From Systems to Descriptive Models

Prof. Daniel A. Menascé
Department of Computer Science
George Mason University
www.cs.gmu.edu/faculty/menasce.html

Copyright Notice

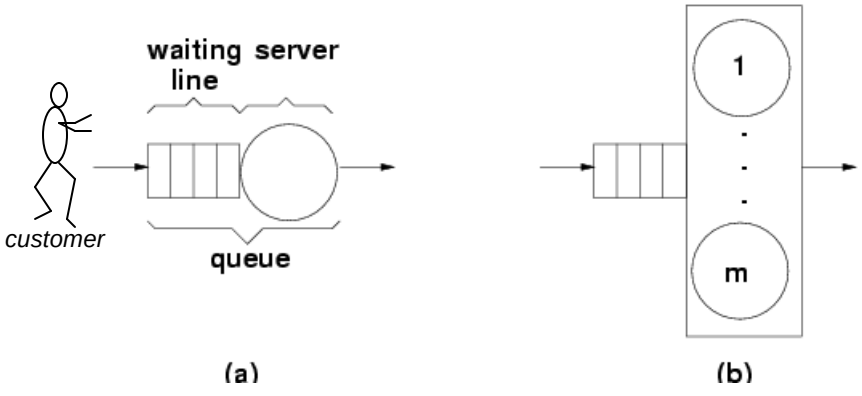
- Most of the figures in this set of slides come from the book "Performance by Design: computer capacity planning by example," by Menascé, Almeida, and Dowdy, Prentice Hall, 2004. It is strictly forbidden to copy, post on a Web site, or distribute electronically, in part or entirely, any of the slides in this file.

Modeling

- Abstraction of a real system.
- Should capture enough details to satisfy goals of the study.
- Types of models:
  - Simulation
  - Analytic
  - Hybrid

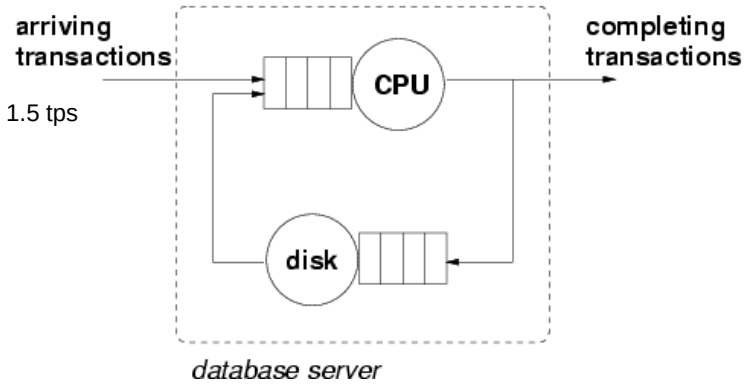
3

Graphical Notation for Queues



4

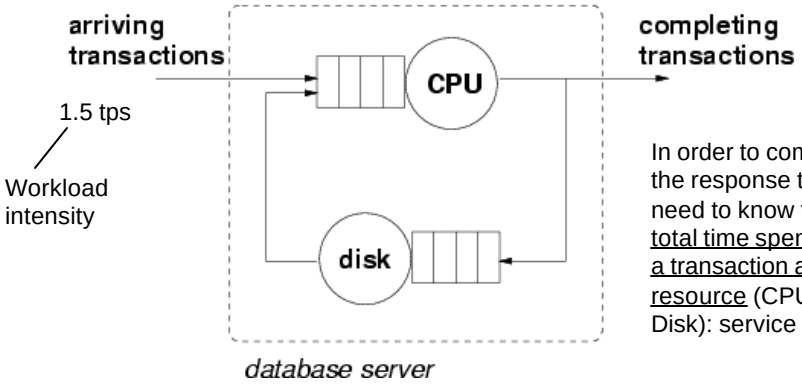
Simple Database Server Example: Open Workload



5

Simple Database Server Example: Open Workload

Queuing Network (QN)



Workload intensity

In order to compute the response time, we need to know the total time spent by a transaction at each resource (CPU and Disk): service demand.

