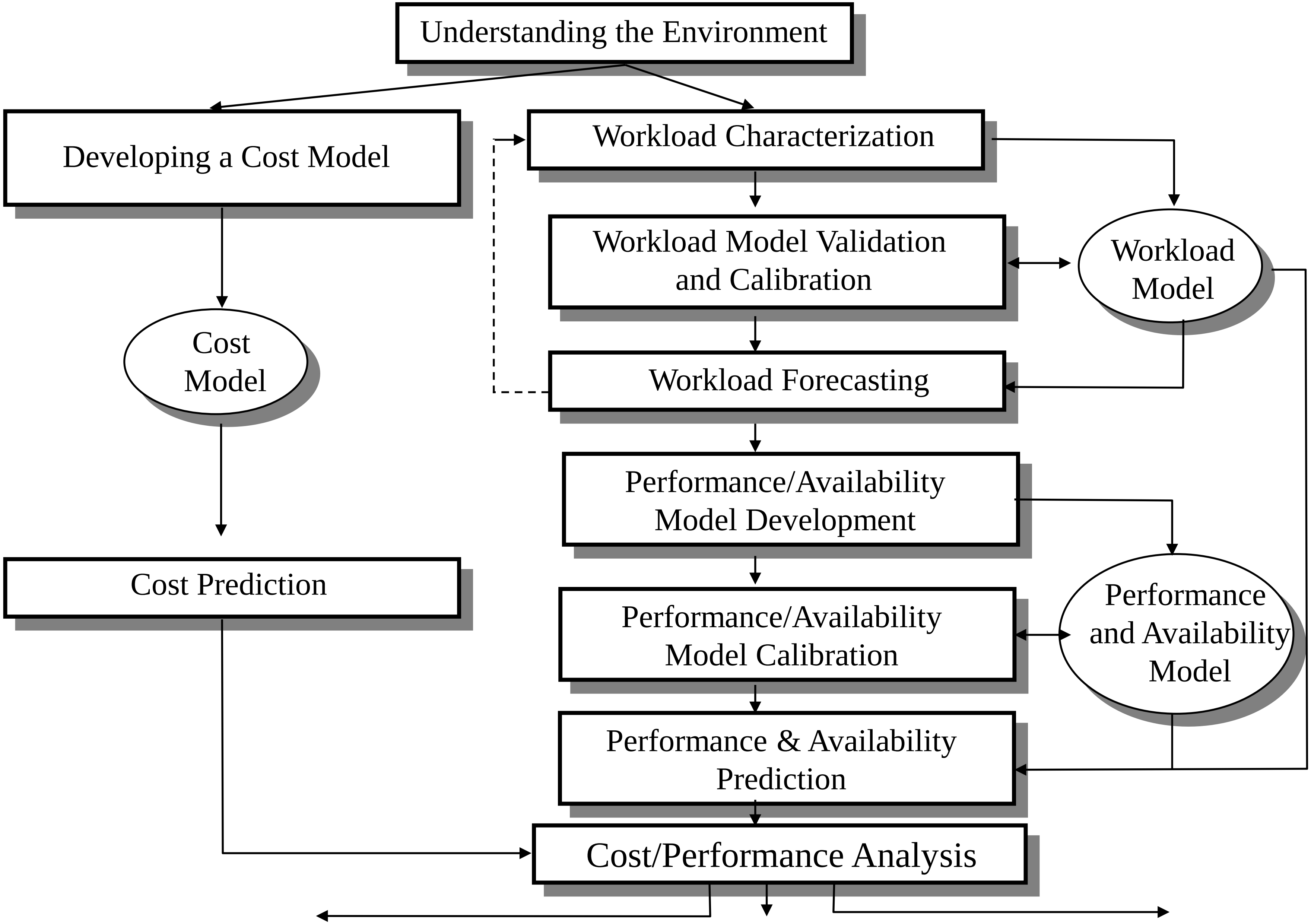
Workload Characterization for the Web

Adapted from Menascé & Almeida.


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

Configuration Plan **Investment Plan** **Personnel Plan**

Adapted from Menasce & Almeida.

Learning Objectives (1)

- Introduce the workload characterization problem.
- Discuss a simple example of characterizing the workload for an intranet.
- Present a workload characterization methodology.

Learning Objectives (2)

- Discuss the following steps:
  - analysis standpoint
  - identification of the basic component
  - choice of the characterizing parameters
  - data collection
  - partitioning the workload
- Characteristics of Web workloads:
  - burstiness
  - heavy-tailed distributions

