Scalability
Describing Performance

- Throughput - batch processing system
  - e.g., the # of records per second to be processed, the total time to run a job on a dataset
- Response time - online system
  - the time between a client sending a request and receiving a response
  - Percentiles can be a good measure
    - Median (50th percentile) - how long users typically have to wait
    - High percentiles or tail latencies - may measure the most valuable customers
    - Often used to define SLO and SLA
  - Queuing delays (head-of-line blocking)
    - a small number of slow requests hold up the processing of subsequent requests
    - the impact to the client is often more significant than the server side

Scalability

Approaches for Coping with Load

- Scale up v.s. scale out
- Shared-nothing architecture
- Elastic v.s. manually called system
- No silver bullets