Capacity Planning Methodology

Learning Objectives

- Discuss the concept of adequate capacity of a system.
- Introduce service level agreements.
- Present a methodology for capacity planning.

Learning Objectives (cont'd)

- Discuss the main steps of the methodology:
  - understanding the environment
  - workload characterization
  - workload forecasting
  - Performance/dependability modeling
  - Performance/dependability prediction
  - cost/performance-dependability analysis.

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.

Technology and Standards

- T&S means, for instance:
  - HW for servers (and for clients)
  - O.S. software
  - LAN, WAN line infrastructure (type, speed)
- Sometimes the choice is based on factors not related to performance:
  - ease of system administration
  - familiarity of personnel with the system
  - number/quality of vendors for HW/SW

Service-Level Agreements (SLA)

- SLAs determine 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

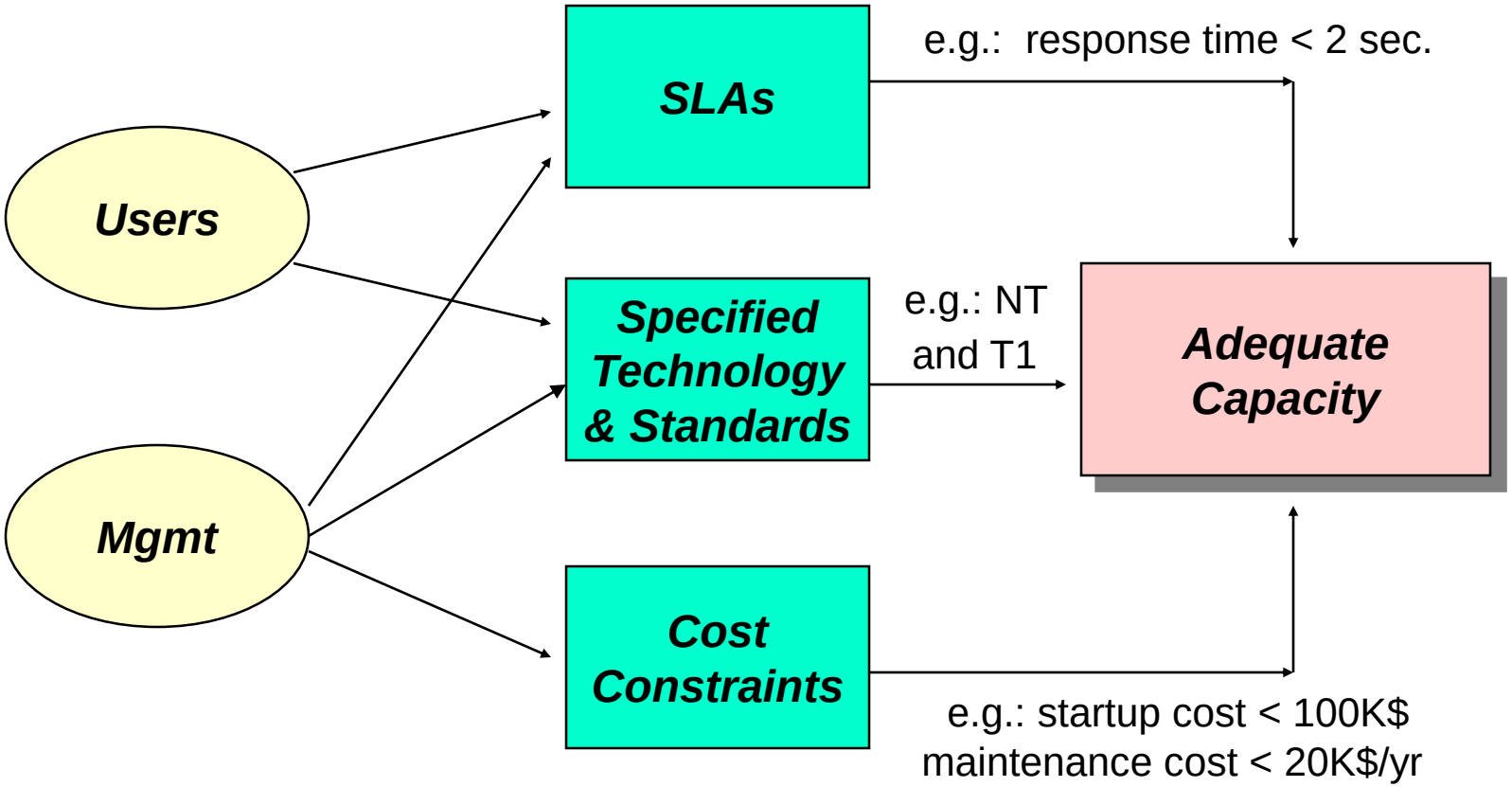
  – set easy-to-measure goals

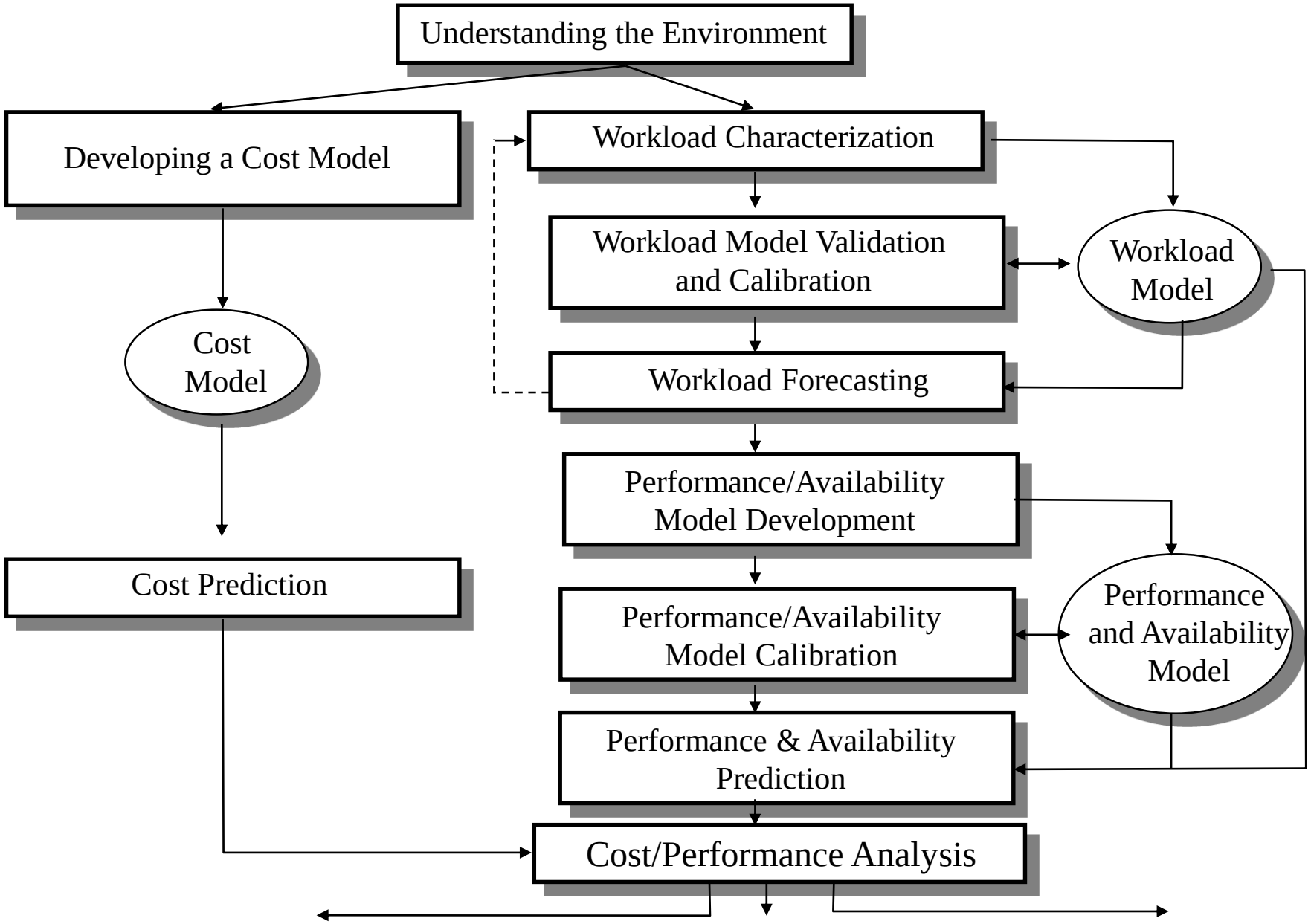
  – tie IT costs to your SLAs

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% of availability and less than 1-sec response time for 90% of the HTTP requests for small documents.

