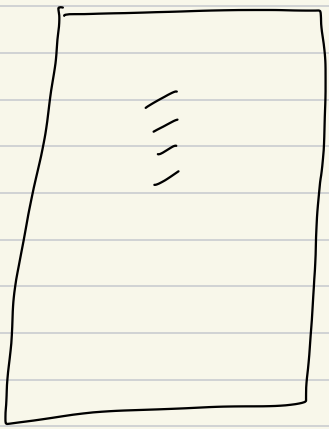
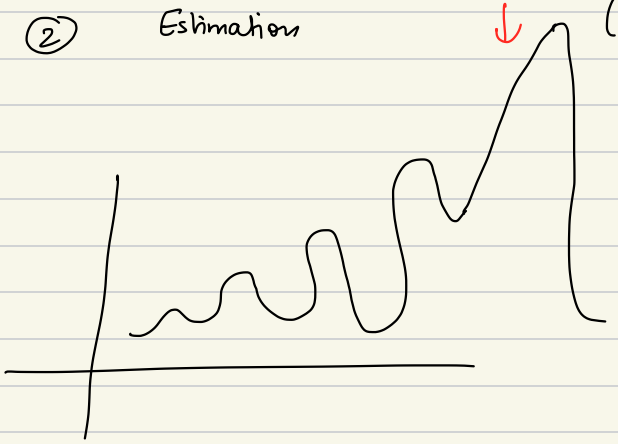
HLD - 2

20/Nov/2023



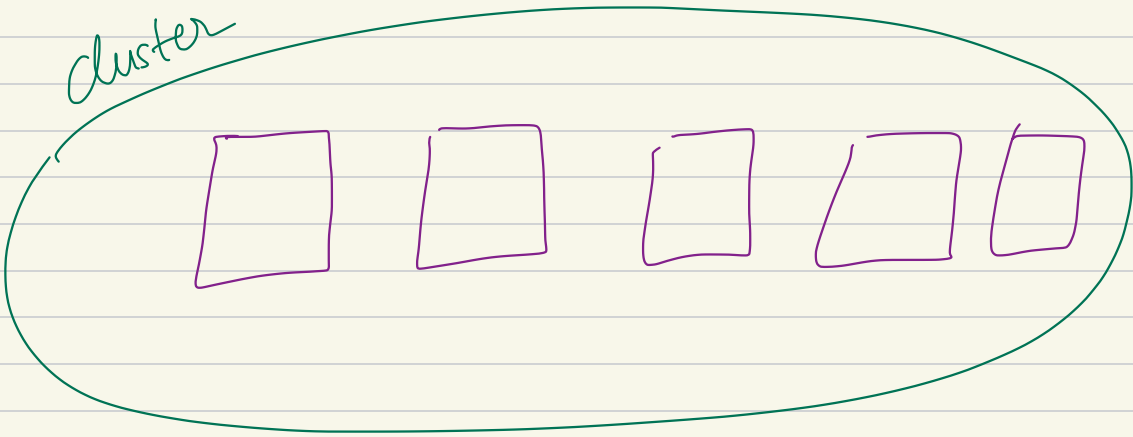
Vertically scaled

1. cost implications ↓
2. Estimation ↓ (scale flexibility is lost)



