Unavailability and Failure

Figure 1: Number of failed nodes over a single month period in a 3000 node production cluster of Facebook.

Unavailability and Failure

- In a 10000-server DC, with 10000-day MTBF machines, one machine will fail everyday on average

- Build fault-tolerant software infrastructure and hide failure-handling complexity from application-level software as much as possible

- Configuration is one of the largest sources of service disruption
  - Others: resource preemption, maintenance, network partition, …

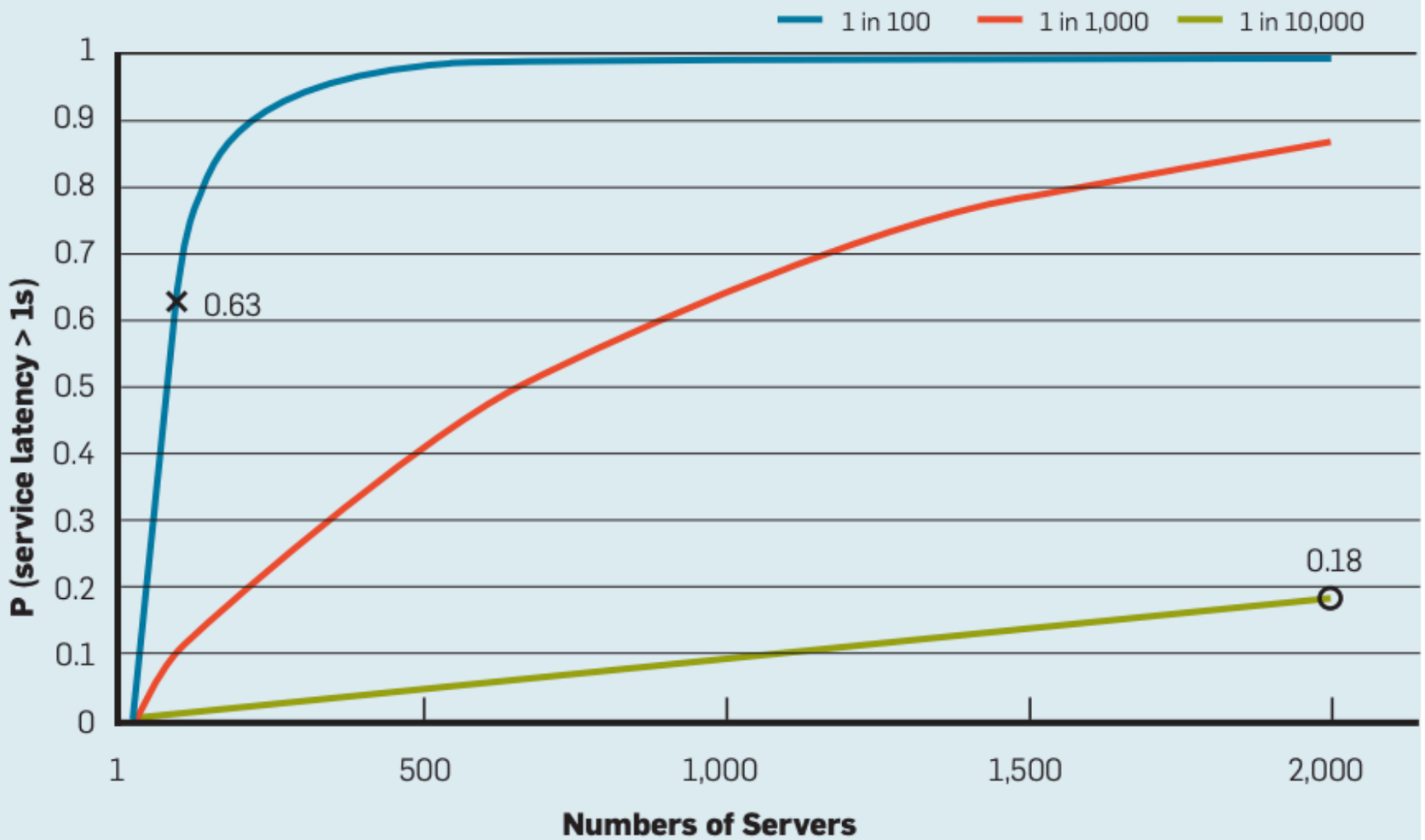
- Storage subsystems are the biggest sources of machine crashes

Related Techniques

Technique	Performance	Availability
Replication	X	X
Erasure coding	X	X
Sharding/partitioning	X	X
Load balancing	X	
Health checks		X
Integrity checks		X
Compression	X	
Eventual consistency	X	X
Centralized controller	X	
Canaries		X
Redundant execution	X	

The Tail at Scale

Probability of one-second service-level response time as the system scales and frequency of server-level high-latency outliers varies.



The Tail at Scale

Table 1. Individual-leaf-request finishing times for a large fan-out service tree (measured from root node of the tree).

	50%ile latency	95%ile latency	99%ile latency
One random leaf finishes	1ms	5ms	10ms
95% of all leaf requests finish	12ms	32ms	70ms
100% of all leaf requests finish	40ms	87ms	140ms