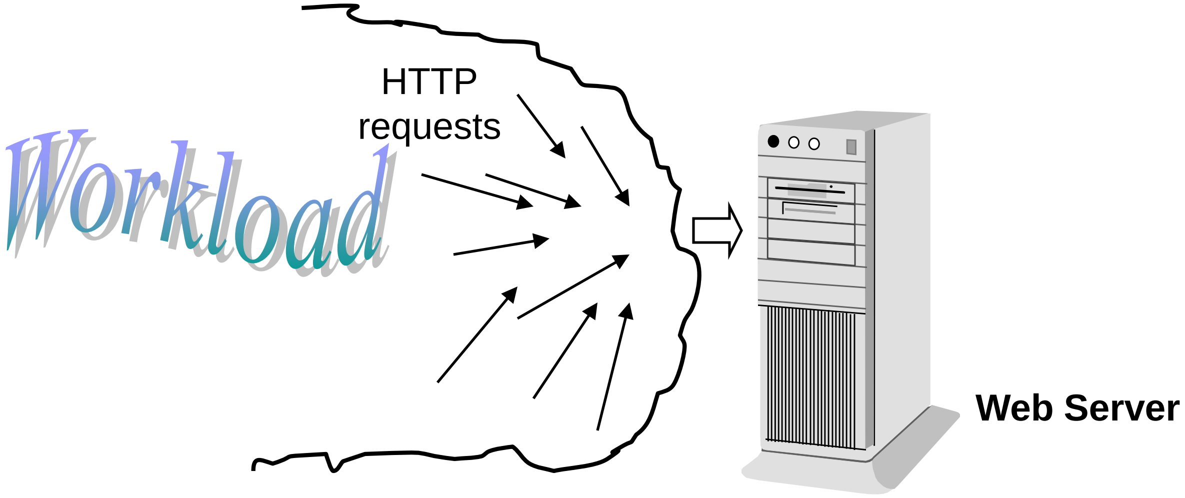
Adapted from Menascé & Almeida.

What is Workload Characterization?

Adapted from Menascé & Almeida.

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.



Workload Characterization

- Depends on the purpose of the study
  - cost x benefit of a proxy caching server
  - impact of a faster CPU on the response time
- Common steps
  - specification of a point of view from the workload will be analyzed;
  - choice of set of relevant parameters;
  - monitoring the system;
  - analysis and reduction of performance data
  - construction of a workload model.

Adapted from Menascé & Almeida.

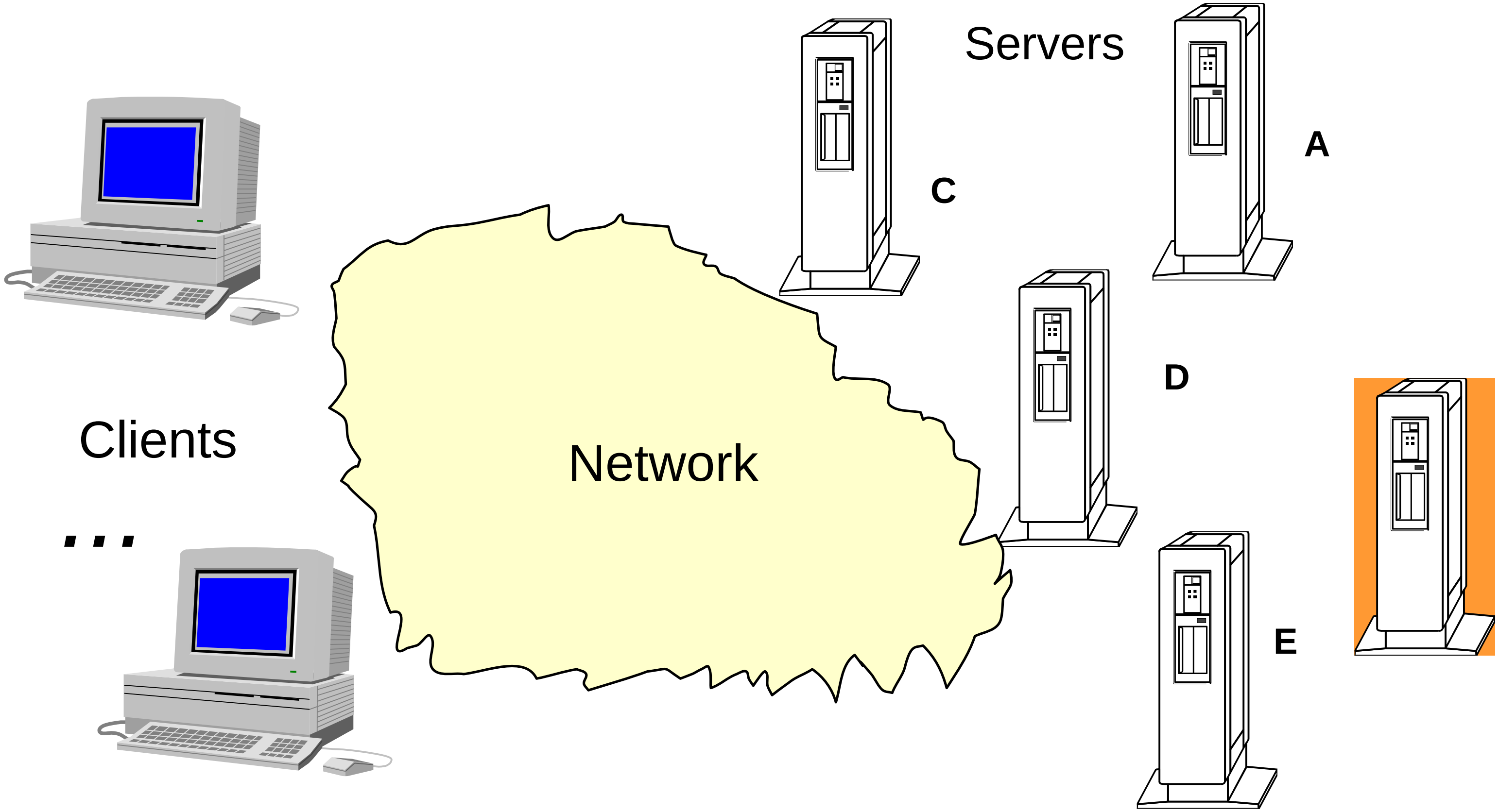
A Simple Example

- A construction and engineering company is planning to roll out new applications and to increase the number of employees that have access to the corporate intranet. The main applications are health human resources, insurance payments, on-demand interactive training, etc.

