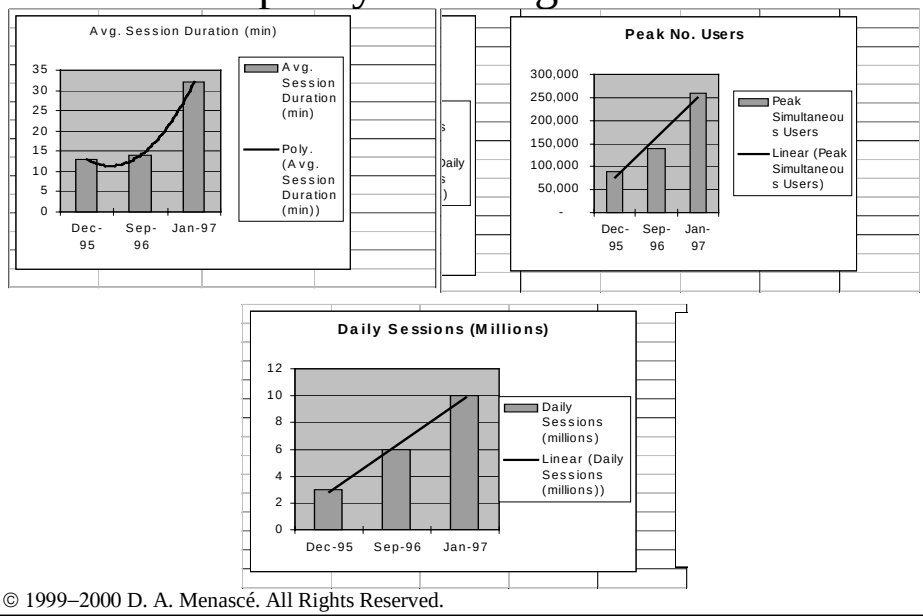
CS 672 – Introduction to Performance Evaluation and Capacity Planning

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

Examples of Performance Problems

- AOL
- Market crash of October 1997: investors lost millions of dollars!
- Holiday season of 1998: e-commerce sites were crowded!
- Online trading surge: http://www.nytimes.com/aponline/f/AP-Internet-Trading-Snarls.html

AOL Demand Increase and Capacity Planning Problem



Performance Problems Threaten E-commerce Success

- "... nascent attempt to reclaim its role as a leading source of information faltered Tuesday when <u>its Web site crashed after being flooded by visitors.</u>
  ... the overwhelming response caught the company unprepared."

  latimes.com October 20, 1999.

Performance Problems Threaten E-commerce Success

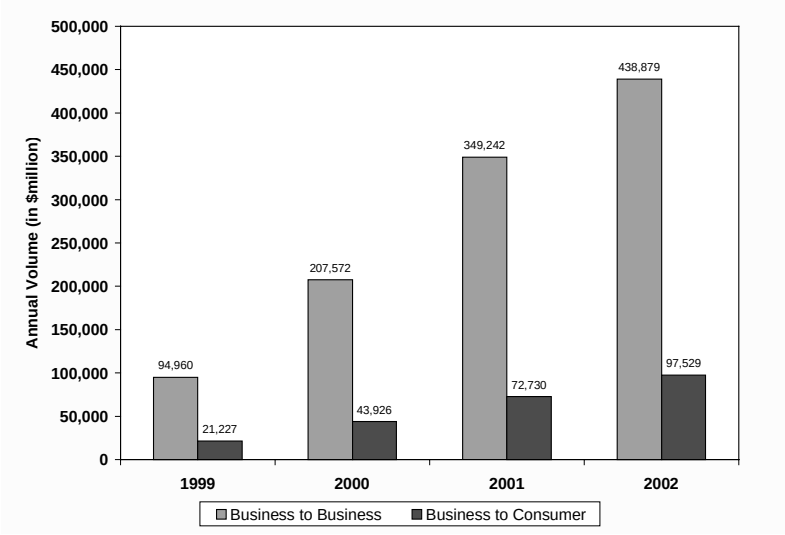
- The site went down around 4 p.m. PT yesterday while experiencing record traffic, according to a company spokeswoman. The site posted this message: "Please check back soon. Due to the high number of visitors currently using our site, we're unable to start your session at this time."

