Performance Engineering Methodology

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

1

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.

2

Typical PE Questions

- Can the insurance claim system meet its performance requirements of sub-second response time when a natural disaster occurs (e.g., a hurricane).
- Is the infrastructure of a government agency scalable and can it cope with the computing demands of the new required online security mechanisms?
- Is the reservation system for cruise lines able to respond to anticipated peak of customer inquiries after a TV ad campaign?

3

PE Larger Questions

- How can one plan, design, develop, deploy, and operate IT services that meet ever increasing demands for performance, availability, reliability, and security?
- Is a given IT system properly designed and sized for a given load condition?

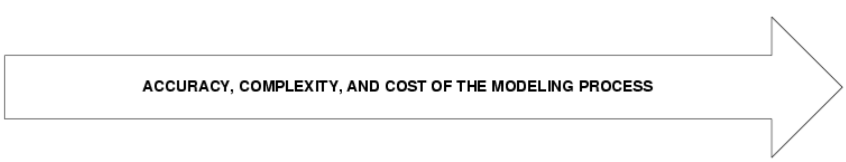
4

PE Activities

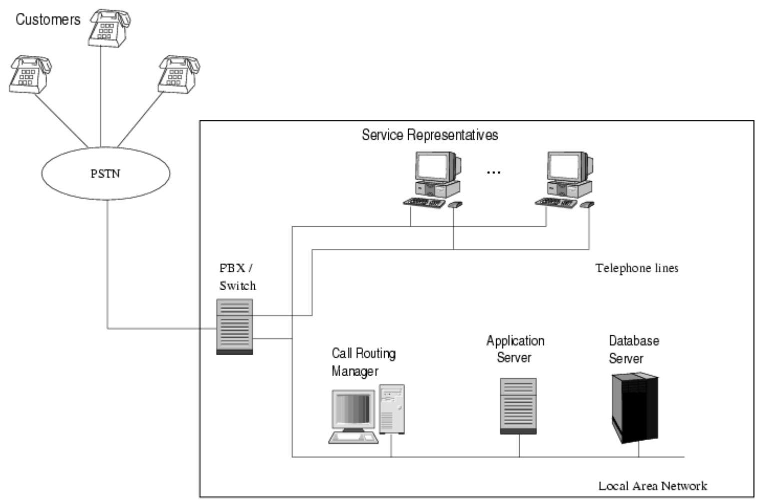
- Understand the key factors that affect a system's performance.
- Measure the system and understand its workload.
- Develop and validate a workload model that captures the key characteristics of the actual workload.
- Develop and validate an analytic model that accurately predicts the system's performance.
- Use the models to predict and optimize the system's performance.

Modeling Process

PHASES	Requirements	Design	Development	Testing	Deployment	Operation	Evolution
MODELS	WORKLOAD MODELS						
	PERFORMANCE MODELS						
	AVAILABILITY, RELIABILITY and COST MODELS						



Motivating Example: a Call Center



7

Call Center

- Goals:
  - Foster better relationships with customers, creating customer loyalty and ensuring quality service.
  - Improve efficiency and service performance.
  - Identify and explore new sales opportunities.
- Main Functions:
  - Order status inquiry
  - Shipment tracking
  - Problem resolution status inquiry
- Requirements: sub-second response time and 24x7 operation.

8

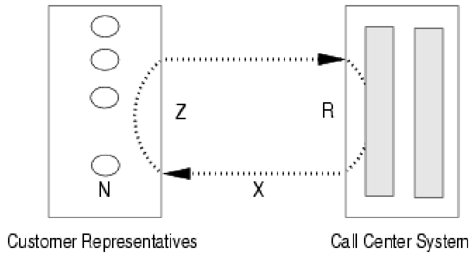
At the Requirements Analysis Phase

- Workload definition:
  - Call center’s view: Arrival rate of phone calls
  - IT system’s view: Functions received from the representatives.
  - DB server view: SQL requests from the application server.
  - LAN view: packet size distribution and interpacket arrival time.

