CS 672
Capacity Planning Methodology

Dr. Daniel A. Menascé

http://www.cs.gmu.edu/faculty/menasce.html

Department of Computer Science

George Mason University

What is Adequate Capacity?

We say that a Web service has adequate capacity if the service-level agreements are continuously met for a specified technology and standards, and if the services are provided within cost constraints.

Service-Level Agreements (SLA)

- SLAs outline what a user of an application can expect in terms of response time, throughput, system availability, and reliability

  - focus on metrics that users can understand
  - set easy-to-measure goals
  - tie IT costs to your SLAs

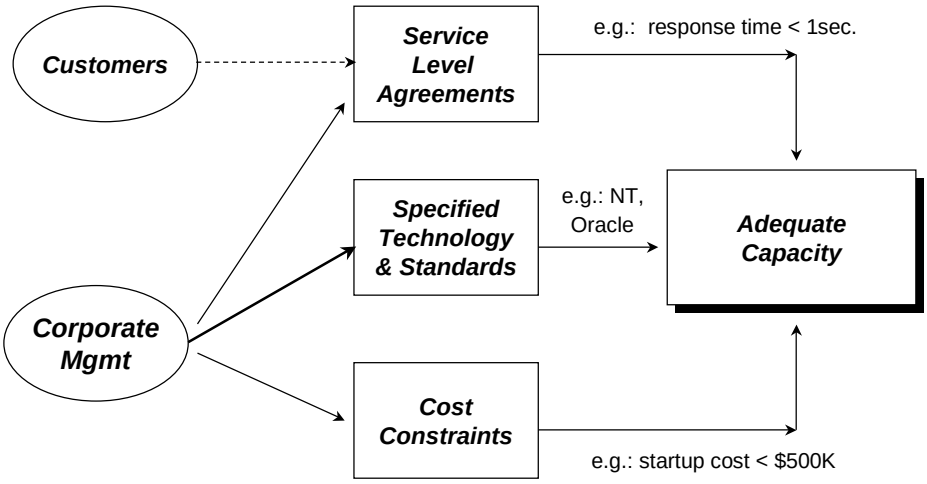
3

Service Level Agreements: examples

- Response time for trivial database queries should not exceed 2 sec.
- We want the same level of availability and response time that we had in the mainframe environment.
- The goal for Web services is 99.99% availability and less than 1-sec response time for 90% of the HTTP requests for small documents.