- Main problem: response time of the human resource system

Adapted from Menascé & Almeida.

A Simple Example (2)



A Simple Example: basic questions

- What is the purpose of the study?
- What workload we want to characterize?
- What is the level of the workload description?
  - High-level characterization in terms of Web applications;
  - Low-level characterization in terms of resource usage.
- How could this workload be precisely described?

Adapted from Menascé & Almeida.

Workload Characterization: concepts and ideas

- Basic component of a workload refers to a generic unit of work that arrives at the system from external sources, e.g.
  - transactions,
  - interactive commands,
  - HTTP requests

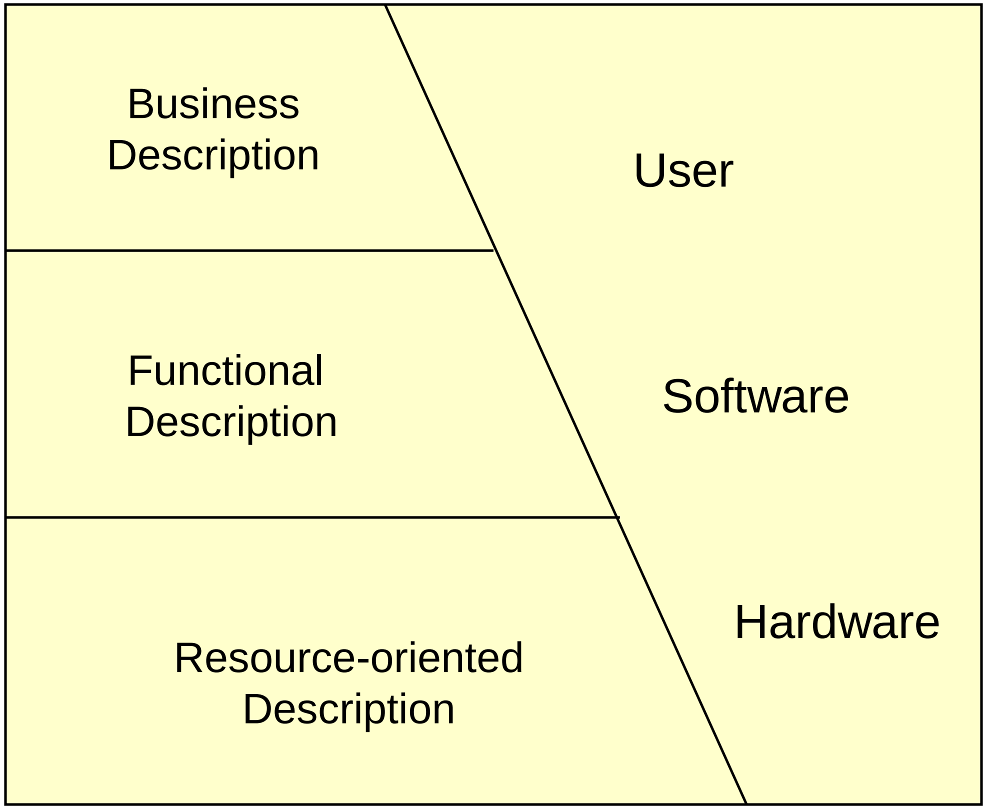
Adapted from Menascé & Almeida.

Workload Characterization: concepts and ideas

- Workload characterization
  - **workload model is a representation that mimics the workload under study.**
- Workload models can be used:
  - selection of systems
  - performance tuning
  - capacity planning

Adapted from Menascé & Almeida.

Workload Description



Business Description

Functional Description

Resource-oriented Description

User

Software

Hardware

Adapted from Menascé & Almeida.

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.
- Measures related to 10 HTTP requests
- Requested documents have different sizes.

