CS 672
Component Level Performance Models of Computer Systems

Dr. Daniel A. Menascé

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

Department of Computer Science

George Mason University

Outline

❑Component-level Models

❑Computing Service Demands

❑Open Queuing Models

❑Closed Queuing Models

❑Examples

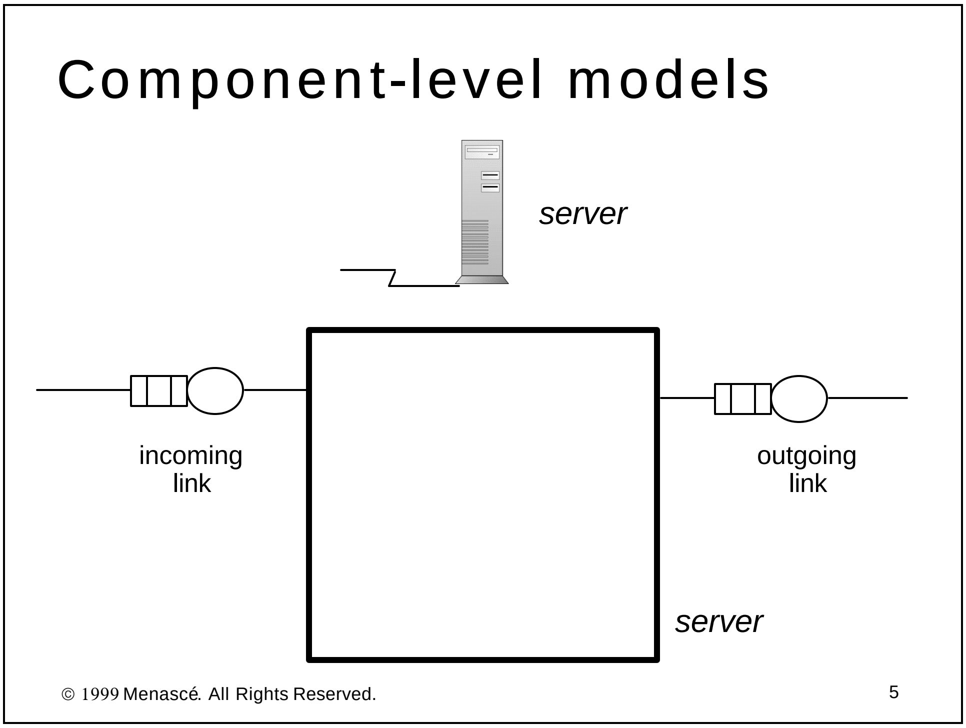
❑Models of E-commerce Servers

Outline

- □Component-level Models
- □Computing Service Demands
- □Open Queuing Models
- □Closed Queuing Models
- □Examples
- □Models of E-commerce Servers

Component-level Models

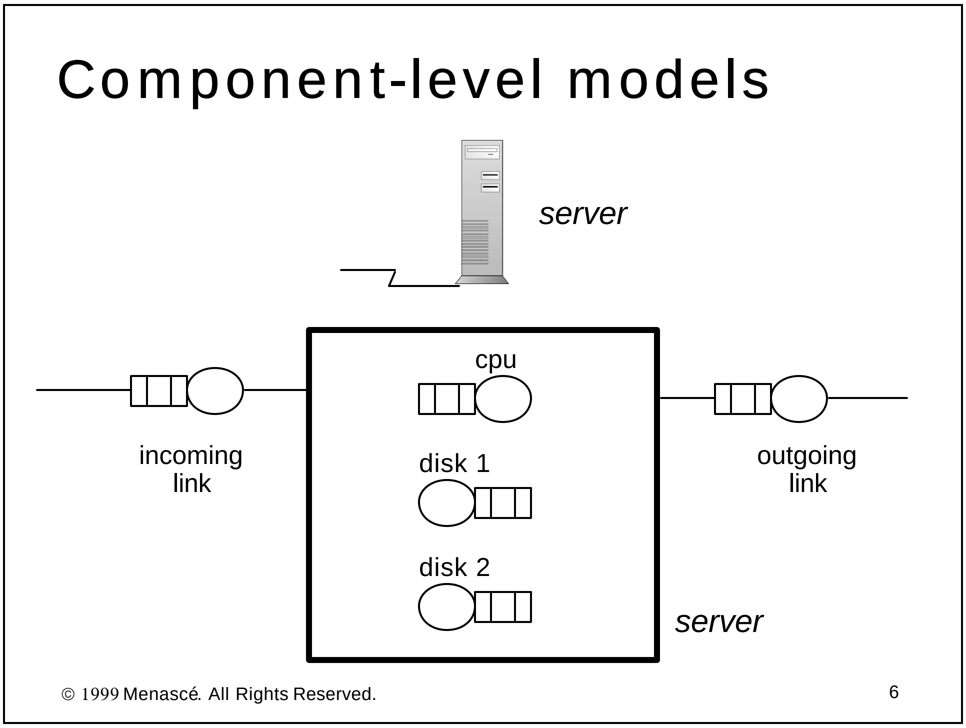
- □The internal components of a server (e.g., processors, disks) as well as network links are modeled explicitly.
- □Changes in server architecture, component upgrades (e.g., use of a faster CPU or faster network connections) can be evaluated with component-level models.



Component-level models

© 1999 Menascé. All Rights Reserved.

5

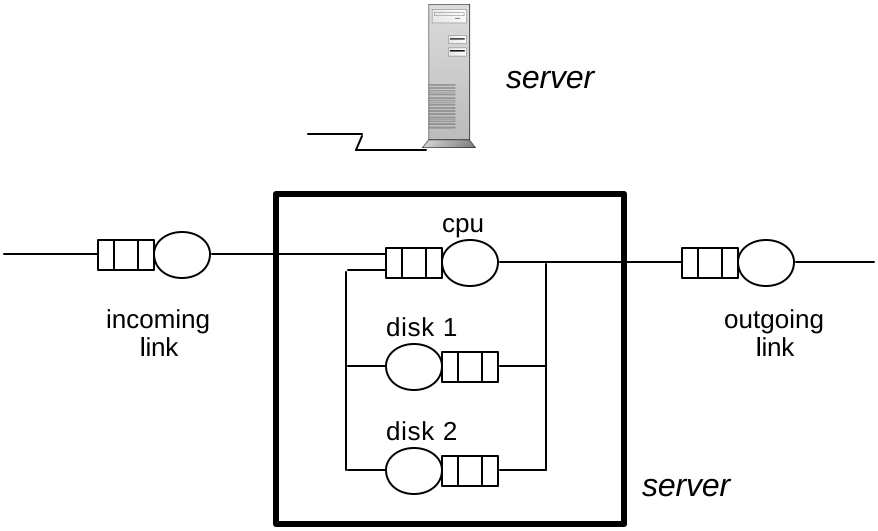


Component-level models

© 1999 Menascé. All Rights Reserved.

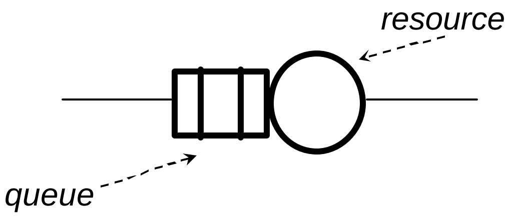
6

Component-level models



Component-level Models

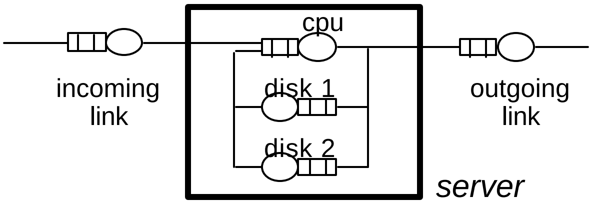
❑Each component is represented by a resource (e.g. CPU, disk, communication link) and a queue of requests waiting for the resource.



Outline

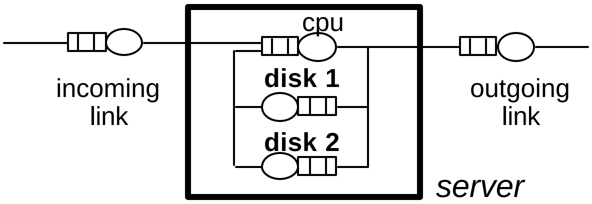
- ❑Component-level Models
- ❑Computing Service Demands
- ❑Open Queuing Models
- ❑Closed Queuing Models
- ❑Examples
- ❑Models of E-commerce Servers

Computing Service Demands



- The average size of a file retrieved per request is 20KBytes.
- The average disk service time per KByte accessed is 10 msec.
- 40% of the files are on disk 1 and 60% on disk 2.
- The speed of the link connecting the server to the Internet is 1.5 Mbps (a T1 link).
- The CPU processing time per request is 2 msec + 0.05 msec per KByte accessed.
- The average size of an HTTP request is 200 bytes.

Disk Service Demand

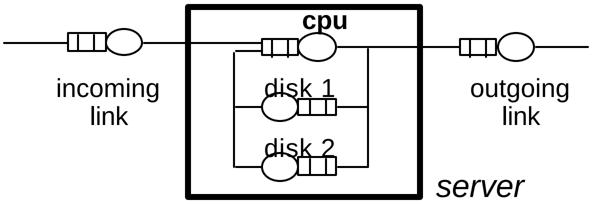


- The average size of a file retrieved per request is 20KBytes.
- The average disk service time per KByte accessed is 10 msec.
- 40% of the files are on disk 1 and 60% on disk 2.

D_{disk1}= 0.4*10 msec/Kbyte * 20KBytes = 80 msec = 0.080 sec

D_{disk2}= 0.6*10 msec/Kbyte * 20KBytes = 120 msec = 0.120 sec

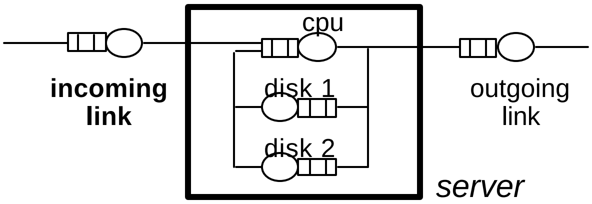
CPU Service Demand



- The average size of a file retrieved per request is 20KBytes.
- The CPU processing time per request is 2 msec + 0.05 msec per KByte accessed.