6

Multiclass DB Example

transaction group	percent of total	avg. CPU time (sec)	avg. no. I/Os
Trivial	45%	0.04	5.5
Medium	25%	0.18	28.9
Complex	30%	1.20	85

Each transaction group is assigned to a customer class in the Queuing Network.

When do we need multiple classes?

- Heterogeneous service demands.
- Different types of workloads.
- Different service level objectives.

Open Class

- Workload intensity specified by an arrival rate (usually independent of the system state).
- Unbounded number of customers in the system.
- Throughput is an input parameter, which is equal to the arrival rate in equilibrium.

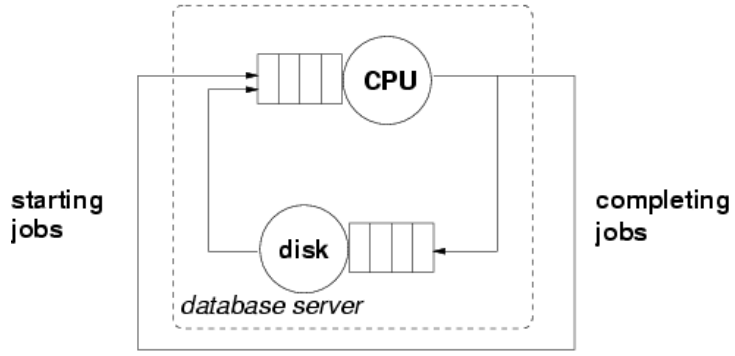
9

Closed Class

- Workload intensity specified by customer population (i.e., concurrency level)
- Bounded and known number of customers in the system.
- Throughput is an output parameter.

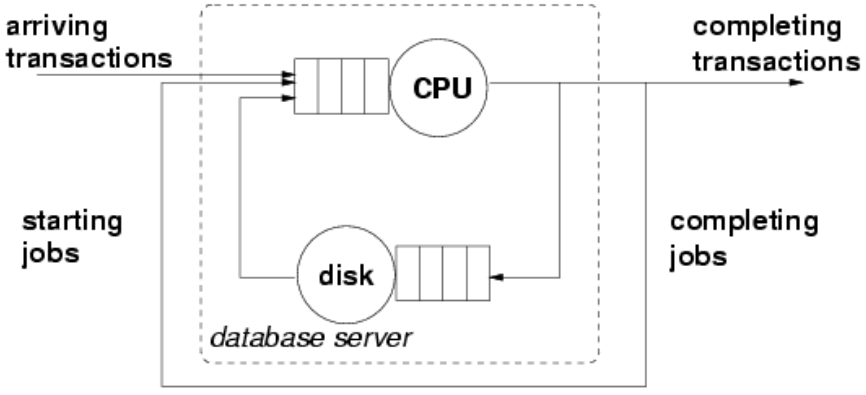
10

Closed Workload Example



11

Mixed Workload Example



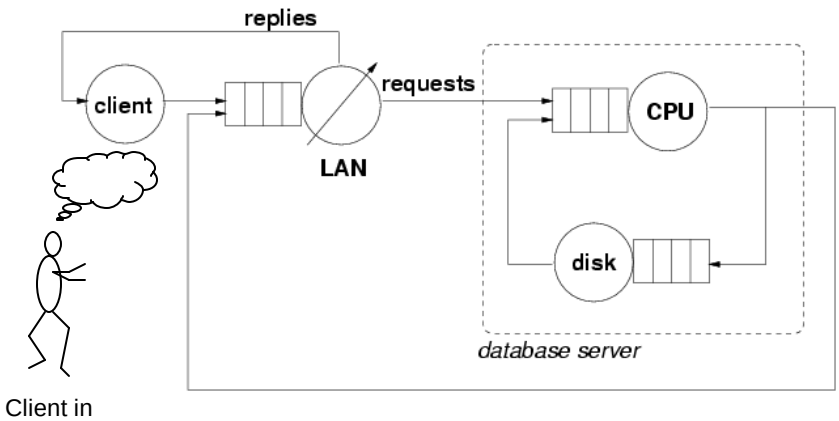
12

Service Level Agreements (SLA)

- Agreement between users and providers of computing services on the levels of various performance metrics.
  - 99.99% availability during 8:00am-11:00pm period and 99.9% at other times
  - Less than 4 sec page download time for requests over non-secure connections less than 6 sec for requests over secure connections
  - Minimum throughput of 2,000 page downloads/sec.