Adequate Capacity



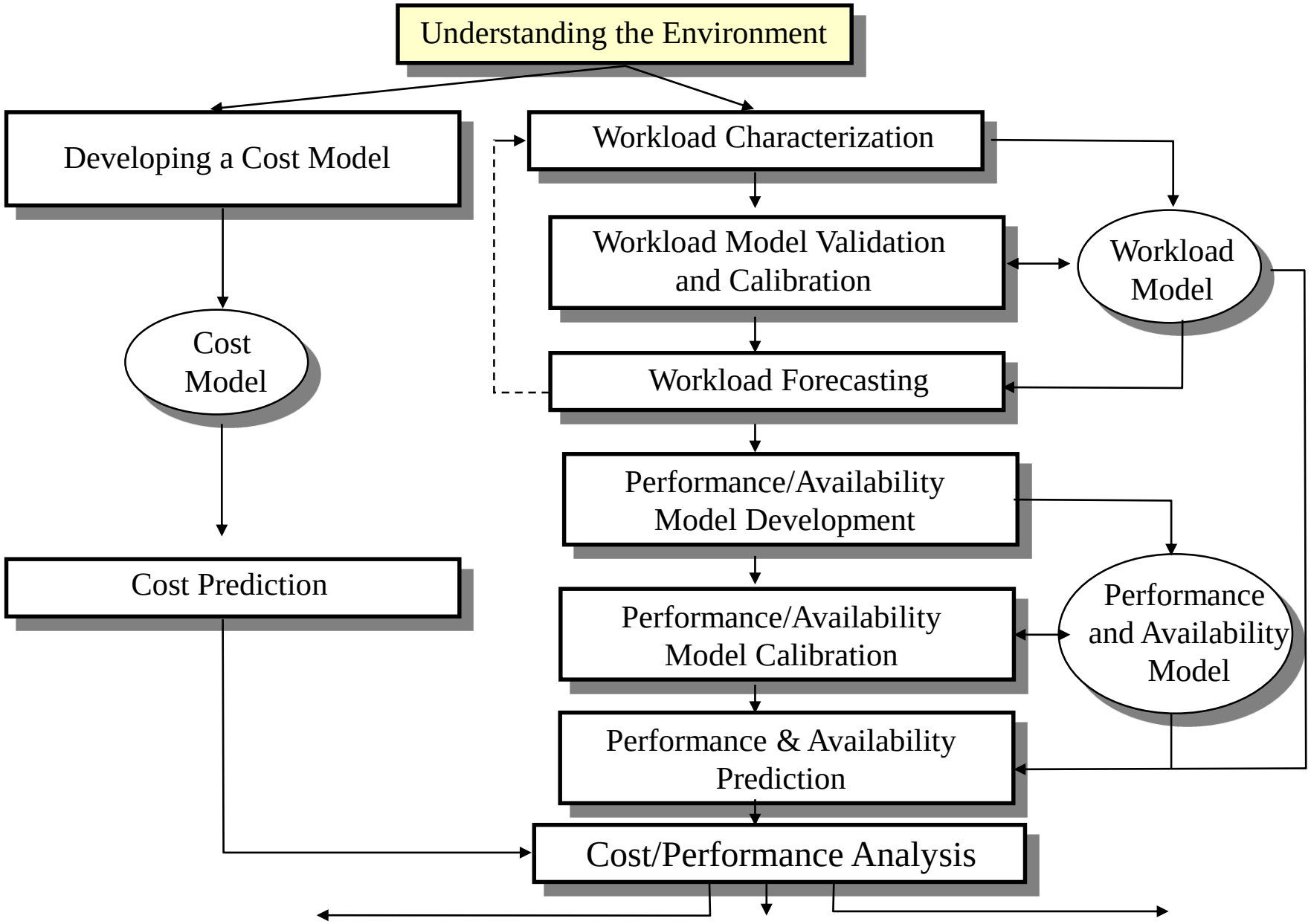

Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

The Three Models

- Workload model:
  - resource demand, load intensity - for each component of a global workload
- Performance/Dependability model:
  - used to predict performance/dependability as function of system description and workload parameters
  - outputs: response times, throughputs, system resources utilizations, queue lengths, availability, reliability, safety, etc.

The Three Models (cont.)

- Performance/Dependability model (cont.):
  - the performance/dependability metrics are matched against SLAs to check if capacity is adequate
- Cost model:
  - accounts for SW, HW, TLC, personnel, training, support expenditures, etc.


Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

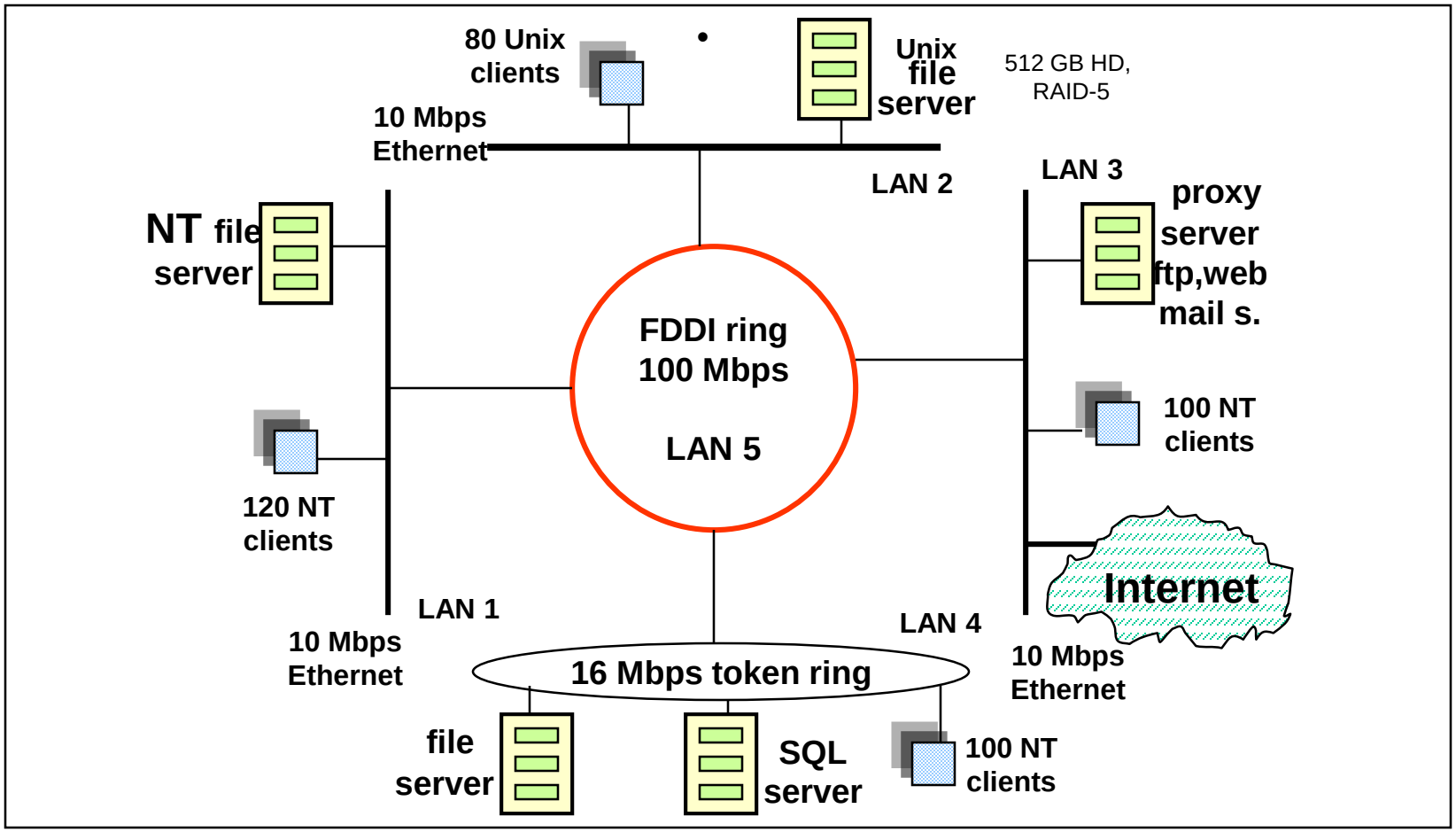
Understanding the Environment

The goal is

- to learn what kind of
  - hardware (clients and servers)
  - software (OS, middleware, applications)
  - network connectivity and protocols

  are present in the environment.
- Identify peak periods, management structures, SLAs