D_{cpu} = 2 msec + 0.05 msec/KByte * 20 Kbyte
= 3 msec = 0.003 sec

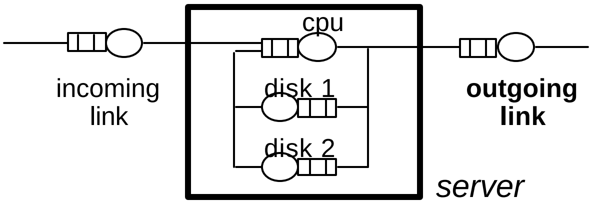
Incoming Link Service Demand



- The speed of the link connecting the server to the Internet is 1.5 Mbps (a T1 link).
- The average size of an HTTP request is 200 bytes.

$D_{IncLink}$ = 200 * 8 bits / 1,500,000 bps = 0.00107 sec

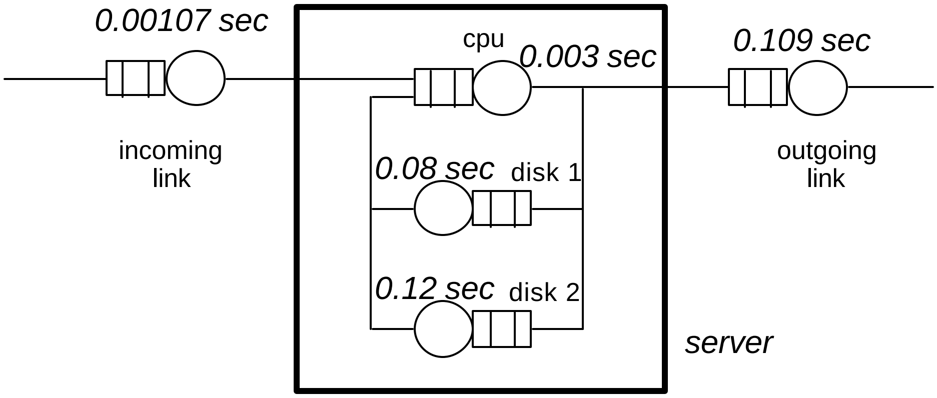
Outgoing Link Service Demand



- The average size of a file retrieved per request is 20KBytes.
- The speed of the link connecting the server to the Internet is 1.5 Mbps (a T1 link).

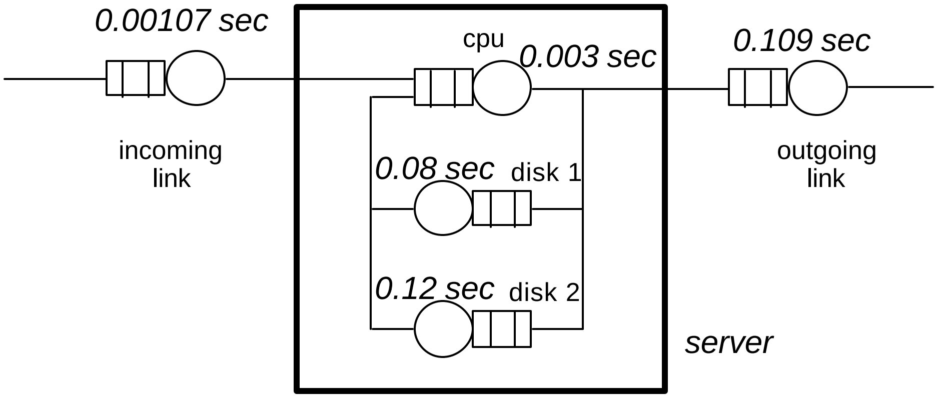
$D_{OutLink}$ = 20 * 1024 * 8 bits / 1,500,000 bps = 0.109 sec

Computing Service Demands



Service demands do not include any queuing time! It is just service time.

Computing Waiting Times



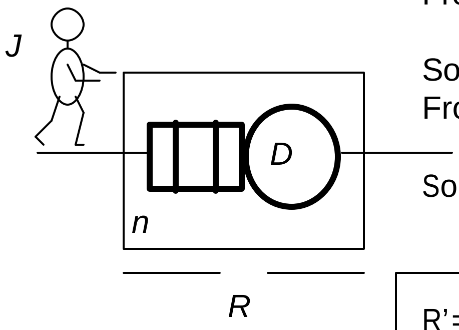
Waiting times depend on the load (arrival rate of requests) and on the service demands.

Outline

- ❑Component-level Models
- ❑Computing Service Demands
- ❑Open Queuing Models
- ❑Closed Queuing Models
- ❑Examples
- ❑Models of E-commerce Servers

Computing Residence Times

- *n transactions seen, on average, by arriving request J.*
- *Each of the n requests found by J need S sec of total service.*
- *So, J has to wait for n * S seconds before being served.*



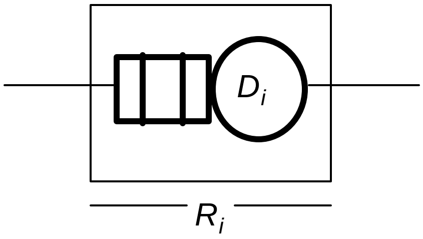
R = S + n*S
From Little's Law: n = R * l

So, R = S + R * l * S
From the Utilization Law: U = l * S

So, R = S + R * U. Then,
R = S / (1 – U)

R'= V*S/(1-U)=D/(1-U)

Computing Residence Times



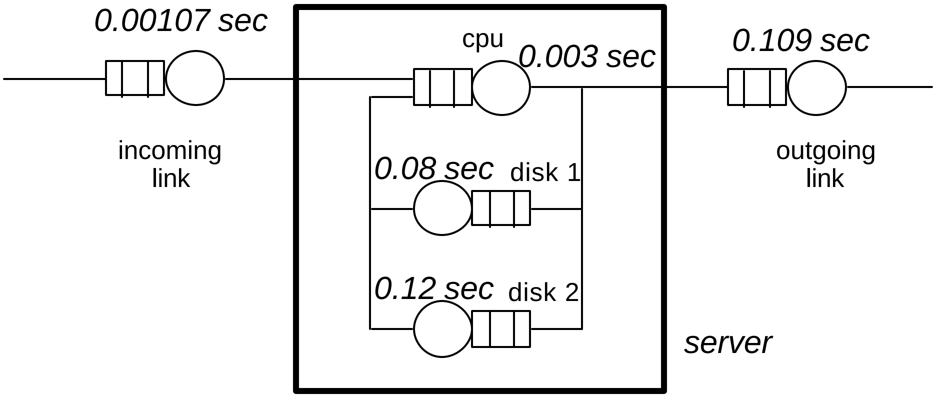
Service demand at resource i

$$R_i' = \frac{D_i}{\underbrace{1-\lambda D_i}}$$

Residence time at resource i

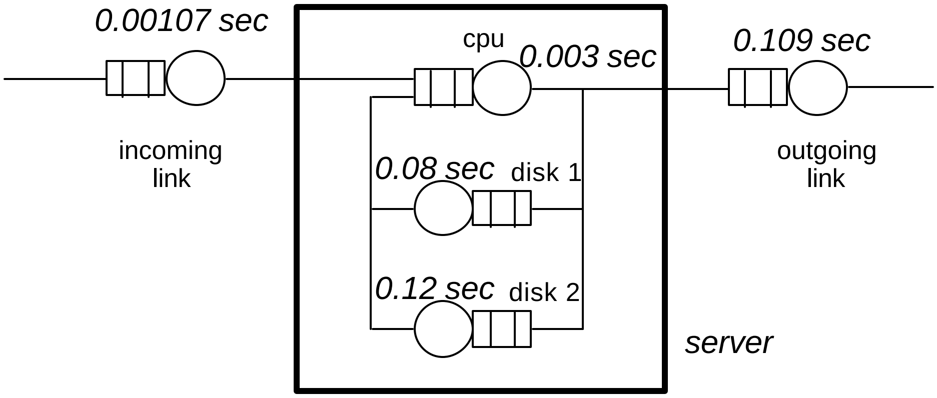
Utilization of resource i (U_i)

Service Demands



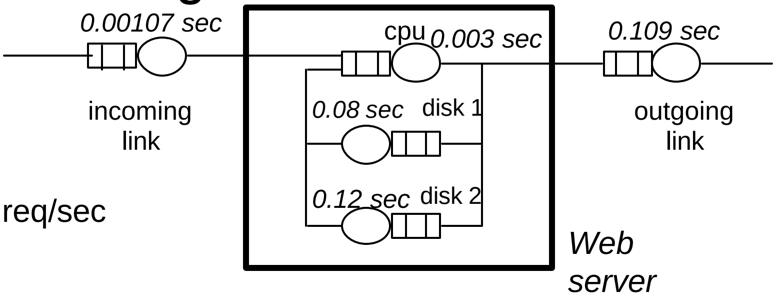
Service demands do not include any queuing time! It is just service time.

Computing Waiting Times



Waiting times depend on the load (arrival rate of requests) and on the service demands.

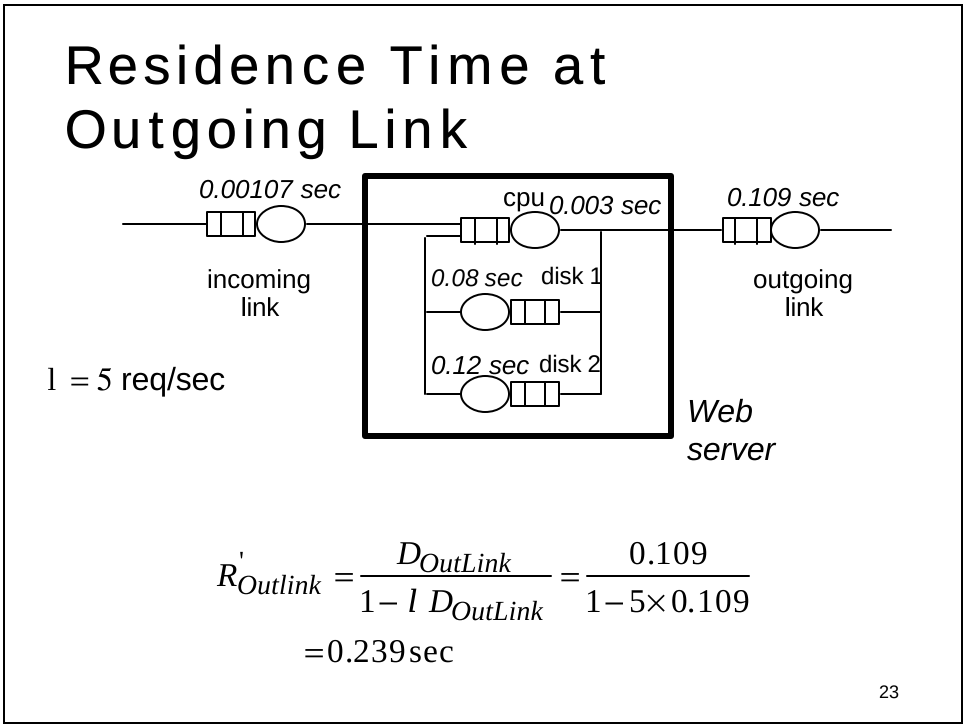
Residence Time at Incoming Link



$l = 5$ req/sec

$$R'_{IncLink} = \frac{D_{IncLink}}{1 - l\, D_{IncLink}} = \frac{0.00107}{1 - 5 \times 0.00107}$$