13

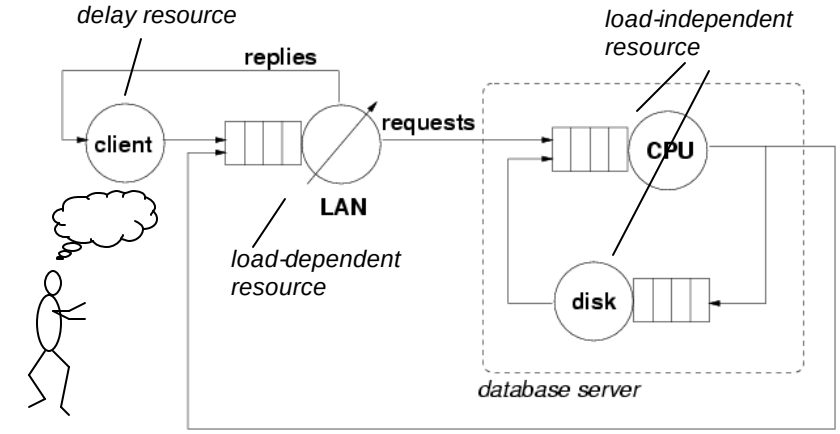
Different Types of Resources



Client in think mode

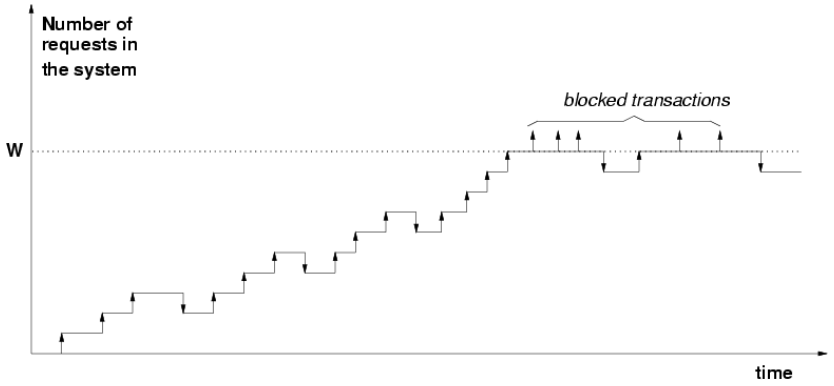
14

Different Types of Resources



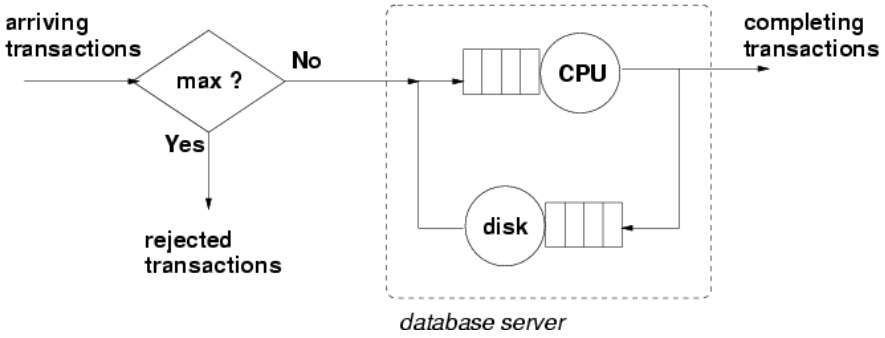
15

Blocking (Admission Control)