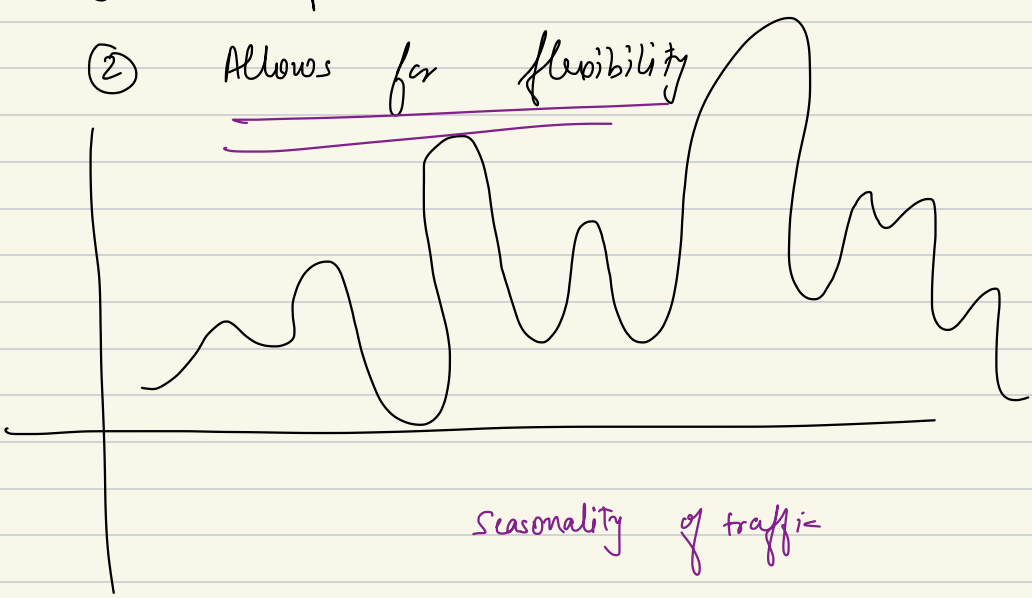
3. SPOF ☹ ↓

HORIZONTAL SCALING



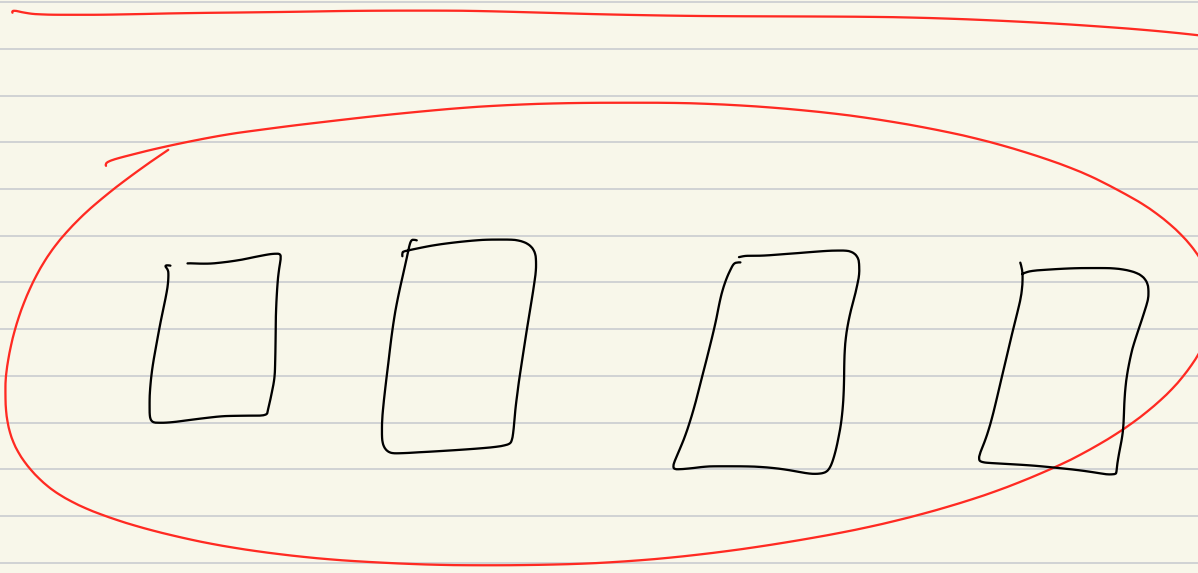
commodity hardware

1. cheaper
2. Allows for flexibility



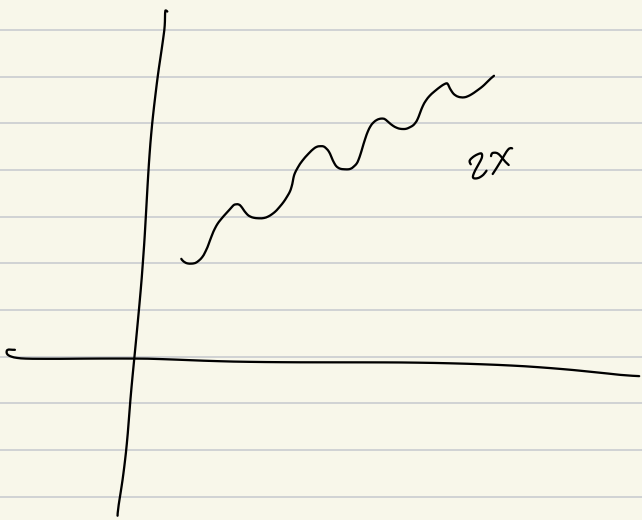
seasonality of traffic

③ No SPOF

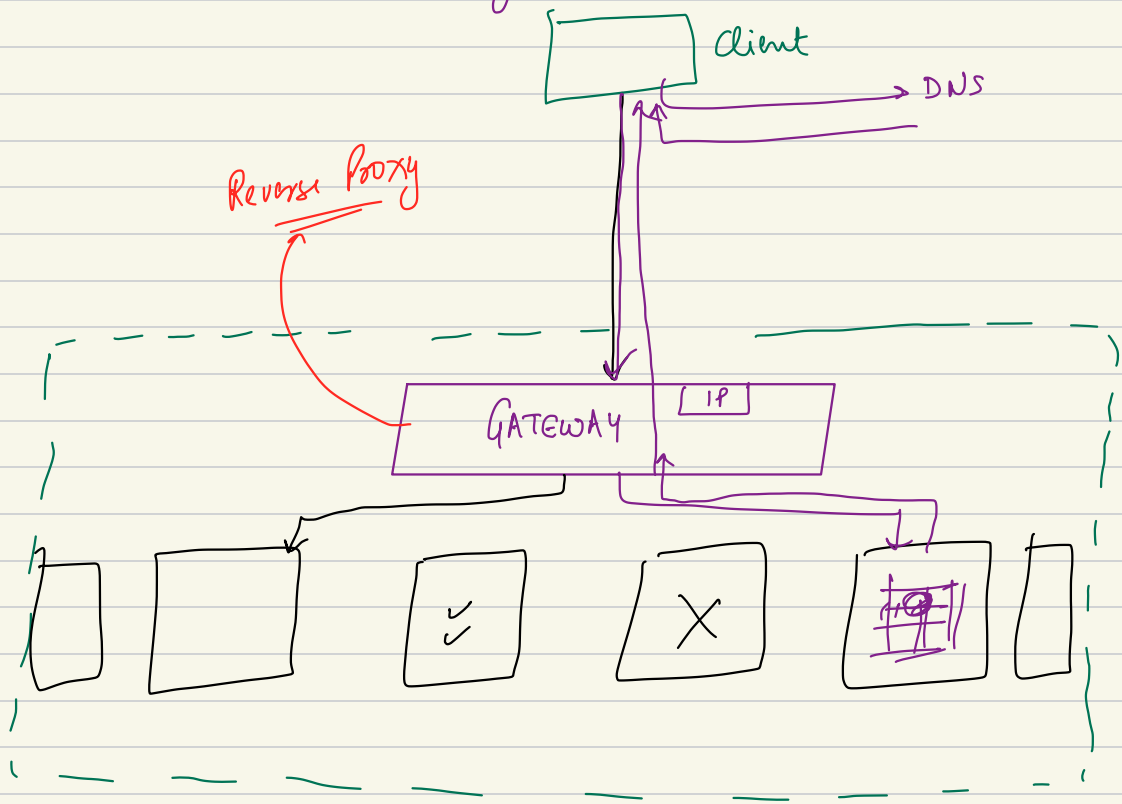


These machines have to communicate with each other over network..

Horizontal scaling becomes tougher as you have to navigate the challenges of machines talking to each other over network