$$= 0.00108\,\text{sec}$$



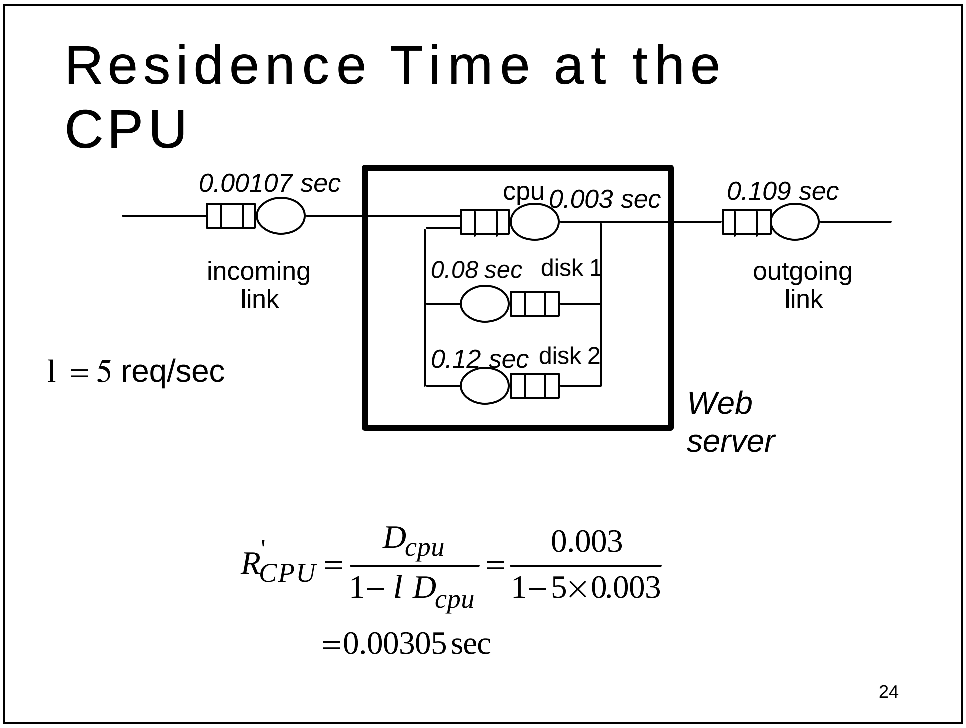
Residence Time at Outgoing Link

l = 5 req/sec

$$R'_{Outlink} = \frac{D_{OutLink}}{1 - l\,D_{OutLink}} = \frac{0.109}{1 - 5 \times 0.109}$$

$$= 0.239 \sec$$

23



Residence Time at the CPU

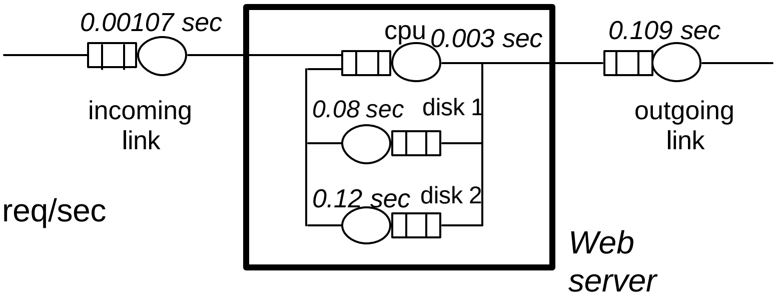
l = 5 req/sec

$$R'_{CPU} = \frac{D_{cpu}}{1 - l\,D_{cpu}} = \frac{0.003}{1 - 5 \times 0.003}$$

$$= 0.00305 \sec$$

24

Residence Time at Disk 1

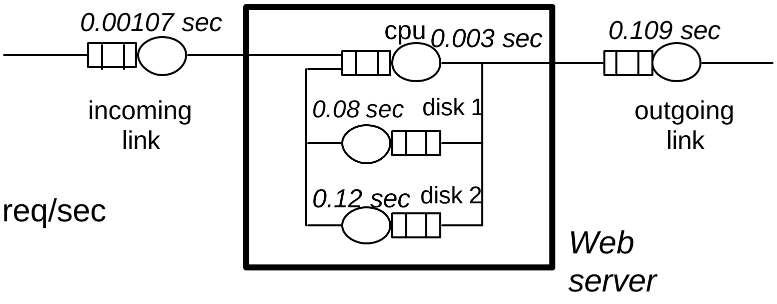


$l = 5$ req/sec

$$R'_{disk1} = \frac{D_{disk1}}{1 - l\ D_{disk1}} = \frac{0.08}{1 - 5 \times 0.08}$$

$$= 0.133 \text{sec}$$

Residence Time at Disk 2



$l = 5$ req/sec

$$R'_{disk\,2} = \frac{D_{disk\,2}}{1 - l\ D_{disk\,2}} = \frac{0.12}{1 - 5 \times 0.12}$$

$$= 0.3 \text{ sec}$$

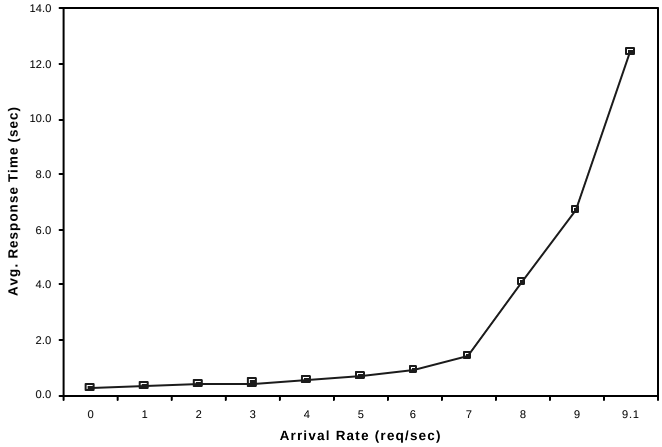
Summary of Results

Resource	Service Demand (sec)	Utilization	Residence Time (sec)
Inc. Link	0.00107	0.54%	0.00108
CPU	0.00300	1.50%	0.00305
Disk 1	0.08000	40.00%	0.13333
Disk 2	0.12000	60.00%	0.30000
Out. Link	0.10900	54.50%	0.23956
	0.31307		0.67702

Sum of service demands

Average Response Time

Response Time vs. Arrival Rate



Open vs. Closed QN Models

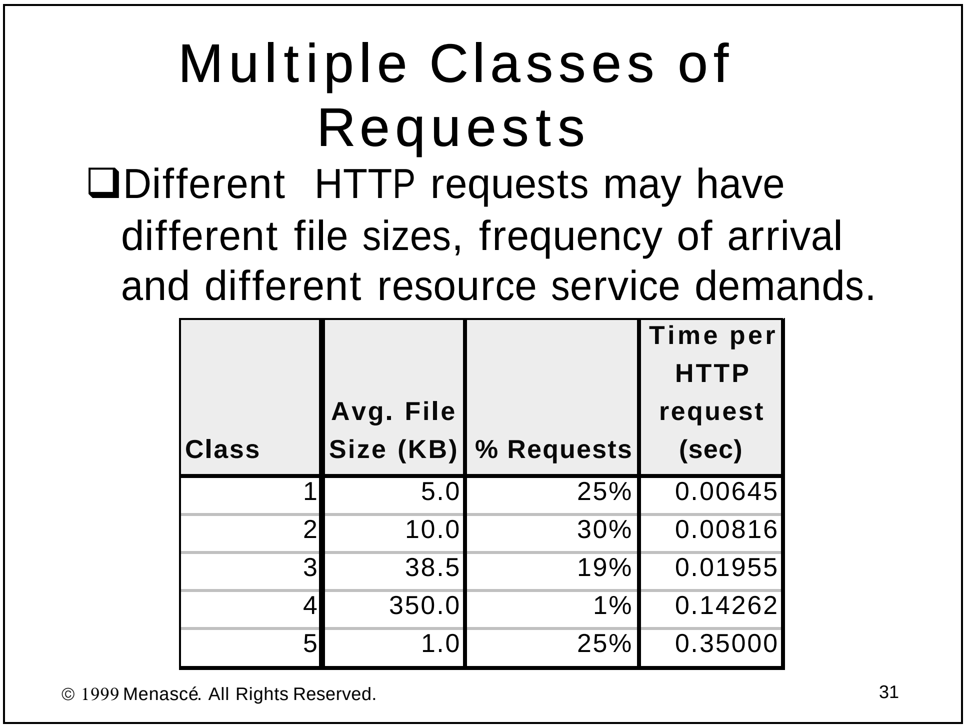
- ❑The models presented so far are open QN models because there is no limit on the number of requests in the system.
- ❑When the number of requests in the system is limited, we need closed QN models.
  - e.g., servers with limited degree of MPL.
  - C/S systems with known number of clients.