16

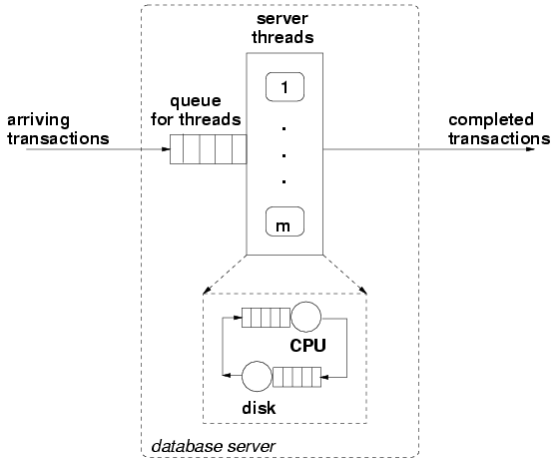
System with Admission Control



Throughput = Arrival Rate x (1 – Probability of Rejection)

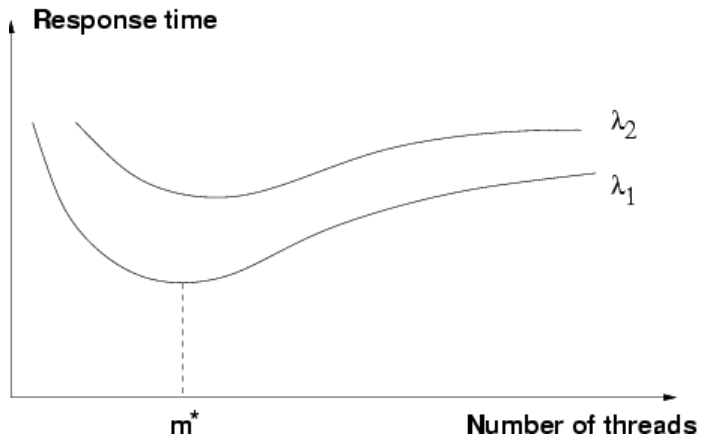
17

Database Server with Software Contention



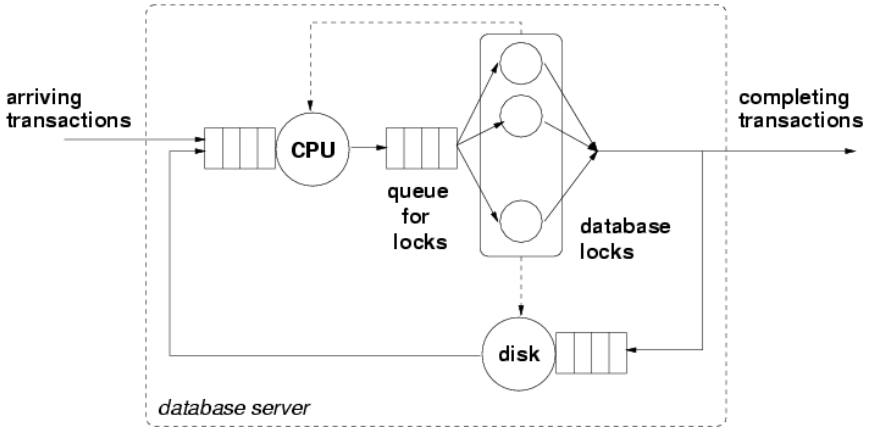
18

Effect of Software Contention



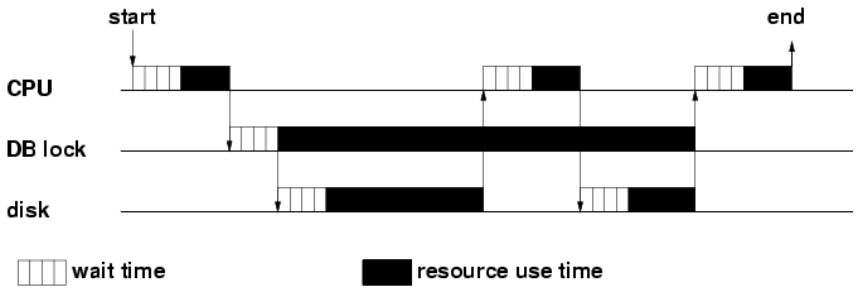
19

A DB Server with Lock Contention



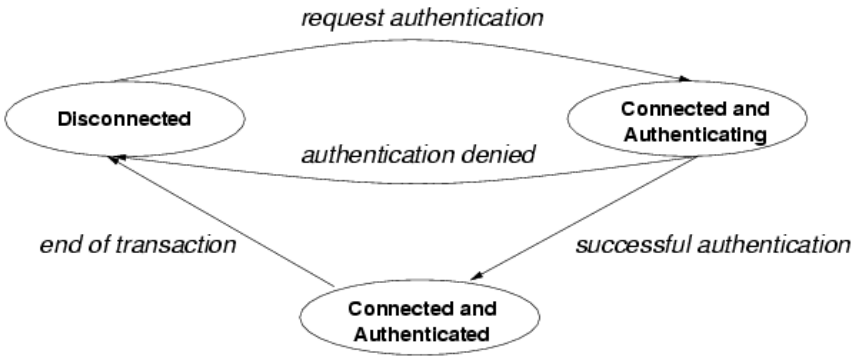
20

Simultaneous Resource Possession Time Axes



21

Class Switching



Authentication can be significantly more CPU-intensive than DB access.

22

Queuing Disciplines

- First Come First Served (FCFS)