At the System Design Phase

- What should the system throughput be to meet sub-second response times?
  - 200 customer service representatives and 80% are working during the peak hour.
  - Average think time of 30 sec.

Call Center Model



Using the Interactive Response Time Law:

$$R = N / X_0 - Z \leq 1 \quad \text{sec}$$

$$\Rightarrow X_0 \geq \frac{N}{Z+R} = \frac{200 \times 0.8}{30+1} = 5.16 \quad \text{functions/ sec}$$

11

At the System Development Phase

- What should be the capacity of the DB server so that the performance goals are met?
  – Each submitted functions requires 2.2 SQL calls on average.
  – From the Forced Flow Law:

$$X_{DB} = V_{DB} \times X_0 \geq 2.2 \times 5.16 = 11.32 \quad \text{tps}$$

12

At the Operation Phase

- Assume DB server is a problem. Response times exceed sub-second goal.
- Measurements during peak hour:
  - 57600 queries/hour
  - Each query needs 50 msec of CPU, performs 4 I/Os on disk 1 and 2 I/Os on disk 2. Each I/O takes 8 msec on average.
  - X_0 = 57600 / 3600 = 16 queries/sec
  - Service demands:
    - Dcpu = 0.05 sec; Ddisk1 = 4 x 0.008 = 0.032 sec; Ddisk2 = 2 x 0.008 = 0.016 sec.

13

At the Operation Phase (cont'd)

- Utilization computations (Service Demand Law):
  - Ucpu = Dcpu x X0 = 0.05 x 16 = 80%
  - Udisk1 = Ddisk1 x X0 = 0.032 x 16 = 51.2%
  - Udisk2 = Ddisk2 x X0 = 0.016 x 16 = 25.6%
- Response Time (Open QN Model)

$$R'_{CPU} = \frac{D_{cpu}}{1-U_{cpu}} = \frac{0.05}{1-0.8} = 0.25 \sec$$

$$R'_{disk1} = \frac{D_{disk1}}{1-U_{disk1}} = \frac{0.032}{1-0.512} = 0.066 \sec$$

$$R'_{disk2} = \frac{D_{disk2}}{1-U_{disk2}} = \frac{0.016}{1-0.256} = 0.022 \sec$$

$$R_0 = R'_{cpu} + R'_{disk1} + R'_{disk2} = 0.388 \sec$$

14

At the Evolution Phase

- Develop Web-based interface. Security requirements mandate that new applications be developed for Web access (authentication, auditing, DB access control mechanisms).

	Local	Web
Arrival Rate(tps)	16	1
Service demands (sec)		
CPU	0.05	0.15
Disk1	0.032	0.20
Disk2	0.016	0.10

Model for Evolution Scenario

Open Multiclass Queuing Networks

This wokbook comes with the books "Performance by Design," "Capacity Planning for Web Services" and "Scaling for E-Business"
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 2004, 2002 and 2000.

No. Queues: 3
No. of Classes: 2

		Classes ®	
Arrival Rates:		16.000	1.000
		Service Demand Matrix	
		Classes ®	
Queues ¯	**Type ¯ (LI/D/MPn)**	Local	Web
CPU	LI	0.05	0.15
Disk 1	LI	0.03	0.20
Disk 2	LI	0.02	0.10

Results for Evolution Scenario

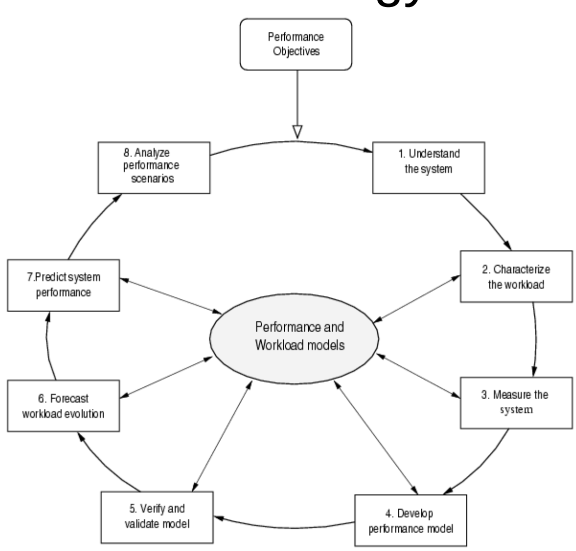
Open Multiclass Queuing Networks - Residence Times

This wokbook comes with the books "Performance by Design," "Capacity Planning for Web Services" and "Scaling for E-Business" by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 2004, 2002 and 2000.