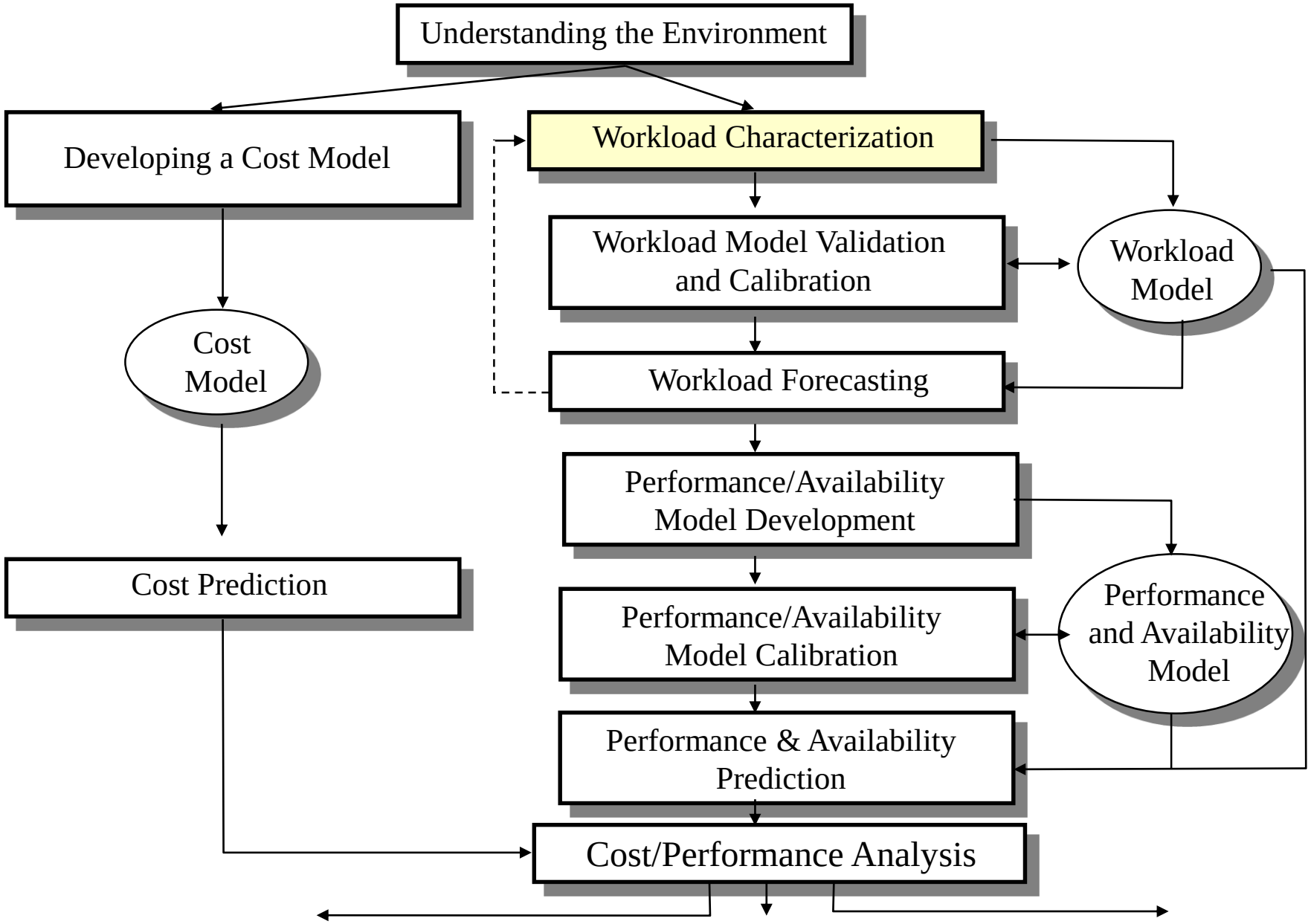
Understanding the Environment: example



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.**

Adapted from Menascé & Almeida.

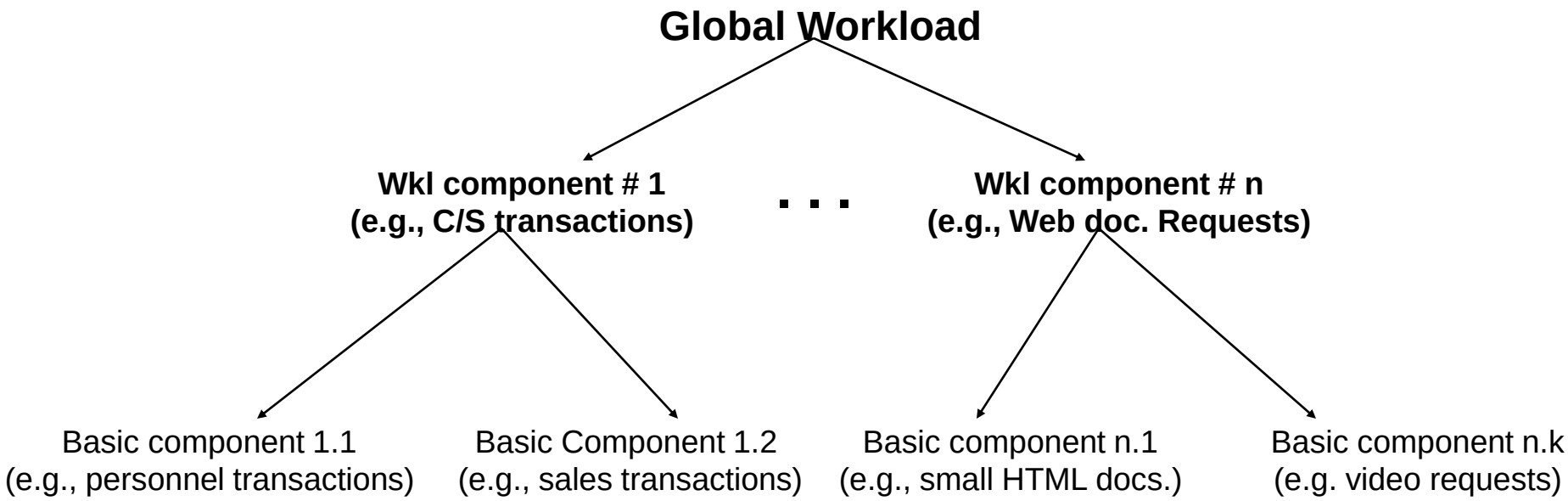

Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

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 (e.g. transaction arrival rate) and service demand parameters at each resource of the system.