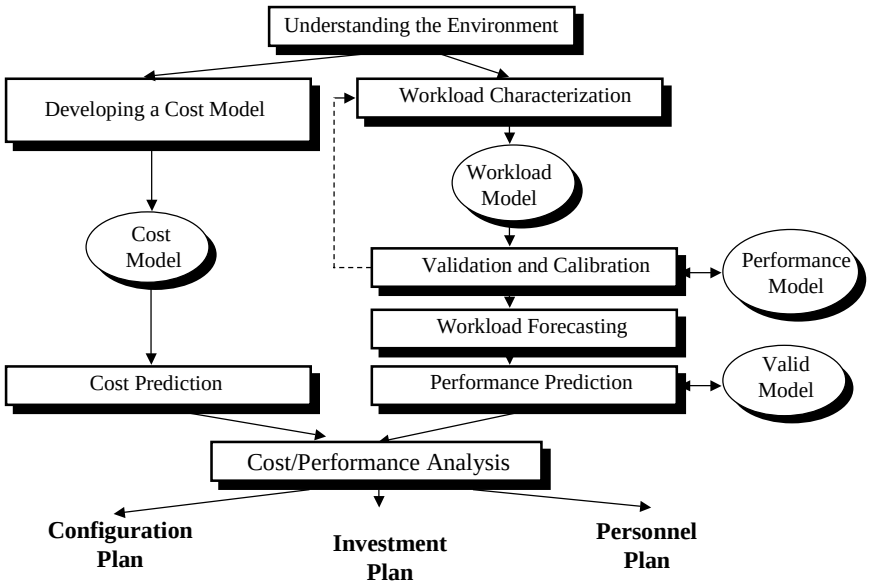
4

Adequate Capacity



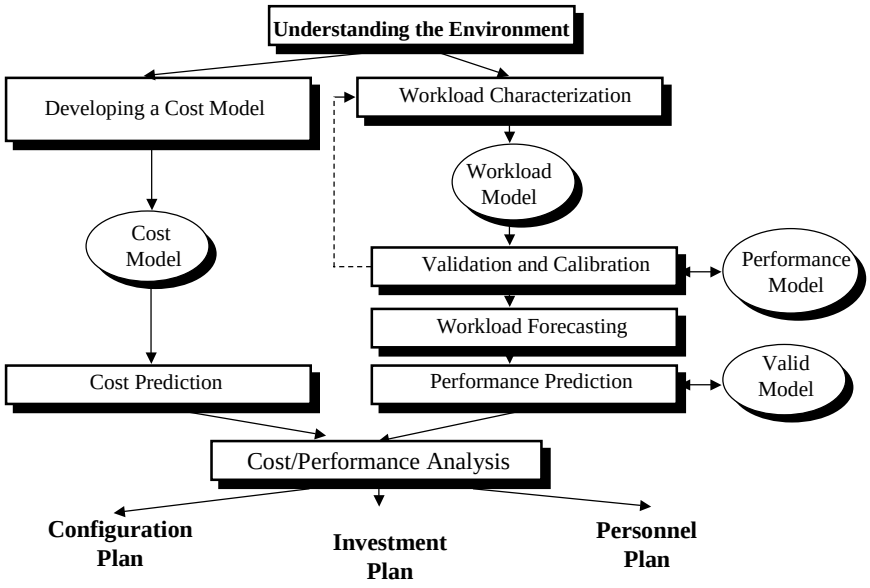
 5

Methodology



 6

Methodology



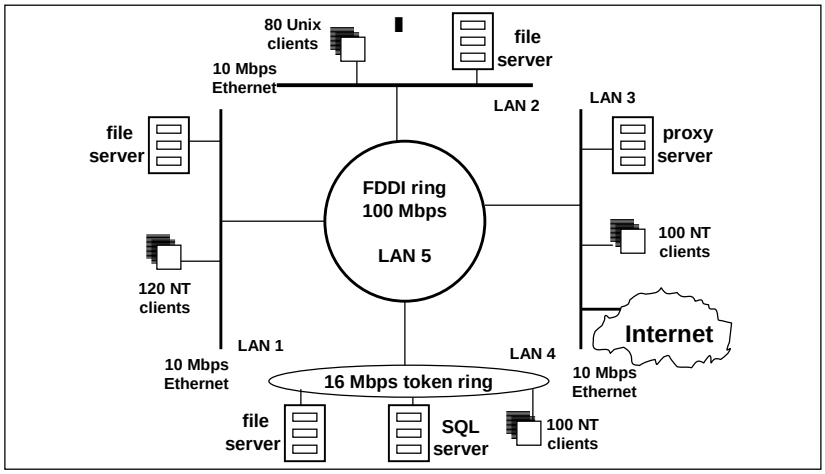
Understanding the Environment

The goal is to learn what kind of

- hardware (clients and servers)
- software (OS, middleware, and applications)
- network connectivity and protocols

are present in the environment.