	Classes ®	
Queues ¯	Local	Web
CPU	1.00000	3.00000
Disk 1	0.11111	0.69444
Disk 2	0.02484	0.15528
Response Time	1.14	3.85

17

Performance Engineering Methodology



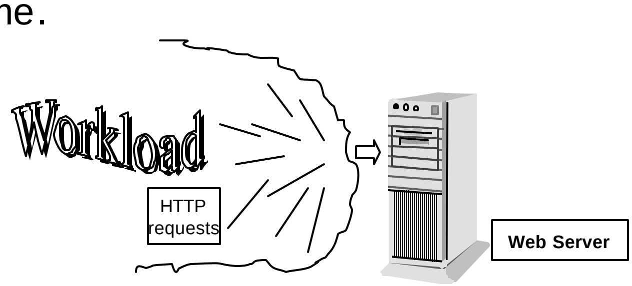
18

What is Workload Characterization?

19

Workload

- The workload of a system can be defined as the set of all inputs that the system receives from its environment during any given period of time.



20

Workload Characterization: concepts and ideas

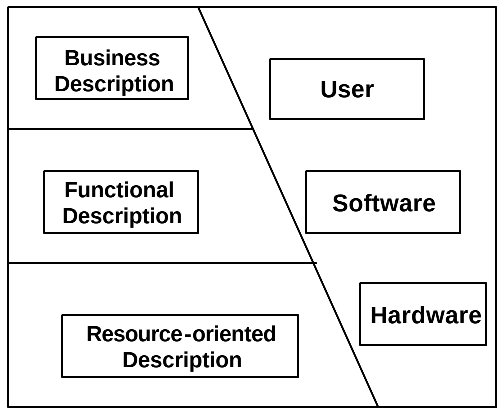
- <u>Basic component</u> of a workload refers to a generic unit of work that arrives at the system from external sources.
  - Transaction,
  - interactive command,
  - process,
  - HTTP request, and
  - depends on the nature of service provided

Workload Characterization: concepts and ideas

- Workload characterization
  - **workload model is a representation that mimics the workload under study.**

- Workload models can be used for:
  - the selection of systems
  - performance tuning
  - capacity planning

Workload Description



Workload Description

- Business characterization: a user-oriented description that describes the load in terms such as number of employees, invoices per customer, etc.
- Functional characterization: describes programs, commands and requests that make up the workload
- Resource-oriented characterization: describes the consumption of system resources by the workload, such as processor time, disk operations, memory, etc.

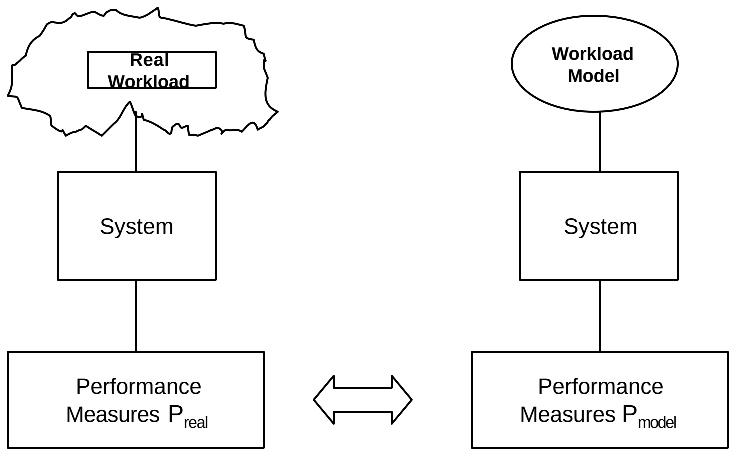
A Web Server Example

- The pair **(CPU time, I/O time)** characterizes the execution of a request at the server.
- Our basic workload: 10 HTTP requests
- First case: only one document size (15KB)
- 10 executions ---> (0.013 sec, 0.09 sec)
- More realistic workload: documents have different sizes.

