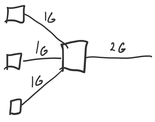
Lecture 9

Friday, 6 October 2023 08:35

overcubscription



overload

---

no uniform definition of scalability in literature

definitions:

1. ability to handle increasing workload without adding resources
2. ability to handle increased workload by repeatedly applying a cost-effective strategy for extending a system's capacity

load scalability — deal with increased load

space scalability — deal with increased memory demand (primary, secondary storage)

space-time scalability — deal with increased demand for number of objects

structural scalability — deal with expanding system without major modifications to its architecture

slide 15, middle figure: network limit $= \frac{16 \text{ servers} \cdot 12.5 \text{ MB/s}}{3 \text{ replicas}} \approx 67 \text{ MB/s}$

scaling has no effect on the sequential part of the system, and a linear effect on the parallel part

Amdahl's law keeps the problem size; the

$$\text{speedup} = \frac{1}{1-p+\frac{p}{N}}$$

with p the portion of the program which can run in parallel and N the number of processors

Gustafson's law observes that parallelism increases in applications when the problem size increases

$$\text{speedup} = S + p \cdot N$$

with S the execution time of the serial portion, p the parallelizable portion of the program, N the number of processors

Amdahl's law: if timing one variable does not work, time another
Gustafson's law: use more resources to solve bigger problems

tactics for scalability:

- limit dependencies
  - no machine has (real) complete information
  - decisions are based on local information
  - failure of one machine does not harm results
  - no assumption of global/shared clock
- techniques
  - hide latency — do something useful while waiting
  - introduce concurrency
  - partition
  - limit communication

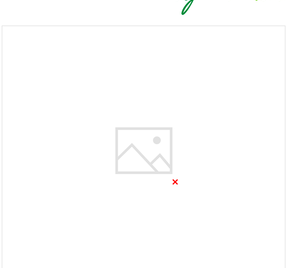
scalability is a prerequisite for elasticity

elasticity is the ability of a system to adapt to workload changes by provisioning and deprovisioning in an autonomic manner

adaptation mechanism: auto scalers

aspects of elasticity:

- timing
  - % of under-/over-provisioned time
  - frequency of adaptation
- latency
  - time to bring up/tear down resources
- accuracy
  - relative deviation of allocated resources from actual demand



jitter $J = \frac{E_S - E_D}{T}$, where E_S is the number of supply changes during interval T and similarly for E_D