Open Model Equations

$$U_i = l \times D_i$$

$$R_i = \frac{D_i}{1 - U_i}$$

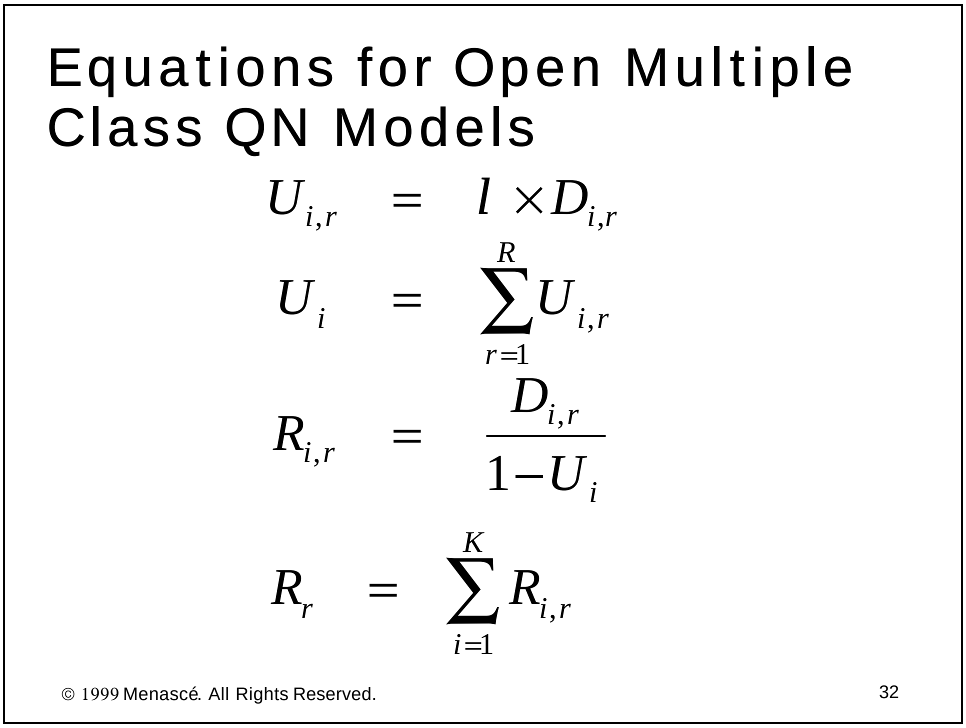
$$U_i < 1 \text{ for all } i$$



Multiple Classes of Requests

- ❑Different HTTP requests may have different file sizes, frequency of arrival and different resource service demands.

Class	Avg. File Size (KB)	% Requests	Time per HTTP request (sec)
1	5.0	25%	0.00645
2	10.0	30%	0.00816
3	38.5	19%	0.01955
4	350.0	1%	0.14262
5	1.0	25%	0.35000



Equations for Open Multiple Class QN Models

$$U_{i,r} = l \times D_{i,r}$$

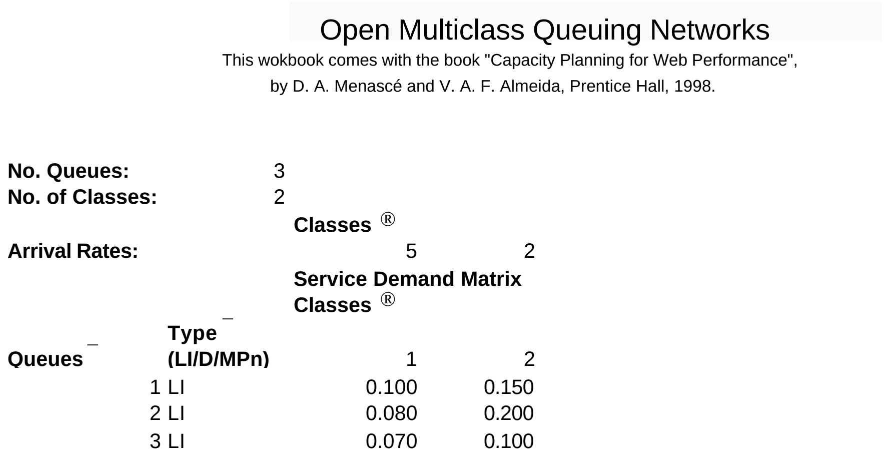
$$U_i = \sum_{r=1}^{R} U_{i,r}$$

$$R_{i,r} = \frac{D_{i,r}}{1 - U_i}$$

$$R_r = \sum_{i=1}^{K} R_{i,r}$$

Multiclass Example

- ❑A Web server has one CPU and two disks. It receives two types of HTTP requests: for small text files and for large images. The average arrival rates are 5 requests/sec for text and 2 requests/sec for images. What is the response times for each type of request?

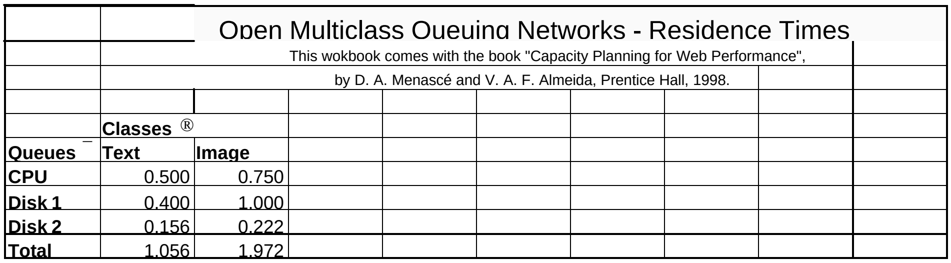


Open Multiclass Queuing Networks

This wokbook comes with the book "Capacity Planning for Web Performance", by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

No. Queues: 3
No. of Classes: 2

		Classes ®	
Arrival Rates:		5	2
		Service Demand Matrix **Classes ®**	
Queues _	**Type _** **(LI/D/MPn)**	1	2
1	LI	0.100	0.150
2	LI	0.080	0.200
3	LI	0.070	0.100

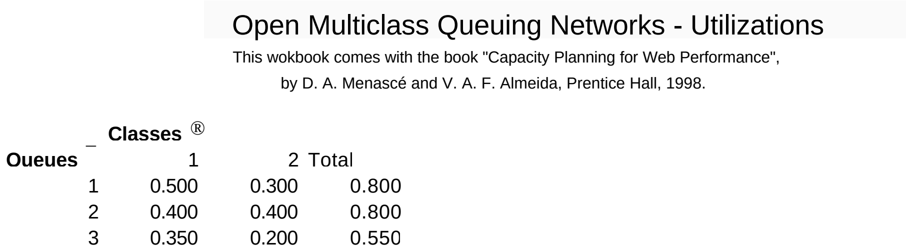


Open Multiclass Queuing Networks - Residence Times

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®	
Queues _	Text	Image
CPU	0.500	0.750
Disk 1	0.400	1.000
Disk 2	0.156	0.222
Total	1.056	1.972



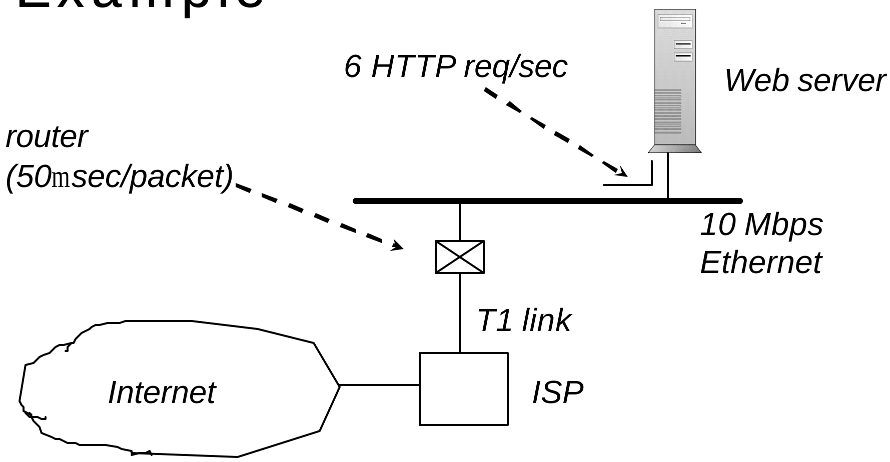
Open Multiclass Queuing Networks - Utilizations

This wokbook comes with the book "Capacity Planning for Web Performance",

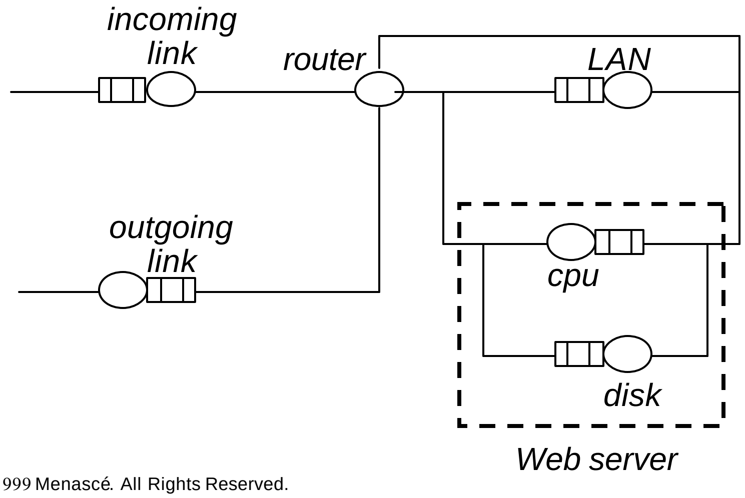
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

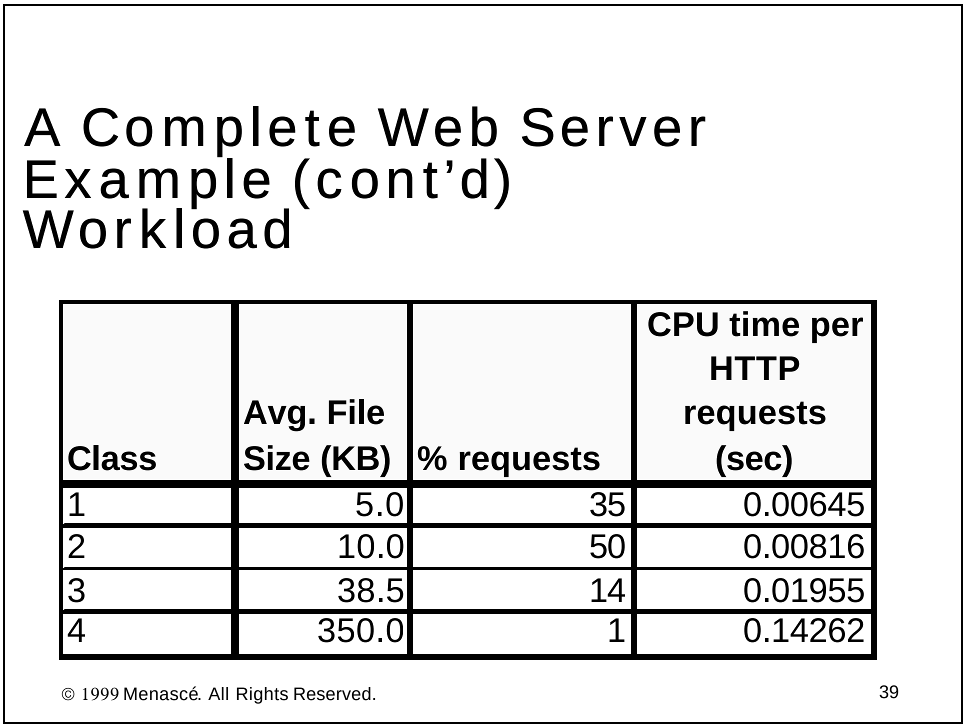
	Classes ®		
Oueues _	1	2	Total
1	0.500	0.300	0.800
2	0.400	0.400	0.800
3	0.350	0.200	0.550