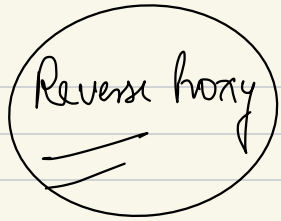

2x

1. Buy a domain
2. Static public IP address
3. DNS entry (domain ↔ IP)

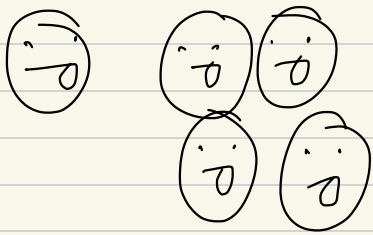


- Gateways usually follow

Active Passive Gatewaying

Gateway ≡ 

≃ (Load Balancer)

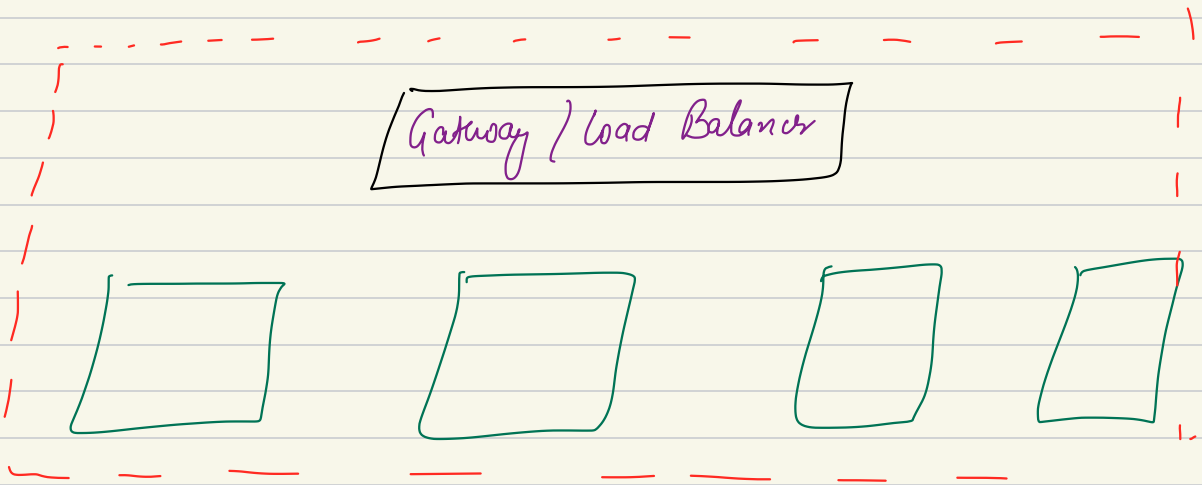


Gateway

Gateway machine can (usually does) also perform the role of Load Balancer and hence some people also call this machine a load balancer