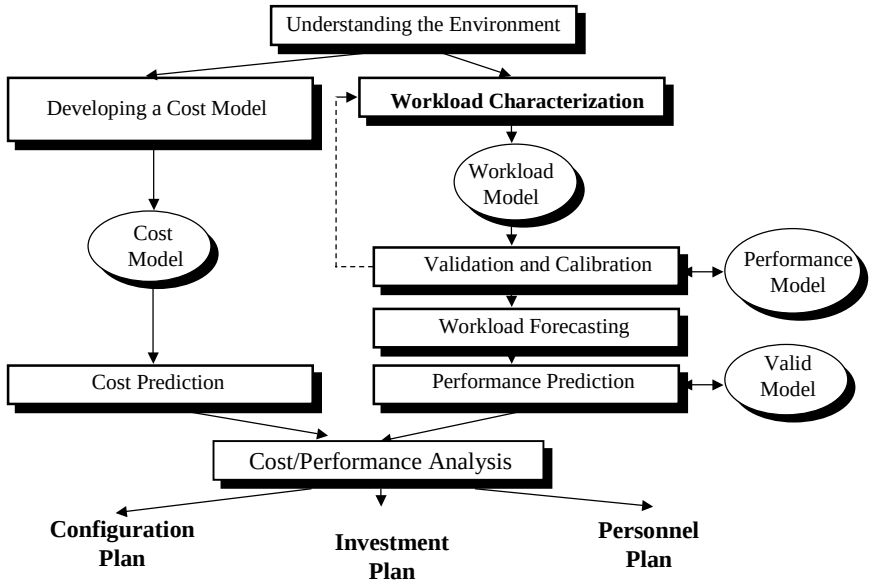
Understanding the Environment: example



Elements in Understanding the Environment

Client platform	**Quantity and type**
Server platform	**Quantity, type, configuration and functions**
Middleware	**Type (e.g. TP monitors)**
DBMS	**Type**
Application	**Main types of applications, criticality, etc.**
Network connectivity	**Network diagrams with LANs, WANs, routers, servers, etc.**
SLAs	**Existing SLAs per application**
Procurement procedures	**Elements of the procurement process, expenditure limits, justification procedures for acquisitions.**

Methodology

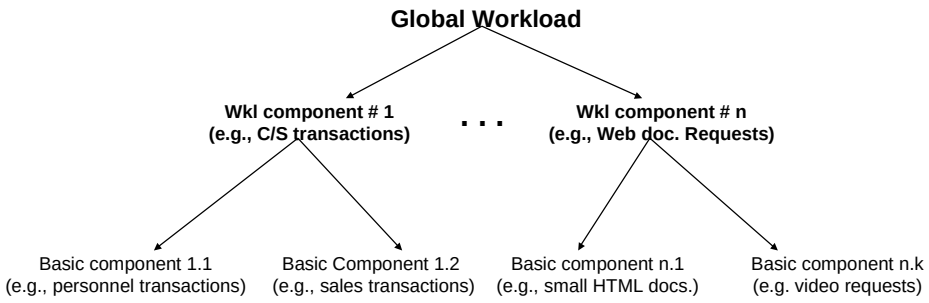


Workload Characterization

Workload characterization is the process of precisely describing the system's global workload in terms of its main components.

The basic components are then characterized by intensity and service demand parameters at each resource of the system.

Workload Characterization Process



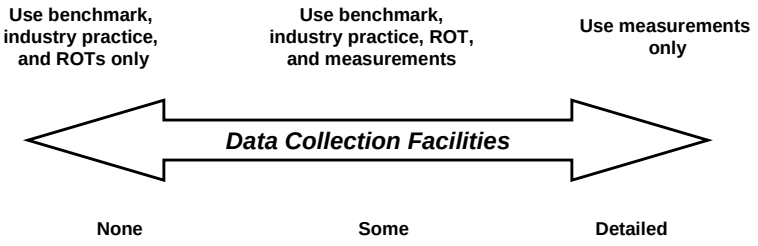
Workload Description: example

Basic Components and Parameters	Type
Sales transaction	--
. Number of transactions submitted per client	WI
. Number of clients	WI
. Total number of I/Os to the Sales DB	SD
. CPU utilization at the DB server	SD
. Avg. messages sent/received by the DB server	SD
Web-based training	--
. Avg. number of training sessions per day	WI
. Avg size of image files retrieved	SD
. Avg. size of http documents retrieved	SD
. Avg number of image files retrieved/session	SD
. Avg. number of documents retrieved/session	SD
. Avg. CPU utilization of the httpd server	SD

SD = service demand
WI = workload intensity

Data Collection Issues

- How to determine the parameter values for each basic component?



Data Collection Issues: example

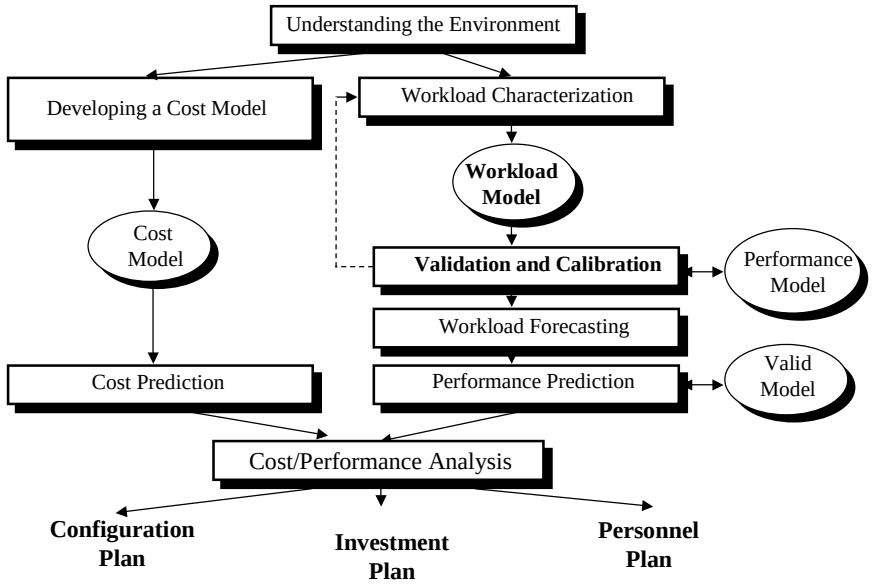
- The server demand at the server for a given application was 10 msec obtained in a controlled environment with a server with a SPECint rating of 3.11.
- What would be the service demand if the server used in the actual system were faster and had a SPECint rating of 10.4?