Workload Characterization Process



Workload Description

- Parameters for a basic component: must usually be derived indirectly (measurement or estimation of other parameters; usage of performance monitors, accounting systems, system log files, etc.)
- Measurements during normal and peak workload periods
- Use of clustering algorithms

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

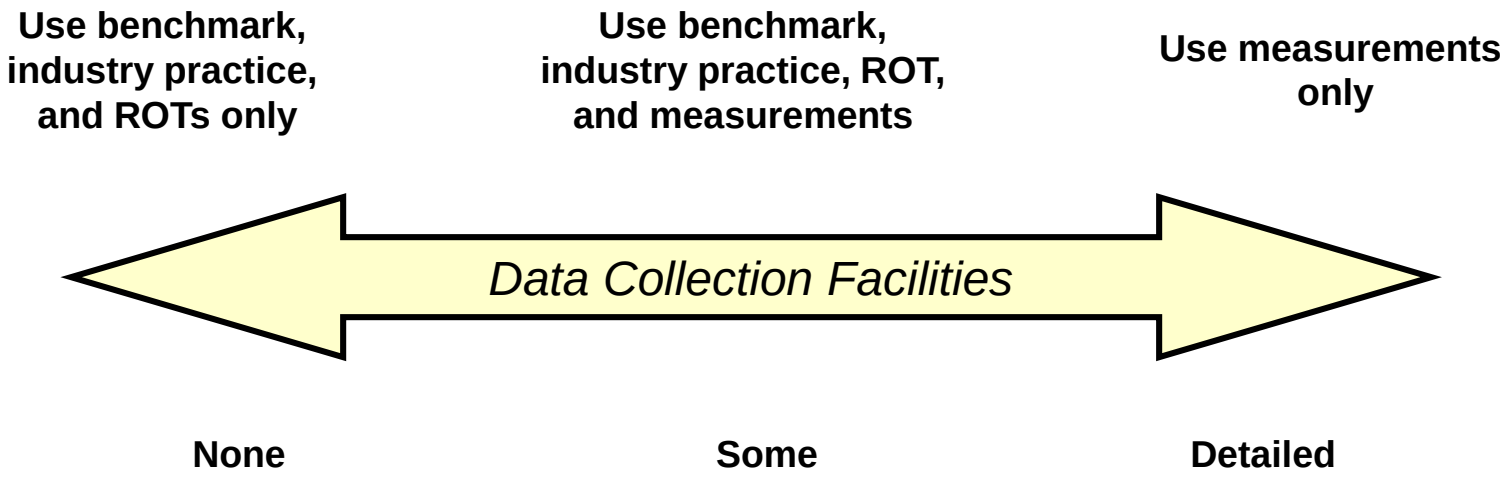
SD = service demand
WI = workload intensity

Workload Parameters

- Workload intensity parameters:
  - *provide a measure of the load placed on system, indicated by number of units of work that contend for system resources*
- Workload service demand parameters:
  - *specify the total amount of service time required by each basic component at each resource*

Data Collection Issues

- How to determine the parameter values for each basic component? -> adequate tools are often unavailable; most tools provide only aggregate data for resource levels. (ROT: Rule of Thumb)



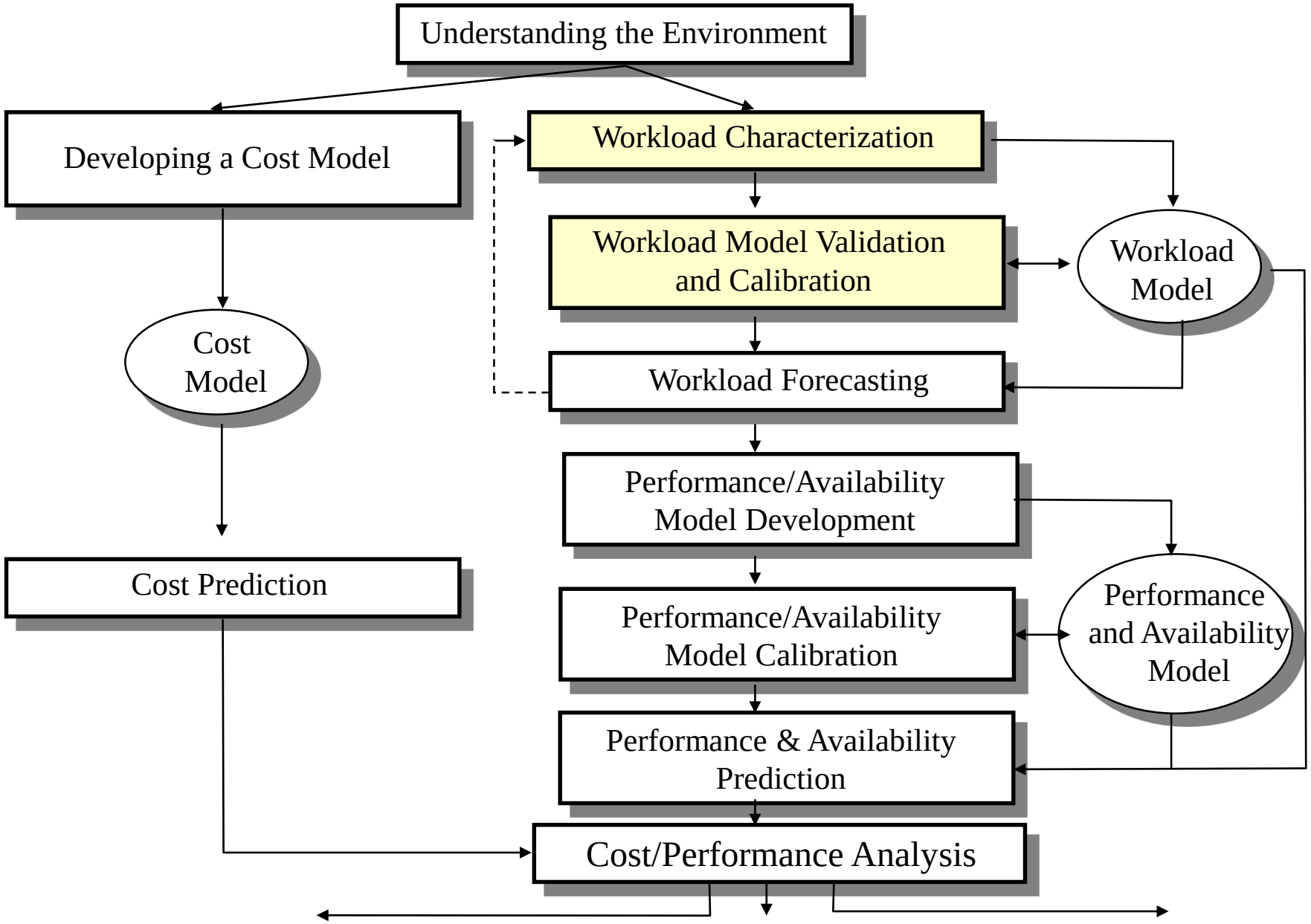
Data Collection Issues: example

- 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?