... and other e-commerce sites have experienced similar intermittent outages as the volume of visitors flooding their sites steadily increases.

CNET News.com October 22, 1999, 11:30 a.m. PT

E-commerce Growth

(source: Giga Information Group)



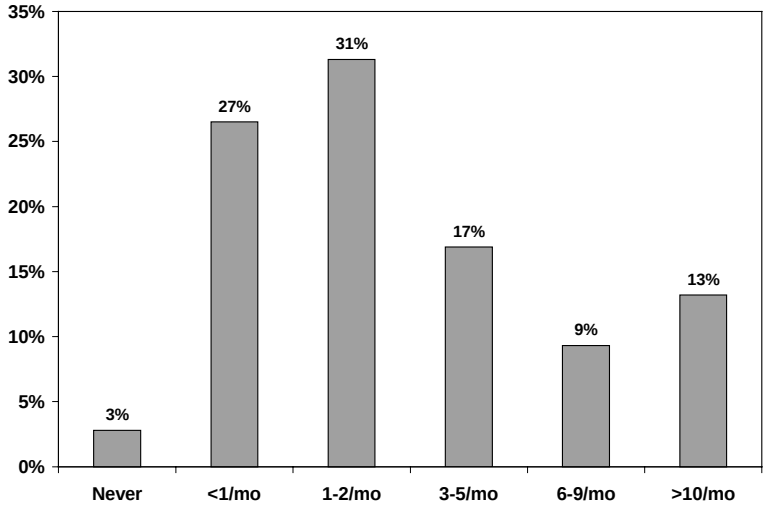
Business in the Internet Age

(Business Week, June 22, 1998)

Type of Business	1997	2001 (forecast)
Business to Business	8.000	183.000
Travel	0.654	7.400
Financial Services	1.200	5.000
PC Hardware & Software	0.863	3.800
Entertainment	0.298	2.700
Ticket Event Sales	0.079	2.000
Books & Music	0.156	1.100
Apparel & Footware	0.092	0.514
Total	**11.342**	**205.514**

How often do you use the Web for shopping for personal reasons?

GVU's 10 th WWW User Survey (October 1998)



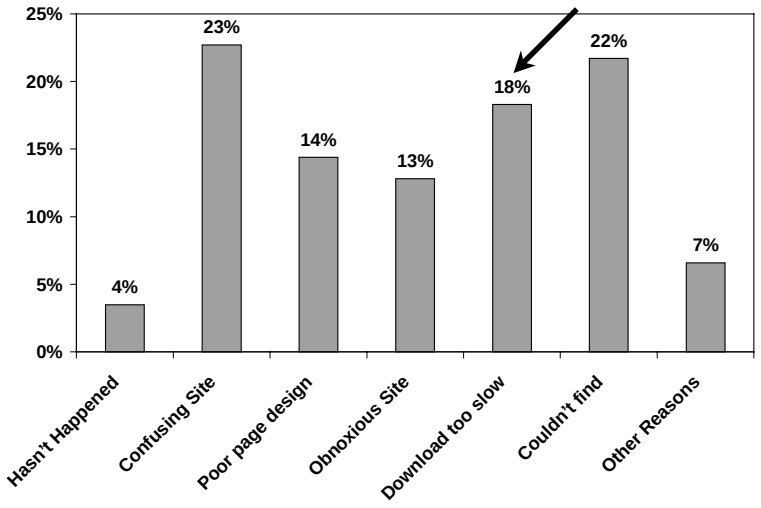
8

Caution Signs Along the Road

(Gross & Sager, Business Week, June 22, 1998, p. 166.)

There will be jolts and delays along the way for electronic commerce: congestion is the most obvious challenge.