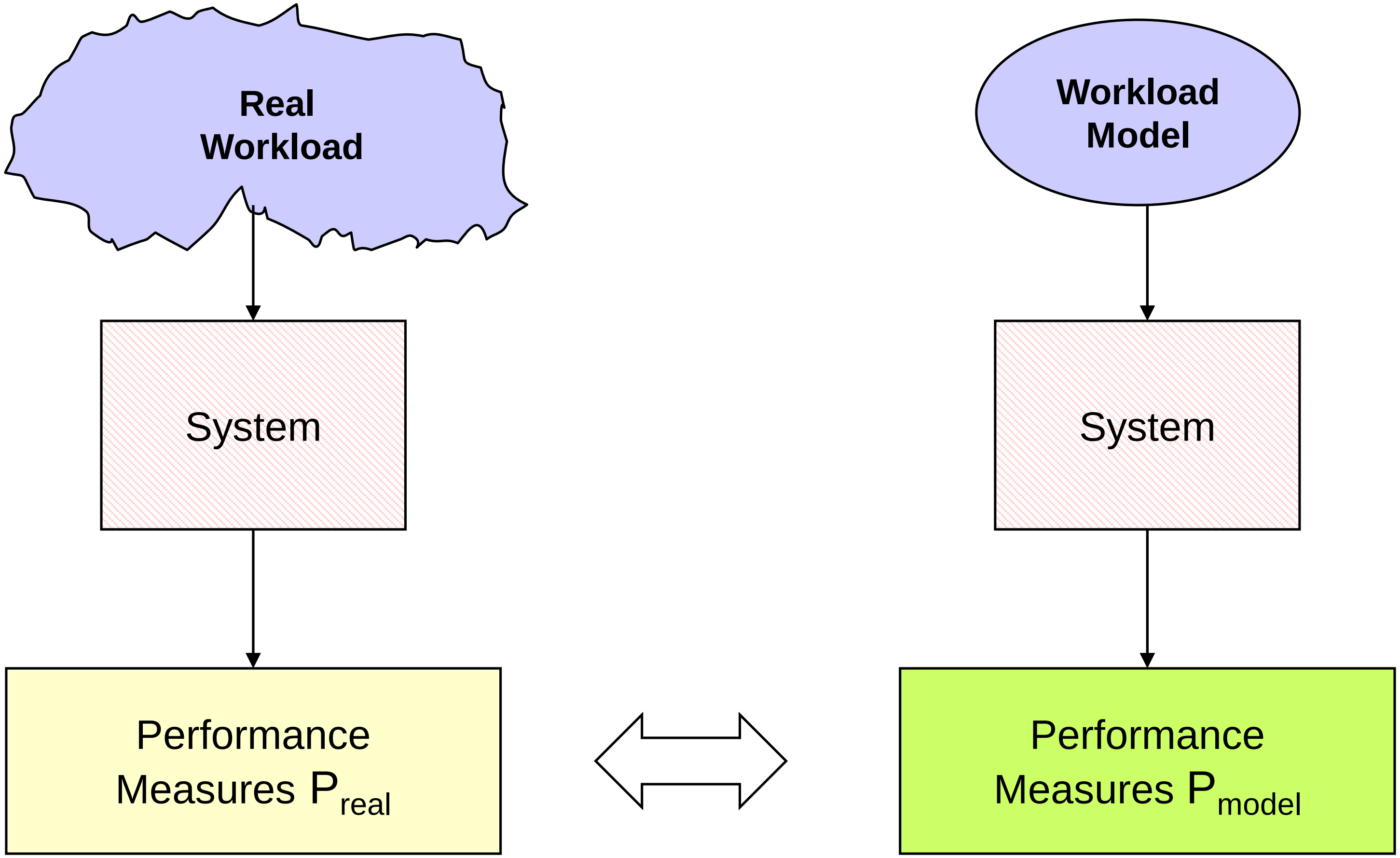
Adapted from Menascé & Almeida.

Execution of HTTP Requests (sec)

Request No.	CPU time	I/O time	Elapsed time
1	0.0095	0.04	0.071
2	0.0130	0.11	0.145
3	0.0155	0.12	0.156
4	0.0088	0.04	0.065
5	0.0111	0.09	0.114
6	0.0171	0.14	0.163
7	0.2170	1.20	4.380
8	0.0129	0.12	0.151
9	0.0091	0.05	0.063
10	0.0170	0.14	0.189
Average	**0.0331**	**0.205**	**0.550**

Adapted from Menascé & Almeida.

Representativeness of a Workload Model


Real Workload
Workload Model
System
System
Performance Measures P_{real}
Performance Measures P_{model}

Adapted from Menascé & Almeida.

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

Execution of HTTP Requests (sec)

Request No.	CPU time	I/O time	Elapsed time
1 small	0.0095	0.04	0.071
2 medium	0.0130	0.11	0.145
3 medium	0.0155	0.12	0.156
4 small	0.0088	0.04	0.065
5 medium	0.0111	0.09	0.114
6 medium	0.0171	0.14	0.163
7 large	0.2170	1.20	4.380
8 medium	0.0129	0.12	0.151
9 small	0.0091	0.05	0.063
10 medium	0.0170	0.14	0.189

Adapted from Menascé & Almeida.

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

Adapted from Menascé & Almeida.

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.

Adapted from Menascé & Almeida.

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)

Adapted from Menascé & Almeida.