A Complete Web Server Example



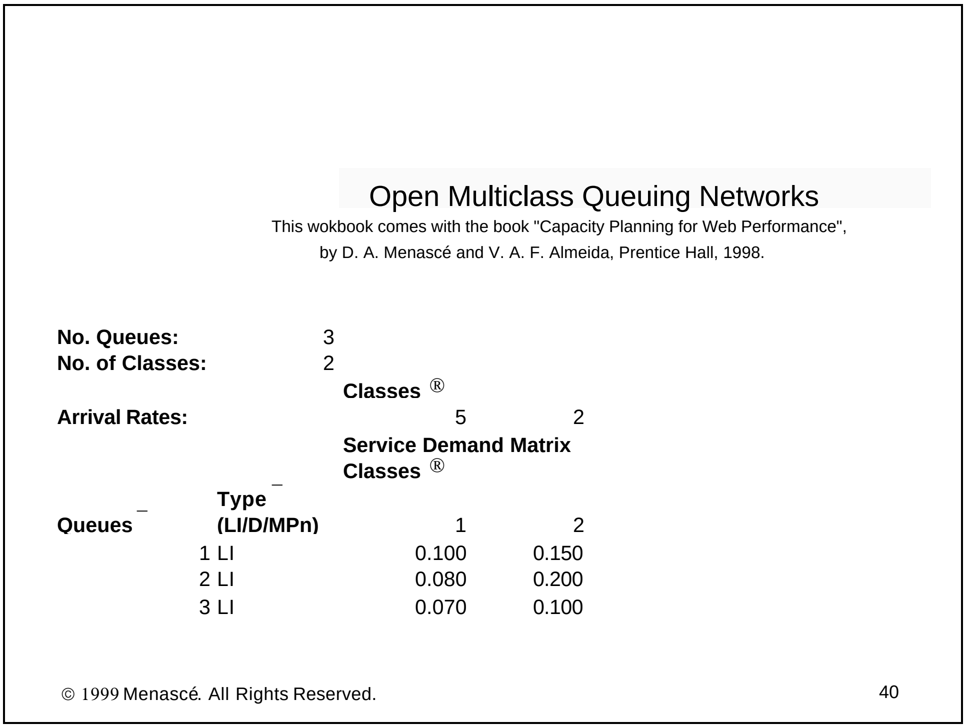
A Complete Web Server Example (cont'd)





A Complete Web Server Example (cont'd) Workload

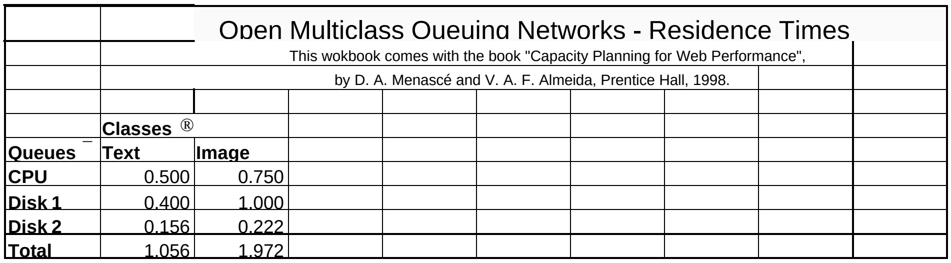
Class	Avg. File Size (KB)	% requests	CPU time per HTTP requests (sec)
1	5.0	35	0.00645
2	10.0	50	0.00816
3	38.5	14	0.01955
4	350.0	1	0.14262



Open Multiclass Queuing Networks

This wokbook comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

No. Queues:		3		
No. of Classes:		2		
		Classes ®		
Arrival Rates:			5	2
		Service Demand Matrix **Classes ®**		
Queues ¯		**Type ¯** **(LI/D/MPn)**	1	2
	1	LI	0.100	0.150
	2	LI	0.080	0.200
	3	LI	0.070	0.100

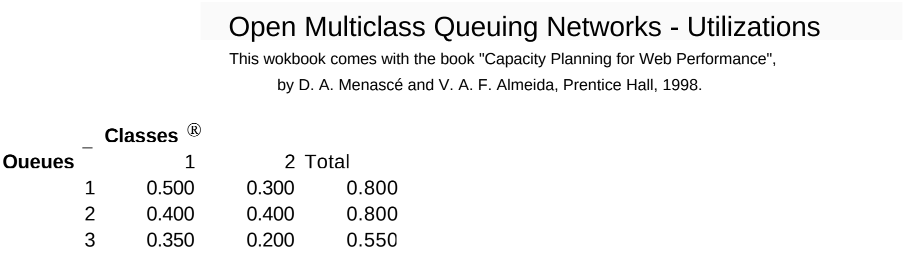


Open Multiclass Queuing Networks - Residence Times

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®	
Queues _	Text	Image
CPU	0.500	0.750
Disk 1	0.400	1.000
Disk 2	0.156	0.222
Total	1.056	1.972



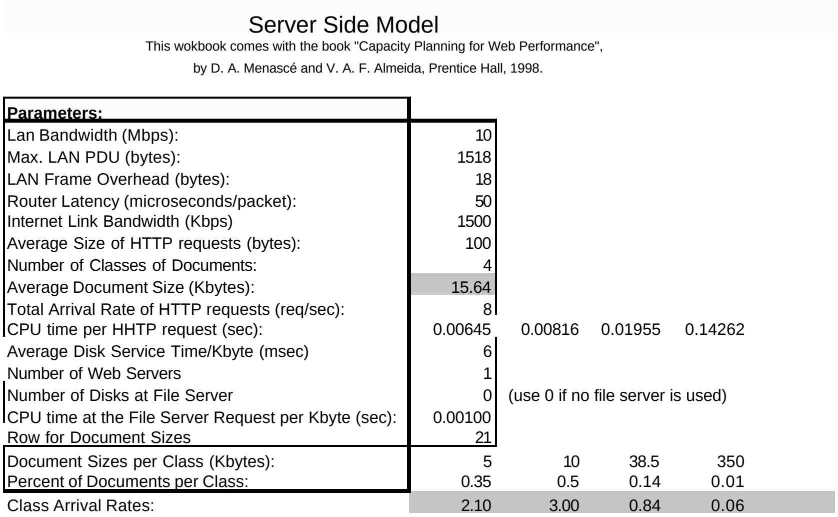
Open Multiclass Queuing Networks - Utilizations

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®		
Oueues _	1	2	Total
1	0.500	0.300	0.800
2	0.400	0.400	0.800
3	0.350	0.200	0.550

A Web Server Example (cont’d)



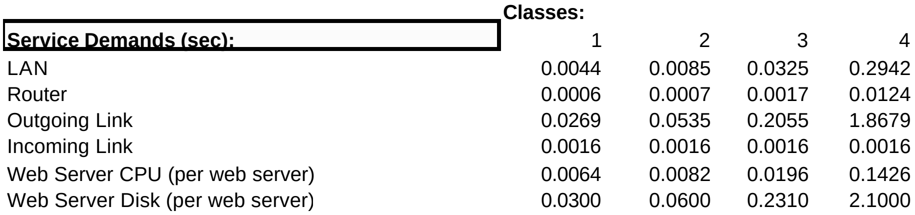
Server Side Model

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

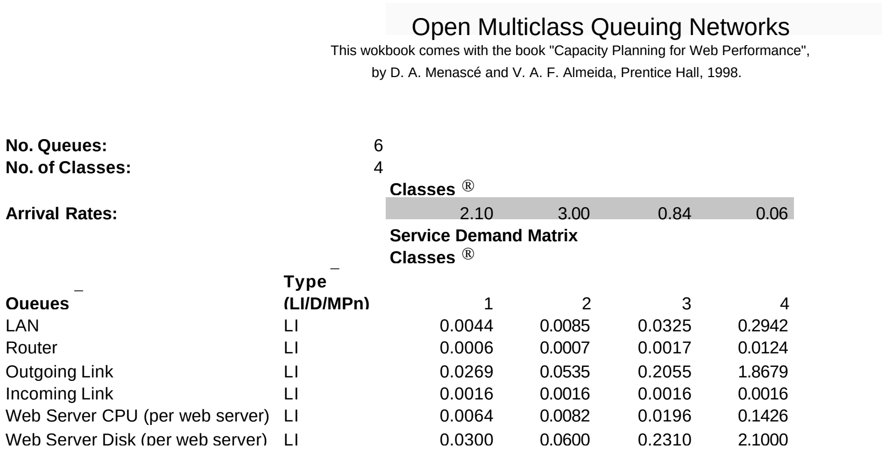
Parameters:					
Lan Bandwidth (Mbps):	10				
Max. LAN PDU (bytes):	1518				
LAN Frame Overhead (bytes):	18				
Router Latency (microseconds/packet):	50				
Internet Link Bandwidth (Kbps)	1500				
Average Size of HTTP requests (bytes):	100				
Number of Classes of Documents:	4				
Average Document Size (Kbytes):	15.64				
Total Arrival Rate of HTTP requests (req/sec):	8				
CPU time per HHTP request (sec):	0.00645	0.00816	0.01955	0.14262	
Average Disk Service Time/Kbyte (msec)	6				
Number of Web Servers	1				
Number of Disks at File Server	0	(use 0 if no file server is used)			
CPU time at the File Server Request per Kbyte (sec):	0.00100				
Row for Document Sizes	21				
Document Sizes per Class (Kbytes):	5	10	38.5	350	
Percent of Documents per Class:	0.35	0.5	0.14	0.01	
Class Arrival Rates:	2.10	3.00	0.84	0.06	

A Web Server Example (cont’d)



