Performance Evaluation and Capacity Planning

Basic concepts

PE & CP. Main textbooks

- R. Jain, The Art of Computer Systems Performance Analysis
- D. A. Menasce et al., Capacity Planning and Performance Modeling. From Mainframes to Client-Server Systems
- W. Feller, An Introduction to Probability Theory and Its Applications Vol. 1 and 2

The Capacity Planning Concept

- Capacity planning is the determination of the predicted time in future when system saturation is going to occur, and the most cost-effective way of delaying system saturation as long as possible.
- Why is saturation occurring?
- Which are the best alternatives
- What is bottleneck (Utilization 100%, Longest delay)

The Capacity Planning Concept

Why Capacity planning is important?

- User dissatisfaction
- External Image of the company
- Productivity decrease
- Budgetary constrains
- Risk of financial losses
- IS environment control

Capacity Planning Methodology

- What is the current installed capacity?
- What services should be provided in the future?
- What quality goals are planned for the services?
- What is the most cost-effective system configuration to handle current and future services and meet the planned quality goals?

Service Levels

- The service level is provided at a given cost. Service level rule the relationship.
- User expects: response time (satisfactory), availability (`uptime` expressed in %), reliability (as much as possible or…?), cost.
- Quantifiers: average, variation, quintiles.

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.)

How to solve performance model

- Formulate states of the system. Form the set of states.
- Formulate possible transitions between states and their flow rates.
- Check literature for the ready solution. If the solution exist, calculate performance metrics for baseline and prediction models.

How to solve performance model

- If not found, built global balance equations and solve them to obtain states probability distribution
- From the solution derive formula for key performance metrics.
- Calculate performance metrics for baseline and prediction models.

Interactive System

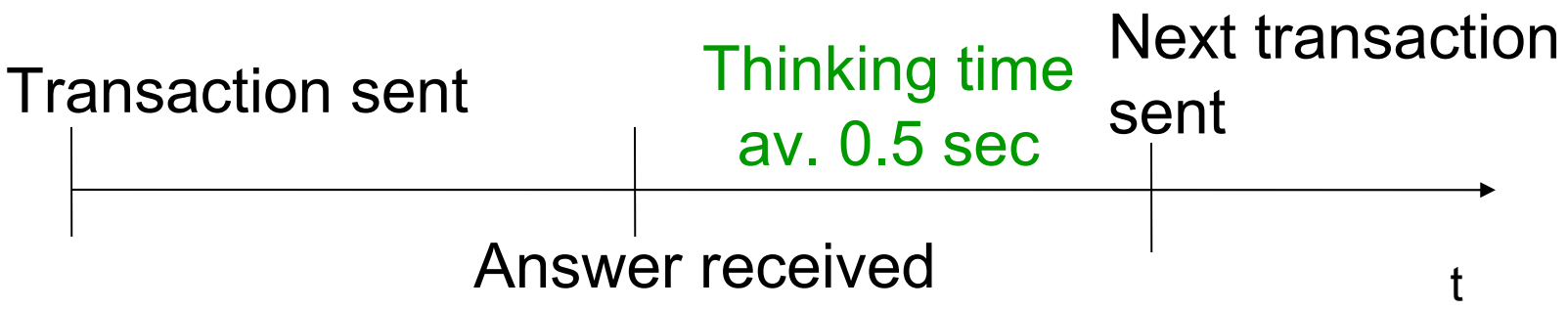
Problem statement

- There are number of user in the transaction system. Average thinking time is 0.5 sec
- The database system completes single request in 1.5 sec, two request 1 sec for each of them and 0.75 rps if three appeal to the database system simultaneously
- A user does not send new request while answer on the previous one arrives.

Interactive System

Questions

- What is current performance of the system?
- What if after training session thinking time will reduce to 0.4 sec?
- What if database system will work 50% faster?