Graph-based models

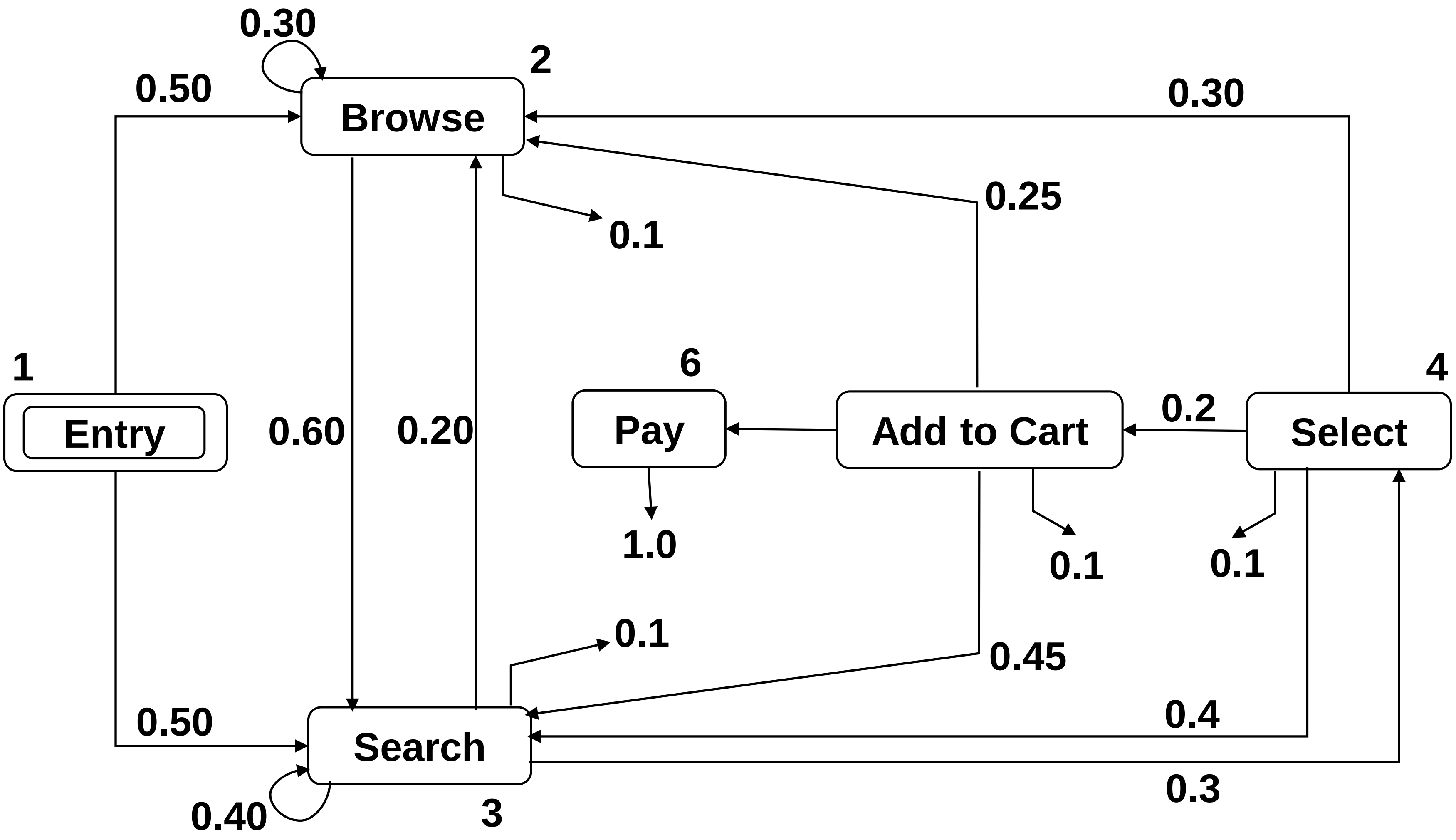
- Customer Behavior Model Graph (CBMG)

  - transitional aspect: how a customer moves from one state to the next

  - temporal aspect: time it takes for the customer to move from one state to the next (*think time*)

Adapted from Menascé & Almeida.

Graph-based models



Adapted from Menascé & Almeida.

Graph-based models

- V_j : average number of visits to the state *j*

- $V_{Add} = V_{Select}$ x 0.2
- $V_{Browse} = V_{Searcht}$ x $0.20 + V_{Select}$ x $0.3 +$
  V_{Add} x $0.25 + V_{Browse}$ x $0.30 + V_{Entry}$ x 0.5

$$V_1 = 1$$

$$V_j = \sum_{k=1}^{n-1} V_k * p_{k,j}$$

Adapted from Menascé & Almeida.

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

Adapted from Menascé & Almeida.

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.

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.

Adapted from Menascé & Almeida.

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?

Partitioning the Workload

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

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

Adapted from Menascé & Almeida.

Workload Partitioning: Internet Applications

Application	Percentage of total traffic
HTTP	**29**
ftp	**20**
SMTP and POP3	**9**
Streaming	**11**
P2P	**14**
Others	**17**

Adapted from Menascé & Almeida.

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**

Adapted from Menascé & Almeida.