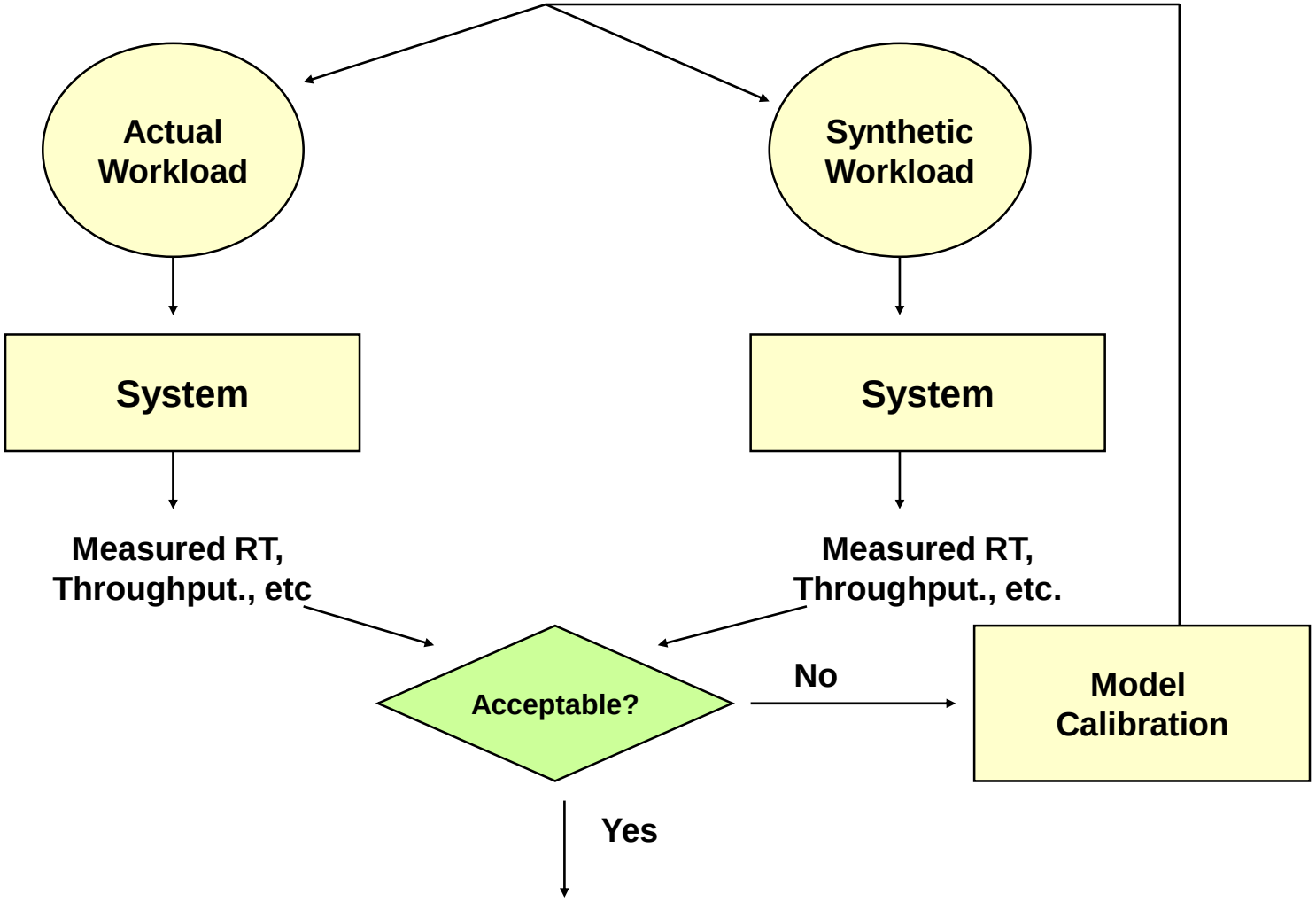
ActualServiceDemand = MeasuredServiceDemand x ScalingFactor

ScalingFactor = ControlledResourceThroughput / ActualResourceThroughput

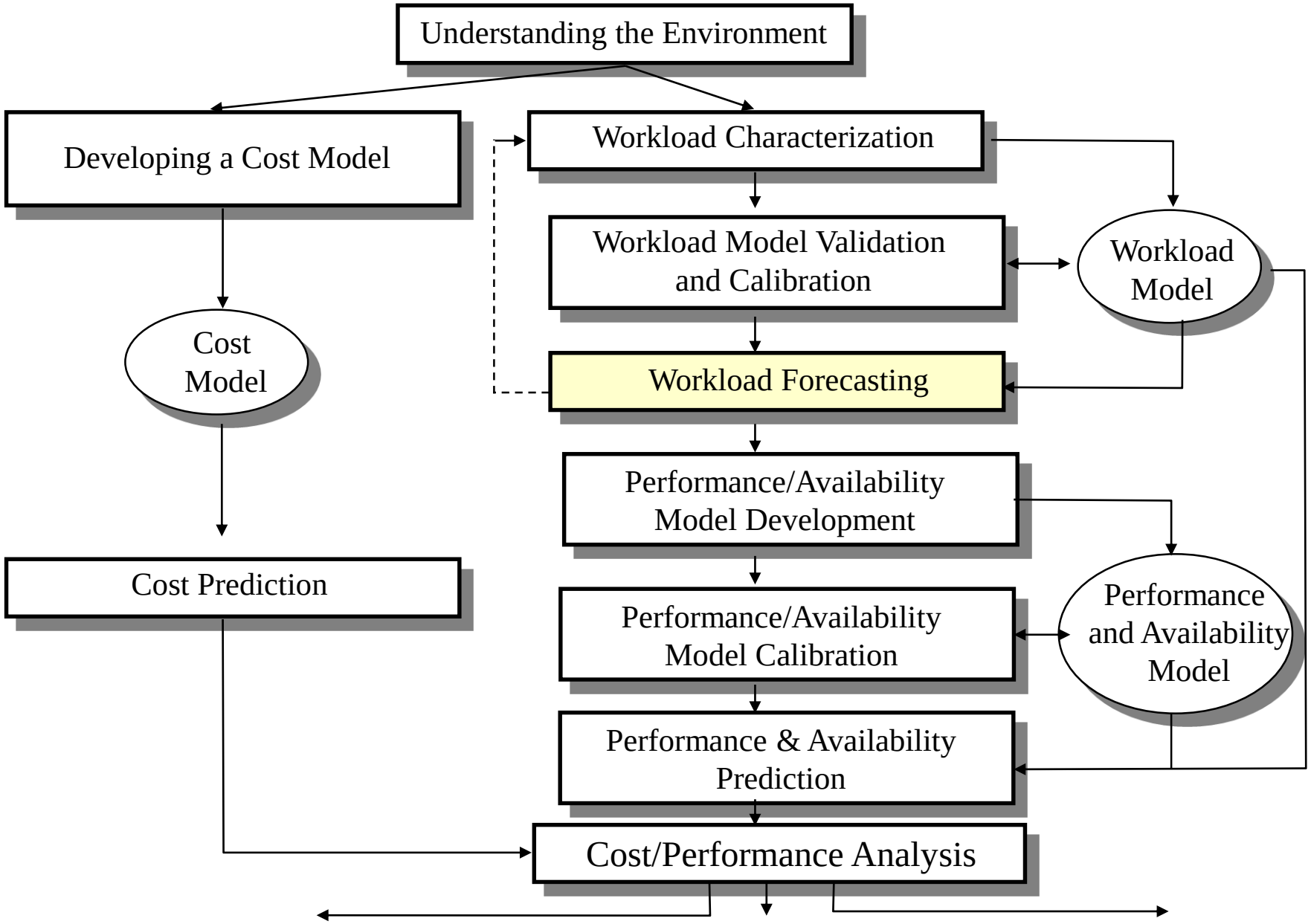
ActualServiceDemand = 10 * (3.11/10.4) = 3.0 msec.


Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

Validating Workload Models



Adapted from Menascé & Almeida.