Execution of HTTP Requests (sec)

Request No.	CPU time (sec)	I/O time (sec)	Elapsed time (sec)
1	0.0095	0.0400	0.0710
2	0.0130	0.1100	0.1450
3	0.0155	0.1200	0.1560
4	0.0088	0.0400	0.0650
5	0.0111	0.0900	0.1140
6	0.0171	0.1400	0.1630
7	0.2170	1.2000	4.3800
8	0.0129	0.1200	0.1510
9	0.0091	0.0500	0.0630
10	0.0017	0.1400	0.1890
Average	0.03157	0.205	0.5497

Representativeness of a Workload Model



27

A Refinement in the Workload Model

- The average response time of 0.55 sec does not reflect the behavior of the actual server.
- Due to the heterogeneity of the its components, it is difficult to view the workload as a single collection of requests.
- Three classes
  - small documents
  - medium documents
  - large documents

28

Execution of HTTP Requests (sec)

Request No.	CPU time (sec)	I/O time (sec)	Elapsed time (sec)
1 small	0.0095	0.0400	0.0710
2 medium	0.0130	0.1100	0.1450
3 medium	0.0155	0.1200	0.1560
4 small	0.0088	0.0400	0.0650
5 medium	0.0111	0.0900	0.1140
6 medium	0.0171	0.1400	0.1630
7 large	0.2170	1.2000	4.3800
8 medium	0.0129	0.1200	0.1510
9 small	0.0091	0.0500	0.0630
10 medium	0.0017	0.1400	0.1890

29

Three-Class Characterization

Type	CPU time (sec)	I/O time (sec)	No of omponents
Small Docs.	0.0091	0.04	3
Medium Docs.	0.0144	0.12	6
Large Docs.	0.2170	1.20	1
Total	0.331	2.05	10

30

Workload Models

- A model should be representative and compact.
- Natural models are constructed either using basic components of the real workload or using traces of the execution of real workload.
- Artificial models do not use any basic component of the real workload.
  - Executable models (e.g.: synthetic programs, artificial benchmarks, etc)
  - **Non-executable models, that are described by a set of parameter values that reproduce the same resource usage of the real workload.**

31

Workload Models

- The basic inputs to analytical models are parameters that describe the service centers (i.e., hardware and software resources) and the customers (e.g. requests and transactions)
  - component (e.g., transactions) interarrival times;
  - service demands
  - execution mix (e.g., levels of multiprogramming)

32

A Workload Characterization Methodology

- Choice of an analysis standpoint
- Identification of the basic component
- Choice of the characterizing parameters
- Data collection
- Partitioning the workload
- Calculating the class parameters

33

Selection of characterizing parameters

- Each workload component is characterized by two groups of information:
- Workload intensity
  - arrival rate
  - number of clients and think time
  - number of processes or threads in execution simultaneously
- Service demands (D_{i1}, D_{i2}, … D_{iK}), where D_{ij} is the service demand of component i at resource j.

34

Data Collection

- This step assigns values to each component of the model.

  - Identify the time windows that define the measurement sessions.
  - Monitor and measure the system activities during the defined time windows.
  - From the collected data, assign values to each characterizing parameters of every component of the workload.

35

Partitioning the workload

- Motivation: real workloads can be viewed as a collection of heterogeneous components.
- Partitioning techniques divide the workload into a series of classes such that their populations are composed of quite homogeneous components.
- What attributes can be used for partitioning a workload into classes of similar components?