Workload Partitioning: Geographical Orientation

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

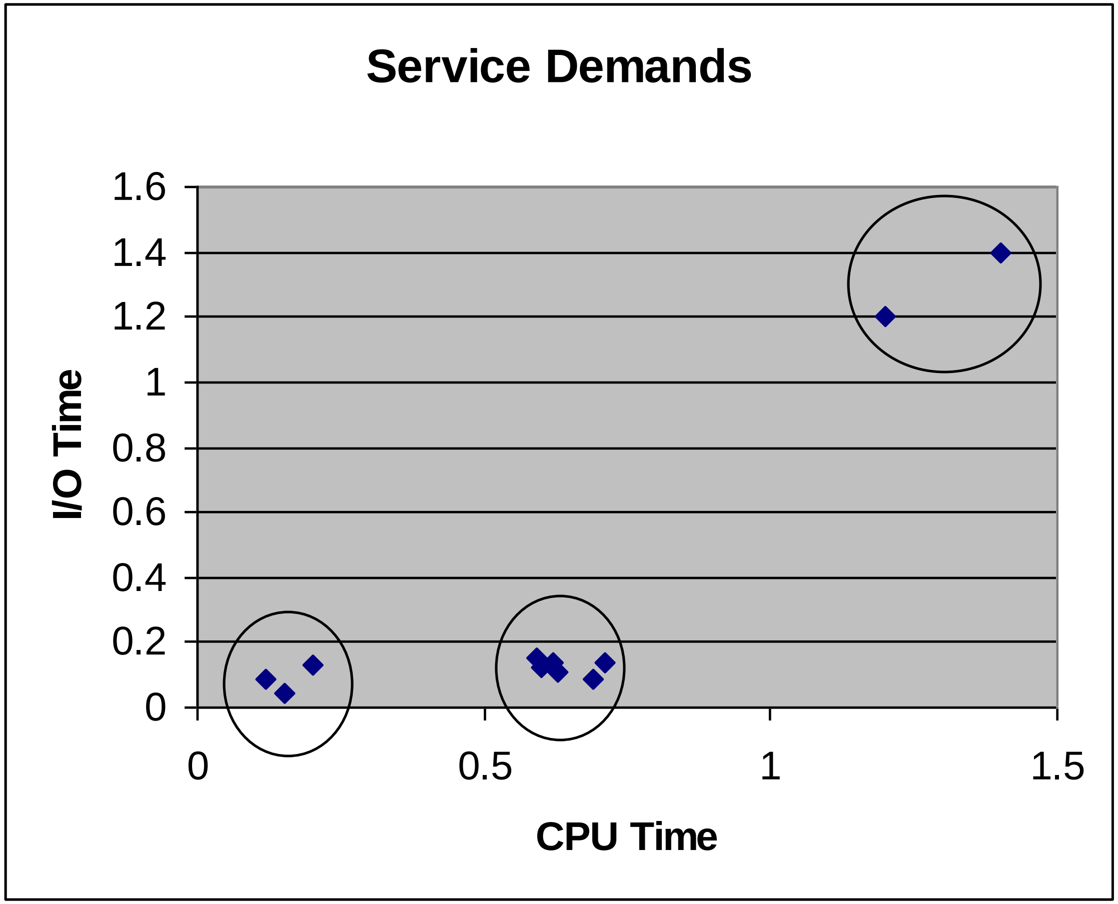
Adapted from Menascé & Almeida.

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.

Adapted from Menascé & Almeida.

Clustering Analysis



Adapted from Menascé & Almeida.

New Phenomena in the Internet and WWW

- Self-similarity - a self-similar process looks bursty across several time scales.
- Heavy-tailed distributions in workload characteristics, that means a very large variability in the values of the workload parameters.

Adapted from Menascé & Almeida.

Power Laws: $y \propto x^{\alpha}$

- Heavy-tailed distribution

$$P[X > x] = kx^{-\alpha} L(x)$$

- Great degree of variability, and a non negligible probability of high sample values
- When α is less then 2, the variance is infinite, when α is less than 1, the mean is infinite.
- Pareto distribution decays slowler than the exponential distribution
- Zipf’s Law describes phenomena where large events are rare, but small ones are quite common
- Popularity of static pages

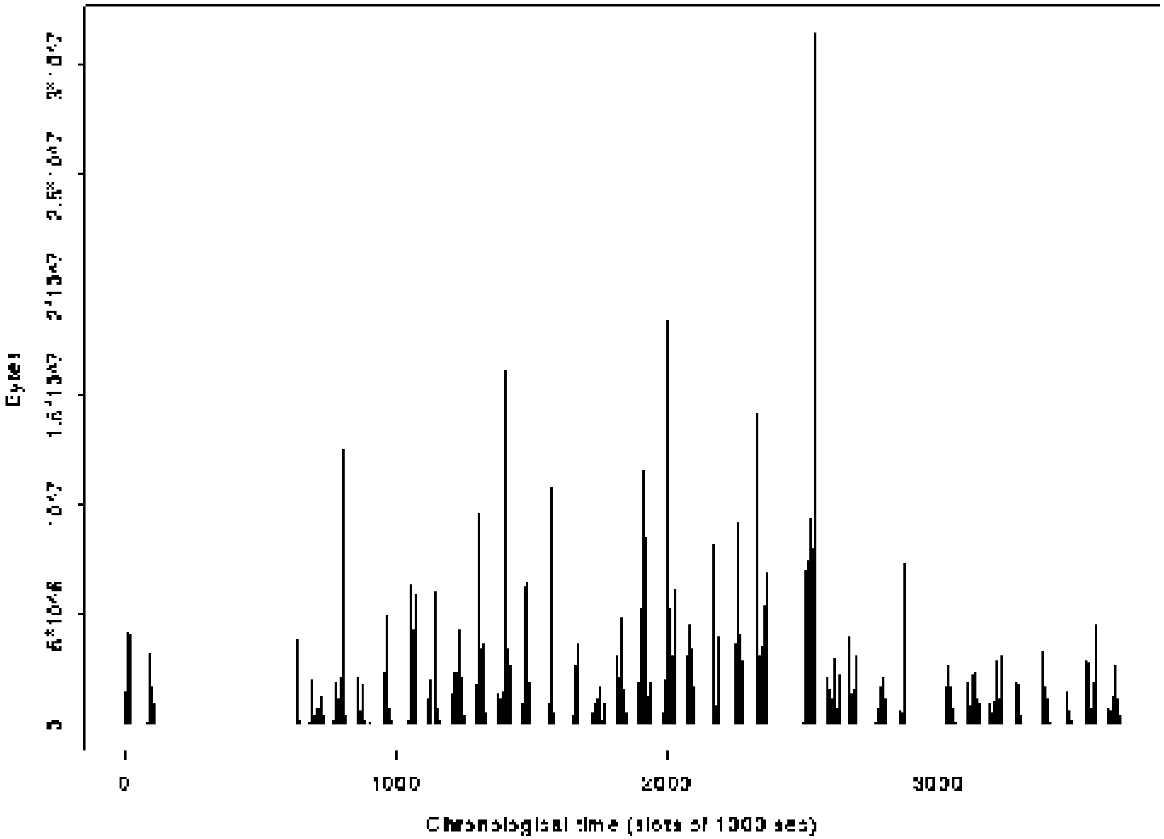
Adapted from Menascé & Almeida.