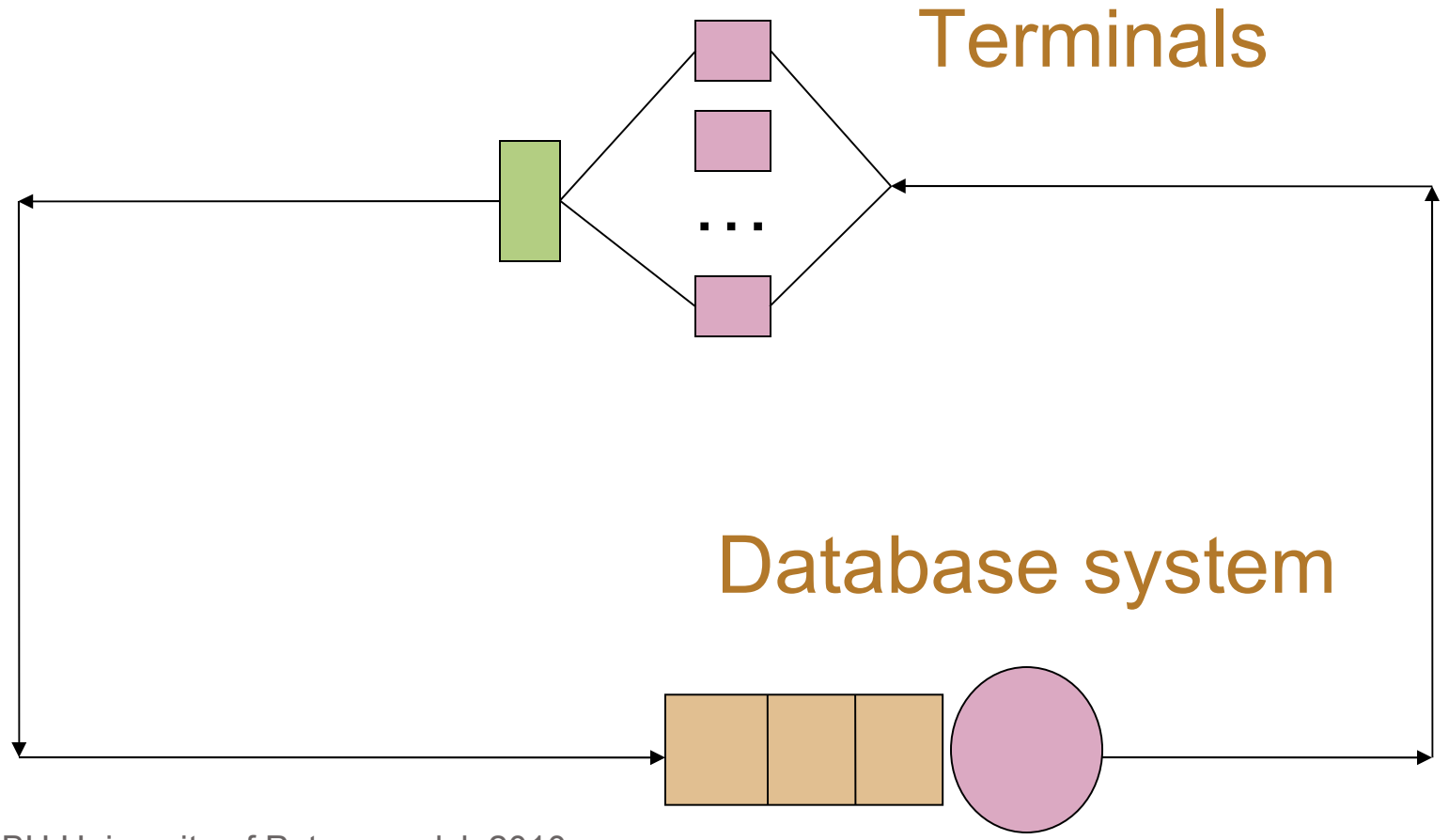


Interactive System

Functional description

- Number of terminals is 3. Workload type — transactions.
- User generates transaction, sends it to the database and waits for the answer.
- User receives the answer, thinks for a while then sends new transaction.
- Transaction arrives at database system and get service immediately
- Service intensity depends on the number of transaction in the database system

Interactive System



Interactive System

The model parameterization

- Workload parameters
  Each user can generate 2 tps
- Service rates: 1.5 sec for transaction means 2/3 tps

Number of req. in database	Average flow rate
0	6 tps
1	4 tps
2	2 tps

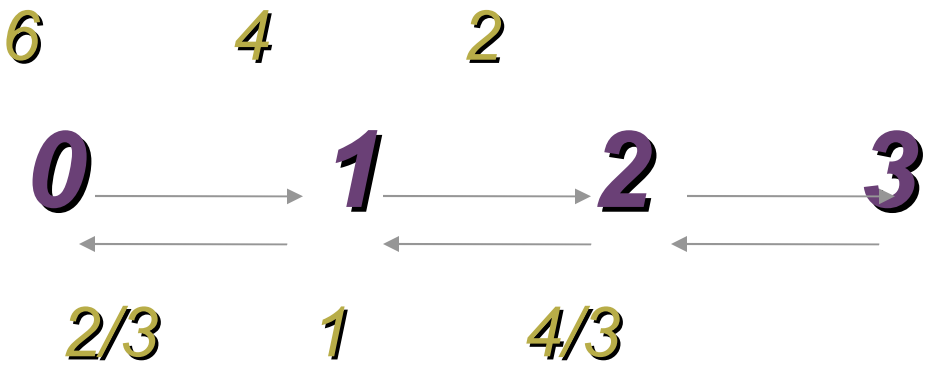
Number of req. in database	Average flow rate
1	2/3 tps
2	1 tps
3	4/3 tps

Interactive System

- n(t) = 0, 1, 2, 3 is number of the customers in the database at time t
- $\{n(t)\}_{t>0}$ is random process with continuous time
- Customers are statistically indistinguishable
- Old history is irrelevant
- $\{n(t)\}_{t>0}$ is Markov process

Interactive System

State transition diagram



Interactive System

Global balance equations

$$2/3P_1=6P_0$$

$$6P_0+1P_2=2/3P_1+4P_1$$

$$4P_1+4/3P_3=1P_2+2P_2$$

$$2P_2=4/3P_3$$

Normalizing condition

$$P_0+P_1+P_2+P_3=1$$

Interactive System

- The solution of the equations are
  P_0=0.01 P_1=0.09 P_2=0.36 P_3=0.54
- Utilization U = 1-P_0 = 1 – 0.01 = 99%
- Throughput
  T = 2/3P_1+1P_2+4/3P_3 = 1.14 tps
- Average queue length
  n_q = 1P_1+2P_2+3P_3 = 2.43 request
- Response time (by Little's low)
  D = n_q /T = 2.43/1.14 = 2.13 sec