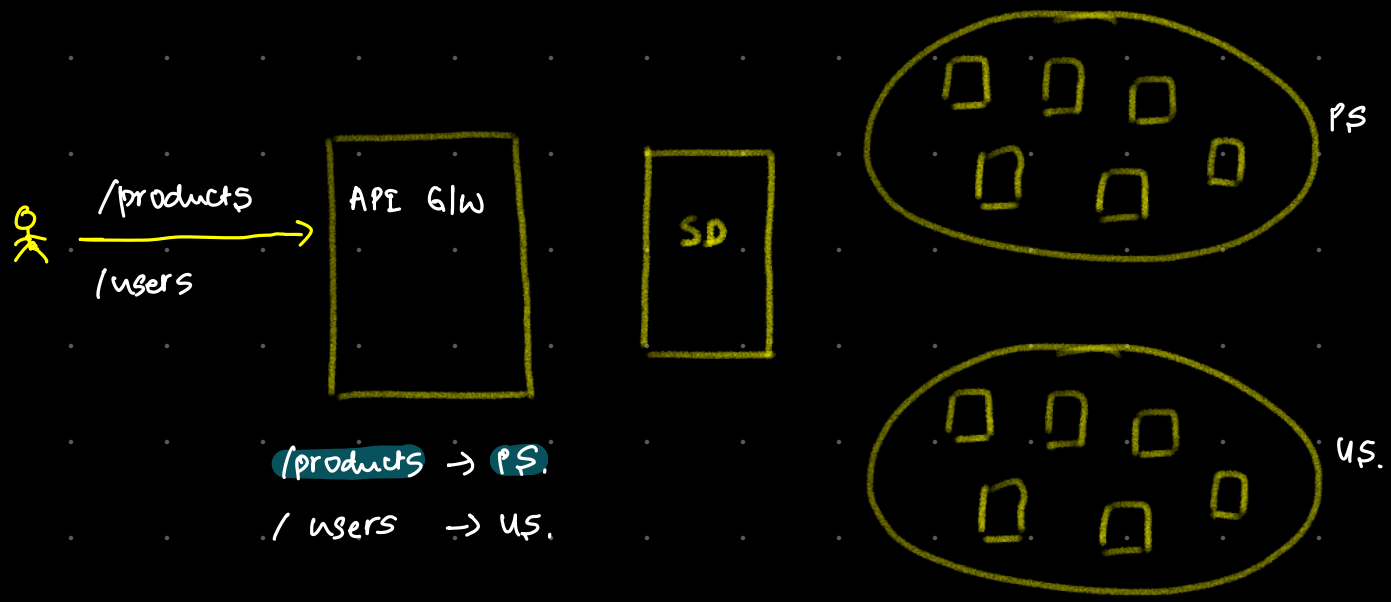
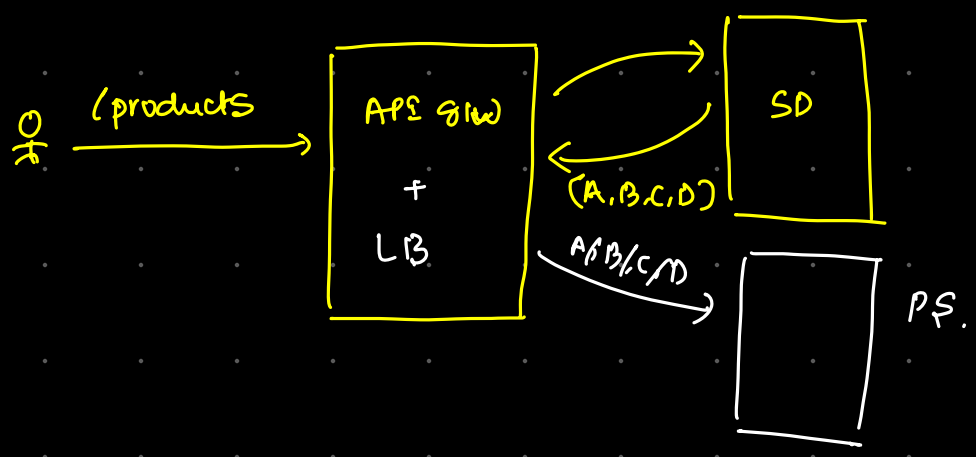


Two options.