Dissatisfying Experiences in E-commerce
GVU's 10th WWW User Survey (October 1998)



10

What are people saying about Web performance…

- **"Tripod's Web site is our business. If it's not fast and reliable, there goes our business.",**

Don Zereski, Tripod's vice-president of Technology (Internet World)

- **"Computer shuts down Amazon.com book sales. The site went down at about 10 a.m. and stayed out of service until 10 p.m."**

The Seattle Times, 01/08/98

What are people saying about Web performance…

- **"Sites have been concentrating on the right content. Now, more of them -- specially e-commerce sites -- realize that performance is crucial in attracting and retaining online customers."**

Gene Shklar, Keynote, The New York Times, 8/8/98

What are people saying about Web performance…

- **"Capacity is King."**

  Mike Krupit, Vice President of Technology, CDnow, 06/01/98

- **"Being able to manage hit storms on commerce sites requires more than just buying more plumbing."**

  Harry Fenik, vice president of technology, Zona Research, LANTimes, 6/22/98

What are people saying about Internet performance…

- **"The capacity crunch is real and will continue for quite some time."**

  Mike O'Dell, Chief Scientist, UUNET

Impacts of Bad Performance

- Bad performance: response time above 8 seconds (eight-second rule).
- $43.5 billion lost each year in e-commerce due to bad performance (Zona Research, April 1999).
- Holiday Season of 1998: over 1/3 of customers gave up due to slowness, 44% turned to conventional stores, 14% moved to another site.

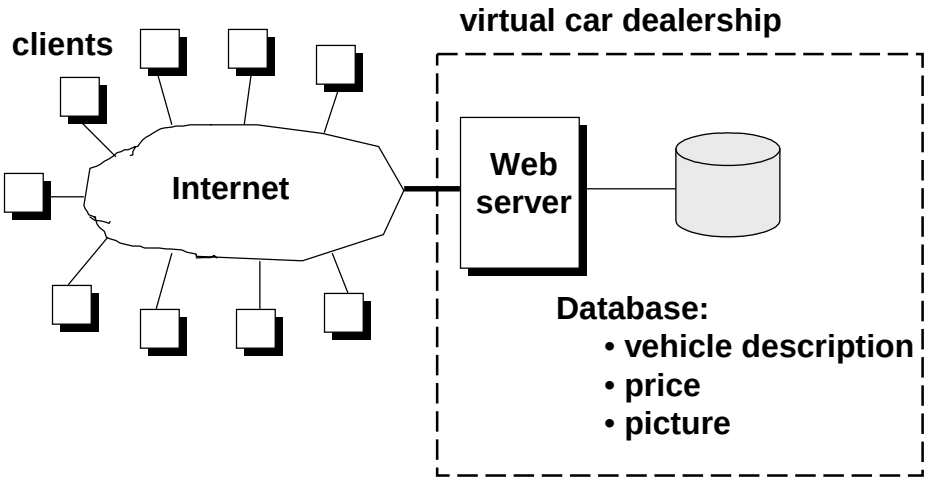
Performance Problems in E-commerce tend to get worse!

- Proliferation of mobile devices
- Easier to use interfaces (VUI, novel browsing paradigms)
- Increasing load placed by software agents
- Impacts of authentication and payment protocols (e.g., SSL, TLS, and SET) on e-commerce site performance.

Important Questions

- What is performance? What are important performance metrics?
- What is capacity?
- What is capacity planning?
- How important are capacity planning and performance evaluation?

Virtual Car Dealer Example



Virtual Car Dealer Example

- Web server request types:
  - retrieve document and images
  - search the DB according to make, model, price range, and distance of dealer from buyer.
  - purchase request.
- Critical request: search
- 5% of searches generate a car sale
- Average sale generates $18,000 in revenues.

Virtual Car Dealer Example

- Service Level Considerations:

Response time (in sec)	Outcome
$4 < r \leq 6$	60% lost
$r > 6$	95% lost

Virtual Car Dealer Example

- Management questions:
  - will the Web server support the load increase while preserving the response time below 4 sec?
  - if not, at which point will its capacity be saturated?
  - how much money could be lost daily if the Web server saturates when the load increases?