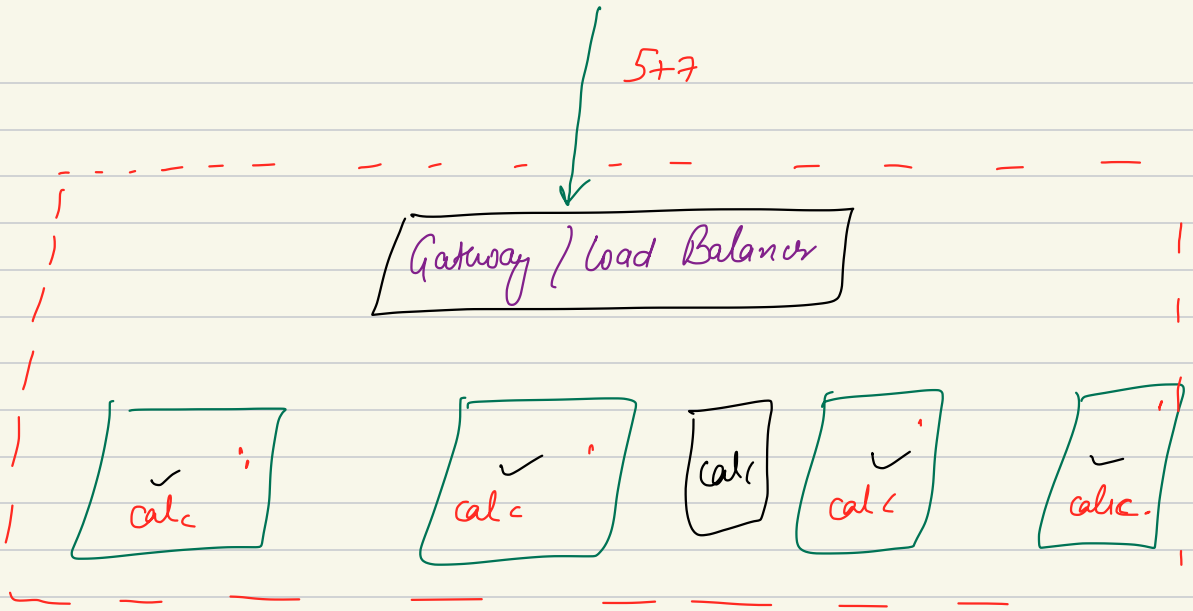
Gateway → First point of contact for the outside world.

Load Balancer → When you have so many incoming requests and you want a machine to distribute those requests amongst different machines of the cluster.



2 kinds of load Balancing

1. stateless load Balancing
2. stateful load Balancing



if each of the machines in the backend cluster are equally well equipped to handle incoming requests.

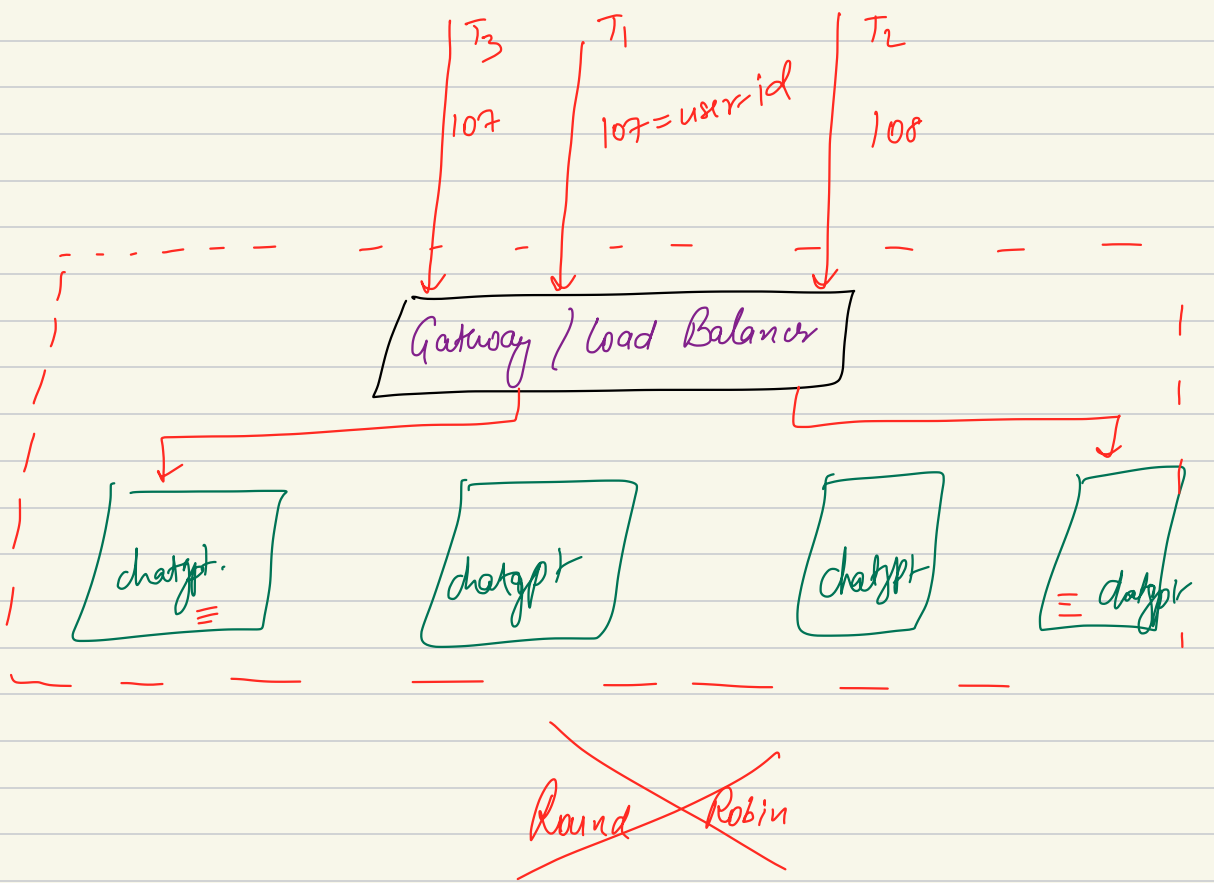
"stateless load Balancer"

