#ASLI ENGINEERING

Throttling and Rate Limitting

SWIPE

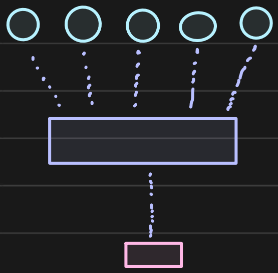
BY

ARPIT BHAYANI

Throttling and Rate limitting

What is throttling?

Throttling is a technique that ensures that the flow of data being sent at the target machine/service/sub-system can be digested at an acceptable rate.



Throttling is more of a defensive measure.

- Throttling could be slowing
- Throttling could be rejecting
- Throtting could be ignoring

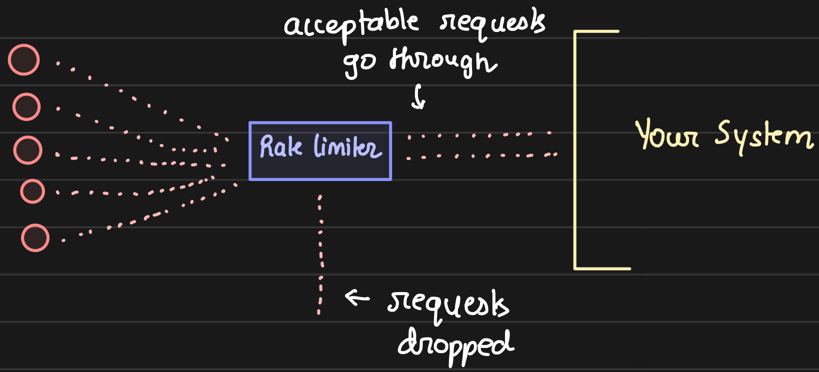
Why do we need throttling in the first place?

1. To prevent system abuse
2. To only allow traffic that could be handled
3. Control consumption cost
4. To prevent cascading failures

Use-cases of throttling...

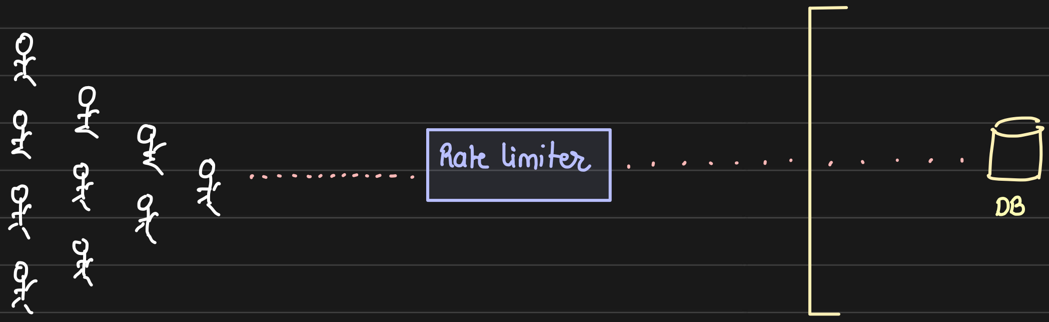
Use-cases of Throttling

1 Prevents catastrophic DDoS attack



2. Gracefully handle a surge of users

eg: if your website/ product went viral

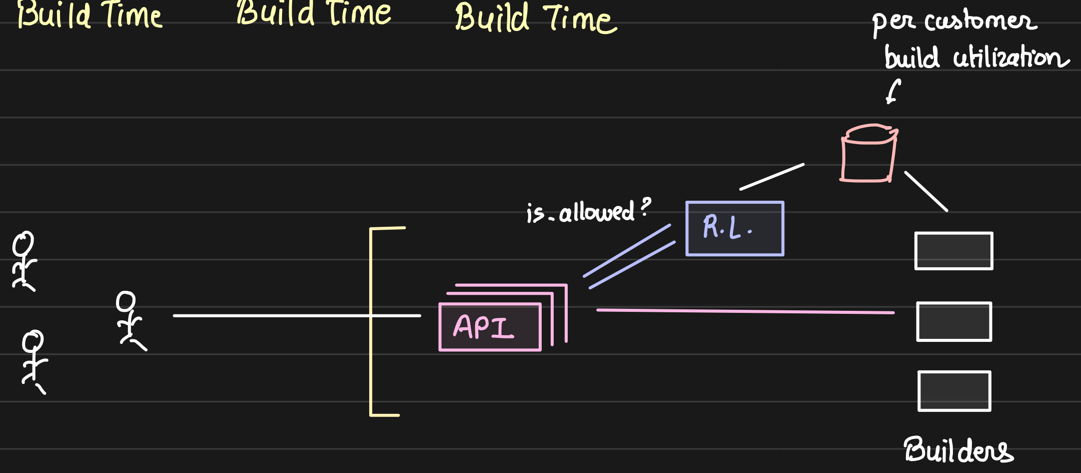


Your infra is not overwhelmed
& continues to be up & serving
a fraction of your users

3. Multi-tiered limits

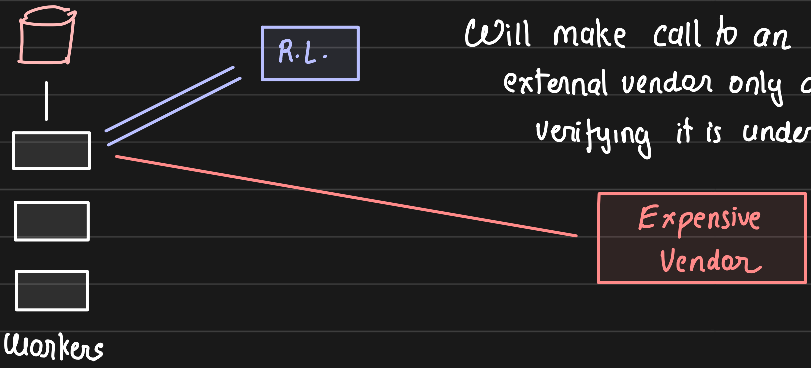
eg: Say you are a CICD company who offers multi-tiered pricing

Tier 1	Tier 2	Tier 3
200 min Build Time	1000 mins Build Time	∞ mins Build Time



4. You not overusing a third-party system

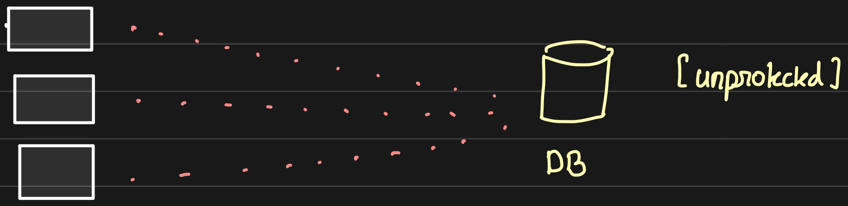
eg: You are consuming an expensive third-party API and their pricing is aggressive



Will make call to an external vendor only after verifying it is under acceptable limits

5. Not overwhelming your own unprotected systems

eg: Hard deleting from DB should be uniformly distributed



Deleting a million rows in one go, can take down your DB, and hence you should streamline the deletions and spread it uniformly