Virtual Car Dealer Results

	Current	Current + 10%	Current+20%	Current+30%
Searches per day	92,448	101,693	110,938	120,182
Response Time	2.86	3.80	5.67	11.28
Percent Sales Lost	0	0	60%	95%
Sales per day	4622	5085	2219	300
Daily revenue ($1K)	$ 83,203	$ 91,524	$ 39,938	$ 5,408
Potential daily revenue ($1K)	$ 83,203	$ 91,524	$ 99,844	$ 108,164
Lost daily revenue ($1K)	$ -	$ 0	**$ 59,906**	**$ 102,756**

Important Concepts in Example

- Workload characterization?
- Workload Growth Forecast?
- Performance metrics?
- Desired Service Levels?

Important Concepts in Example

- Workload characterization?
  - retrieve, search, and buy transactions
  - critical transaction: search
  - arrival rate: 92,448 transactions/day

Important Concepts in Example

- Workload growth forecast?
  - 10%, 20%, 30% increase in the workload intensity.
  - no change expected in workload nature.

Important Concepts in Example

- Performance metrics of interest?
  - response time for search transactions.
  - How can be the response time measured?

Important Concepts in Example

- Performance metrics of interest?
  - Response time for search transactions.
  - How can the response time be measured?
    - at the server. What are the components?
    - at the browser. Then, Internet and client LAN delays are part of response time.
    - Measuring end-user response time for e-commerce sites: www.keynote.com

Important Concepts in Example

- Service levels?
  - response time for search transactions < 4 sec.

Intranet Performance Example

- Major airplane manufacturer with 60,000 employees is implementing an intranet to support:
  - corporate training,
  - help desk support,
  - dissemination of internal corporate news, and
  - handling of personnel forms and memos.

Intranet Performance Example

- Help desk application:
  - dedicated Web server with a FAQs DB about common hardware/software problems and solutions,
  - submission of problem descriptions via forms,
  - requests for status on previous claims.

Intranet Performance Example

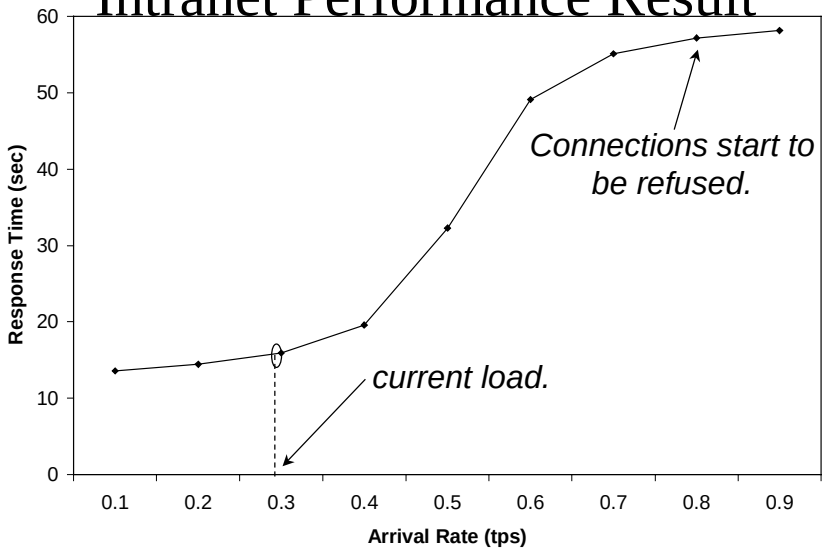
- 10% of employees submit requests to the help desk application every day
- 70% of these requests fall in the 10:00AM 12:00PM period and from 2:00PM to 4:00PM.
- arrival rate =

  (60,000 * 0.1 * 0.7) / (4 * 3,600) = 0.29 request/sec.

Intranet Performance Example

- The OS on all clients will be changed:
  - likely to create big surge in number of requests to the help desk application.
- Management question:
  - how will the help desk application response time vary with the arrival rate?

