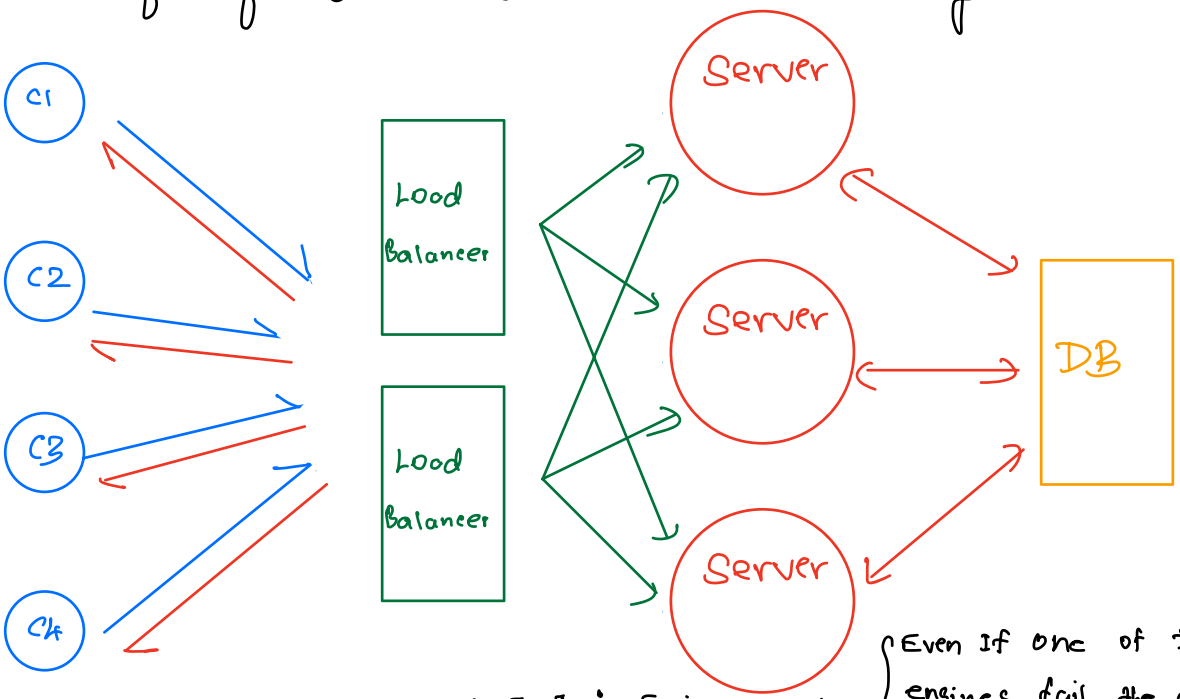
Redundancy is the process of multiplying certain parts of the system in an effort to make it more reliable.

Eg :



In the above system if the server fails then the entire system would come to a hault. This is avoided by having multiple servers. Since we are having multiple servers we should also have load balancer. If we have one

load balancer then it would become a single point of failure. So we are having two LB

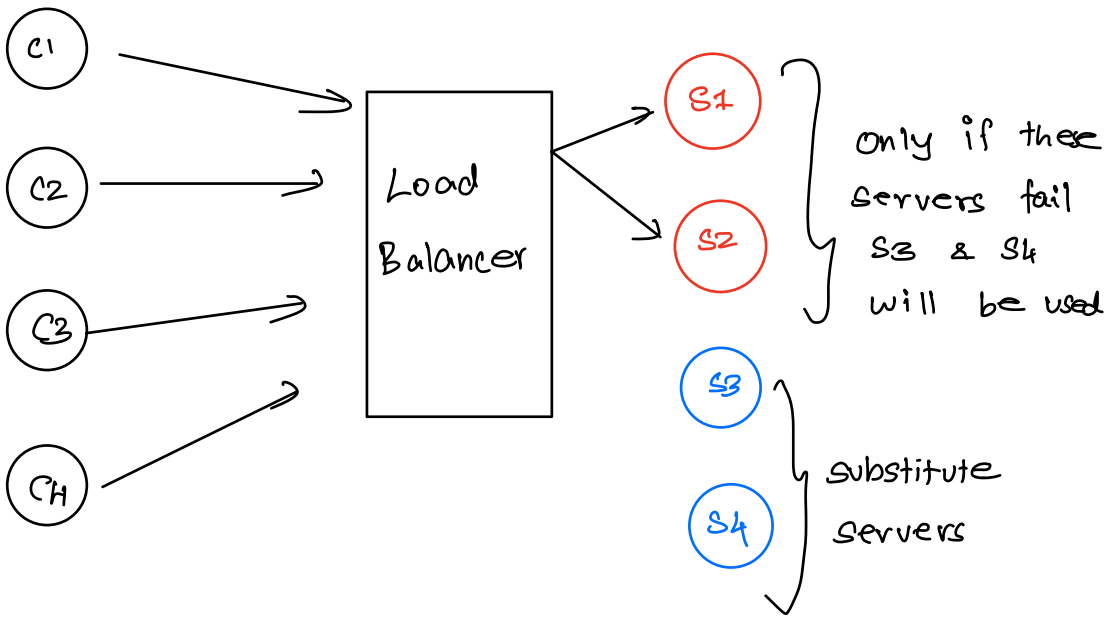


→ Eg: Twin Engine Aeroplane { Even if one of the engines fail the aeroplane can still land safely

Here we are using **Passive Redundancy** where anticipating a failure we add in some extra resources. Eg: In the place of 1 server we use 3 servers (S1, S2, S3). **All** the **three servers** will be **in use**. If one of the servers fail (say S2) then all the load will be redirected to the servers S1 & S3

In an **Active Redundancy** only one or a few machines will be handling all the traffic

and only if the traffic handling machine fails the substitute machines will be used.



Replication :

Replication means sharing information to ensure consistency among the resources. The primary resource gets all the updates which ripples through to the replica resources

Eg :