ActualServiceDemand = MeasuredServiceDemand x ScalingFactor

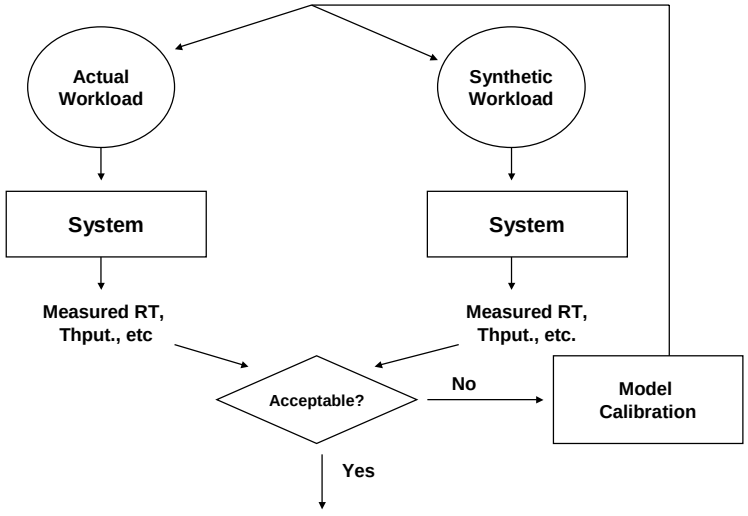
ScalingFactor = ControlledResourceThroughput / ActualResourceThroughput

ActualServiceDemand = 10 * (3.11/10.4) = 3.0 msec.