Intranet Performance Result



Important Concepts in Example

- Workload characterization?
- Workload Growth Forecast?
- Performance metrics?
- Desired Service Levels?

Important Concepts in Example

- Workload characterization?
  - help desk application: searches to a FAQs DB, submission of problems via forms, requests for status on previous claims
  - critical transaction: help desk support requests
  - arrival rate: 0.29 tps during peak period.

Important Concepts in Example

- Workload growth forecast?
  - surge in requests to help desk due to new OS installed on client machines.

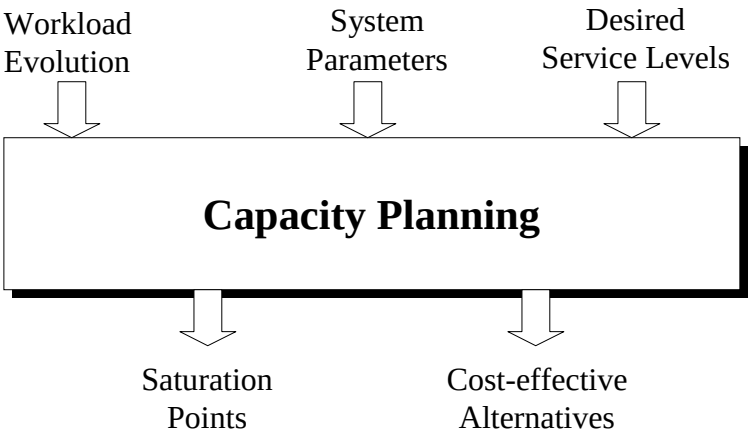
Important Concepts in Example

- Performance metrics of interest?
  - response time for help desk requests.
  - connection rejection probability

Important Concepts in Example

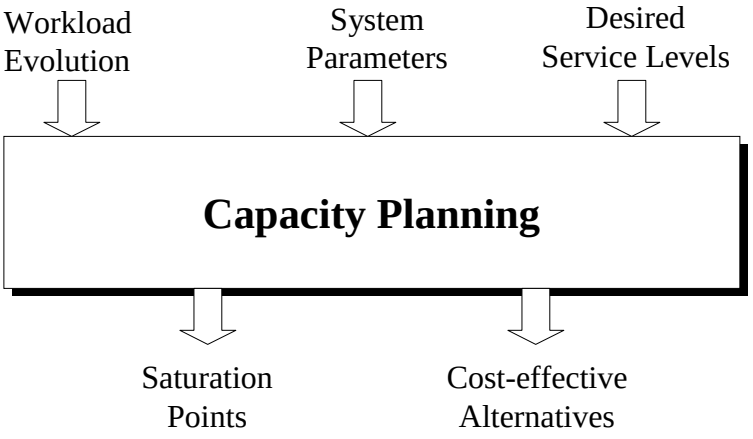
- Service levels?
  - response time for search transactions < 12 sec.

Capacity Planning Input and Output Variables



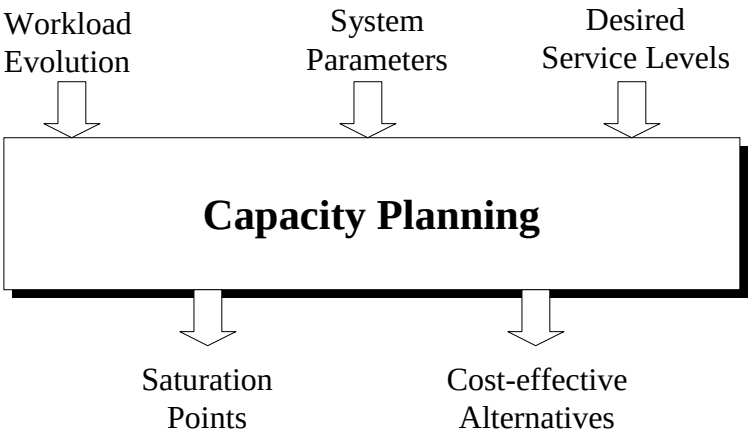
Capacity Planning Input and Output Variables