Algo:

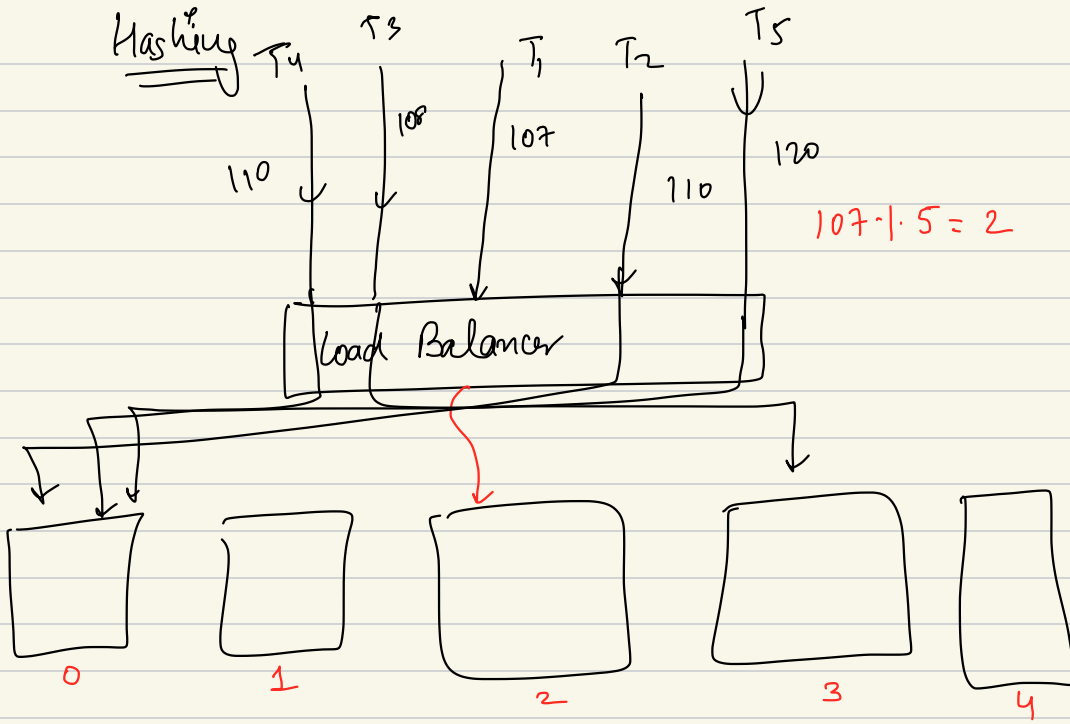
① Round Robin ✓

② Random

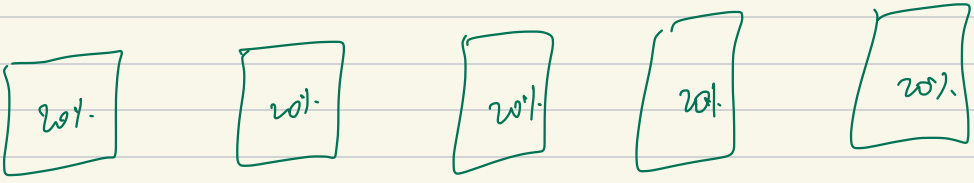
Stateful Load Balancing

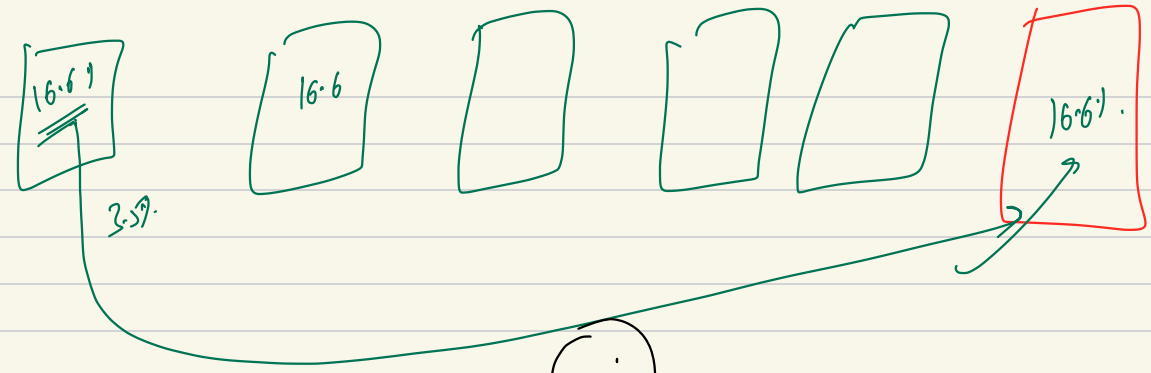


~~Round Robin~~



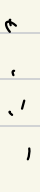
Simple Hashing ☺ ☺





Hashing

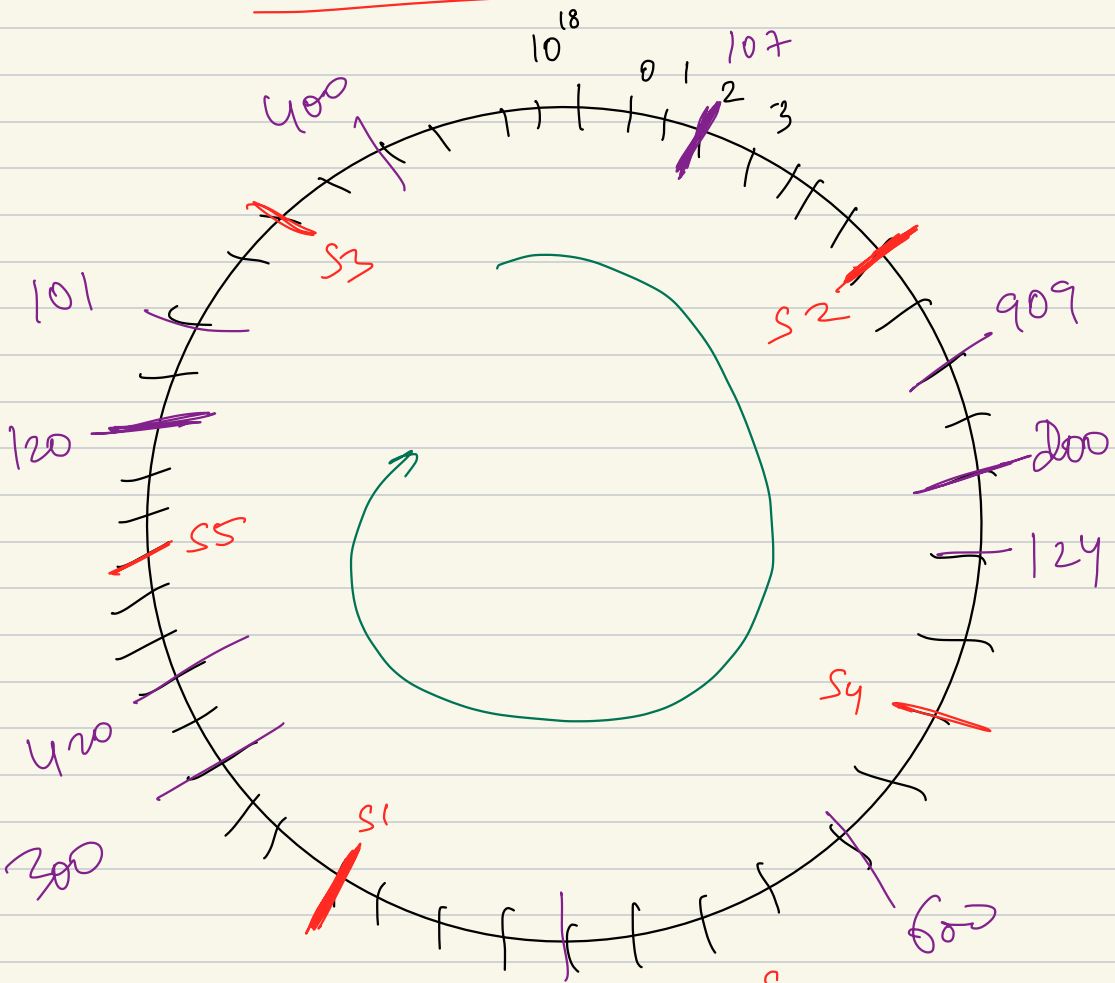
	%5	%6
101	1	5
102	2	0
103	3	1
104	4	2



//

Hashing at first seemed a promising approach but on closer look it becomes clear that simple hashing will NOT be able to optimally handle increase / decrease in # machines ☺

CONSISTENT HASHING



5 servers

S_1
S_2
S_3
S_4
S_5

H_S (server_id) = [output $0 - 10^{18}$]