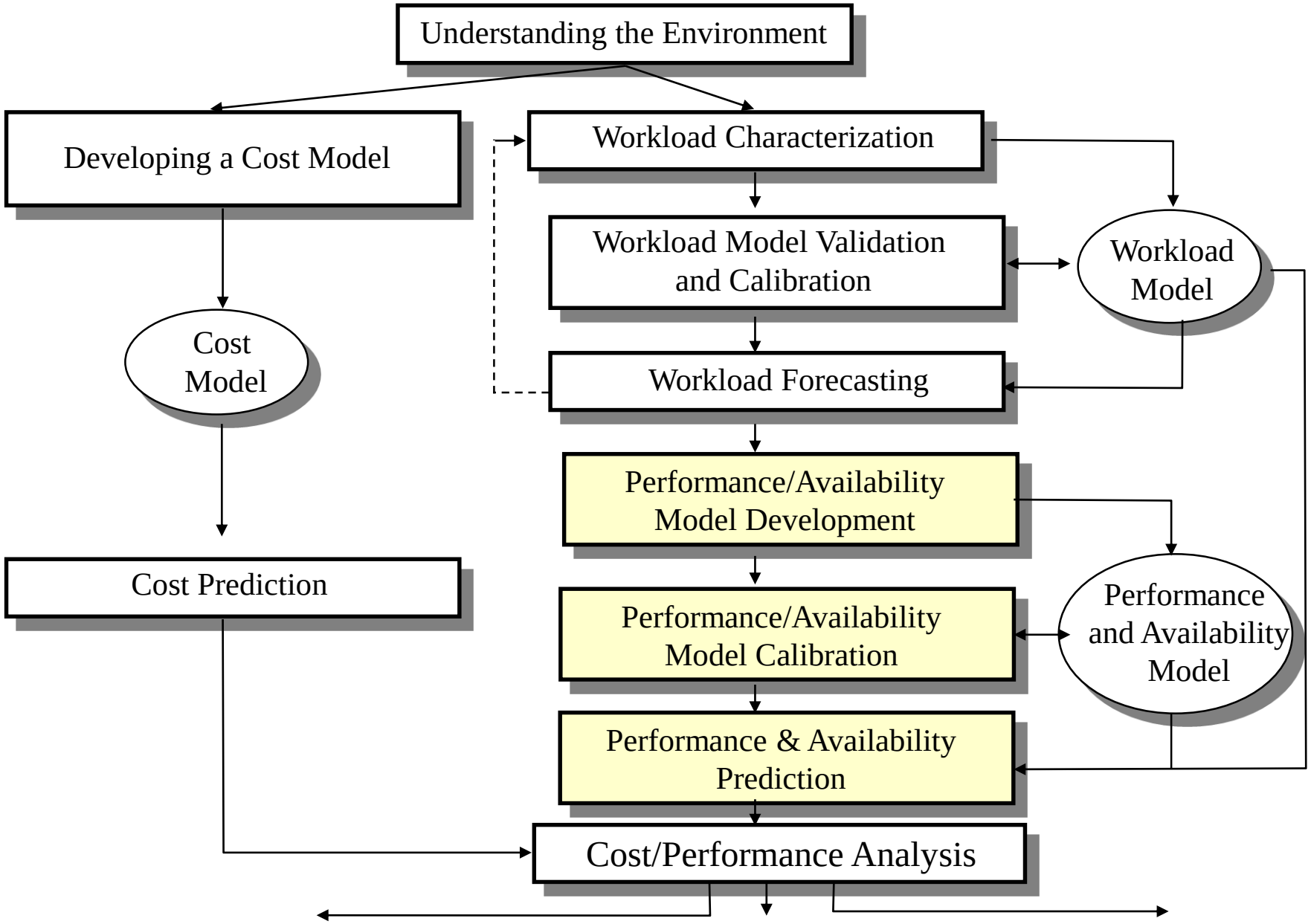

Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

Workload Forecasting

- WL.F. is the process of predicting how system workloads will vary in the future; examples:
  - How will the number of e-mail messages handled per day by the server vary over the next 6 months?
  - How will the number of hits to the corporate intranet's Web server vary over time?

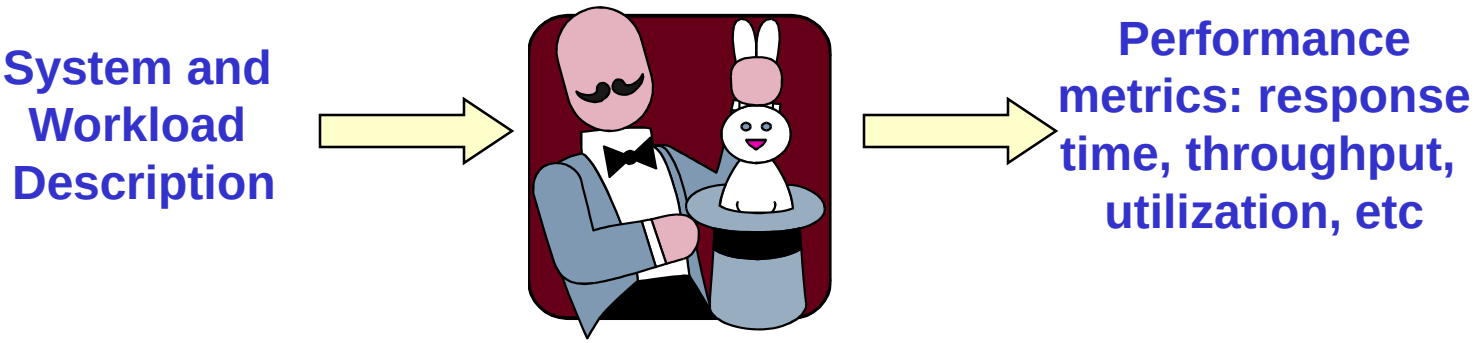
Workload Forecasting (cont'd)

- 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, linear regression.


Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

Performance Modeling and Prediction

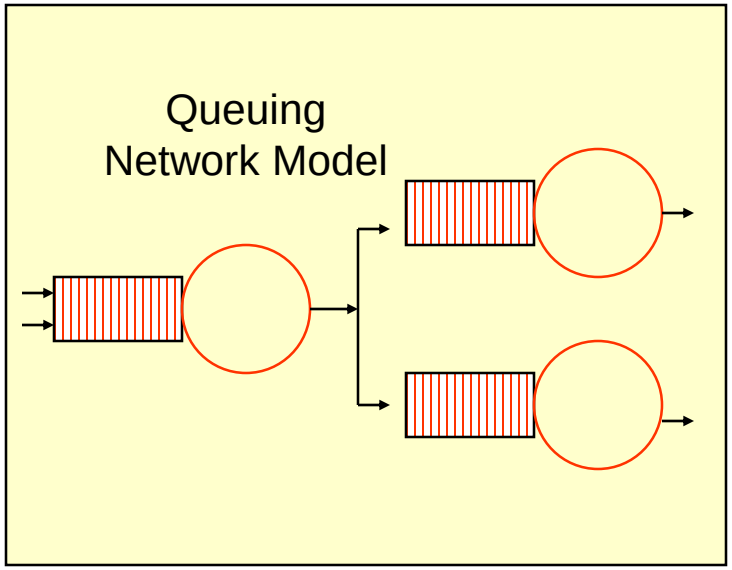
- Is the process of estimating performance measures of a computer system for a given set of parameters.
- How are performance measures estimated?



Estimating performance measures

System Description

- System parameters
- Resources parameters
- Workload parameters
  - service demands
  - workload intensity
  - burstiness



Performance Measures

- Response time
- Throughput
- Utilization
- Queue length

Parameters (Affecting Performance) (I)

- System parameters examples:
  - load-balancing disciplines for WEB serv.s
  - network protocols
  - max. numb. of connections supported
  - max. numb.of threads supported by DBMS
- Resource parameters examples:
  - disk seek times, latency, transfer rate
  - network bandwidth; router latency
  - CPU speed

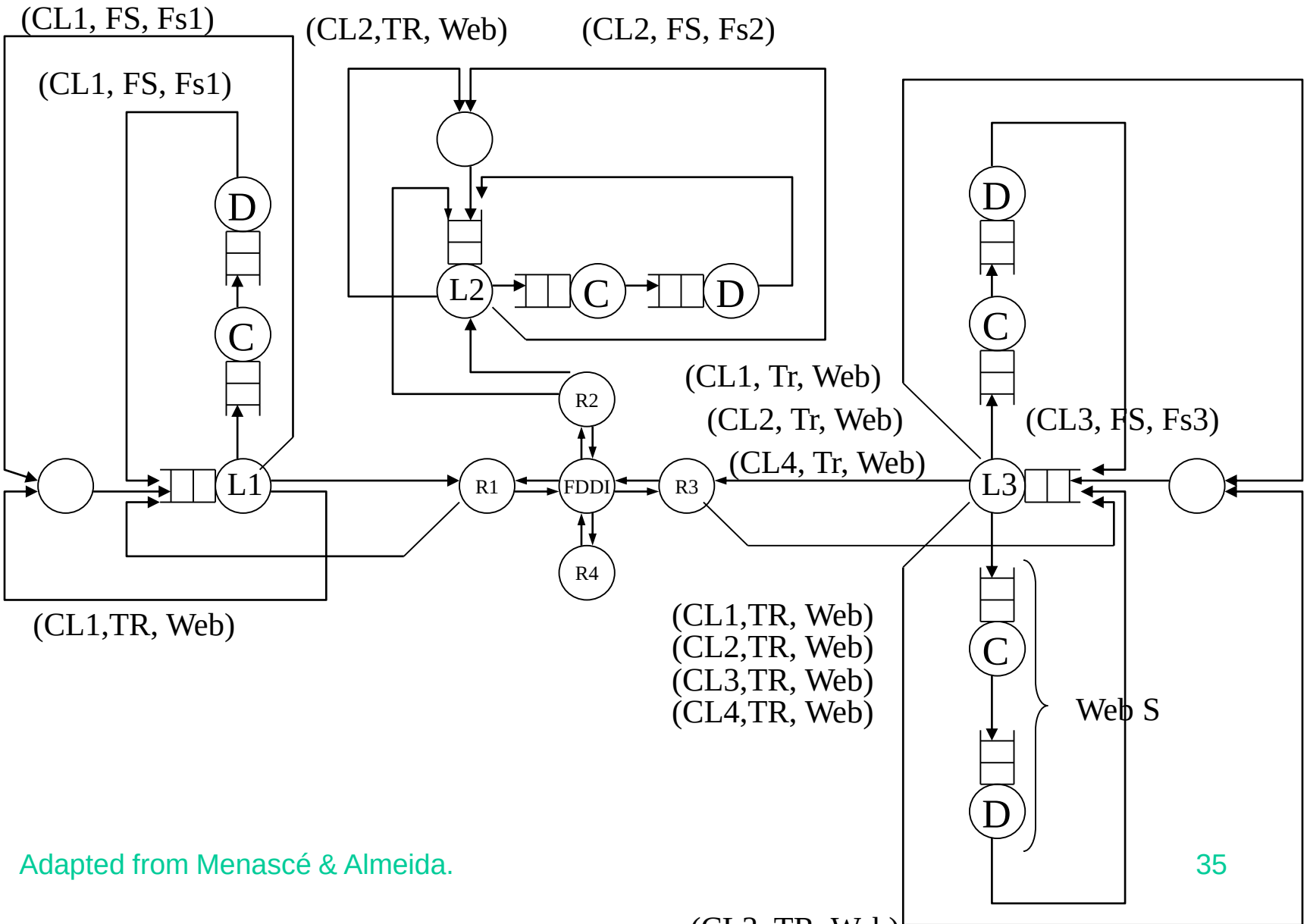
Parameters (Affecting Performance) (II)