- Priority Queuing (FCFS breaks the tie):
  - Non-preemptive
  - Preemptive resume
  - Preemptive repeat.
- Round-Robin (RR)
- Processor Sharing (PS)
- Last Come First Served-Preemptive Resume (LCFS-PR)
- Shortest Job First (SJF)

23

QN Models

- K queues (i = 1, …, K)
- R classes (r = 1, …, R)
- Input parameters:
  - Workload intensity

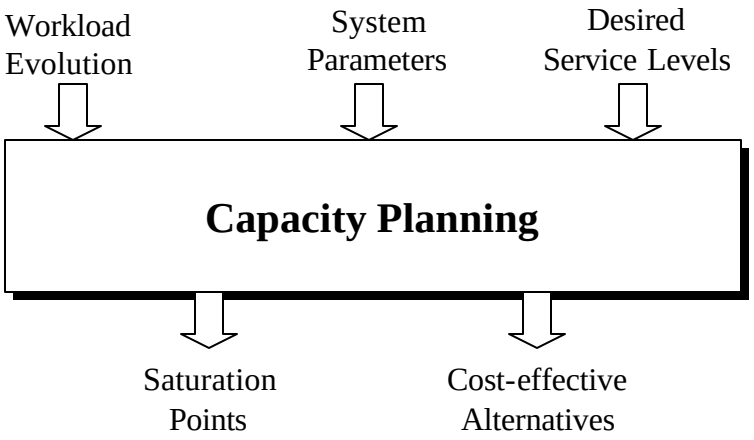
$$\vec{l} = (l_1, ..., l_r, ..., l_R)$$ for open classes

$$\vec{N} = (N_1, ..., N_r, ..., N_R)$$ for closed classes

  - Service demands (D_{ir})

24

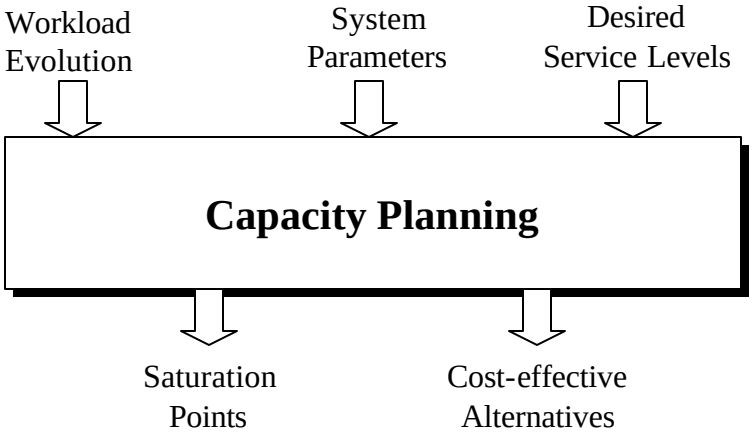
Capacity Planning Input and Output Variables



