(Video 28: Percentile Tail Latency Explained)

When some senior engineer says -

My 95% = 30ms, My 99% = 100ms, 99.9999% = 1000ms.

What do these terms mean?

Lets say a million requests are coming to your application, and You say to Your customers.

1) Minimum latency → 1ms { Doesn't mean anything, there could be lot of requests whose latency is 1sec. }

2) Maximum Latency → 300ms → { Still Does not mean anything. What if most of the requests are 299 ms }. → User cannot rely on this.

3). Average latency → (1000ms) → { lets say all our requests were (300 − 100) ms, but only 2 requests which took 10 seconds, they polluted the whole data }.

(Percentile)

When they say 95th percentile latency is 30ms.

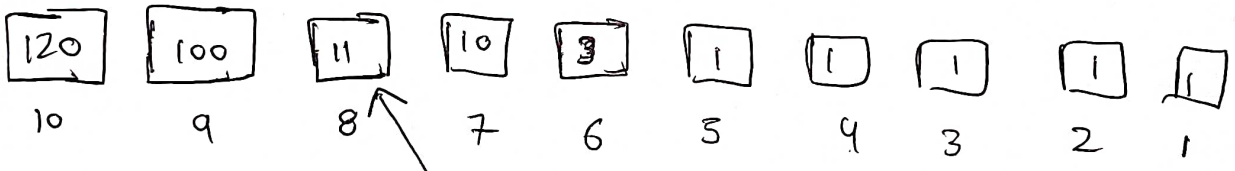
It means if total 100 requests were sent, then

95 of them were below 30ms.

Basically if you sort the latencies in ascending order,

then {95th one will have a latency ≤ 30ms}

Lets say you are doing analysis on 10 requests you have



If someone asks for 75 percentile, we do

$\left(\frac{75}{100} \times \text{total No of request}\right) = 7.5\uparrow$ Round it off to next decimal

$= 8.$

75 percentile of requests are responding in less than 11ms.

Similarly 99.99% $\Rightarrow \frac{99.99}{100} \times 10 = 9.99\ldots\uparrow = 10.$

99.99 percentile of requests are responding in less than 120ms