	Classes:			
Service Demands (sec):	1	2	3	4
LAN	0.0044	0.0085	0.0325	0.2942
Router	0.0006	0.0007	0.0017	0.0124
Outgoing Link	0.0269	0.0535	0.2055	1.8679
Incoming Link	0.0016	0.0016	0.0016	0.0016
Web Server CPU (per web server)	0.0064	0.0082	0.0196	0.1426
Web Server Disk (per web server)	0.0300	0.0600	0.2310	2.1000

A Web Server Example (cont’d)



Open Multiclass Queuing Networks

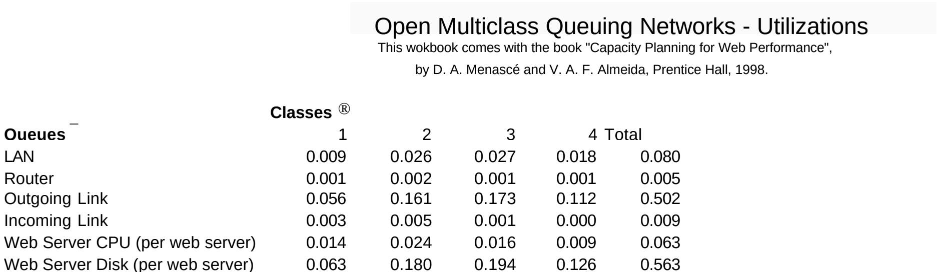
This wokbook comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

No. Queues: 6

No. of Classes: 4

		Classes ®			
Arrival Rates:		2.10	3.00	0.84	0.06
		Service Demand Matrix			
		Classes ®			
	Type				
Oueues	**(LI/D/MPn)**	1	2	3	4
LAN	LI	0.0044	0.0085	0.0325	0.2942
Router	LI	0.0006	0.0007	0.0017	0.0124
Outgoing Link	LI	0.0269	0.0535	0.2055	1.8679
Incoming Link	LI	0.0016	0.0016	0.0016	0.0016
Web Server CPU (per web server)	LI	0.0064	0.0082	0.0196	0.1426
Web Server Disk (per web server)	LI	0.0300	0.0600	0.2310	2.1000

A Web Server Example (cont’d)

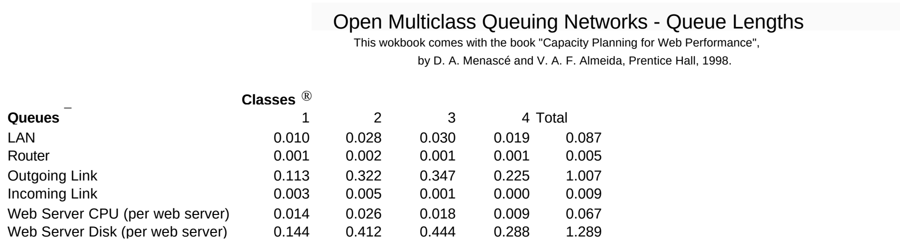


Open Multiclass Queuing Networks - Utilizations

This wokbook comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®				
Oueues	1	2	3	4	Total
LAN	0.009	0.026	0.027	0.018	0.080
Router	0.001	0.002	0.001	0.001	0.005
Outgoing Link	0.056	0.161	0.173	0.112	0.502
Incoming Link	0.003	0.005	0.001	0.000	0.009
Web Server CPU (per web server)	0.014	0.024	0.016	0.009	0.063
Web Server Disk (per web server)	0.063	0.180	0.194	0.126	0.563

A Web Server Example (cont'd)



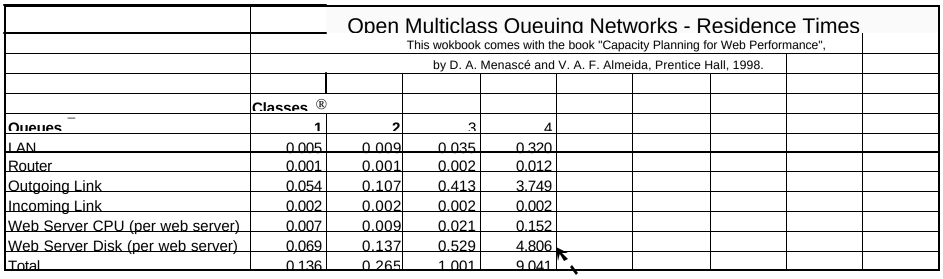
Open Multiclass Queuing Networks - Queue Lengths

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®				
Queues ¯	1	2	3	4	Total
LAN	0.010	0.028	0.030	0.019	0.087
Router	0.001	0.002	0.001	0.001	0.005
Outgoing Link	0.113	0.322	0.347	0.225	1.007
Incoming Link	0.003	0.005	0.001	0.000	0.009
Web Server CPU (per web server)	0.014	0.026	0.018	0.009	0.067
Web Server Disk (per web server)	0.144	0.412	0.444	0.288	1.289

A Web Server Example (cont'd)



Open Multiclass Queuing Networks - Residence Times

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®			
Queues ¯	**1**	**2**	3	4
LAN	0.005	0.009	0.035	0.320
Router	0.001	0.001	0.002	0.012
Outgoing Link	0.054	0.107	0.413	3.749
Incoming Link	0.002	0.002	0.002	0.002
Web Server CPU (per web server)	0.007	0.009	0.021	0.152
Web Server Disk (per web server)	0.069	0.137	0.529	4.806
Total	0.136	0.265	1.001	9.041

major contribution to response time

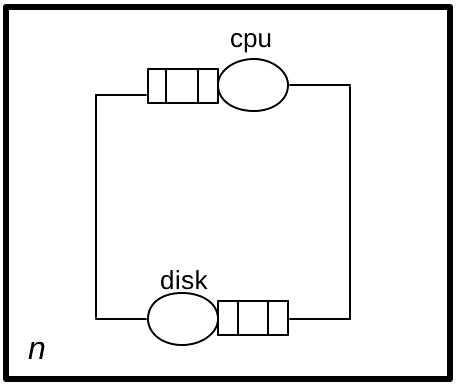
Outline

- Component-level Models
- Computing Service Demands
- Open Queuing Models
- Closed Queuing Models
- Examples
- Models of E-commerce Servers

Closed QN Model: example

- A Web server has one CPU and one disk. Assume that *5* requests are in execution concurrently. Each request takes 3 msec of CPU and 10 msec of disk time. What is the throughput and response time of the Web server?

Closed QN Model



Web server

Closed QN Model: Mean Value Analysis (MVA)

$$R_i\ (n) = S_i + S_i \times \overline{n}_i^{A}(n)$$

"My response time is equal to my service time plus my waiting time (i.e, the service time of all those who arrived ahead of me)."

Arrival theorem:

$$\overline{n}_i^{A}(n) = \overline{n}_i(n-1)$$

"I cannot find myself in the queue, thus the n-1."

Avg. # people I find in the queue.

Avg. # people in the queue.

So:

$$R_i\ (n) = S_i[1+\overline{n}_i\ (n-1)]$$

Notation:
(n) means "a function of n."

Closed QN Model: Mean Value Analysis (MVA)

But:

Avg. # visits *Avg. response time per visit*

$$R_i^{'}(n) = V_i R_i(n) = V_i S_i[1+\overline{n}_i\ (n-1)]$$

"The residence time is equal to the response time per visit times the average number of visits to resource i per transaction."

Finally, we get equation (1) of MVA: *Avg. service demand.*

$$\boxed{R_i^{'}(n) = D_i[1+\overline{n}_i\ (n-1)]}$$

Closed QN Model: Mean Value Analysis (MVA)

no. of resources

Applying Little's Law to the entire system:

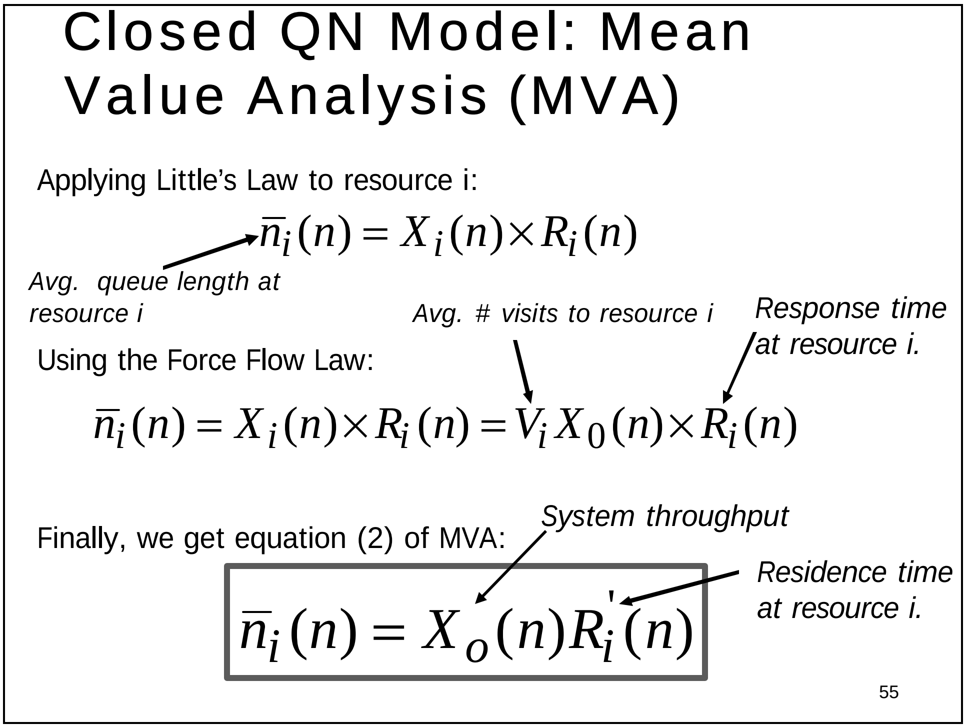
$$n = X_0(n)\times R_o(n) = X_0(n)\times \sum_{i=1}^{K} R_i^{'}(n)$$

Remember that the response time is the sum of all residence times?