- *intensity*
- *nature*



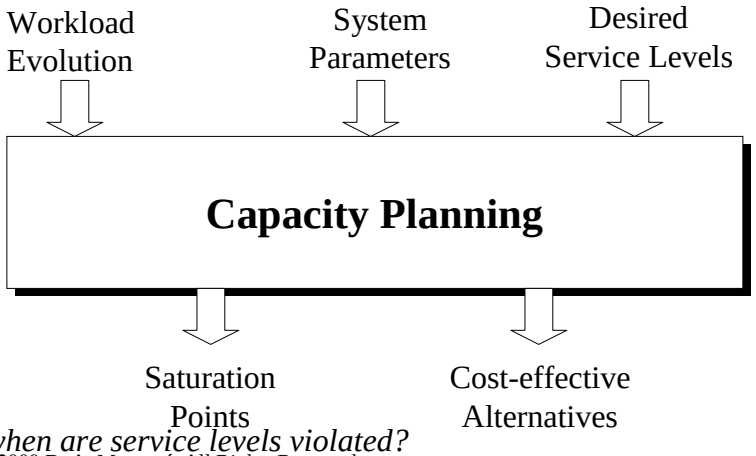
Capacity Planning Input and Output Variables

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



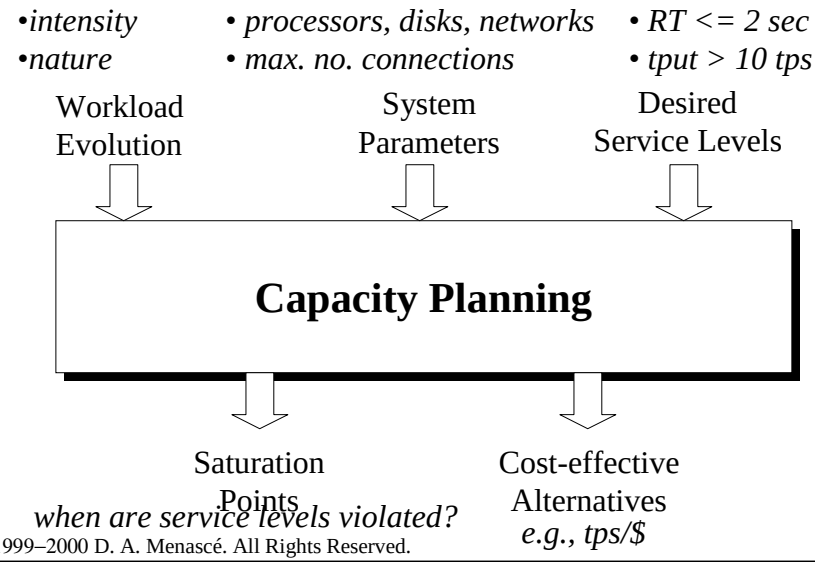
Capacity Planning Input and Output Variables

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



when are service levels violated?

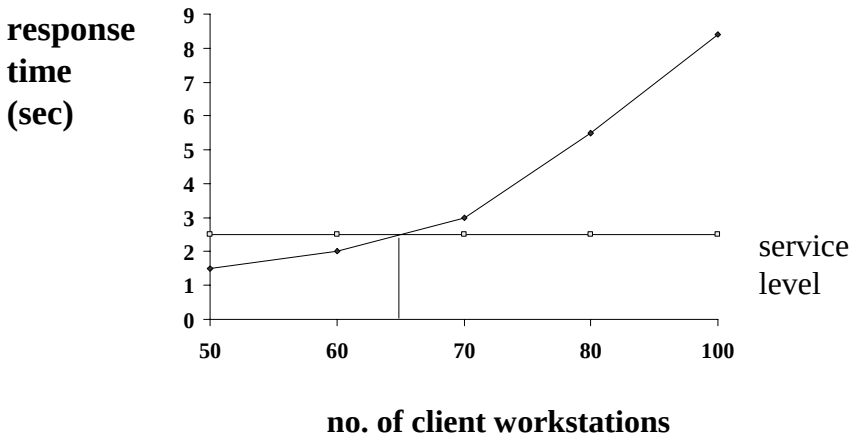
Capacity Planning Input and Output Variables