WWW Traffic Burst

Bytes

10^7

10^6



Chronological time (slots of 1000 sec)

Adapted from Menascé & Almeida.

Incorporating New Phenomena in the Workload Characterization

Burstiness Modeling

- burstiness in a given period can be represented by a pair of parameters (a,b)
    - **a** is the ratio between the maximum observed request rate and the average request rate during the period.
    - **b** is the fraction of time during which the instantaneous arrival rate exceeds the average arrival rate.

Burstiness Modeling

- Consider an HTTP LOG composed of L requests to a Web server.
- τ: time interval during which the requests arrive
- λ: average arrival rate, $\lambda = L / \tau$
- The time interval τ is divided into n equal subintervals of duration τ / n called epochs
- Arr(k) number of HTTP requests that arrive in epoch k
- λ_k arrival rate during epoch k

Burstiness Modeling

- Arr^+ total number of HTTP requests that arrive in epochs in which $\lambda_k > \lambda$
- $b = (\text{number of epochs for which } \lambda_k > \lambda) / n$
- above-average arrival rate, $\lambda^+ = Arr^+ / (b*\tau)$
- $a = \lambda^+ / \lambda = Arr^+ / (b*L)$

Adapted from Menascé & Almeida.

Burstiness Modeling: an example

- Example: Consider that 19 requests are logged at a Web server at instants:

1 3 3.5 3.8 6 6.3 6.8 7.0 10 12 12.2 12.3 12.5
12.8 15 20 30 30.2 30.7

- What are the burstiness parameters?

Adapted from Menascé & Almeida.

Burstiness Modeling: an example

- Let us consider the number of epochs n=21

- Each epoch has a duration of τ / n = 31 /21 = 1.48
- The average arrival rate λ = 19/31 = 0.613 req./sec
- The number of arrivals in each of the 21 epochs are: 1, 0, 3, 0, 4, 0, 1, 0, 4, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 4
- Thus, λ_1 = 1/1.48 = 0.676, that exceeds the avg. λ = 0.613
- In 8 of the 21 epochs, λ_k exceeds λ
- b = 8 / 21 = 0.381
- a = Arr^+ / (b*L) = 19 / (0.381 * 19) = 2.625

Adapted from Menascé & Almeida.

The Impact of Burstiness

- As shown in some studies, the maximum throughput of a Web server decreases as the burstiness factors increase.

- How can we represent in performance models the effects of burstiness?

- We know that the maximum throughput is equal to the inverse of the maximum service demand or the service demand of the bottleneck resource.

The Impact of Burstiness

- To account for the burstiness effect, we write the service demand of the bottleneck resource as:
  - $D = D_f + \alpha \times b$
  - D_f is the portion of the service demand that does not depend on burstiness
  - α is a factor used to inflate the service demand according to burstiness factor b. It is given by:
  - $\alpha = (U_1/X^1_0 - U_2/X^2_0)/(b_1-b_2)$
  - The measurement interval is divided into 2 subintervals $\Im_1$ and $\Im_2$ to obtain U_i, X^i_0, and b_i

Adapted from Menascé & Almeida.

The Impact of Burstiness: an example

- Consider the HTTP LOG of the previous slides. During 31 sec in which the 19 requests arrived, the CPU was found to be the bottleneck. What is the burstiness adjustment that should be applied to the CPU service demand to account for the burstiness effect on the performance of the Web server?