$H_S(S_1) = 10^{14}$

$H_S(S_2) = 100$

$H_S(S_3)$

$H_S(S_4)$

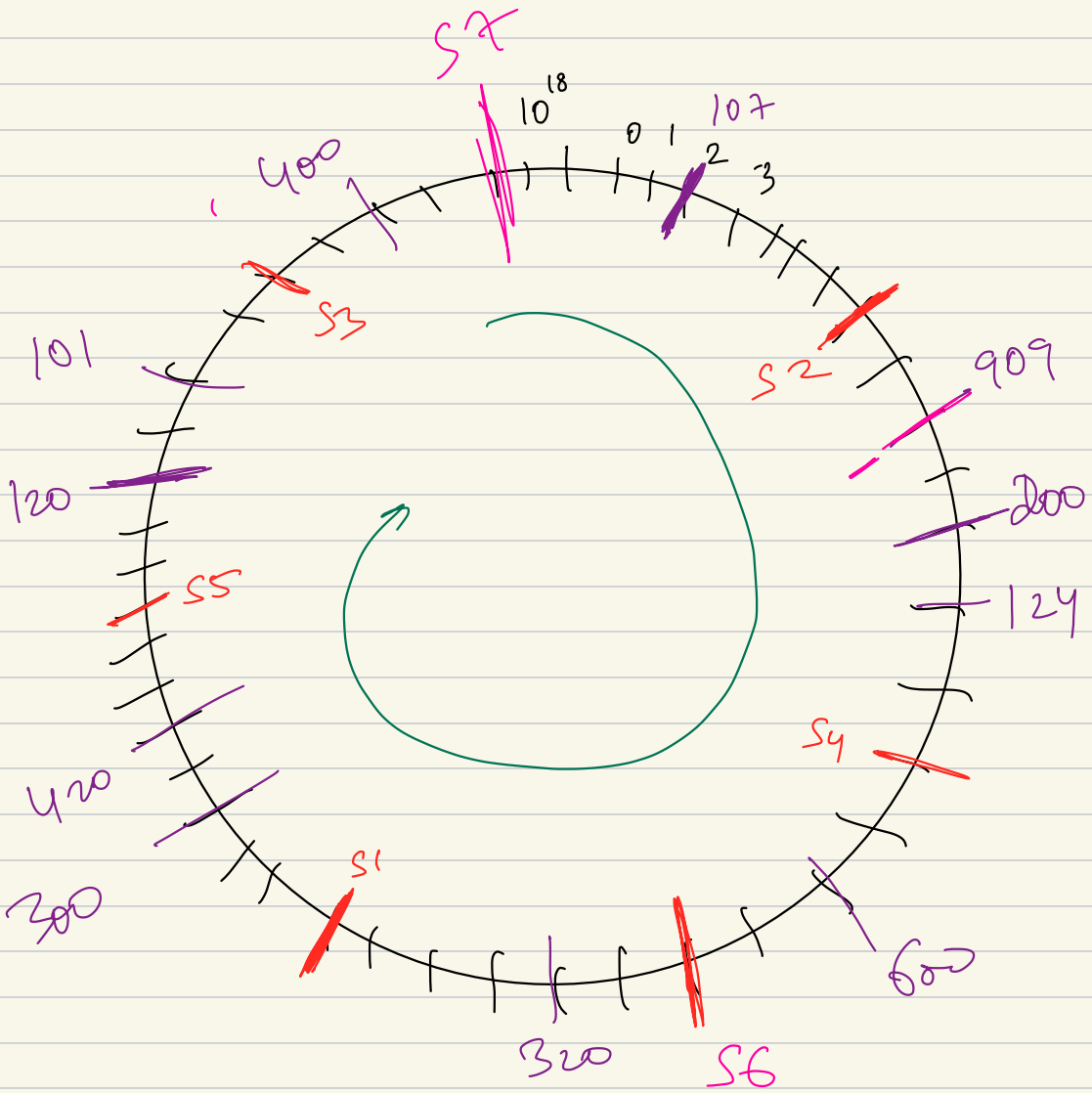
$H_S(S_5)$

H_R (request_id) = [$0 - 10^{18}$]

Sharding_id

req_id

T_1	107	$H_R(107) =$ value
T_2	120	$H_R(120) =$ value 2
T_3	200	$H_R(200) =$ value 3
T_{100}	107	$H_R(107) =$ value