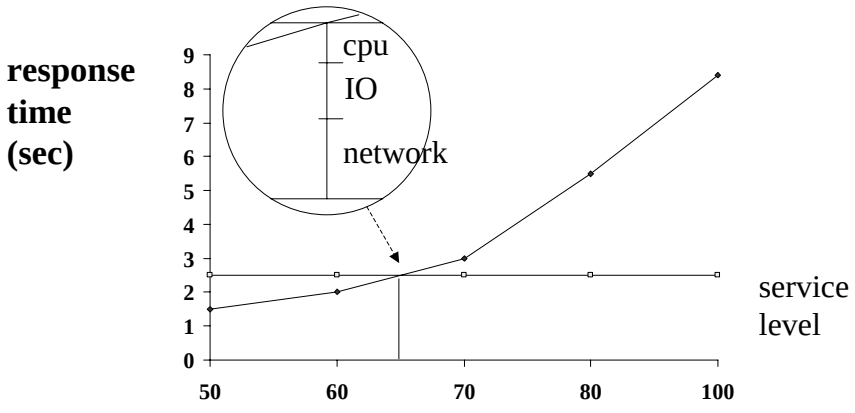
Capacity Planning Definition

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



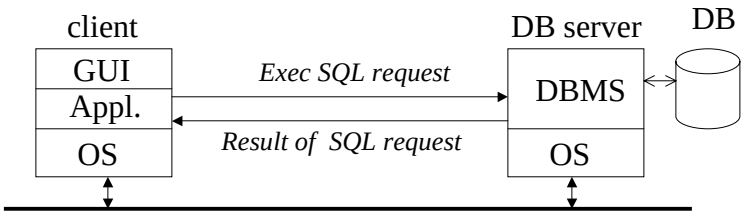
Capacity Planning Concept



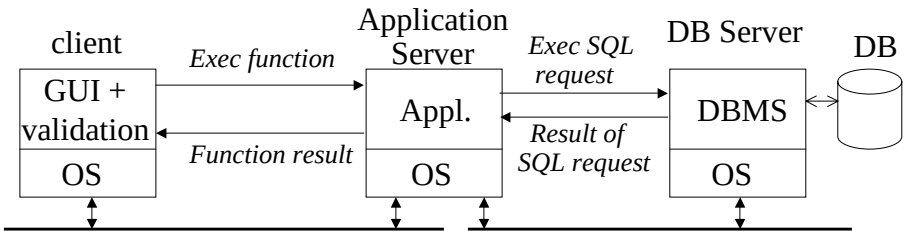
Typical Capacity Planning Questions

- 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?

Two-tier C/S architecture



Three-tier C/S architecture



Typical Capacity Planning Questions

- Situation: migrating from a mainframe based to a C/S system.
- Questions:
  - What should be the configuration of the application servers?
  - Should the DB be replicated and how?
  - Which DB replication strategy should be used?

Typical Capacity Planning Questions

- Situation: redesigning an e-commerce Web site and adding a lot of multimedia content.
- Questions:
  - Will the bandwidth of the link to the ISP support more multimedia content?
  - Will the IO bandwidth be enough to provide adequate response time to multimedia requests?
  - What is the most appropriate configuration for the Web site?

Typical Capacity Planning Questions

- Situation: redesigning an e-commerce Web site and adding a lot of multimedia content.
- Questions:
  - Should I use many small boxes or a few large boxes to support the load?
  - What type of load balancing scheme should be used?

Importance of Performance Evaluation and Capacity Planning

- Risk of financial losses
- External image of the company
- User dissatisfaction
- Productivity decrease
- Procurement cycle and budgetary constraints.

Common Mistakes in Capacity Planning

- Performance varies linearly with the workload!
- Just throw more iron and the problem is solved.
- Incorrect data gathering procedures: garbage in ⇨ garbage out.