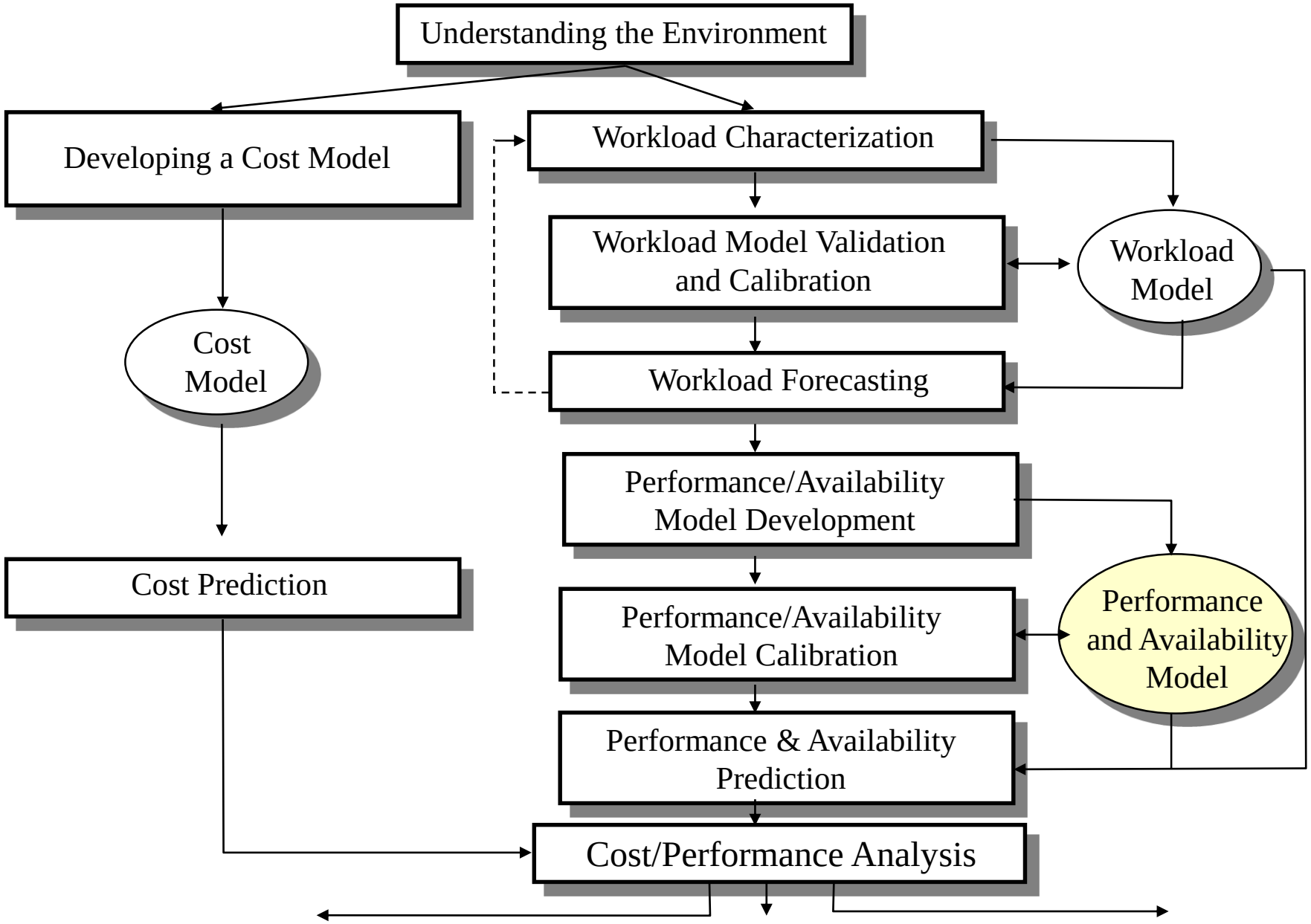
- Workload parameters examples:
  - WL intensity parameters:
    - numb. of hits/day of Web proxy
    - numb. of requests/second to file server
    - numb. of sales transactions to DB server
    - numb. of clients running scientific applications
  - WL service demand parameters:
    - CPU time of transactions at DB server
    - total transmission of replies from DB server
    - total I/O time at Web proxy for images and video clips

Queuing Network Models

- Performance prediction requires use of models
- Two types:
  - analytical models: set of formulas and/or computational algorithms ->studied in this course
  - simulation models: computer programs, all resources and the dataflow are simulated
- Both types must consider contention queues
- The various queues are interconnected: network of queues

QN model


(CL1, FS, Fs1)
(CL2,TR, Web)
(CL2, FS, Fs2)
(CL1, FS, Fs1)
D
C
L2
C
D
D
C
(CL1, Tr, Web)
R2
(CL2, Tr, Web)
(CL3, FS, Fs3)
(CL4, Tr, Web)
L1
R1
FDDI
R3
L3
R4
(CL1,TR, Web)
(CL1,TR, Web)
(CL2,TR, Web)
(CL3,TR, Web)
(CL4,TR, Web)
C
Web S
D