36

Partitioning the Workload

- Resource usage
- Applications
- Objects
- Geographical orientation
- Functional
- Organizational units
- Mode

37

Workload Partitioning: Resource Usage

Transaction Classes	Frequency	Maximum CPU time (msec)	Maximum I/O time (msec)
Trivial	40%	8	120
Light	30%	20	300
Medium	20%	100	700
Heavy	10%	900	1200

38

Workload Partitioning: Internet Applications

Application Classes	KB Transmitted
WWW	**4,216**
ftp	**378**
telnet	**97**
Mbone	**595**
Others	**63**

39

Workload Partitioning: Document Types

Document Class	Percentage of Access (%)
HTML (html file types)	**30**
Images (e.g., gif or jpeg)	**40**
Sound (e.g., au or wav)	**4.5**
Video (e.g., mpeg, avi or mov)	**7.3**
Dynamic (e.g., cgi or perl)	**12.0**
Formatted (e.g., ps, dvi or doc)	**5.4**
Others	**0.8**

40

Workload Partitioning: Geographical Orientation

Classes	Percentage of Total Requests
East Coast	**32**
West Coast	**38**
Midwest	**20**
Others	**10**

41

Calculating the class parameters

- How should one calculate the parameter values that represent a class of components?
  - Averaging: when a class consists of homogeneous components concerning service demands, an average of the parameter values of all components may be used.
  - Clustering of workloads is a process in which a large number of components are grouped into clusters of similar components.

42

Calculating Class Parameters

- Homogeneous Workload:
  - compute arithmetic mean
  - Workload: {(Di1, Di2, …, DiK) | i = 1, …, p}
  - Workload Charaterization:
    - (D1, D2, …, DK) where
    - $Dj = 1/p \sum_{i=1}^{p} Dij$

43

Calculating Class Parameters

- Heterogeneous Workload:
  - use clustering analysis to determine groups of "similar" workloads.
  - Use averaging within each group.
  - Clustering analysis algorithms: minimal spanning tree and k-means.

44

Parameter Transformation

- Preventing extreme values of parameters from distorting distribution use linear transformation:

- Dt = (measured D - minimum{Di}) / (maximum{Di} - minimum{Di})

45

Workload Sample

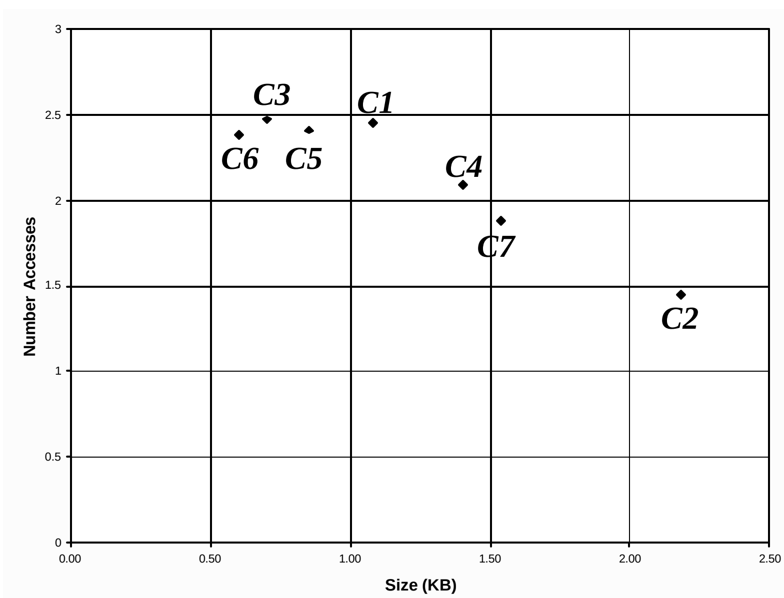
Document	Size (KB)	No. Accesses
1	12	281
2	150	28
3	5	293
4	25	123
5	7	259
6	4	241
7	35	75