(Client side LB)

① API g/w. → service discovery → LB @ API g/w → service Instr

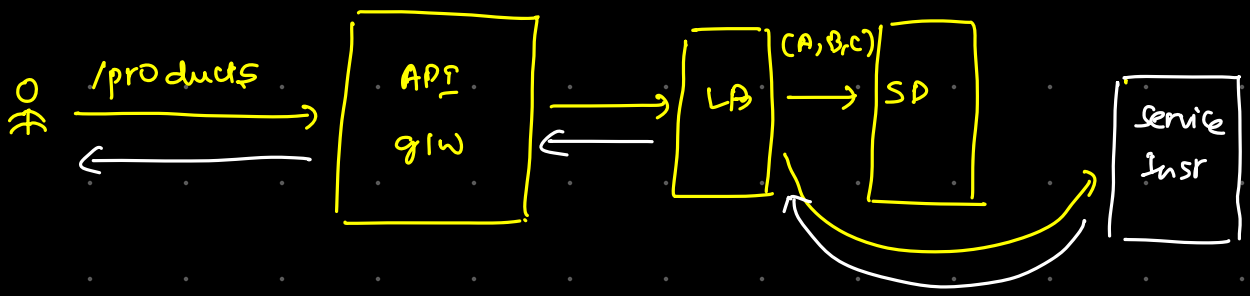
[A, B, C, D, E] [C]



Server side LB.

② API g/w → LB of service → service discovery → service Instr

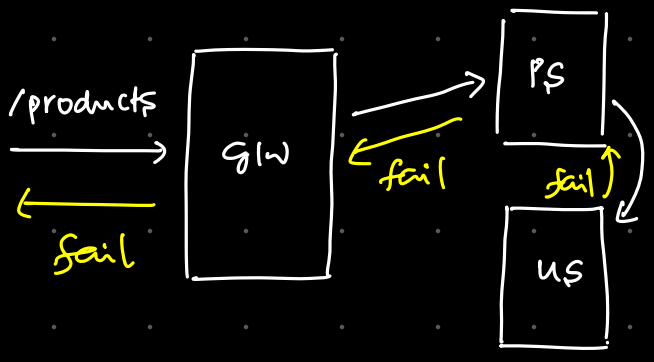
[A, B, C, D, E]



On.

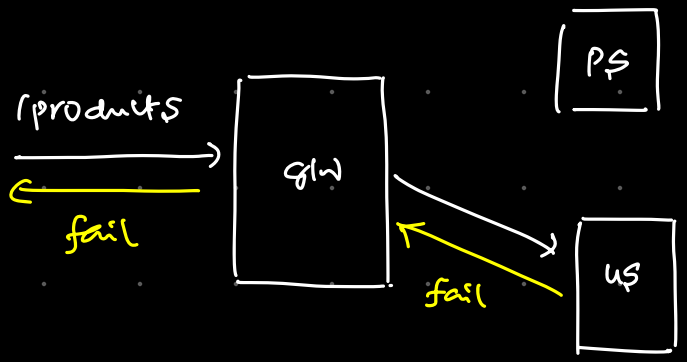
1. Can API gw take care of authentication.

① API gw not doing auth.



auth after API gw.

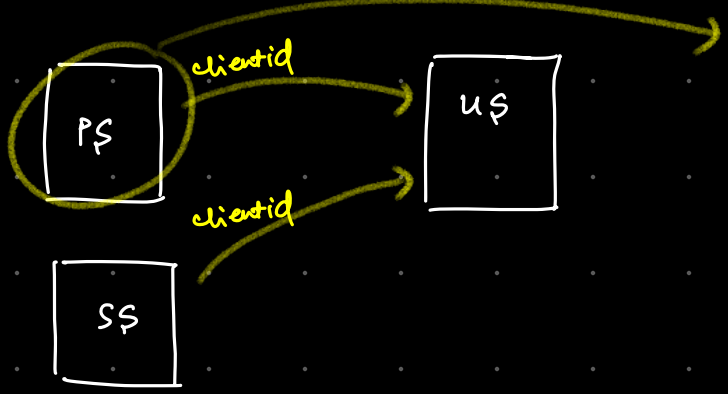
① API gw not doing auth



We're filtering the requests @ gw itself.

Throttling

limiting the no. of requests.



> 100 req / second.
Throttling.

API gw can also be used for throttling.

API gw in action

Monitoring

100 instances in each service

around 20-30 services.

You would want to monitor:-

1. CPU utilization
2. Disk utilization
3. RAM utilization
4. latency of API's [How much time API takes to respond].
5. Requests per second / RPS / QPS.

Actuator

It exposes some endpoints where you see multiple details about resources and other things.

localhost:8080/actuator

① /metrics : memory usage, CPU usage etc

② /health : health info about the app^n.

Prometheus

This will hit the metric api's periodically and capture the data in a time series format.

Grafana

Queries prometheus to receive metrics and visulalizes it through customisable dashboards.