25

Capacity Planning Input and Output Variables

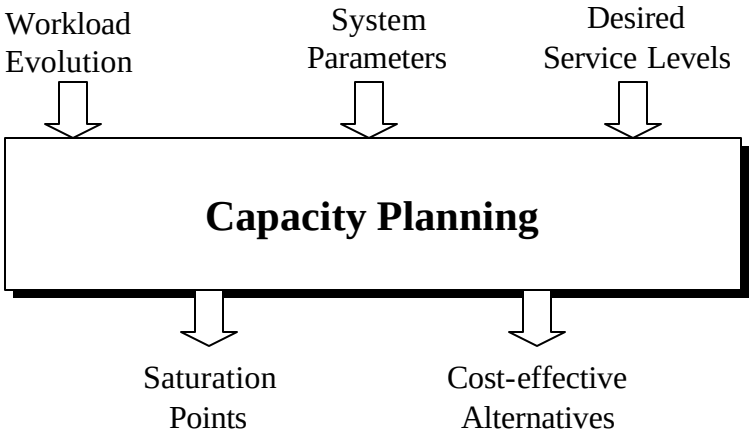
- *intensity*
- *nature*



26

Capacity Planning Input and Output Variables

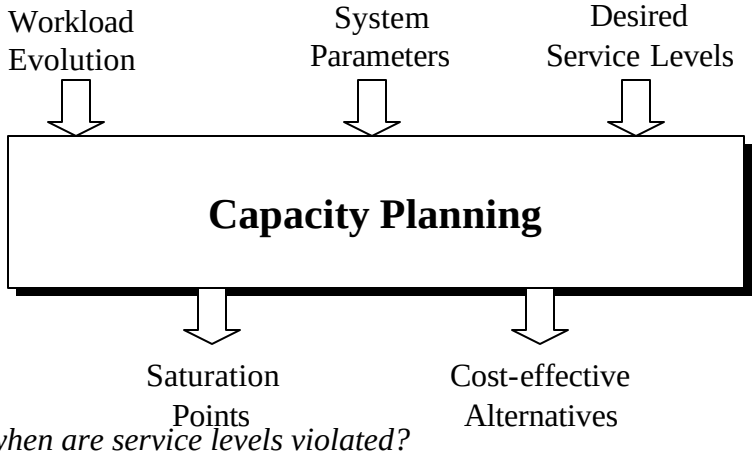
•*intensity* • *processors, disks, networks*
•*nature* • *max. no. connections*



27

Capacity Planning Input and Output Variables

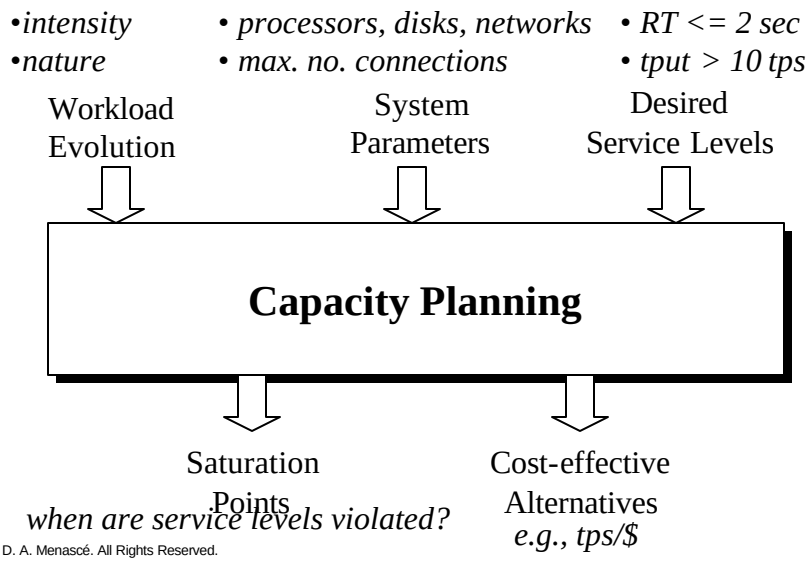
•*intensity* • *processors, disks, networks* • *RT <= 2 sec*
•*nature* • *max. no. connections* • *tput > 10 tps*