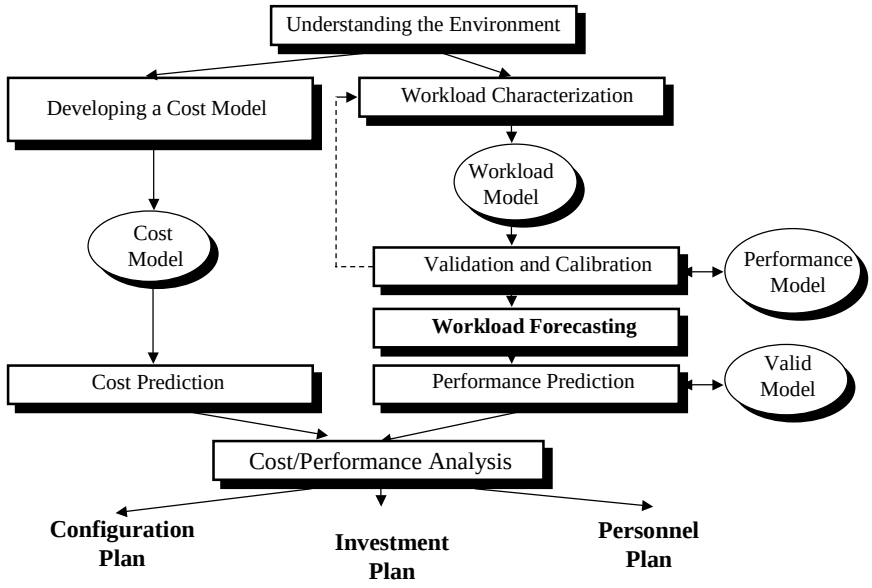
Methodology



Validating Workload Models



Methodology



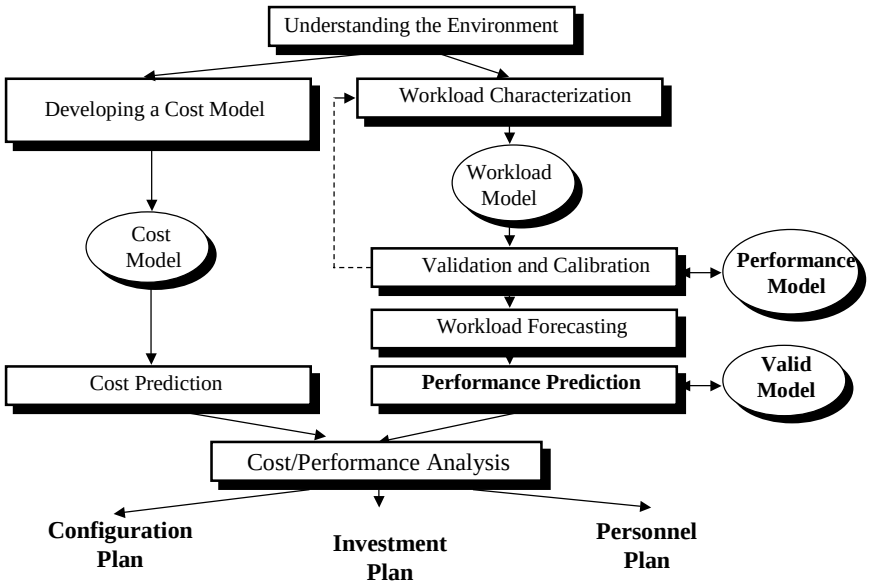
Workload Forecasting

- How will the number of search requests to the company's online catalog vary over the next 6 months?
- How will the number of hits to the corporate intranet's Web server vary over time?

Workload Forecasting (cont'd)

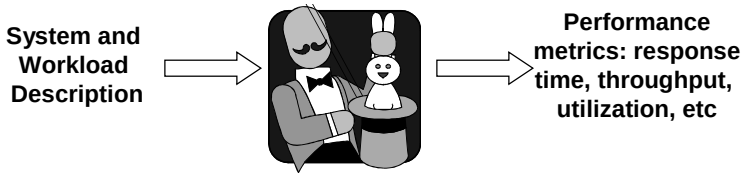
- Answering these questions involves:
  - evaluating the organization's workload trends;
  - analyzing historical usage data;
  - analyzing business or strategic plans;
  - mapping plans into business processes (e.g., paperwork reduction will add 50% more e-mail).
- Workload forecasting techniques: moving averages, exponential smoothing, etc.

Methodology

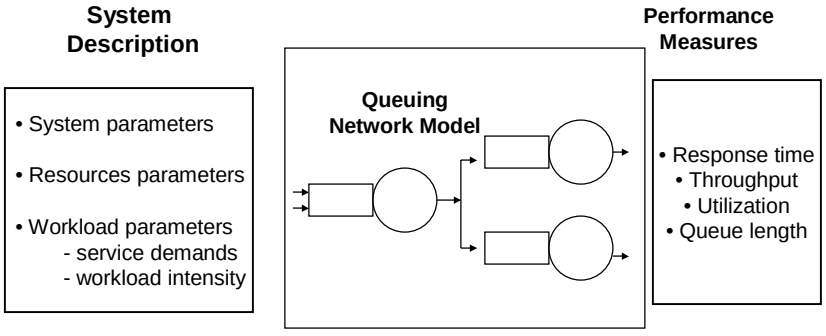


Performance Modeling and Prediction

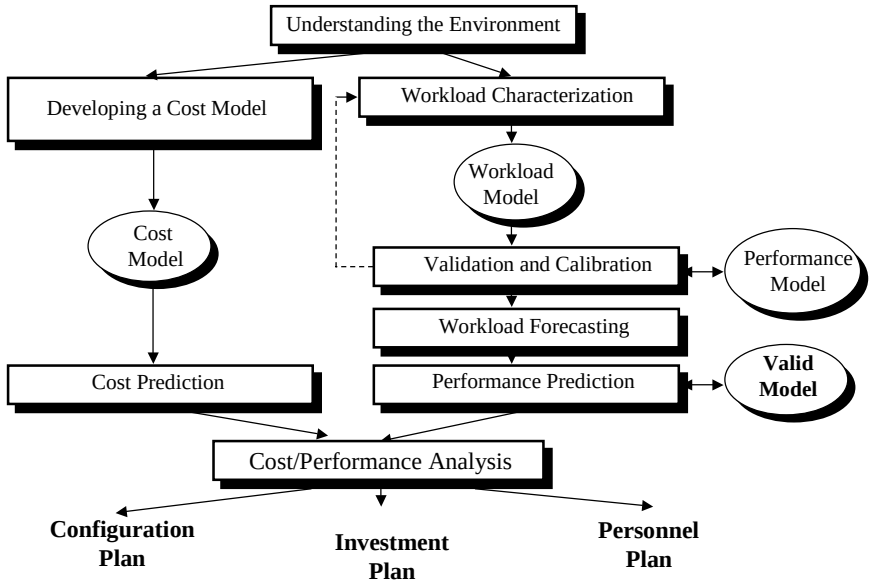
- How are performance measures estimated?