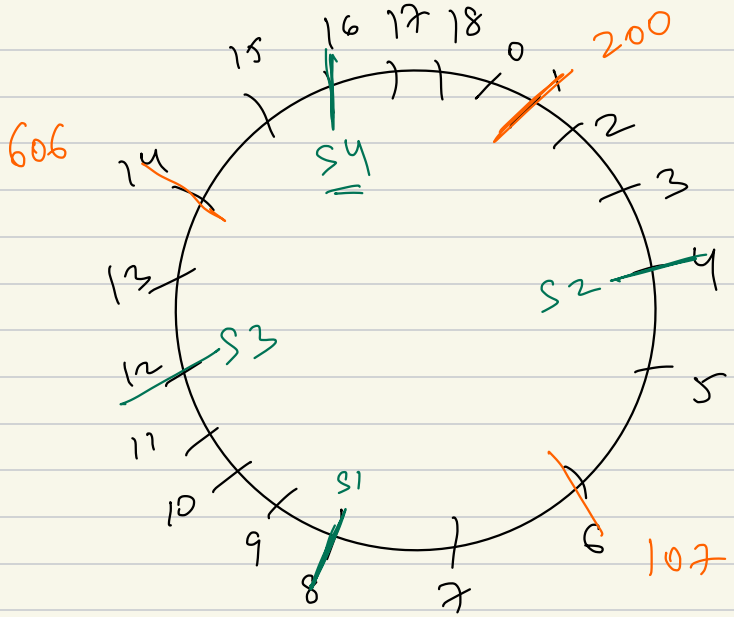
S6 gets added

$H_S(S_6) = \text{value}$

$H_S \equiv \left[7x^2 + 19x + 104x^3 - x^{24}\right] \% \cdot 10^{18}$

$H_R \equiv \left[\qquad\qquad\qquad\qquad\qquad\qquad\right] \% \cdot 10^{18}$

$\equiv$



add S4

H_s (server 1) = 8

H_s (server 2) = 4

H_s (server 3) = 12

H_s (server 4) = 16

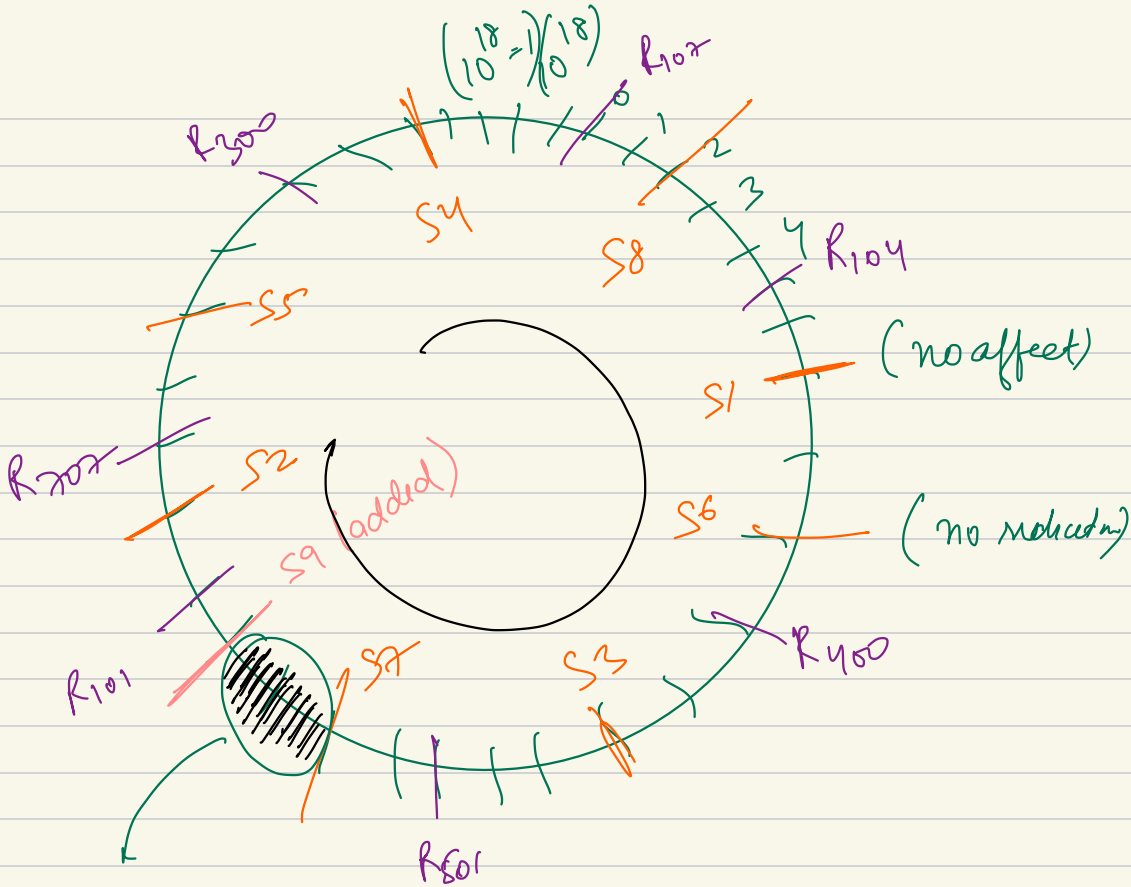
H_R (107) = 6

H_R (200) = 1

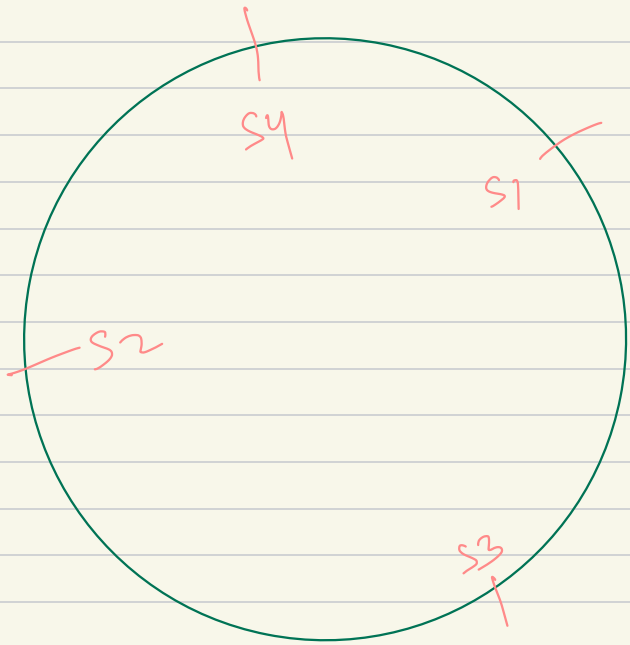
H_R (606) = 14

only 606 needs to move!!!

S2 → S4



only s2 lost load!!!



Remove S4

S2 same

S3 same

S1 will be load