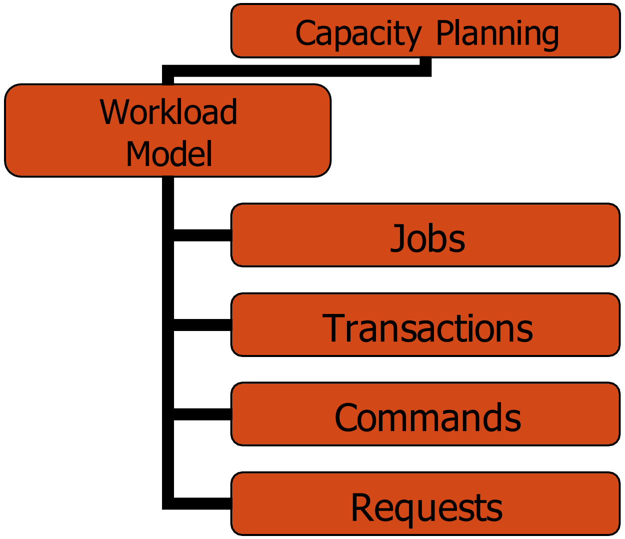
Performance Prediction

Workload models and Performance models

Workload Characterization

- Describe workload precisely
- Identification of the basic components (the job, the transaction, the command, the request)
- Characterization yields workload model



Workload Characterization

- Highest level:
  functional characterization (the programs or applications)
  Needs high-level information about resource requirements
- Physical level:
  Resource-oriented characterization (the resource consumption by the workload)

Workload Characterization

- Workload model is a representation that mimics the real workload.
- It comes from observations and brings compactness
- Synthetic models. Natural synthetic (benchmark) and hybrid synthetic
- Artificial models. Executable (suit of programs) and non-executable (set of parameters)

Workload Characterization

- Non-executable models for PE
  - Program inter arrival time
  - Service demand
  - Program size
  - Execution mix

Workload Characterization

- Frequency distribution of the requests
- Request inter arrival time distribution
- File referencing behavior
- Size of reads and writes which have influence

Workload Characterization

- Input parameters for analytical models
  - Workload intensity
  - Service demand
  - Basic components
- Simplifying assumptions (homogeneity, Poisson flow, etc.)
- Internet traffic as a workload

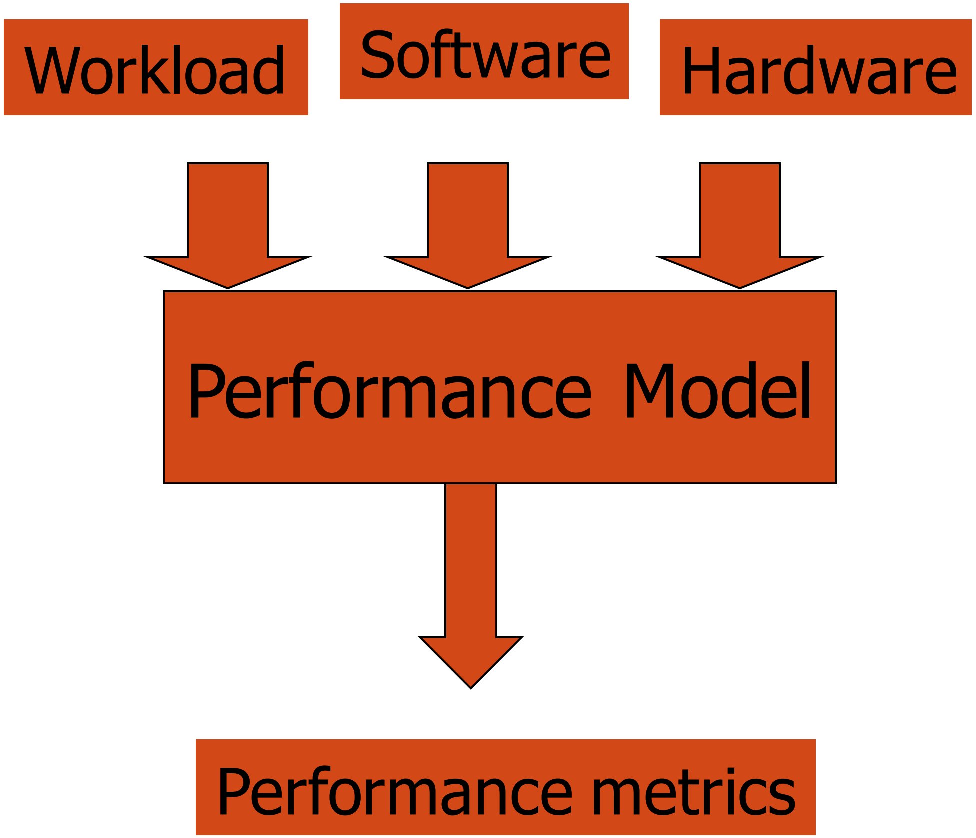
Performance Models

- Model is the presentation of the system
- Functional models (verbal description)
- Analytical model (set of equations)