when are service levels violated?

28

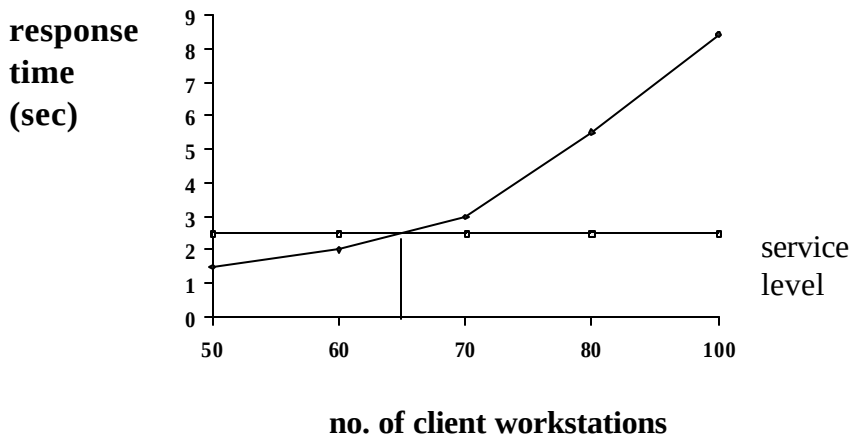
Capacity Planning Input and Output Variables



Capacity Planning Definition

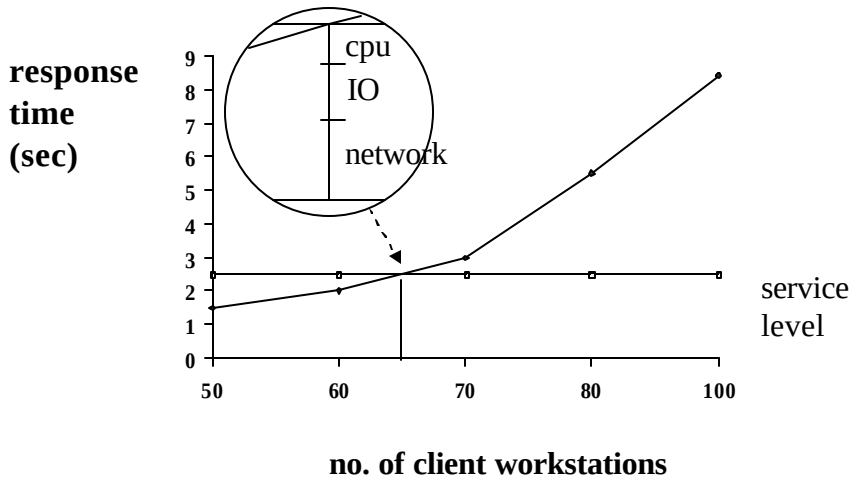
Capacity Planning is the process of *predicting* when the *service levels* will be violated as a function of the *workload evolution*, as well as the determination of the most cost-effective way of delaying system saturation.

Capacity Planning Concept



31

Capacity Planning Concept



32

Typical Capacity Planning Questions

- Situation: migrating from a mainframe based to a C/S system.
- Questions:
  - how many clients will the new system support with acceptable response time?
  - How many servers and how should they be configured to handle the load?
  - Should I use a two-tier or a three-tier architecture?

33