Finally, we get equation (2) of MVA:

System throughput

$$\boxed{X_0(n) = n / \sum_{i=1}^{K} R_i^{'}(n)}$$



Closed QN Model: Mean Value Analysis (MVA)

Applying Little's Law to resource i:

$$\bar{n}_i(n) = X_i(n) \times R_i(n)$$

Avg. queue length at resource i

Using the Force Flow Law:

Avg. # visits to resource i

Response time at resource i.

$$\bar{n}_i(n) = X_i(n) \times R_i(n) = V_i X_0(n) \times R_i(n)$$

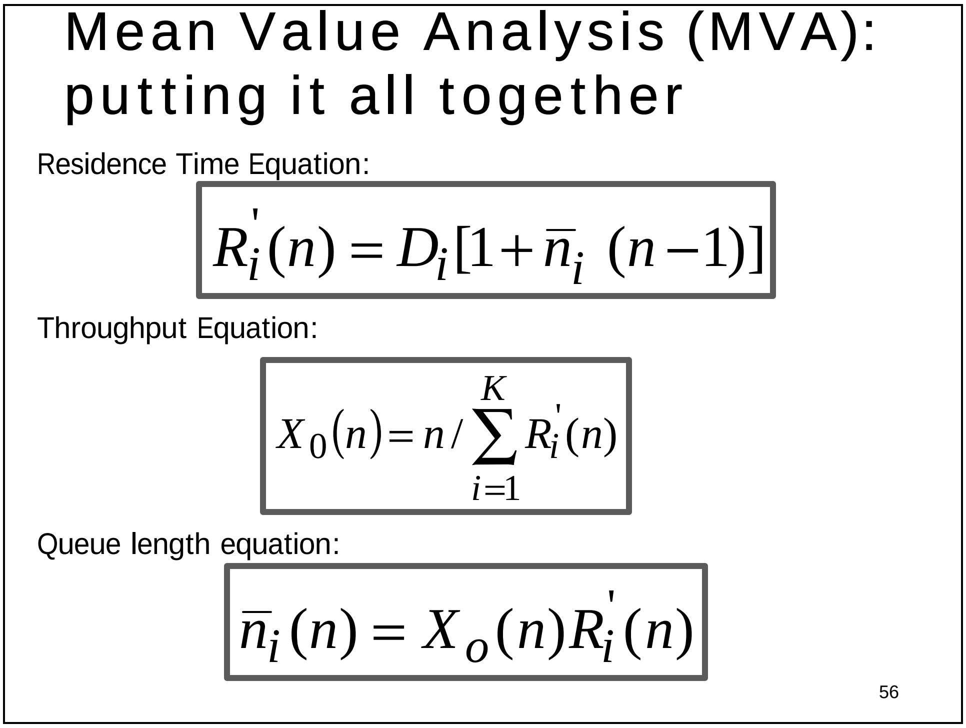
Finally, we get equation (2) of MVA:

System throughput

Residence time at resource i.

$$\boxed{\bar{n}_i(n) = X_o(n) R_i^{'}(n)}$$

55



Mean Value Analysis (MVA): putting it all together

Residence Time Equation:

$$\boxed{R_i^{'}(n) = D_i[1 + \bar{n}_i\ (n-1)]}$$

Throughput Equation:

$$\boxed{X_0(n) = n / \sum_{i=1}^{K} R_i^{'}(n)}$$

Queue length equation:

$$\boxed{\bar{n}_i(n) = X_o(n) R_i^{'}(n)}$$

56

MVA Example Revisited

- ❑$D_{\textbf{cpu}}$ = 3 msec; $D_{\textbf{disk}}$ = 10 msec.
- ❑n = 0, 1, 2, 3, 4 , 5
- ❑Look at the MVA equations and think how you would use them to solve the problem? [Hint: the queue length at all resources is 0 when n = 0]. In other words:

$$\overline{n}_i(0) = 0$$

57

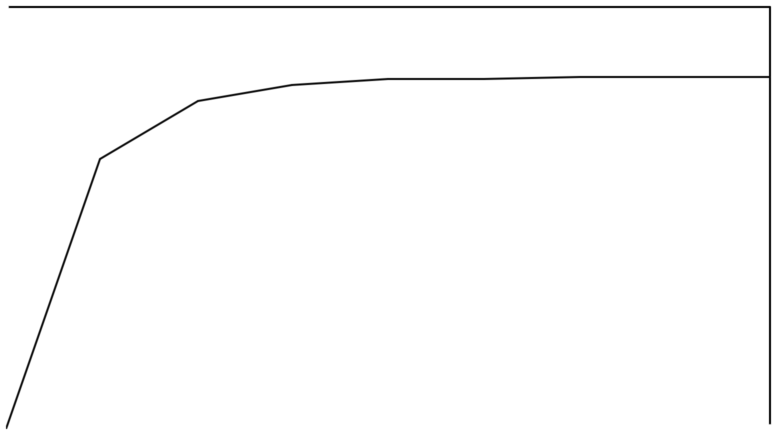
Solution to the Closed QN Model

n	Rcpu	Rdisk	Ro	Xo	ncpu	ndisk
0			0.000	0.000	0.000	0.000
1	3.000	10.000	13.000	0.077	0.231	0.769
2	3.692	17.692	21.385	0.094	0.345	1.655
3	4.036	26.547	30.583	0.098	0.396	2.604
4	4.188	36.041	40.229	0.099	0.416	3.584
5	4.249	45.836	50.085	0.100	0.424	4.576
6	4.273	55.758	60.031	0.100	0.427	5.573
7	4.281	65.730	70.011	0.100	0.428	6.572
8	4.284	75.720	80.004	0.100	0.428	7.572

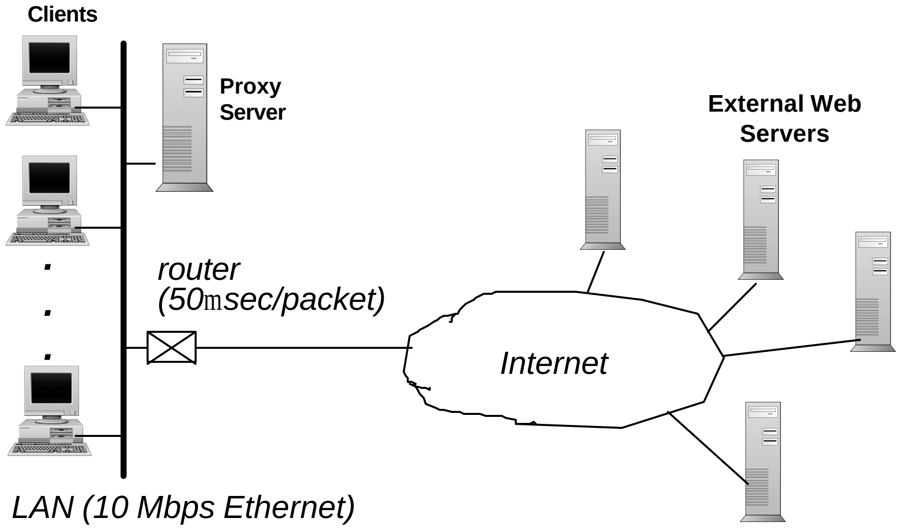
Can be easily computed using a spreadsheet!

58

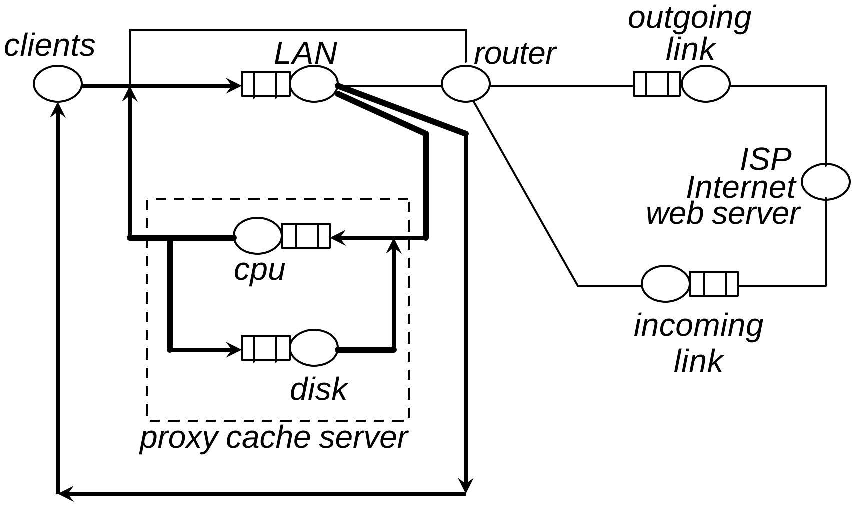
Solution to the Closed QN Model



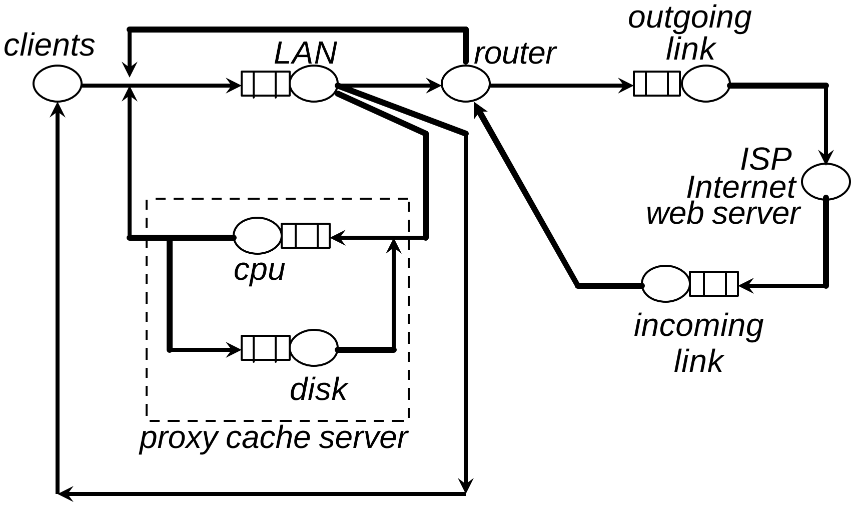
An Intranet Example with Proxy Cache Server