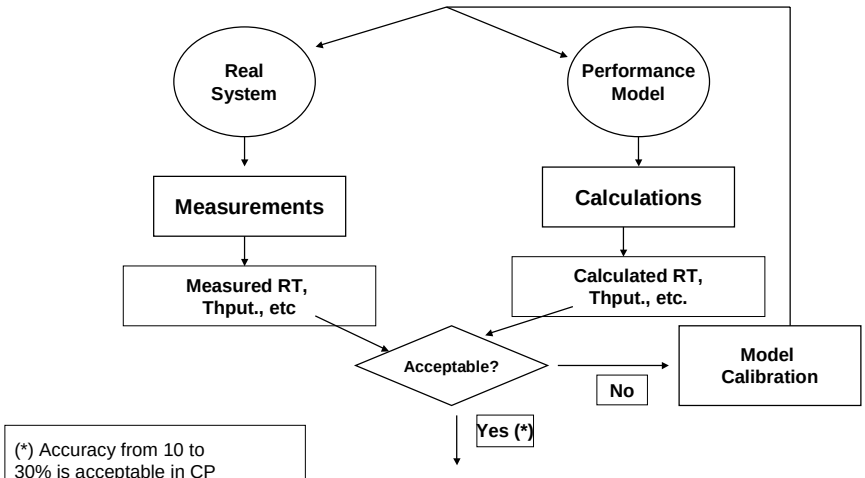
Estimating performance measures



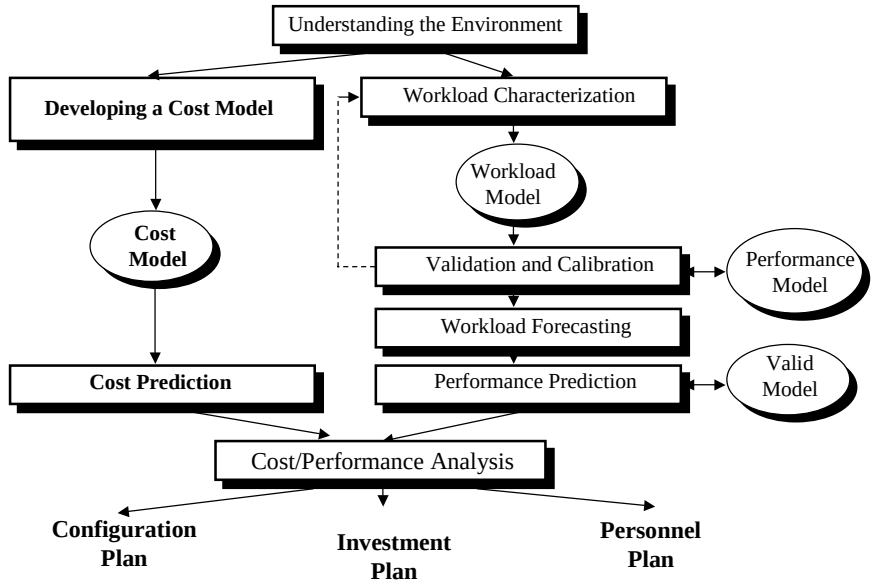
Methodology



Validating Performance Models



Methodology



Cost Model

- A capacity planning methodology requires the identification of major sources of cost as well as the determination of how cost will vary with system size and architecture.
  - Startup costs
  - Operating costs

Cost Model: categories

- Hardware costs: client and server machines, backend mainframes, disks, routers, bridges, cabling, maintenance, etc.
- Software costs: operating systems, middleware, DBMS, mail processing software, office automation, applications, etc.
- Telecommunication costs: WAN services, ISP, etc.
- Support costs: salaries and benefits of all system administrators, help desk support, network people, web page designers, etc