46

Workload Sample: logarithmic transformation of parameters

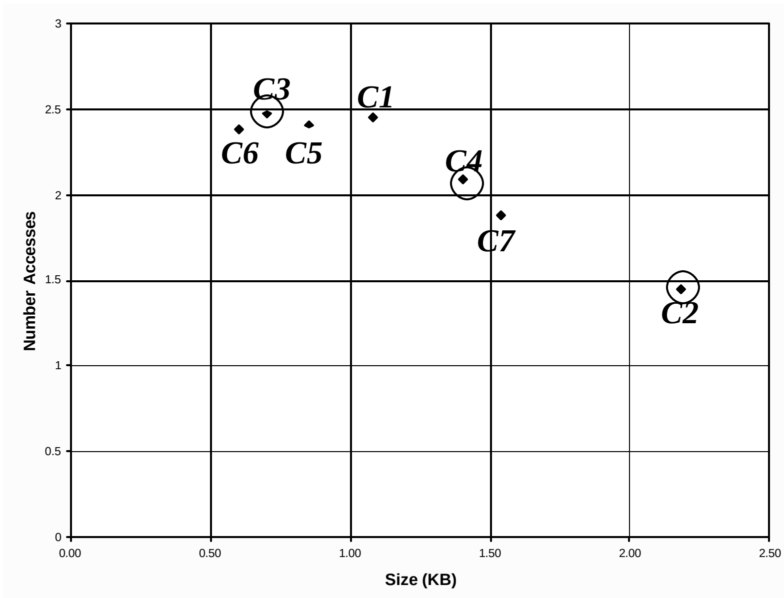
Document	Size (KB)	No. Accesses
1	1.08	2.45
2	2.18	1.45
3	0.70	2.47
4	1.40	2.09
5	0.85	2.41
6	0.60	2.38
7	1.54	1.88

