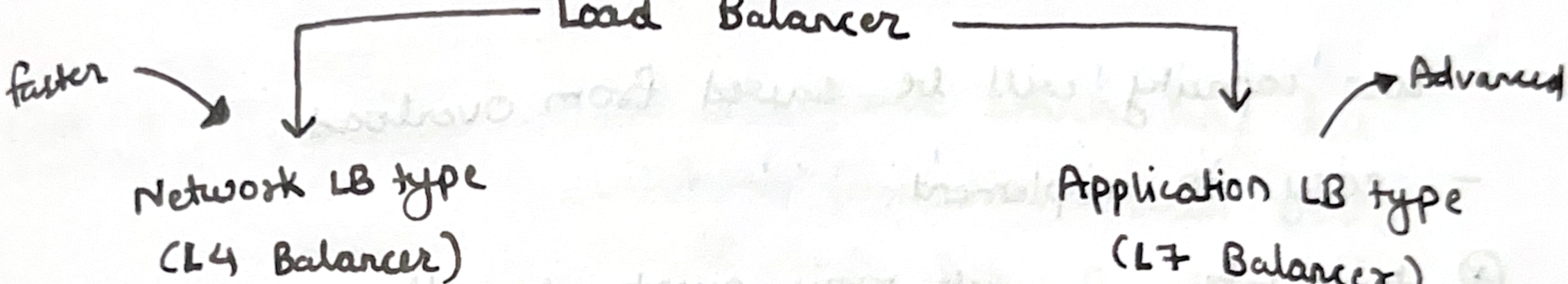
Load Balancer

⇒ distribute traffic (incoming) among multiple servers, as evenly as possible.



⇒ can read ports, IPs (source & dest) and takes decision on basis of that

⇒ faster in working.

⇒ can read data, Header, cookies, response Session and take decision

⇒ capablity to do caching (as it knows response)

⊛ Static load balancers :

1) Round Robin : takes each server turn by turn

server: 1 → 2 → 3 → 1 → 2 → 3 ....

req: 1st 2nd 3rd 4th 5th 6th....

- very easy to implement
- equal load to all servers

⊛ each server will be treated same inspite of capacity

- server can go down with overload of req. (Low capacity)

2) weighted Round Robin: each server is assigned a weight and request are assigned in round robin but ⊛ first capacity of server is exausted then moves to next server

1 → 3

2 → 1

server: 1 → 1 → 1 → 2 → 1 → 1

req: 1st 2nd 3rd 4th 5th 6th

- Low capacity will be saved from overload
- easy to implement

⊛ time taking requests may overburden the Low capacity server.

3) IP Hash: using Source IP, we compute the hash; and based on hash we move it to sevent server.

- good for usecase, where same IP client need to connect to same server,

- ⊛ if client request is coming via PROXY, it would overwhelm the server.
- ⊛ cannot ensure equal distribution.

dynamic LB:

1) <u>Least Connection</u>: check which server has least active connection for any new request
   - Less chance of overburdening a server, when each server has equal capacity
   - ⊛ TCP connection can be active without any traffic
   - ⊛ unequal capacity can lead to a chance of overburdening.

2) <u>weighted least connection</u>: Calculate Ratio of active connection to its weight, server with minimum ratio gets the request.

3) <u>Least Response time</u>: Picks the server which has less (Active Connection * Least TTFB)
   if clash follow Round Robin
   Time to first Byte ←
   [time interval bet" sending a req. and receiving response from server]

Network (L4) LB

Static

- Round Robin
- Weighted RR
- IP Hash

Dynamic

- least conn.
- weighted least conn.
- least resp. time