From Systems to Descriptive Models

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

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.
Modeling

- Abstraction of a real system.
- Should capture enough details to satisfy goals of the study.
- Types of models:
  - Simulation
  - Analytic
  - Hybrid

Graphical Notation for Queues

(a)

(b)
Simple Database Server Example: Open Workload

In order to compute the response time, we need to know the total time spent by a transaction at each resource (CPU and Disk): service demand.
Multiclass DB Example

<table>
<thead>
<tr>
<th>transaction group</th>
<th>percent of total</th>
<th>avg. CPU time (sec)</th>
<th>avg. no. I/Os</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trivial</td>
<td>45%</td>
<td>0.04</td>
<td>5.5</td>
</tr>
<tr>
<td>Medium</td>
<td>25%</td>
<td>0.18</td>
<td>28.9</td>
</tr>
<tr>
<td>Complex</td>
<td>30%</td>
<td>1.20</td>
<td>85</td>
</tr>
</tbody>
</table>

Each transaction group is assigned to a customer class in the Queuing Network.

When do we need multiple classes?

- Heterogeneous service demands.
- Different types of workloads.
- Different service level objectives.
Open Class

• Workload intensity specified by an arrival rate (usually independent of the system state).
• Unbounded number of customers in the system.
• Throughput is an input parameter, which is equal to the arrival rate in equilibrium.

Closed Class

• Workload intensity specified by customer population (i.e., concurrency level)
• Bounded and known number of customers in the system.
• Throughput is an output parameter.
Closed Workload Example

Mixed Workload Example
Service Level Agreements (SLA)

- Agreement between users and providers of computing services on the levels of various performance metrics.
  - 99.99% availability during 8:00am-11:00pm period and 99.9% at other times
  - Less than 4 sec page download time for requests over non-secure connections less than 6 sec for requests over secure connections
  - Minimum throughput of 2,000 page downloads/sec.

Different Types of Resources
Different Types of Resources

Delay resource

Load-dependent resource

Load-independent resource

Blocking (Admission Control)
System with Admission Control

Throughput = Arrival Rate \times (1 - \text{Probability of Rejection})

Database Server with Software Contention
Effect of Software Contention

![Graph showing response time vs. number of threads]

A DB Server with Lock Contention

![Diagram of a DB server with lock contention]

© 2004 D. A. Menascé. All Rights Reserved.
Simultaneous Resource Possession Time Axes

Class Switching

Authentication can be significantly more CPU-intensive than DB access.
Queuing Disciplines

- First Come First Served (FCFS)
- Priority Queuing (FCFS breaks the tie):
  - Non-preemptive
  - Preemptive resume
  - Preemptive repeat.
- Round-Robin (RR)
- Processor Sharing (PS)
- Last Come First Served-Preemptive Resume (LCFS-PR)
- Shortest Job First (SJF)

QN Models

- K queues \((i = 1, \ldots, K)\)
- R classes \((r = 1, \ldots, R)\)
- Input parameters:
  - Workload intensity
    \[ \tilde{\lambda} = (\lambda_1, \ldots, \lambda_r, \ldots, \lambda_R) \]
    for open classes
  - \[ \tilde{N} = (N_1, \ldots, N_r, \ldots, N_R) \]
    for closed classes
  - Service demands \((D_{ir})\)
Capacity Planning Input and Output Variables

Workload Evolution  System Parameters  Desired Service Levels

Capacity Planning

Saturation Points  Cost-effective Alternatives

Capacity Planning Input and Output Variables

• intensity
• nature

Workload Evolution  System Parameters  Desired Service Levels

Capacity Planning

Saturation Points  Cost-effective Alternatives
Capacity Planning Input and Output Variables

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

Workload Evolution

System Parameters

Desired Service Levels

Saturation Points

Cost-effective Alternatives

**Capacity Planning**

when are service levels violated?
Capacity Planning Input and Output Variables

- **Workload Evolution**
  - intensity
  - nature

- **System Parameters**
  - processors, disks, networks
  - max. no. connections

- **Desired Service Levels**
  - $RT <= 2$ sec
  - $tput > 10$ tps

- **Cost-effective Alternatives**
  - e.g., $tp/s$

**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.
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?