Recitation 6: Tail at Scale

Question

What do we mean when we talk about "latency" in a system?

Answer: Latency in a system refers to the time delay between when a request is made and when the response is received. In a distributed system, this delay can be caused by network congestion, server processing time, and resource contention.

Question

Why does latency, and in particular the tail of the latency distribution in a system, matter? Who is impacted by it?

Answer: The tail of the latency distribution refers to the highest response times observed. High tail latency is particularly problematic because even if the average latency is low, a small percentage of requests experiencing extreme delays can significantly impact user experience. It matters because in large-scale applications e.g. search engines or financial trading platforms, slow responses can lead to a decreased user base or even cause financial losses.

Question

How can a distributed system decrease this tail latency? Give one example. Remember to answer in your own words, not using text from the paper.

Answer: To reduce tail latency, distributed systems often use techniques like redundant requests, where a request is sent to multiple servers, and the first response received is used. This approach helps mitigate the impact of slow servers or network delays. For instance, Google's search infrastructure employs redundant queries to ensure that the slowest server does not become a bottleneck, improving overall system responsiveness.