47



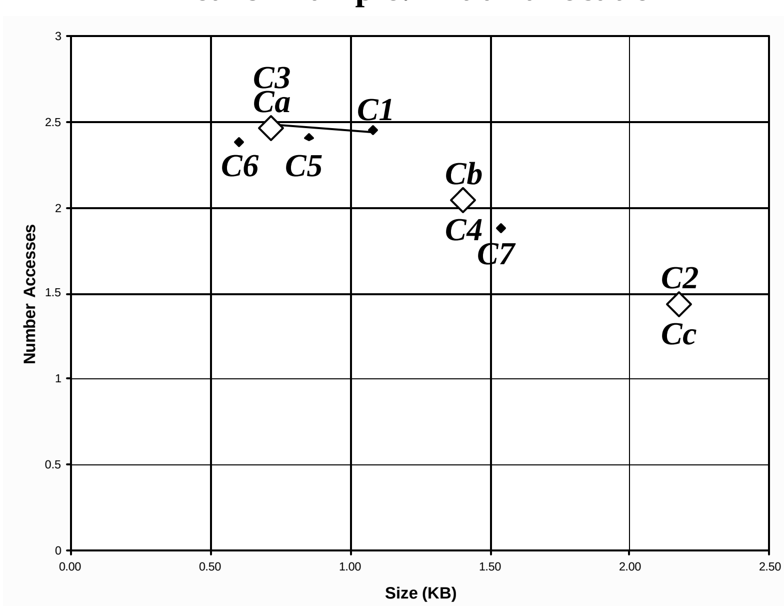
48

K-means Example: initial allocation



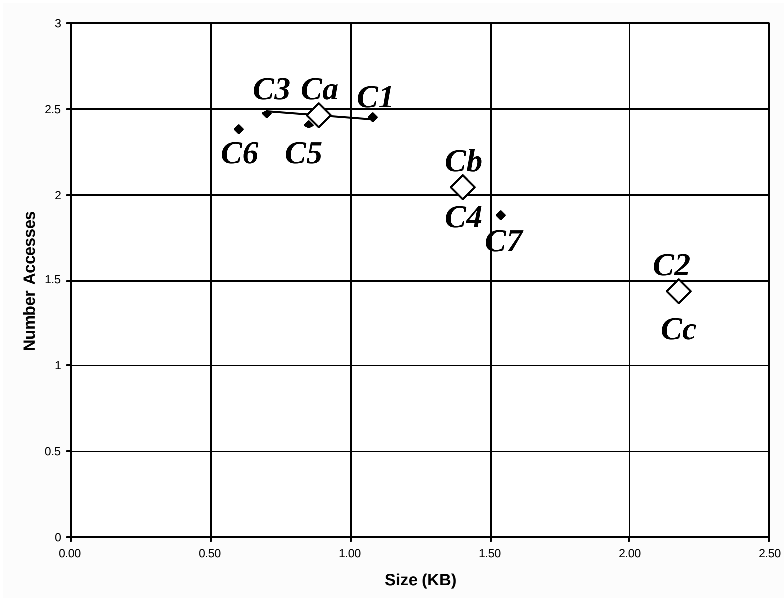
49

K-means Example: initial allocation



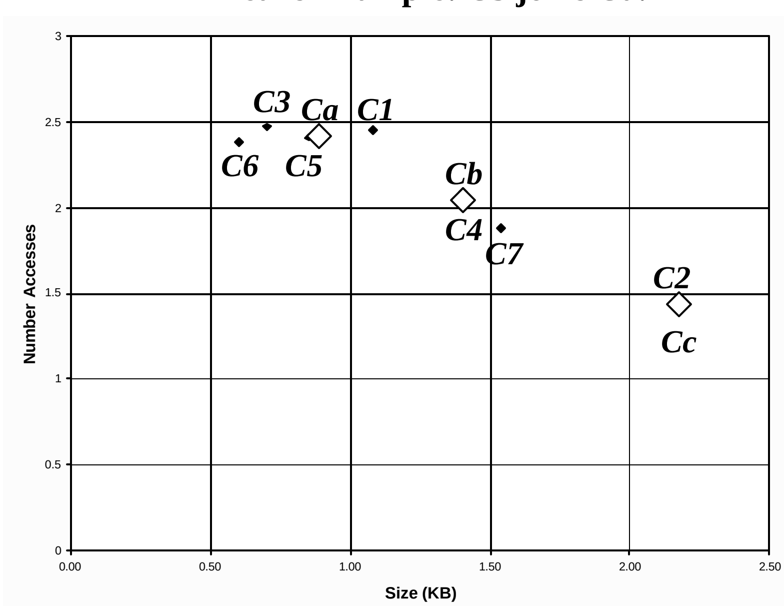
50

K-means Example: C1 joins Ca.



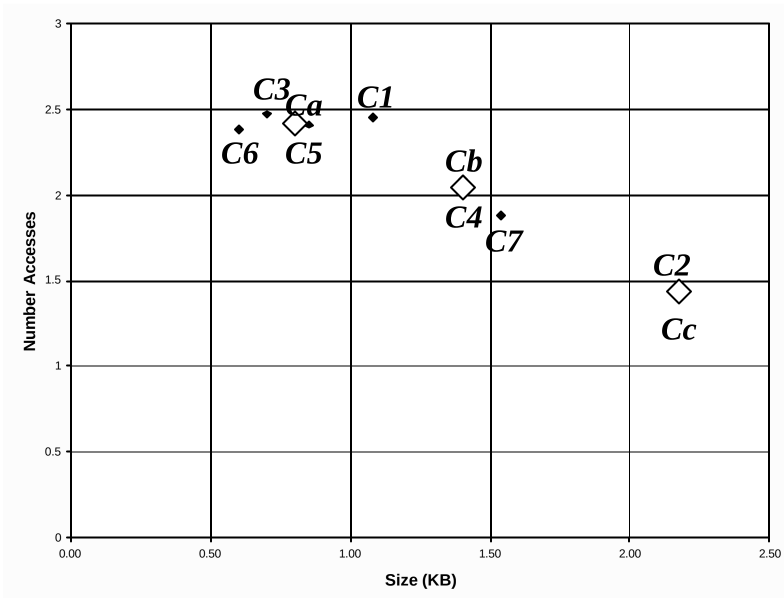
51

K-means Example: C5 joins Ca.