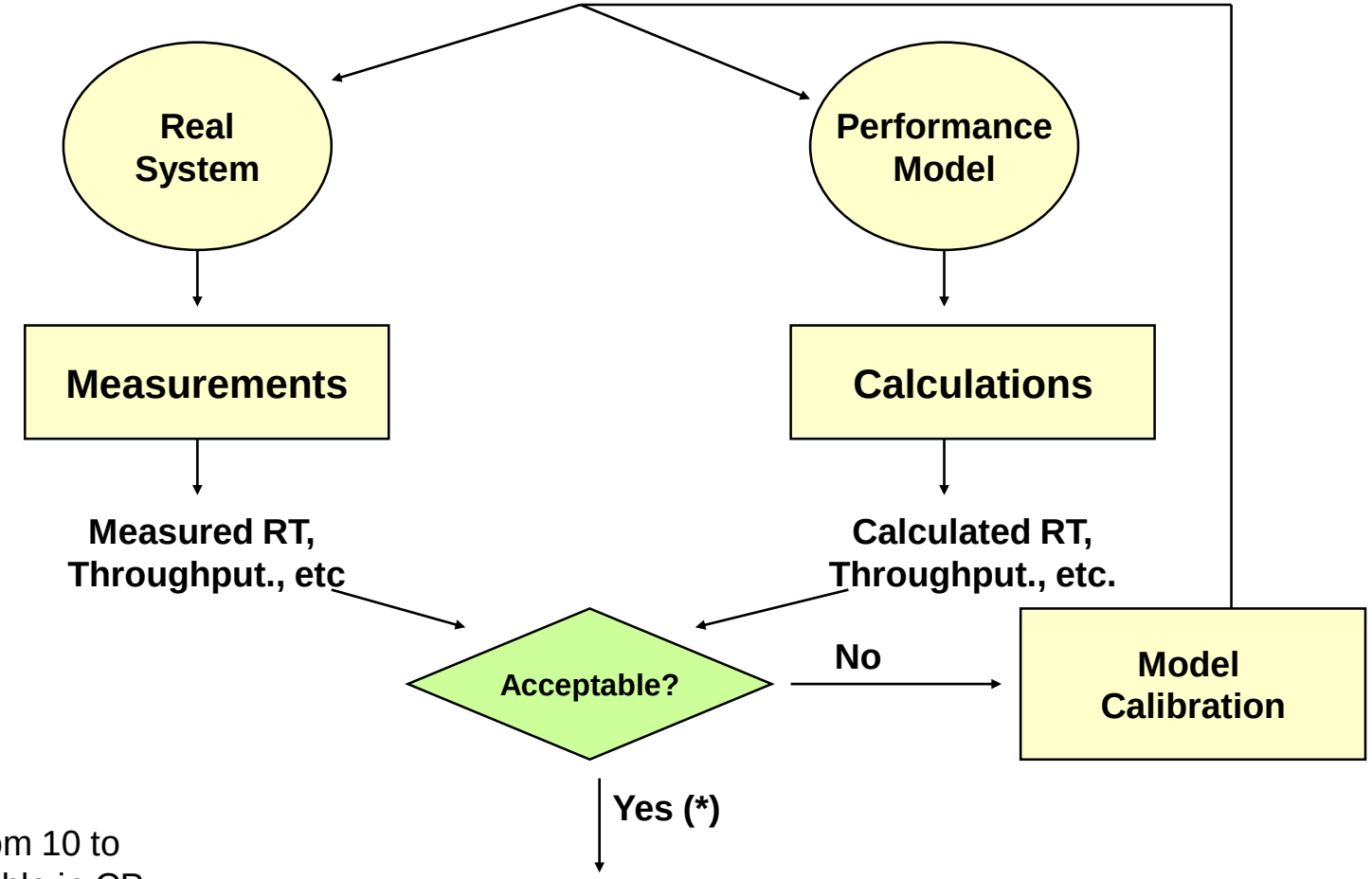


Adapted from Menasce & Almeida.

Performance Model Validation

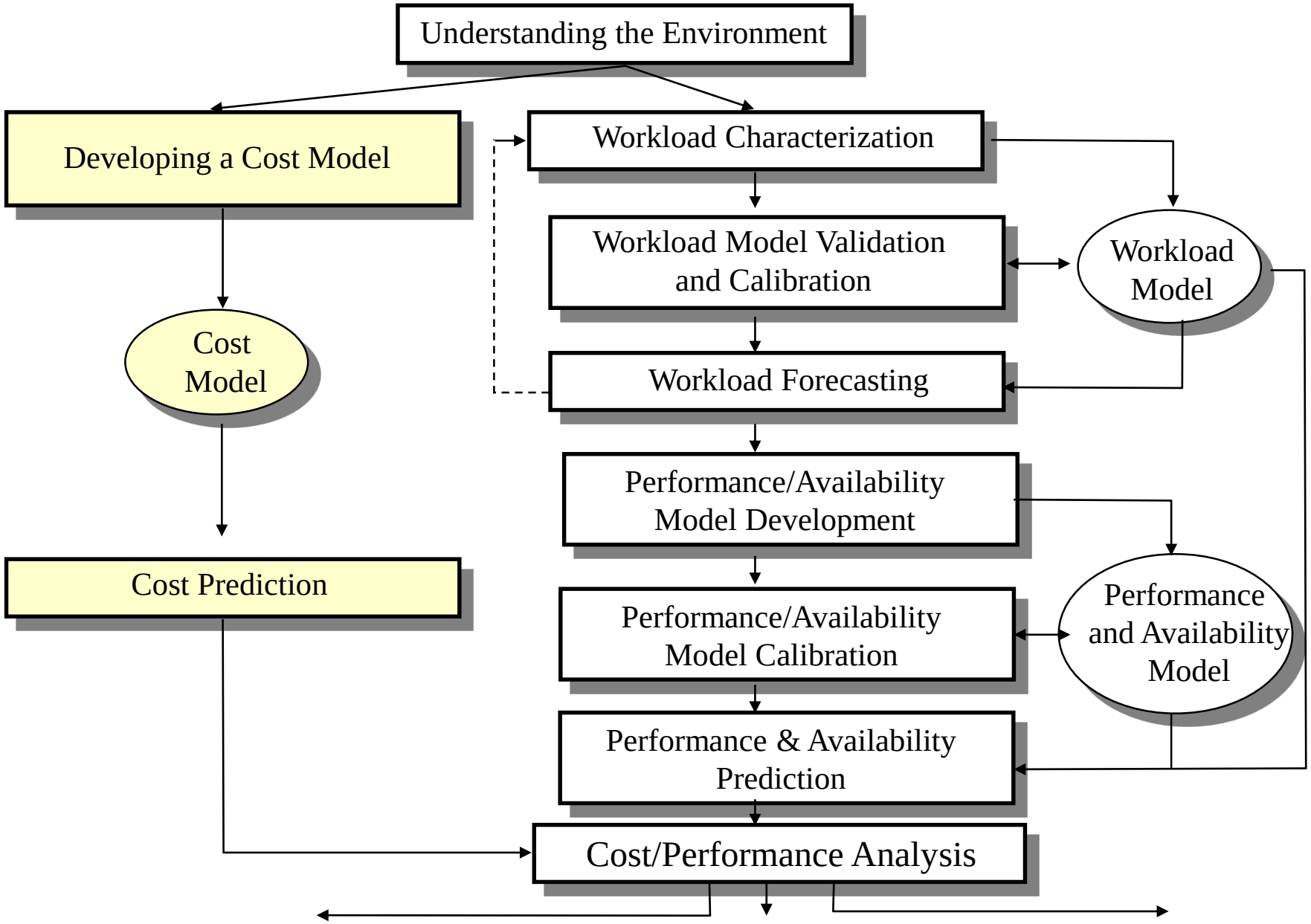
- A performance model is said to be valid if the performance metrics calculated by the model match the actual system, within an acceptable error margin. Usually 10 to 30% are acceptable in Capacity Planning.

Validating Performance Models



(*) Accuracy from 10 to 30% is acceptable in CP

Adapted from Menascé & Almeida.


Understanding the Environment
Developing a Cost Model
Workload Characterization
Workload Model Validation and Calibration
Workload Model
Cost Model
Workload Forecasting
Performance/Availability Model Development
Cost Prediction
Performance/Availability Model Calibration
Performance and Availability Model
Performance & Availability Prediction
Cost/Performance Analysis
Configuration Plan
Investment Plan
Personnel Plan

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.
- Cost categories:
    - Startup costs
    - Operating costs

Cost Model: categories

- Hardware costs: client and server machines, disks, routers, bridges, cabling, UPS, maintenance, vendor maintenance/technical support, etc.
- Software costs: operating systems, middleware, DBMS, mail processing software, office automation, antivirus, applications, etc.
- Telecommunication costs: WAN services, ISP, etc.
- Support costs: salaries and benefits of all system administrators, help desk support, personnel training, network people, etc. –

Personnel costs: 60-70% of total costs

Cost/Performance Analysis

- Cost and performance models: used to assess possible scenarios, e.g.
  - mirror Web server to balance load?
  - replace Web server with faster one?
  - Move to a 3-tier architecture?
- For each scenario, predict performances and costs
- From comparison of scenarios, get
  - configuration plan
  - investment plan
  - personnel plan
- Assess payback: ROI (Return on Investment), company's image strategy, shorter time-to-market, etc.

Summary

- Concept of adequate capacity
- Service Level Agreement (SLA)
- Framework of a methodology for capacity planning:
  - workload characterization
  - workload forecasting
  - performance modeling and prediction
  - model validation
  - cost model