Performance Models

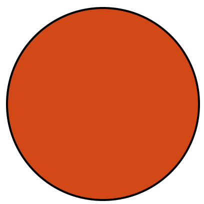
- Workload Parameters (arrival rate, number of terminals etc.)
- Software parameters (level of multiprogramming? Priorities etc.)
- Hardware parameters (CPU frequency? Disk speed, channels throughput etc.)

Performance Models

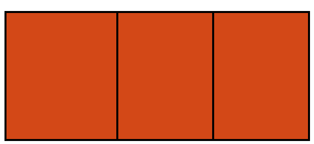


Performance Models

Elements of the graphical presentation



Server (Device)



Queue

Direction of the customer’s movement

Performance Models

- Server is active device, i.e. CPU, disk, data link etc.
- Passive devices, i.e. memory, other types of data storage

Simple Server System

Problem Statement

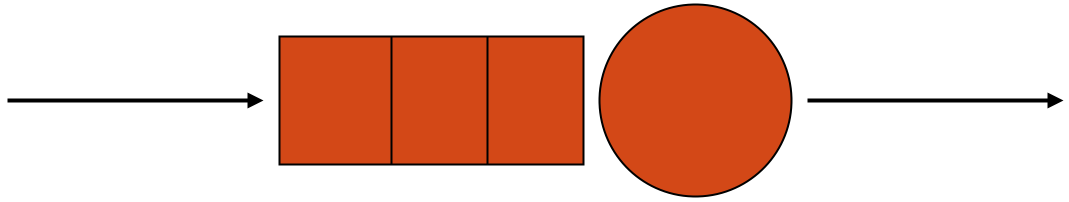
- Average of 30 tps arrive at a system. Each transaction needs 20 msec of processing from the CPU.

Questions:

- What is current performance of the system?
- What if workload increases by 50% more?
- What if processor then is upgraded by 50 %?

Simple Server System

- Single-server waiting queue



Simple Server System

The model parameterization

- Workload parameter (arrival rate)

  $\lambda = 30$ tps

- Service parameter (service rate)

$$\mu = \frac{1 \text{ transaction}}{20 \text{ msec}} = 50 \text{ tps}$$

Simple Server System

Key performance metrics

Metrics of manager

- Utilization (system is busy, unit is %)
- Throughput (customers served by the system per time unit)

Metrics of customer

- Average queue length
- Response time

Simple Server System

- Utilization. System is busy in all states excepted P_0
  $U = P_1 + P_1 + P_2 + \ldots = 1 - P_0 =$
  $1 – (1 - \rho) = \lambda \backslash \mu$
- Throughput. Zero if empty and μ if busy
  $T = 0P_0 + \mu P_1 + \mu P_2 + \ldots = \mu(1 - P_0) =$
  $\mu[1 - (1 - \rho)] = \mu\rho = \lambda$

Simple Server System

Average (expectation) queue length

$$\overline{n}_q = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n =$$

$$(1-\rho)\sum_{n=0}^{\infty} n\rho^n = \left(1-\frac{\lambda}{\mu}\right)\frac{\frac{\lambda}{\mu}}{\left(1-\frac{\lambda}{\mu}\right)^2} = \frac{\lambda}{\mu-\lambda}$$

Response time

$$D = \frac{1}{\mu} \times \frac{\lambda}{\mu - \lambda} + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

Simple Server System

Answer the questions

- Baseline model: U = 60%, T= 30cps,
  n_q= 1.5, D = 1/20 sec.
- Prediction model: Workload increases 50% more. Now λ = 45 tps. This means U = 90%, T=45 cps, n_q= 9,
  D = 1/5 sec
- Thus if workload increases 50% more, utilization increases 50% more
  throughput increases 50 % more
  average queue length increases 500% more
  response time increases 300% more

Simple Server system

Little's low (Little 1961)

Mean number in the system=

Arrival rate x mean response time

- The low could be applied to the whole system or to its subsystems