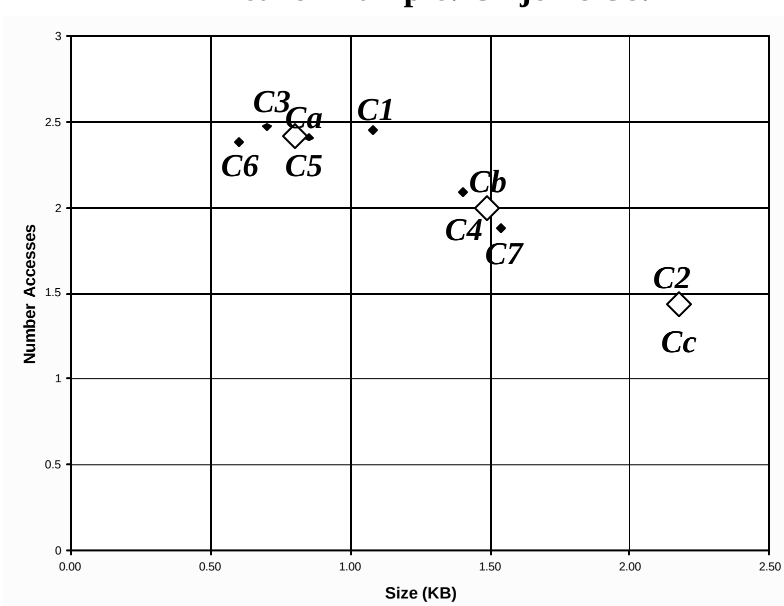
52

K-means Example: C6 joins Ca.



53

K-means Example: C7 joins Cb.



54

Result of Workload Characterization

Type	Class	Size (KB)	No. Accesses	No. Components
Small	C1356	8.19	271.51	4
Medium	C47	29.58	96.05	2
Large	C2	150.00	28.00	1

Clustering Analysis

- The Euclidean distance between points

$$w_i = (D_{i1}, ..., D_{iM})$$
$$w_j = (D_{j1}, ..., D_{jM})$$

- is

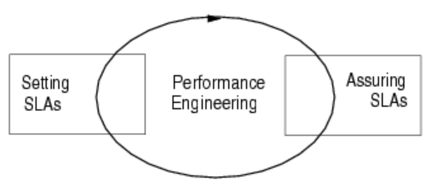
$$d_{ij} = \sqrt{\sum_{n=1}^{M} \left(D_{in} - D_{jn}\right)^2}$$

k-means Clustering

1. Set the number of clusters to k.
2. Choose k starting points as initial estimates of the k clusters.
3. Examine each point and allocate it to the closest centroid. Recompute the centroid's coordinates (avg. of all cluster's points coordinates).
4. Repeat step 3 until no points change allocation or until a max number of passes is performed.

57

PE and SLAs



Examples of Service Level Agreements:

- The system throughput should be greater than 1,000 query transactions per second with at least 90% of the transactions responding in less than 2 seconds.
- The application server should be available at least 99.9% of the time during the business hours of week days.
- The response time of the patient information system should not exceed 1 sec for local users.

SLAs should be associated with the cost of providing a certain level of service.

58

Total Cost of Ownership (TCO)

- Hardware costs (purchase and/or leasing expenses)
- Software costs
- Communication costs
- Management costs
- Support costs
- Facilities costs
- Downtime costs

TCO Example

- Basic cost of System Y = $300,000
- Basic cost of System Z = $350,000
- Throughput of System Y = 220 tps
- Throughput of System Z = 230 tps
- System Y expected downtime = 38 hrs/3 yrs
- System Z expected downtime = 21 hrs/3 yrs
- Call center charges $5 per call from customers.
- Avg. call rate = 1,000 calls/hr
- Avg. cost per hr of downtime = $5,000
- Total cost = Basic Cost + Downtime Cost
- Cost of System Y = $300,000 + 38 * 1,000 * 5 = $490,000
- Cost of System Z = $350,000 + 21 * 1,000 * 5 = $455,000

System Z is preferable over System Y.