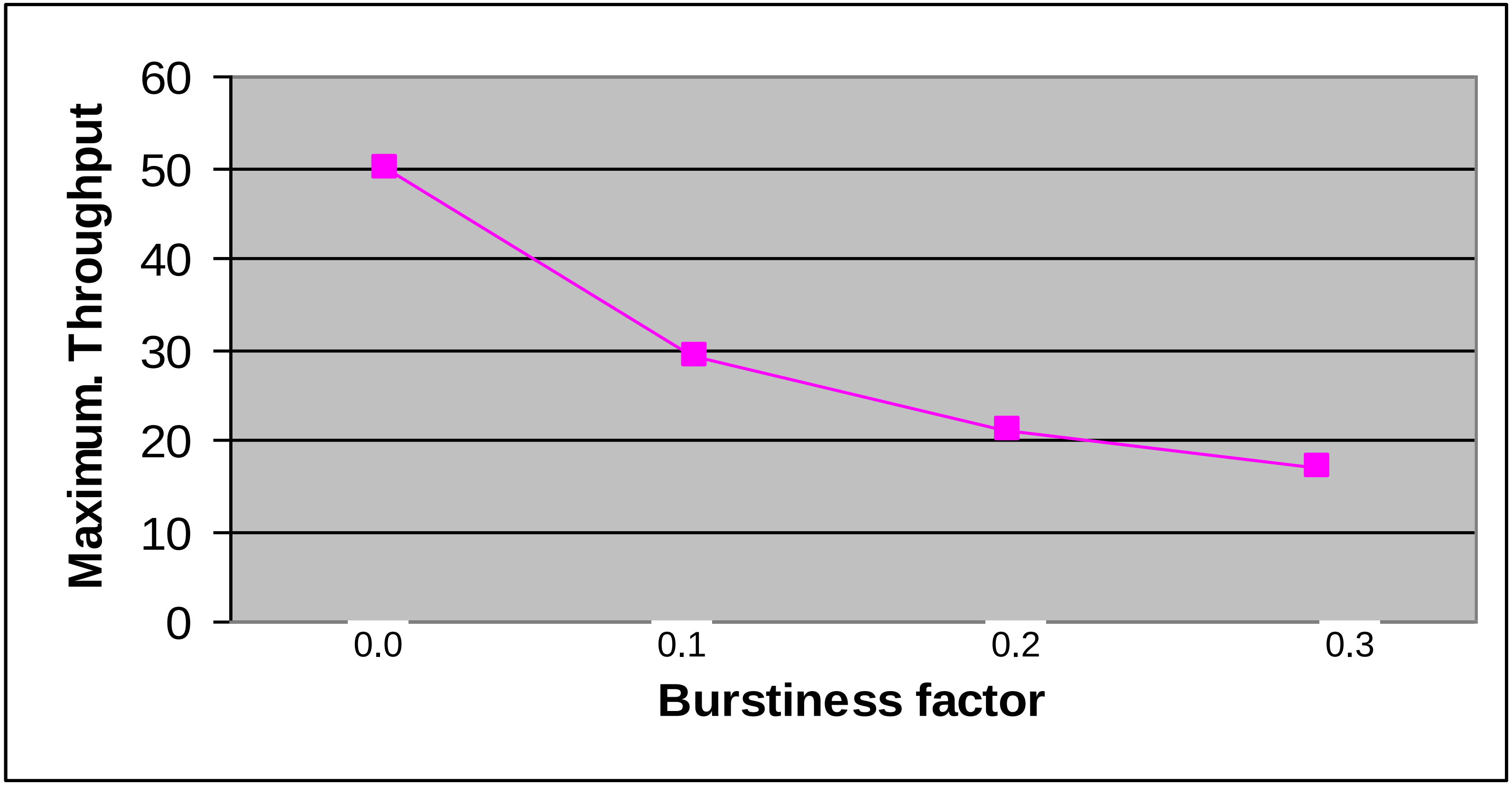
- The number of requests during each 15.5 sec subinterval is 14 and 5, respectively.

- The measured CPU utilization in each interval was 0.18 and 0.06

The Impact of Burstiness: an example (2)

- The throughput in each interval is:
  - $X^1_0 = 14/15.5 = 0.903$
  - $X^2_0 = 5/15.5 = 0.323$
- Using the previous algorithm:
  - $b_1 = 0.273$, $b_2 = 0.182$
  - $\alpha = (0.18/0.903 - 0.06/0.323)/(0.273\text{-}0.182) = 0.149$
  - the adjustment factor is: $\alpha \times b = 0.149 \times 0.381 = 0.057$
- Assuming Df = 0.02 sec, we are able to calculate the maximum server throughput as a function of the burstiness factor (b).

The Impact of Burstiness: an example (2)



Adapted from Menascé & Almeida.

Incorporating New Phenomena in the Workload Characterization

Accounting for Heavy Tails in the Model

- Due to the large variability of the size of documents, average results for the whole population would have very little statistical meaning.

- Categorizing the requests into a number of classes, defined by ranges of document sizes, improves the accuracy and significance of performance metrics.

- Multiclass queuing network models, with classes associated with requests for docs of different size.

Adapted from Menascé & Almeida.

Accounting for Heavy Tails: an example (1)

- The HTTP LOG of a Web server was analyzed during 1 hour. A total of 21,600 requests were successfully processed during the interval.
- Let us use a multiclass model to represent the server.
- There are 5 classes in the model, each corresponding to the 5 file size ranges.

Accounting for Heavy Tails: an example (2)

- File Size Distributions.

Class	File Size Range (KB)	Percent of Requests
1	Size < 5	25
2	5 ≤ size ≤ 50	40
3	50 ≤ size ≤ 100	20
4	100 ≤ size ≤ 500	10
5	size ≥ 500	5

Accounting for Heavy Tails: an example (3)

- The arrival rate for each class r is a fraction of the overall arrival rate $\lambda = 21{,}600/3{,}600 = 6$ requests/sec.
  - $\lambda_1 = 6 \times 0.25 = 1.5$ req./sec
  - $\lambda_2 = 6 \times 0.40 = 2.4$ req./sec
  - $\lambda_3 = 6 \times 0.20 = 1.2$ req./sec
  - $\lambda_4 = 6 \times 0.10 = 0.6$ req./sec
  - $\lambda_5 = 6 \times 0.05 = 0.3$ req./sec

Adapted from Menascé & Almeida.

Summary

- Workload Characterization
  - what is it?
  - basic concepts
  - workload description and modeling
  - representativeness of a workload model
- Methodology (1)
  - Choice of an analysis standpoint
  - Identification of the basic component
  - Choice of the characterizing parameters
  - Data collection

Adapted from Menascé & Almeida.

Summary

- Methodology (2)
  - Partitioning the workload
  - Calculating the class parameters
    - Averaging
    - Clustering techniques and algorithms
- New Phenomena in the Internet and WWW
  - Burstiness
  - Heavy-tailed distributions

Adapted from Menascé & Almeida.