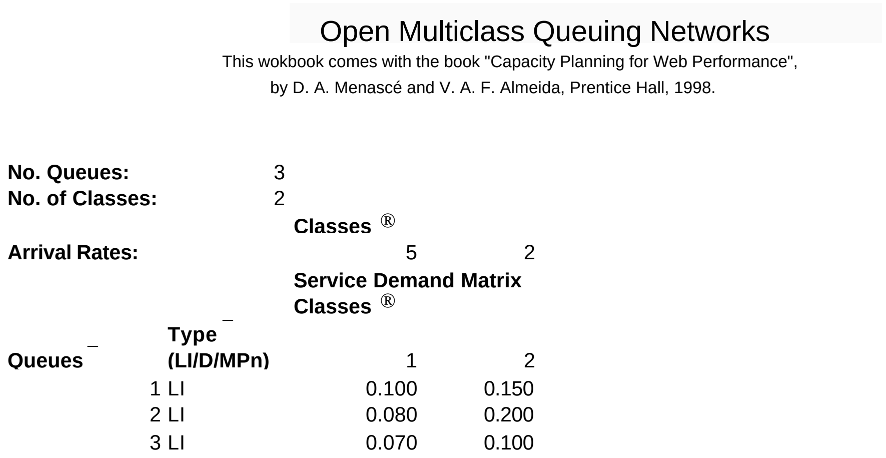


An Intranet Example (cont'd) Cache Hit



An Intranet Example (cont'd) Cache Miss

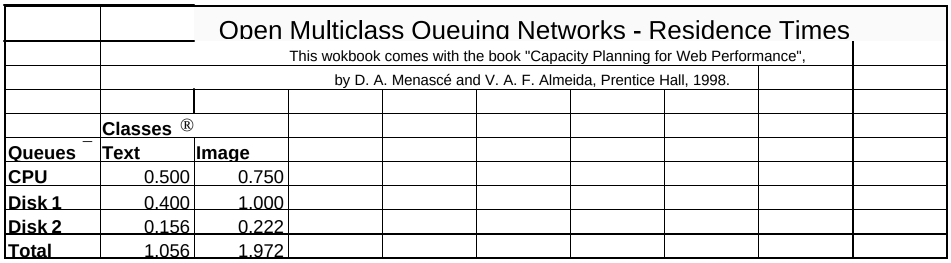




Open Multiclass Queuing Networks

This wokbook comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

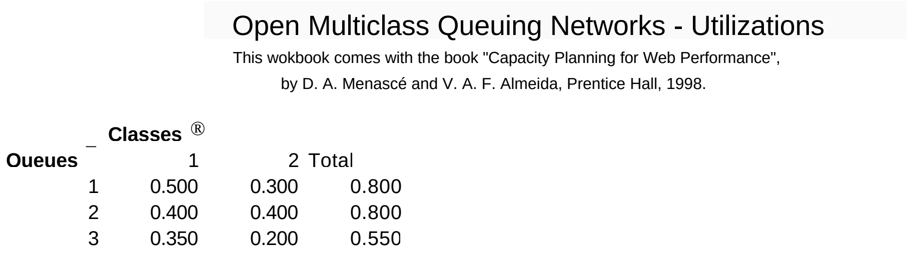
No. Queues:		3		
No. of Classes:		2		
			Classes ®	
Arrival Rates:			5	2
			Service Demand Matrix	
			Classes ®	
	Type ¯			
Queues ¯	**(LI/D/MPn)**		1	2
1	LI		0.100	0.150
2	LI		0.080	0.200
3	LI		0.070	0.100



Open Multiclass Queuing Networks - Residence Times

This wokbook comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®	
Queues ¯	**Text**	**Image**
CPU	0.500	0.750
Disk 1	0.400	1.000
Disk 2	0.156	0.222
Total	1.056	1.972



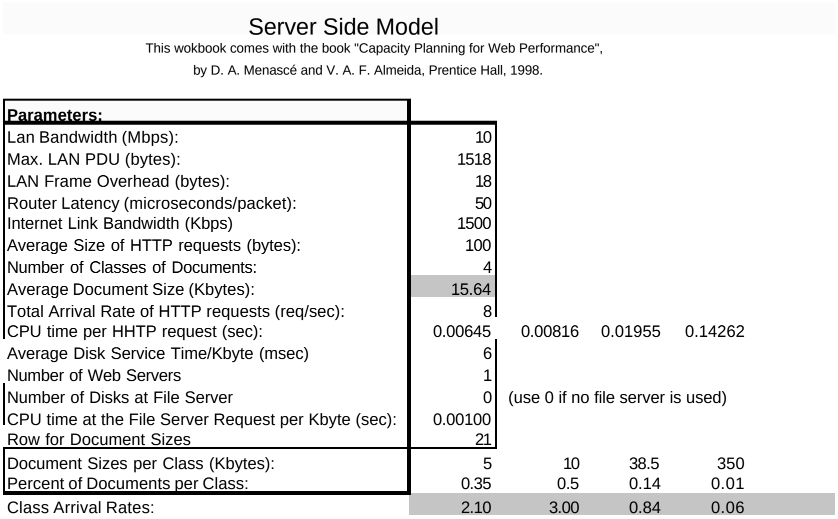
Open Multiclass Queuing Networks - Utilizations

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	_ Classes ®		
Oueues	1	2	Total
1	0.500	0.300	0.800
2	0.400	0.400	0.800
3	0.350	0.200	0.550

A Web Server Example (cont’d)



Server Side Model

This wokbook comes with the book "Capacity Planning for Web Performance",

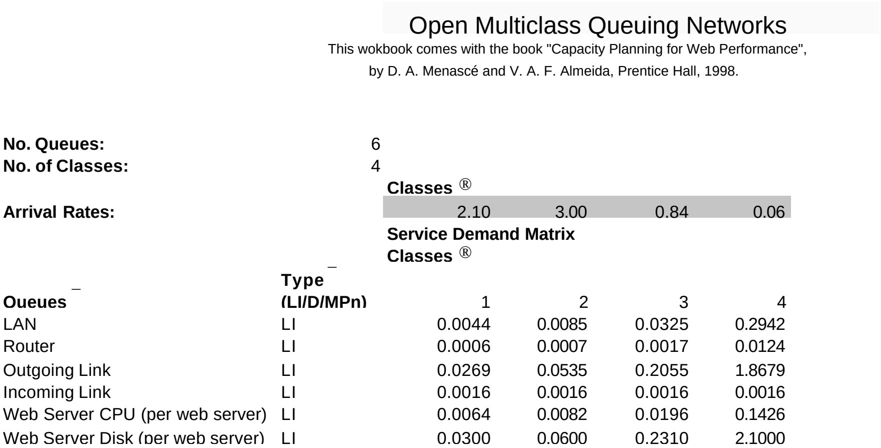
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

Parameters:				
Lan Bandwidth (Mbps):	10			
Max. LAN PDU (bytes):	1518			
LAN Frame Overhead (bytes):	18			
Router Latency (microseconds/packet):	50			
Internet Link Bandwidth (Kbps)	1500			
Average Size of HTTP requests (bytes):	100			
Number of Classes of Documents:	4			
Average Document Size (Kbytes):	15.64			
Total Arrival Rate of HTTP requests (req/sec):	8			
CPU time per HHTP request (sec):	0.00645	0.00816	0.01955	0.14262
Average Disk Service Time/Kbyte (msec)	6			
Number of Web Servers	1			
Number of Disks at File Server	0	(use 0 if no file server is used)		
CPU time at the File Server Request per Kbyte (sec):	0.00100			
Row for Document Sizes	21			
Document Sizes per Class (Kbytes):	5	10	38.5	350
Percent of Documents per Class:	0.35	0.5	0.14	0.01
Class Arrival Rates:	2.10	3.00	0.84	0.06

A Web Server Example (cont'd)

	Classes:			
Service Demands (sec):	1	2	3	4
LAN	0.0044	0.0085	0.0325	0.2942
Router	0.0006	0.0007	0.0017	0.0124
Outgoing Link	0.0269	0.0535	0.2055	1.8679
Incoming Link	0.0016	0.0016	0.0016	0.0016
Web Server CPU (per web server)	0.0064	0.0082	0.0196	0.1426
Web Server Disk (per web server)	0.0300	0.0600	0.2310	2.1000

A Web Server Example (cont'd)

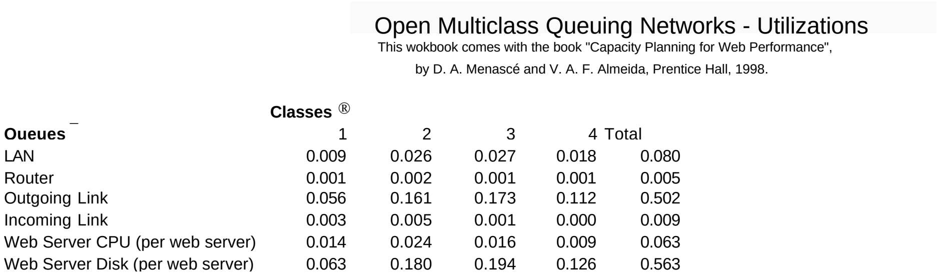


Open Multiclass Queuing Networks

This wokbook comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

No. Queues:		6			
No. of Classes:		4			
		Classes ®			
Arrival Rates:		2.10	3.00	0.84	0.06
		Service Demand Matrix			
		Classes ®			
Oueues	**Type (LI/D/MPn)**	1	2	3	4
LAN	LI	0.0044	0.0085	0.0325	0.2942
Router	LI	0.0006	0.0007	0.0017	0.0124
Outgoing Link	LI	0.0269	0.0535	0.2055	1.8679
Incoming Link	LI	0.0016	0.0016	0.0016	0.0016
Web Server CPU (per web server)	LI	0.0064	0.0082	0.0196	0.1426
Web Server Disk (per web server)	LI	0.0300	0.0600	0.2310	2.1000

A Web Server Example (cont’d)



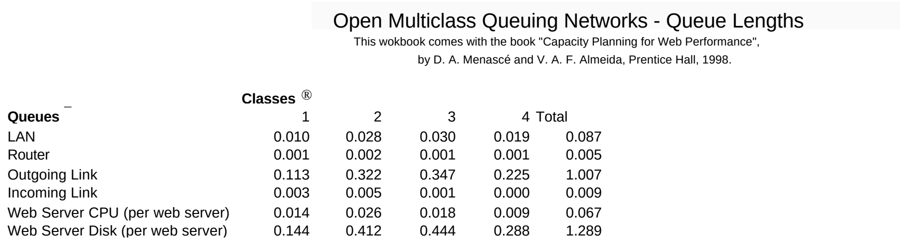
Open Multiclass Queuing Networks - Utilizations

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®				
Oueues –	1	2	3	4	Total
LAN	0.009	0.026	0.027	0.018	0.080
Router	0.001	0.002	0.001	0.001	0.005
Outgoing Link	0.056	0.161	0.173	0.112	0.502
Incoming Link	0.003	0.005	0.001	0.000	0.009
Web Server CPU (per web server)	0.014	0.024	0.016	0.009	0.063
Web Server Disk (per web server)	0.063	0.180	0.194	0.126	0.563

A Web Server Example (cont’d)



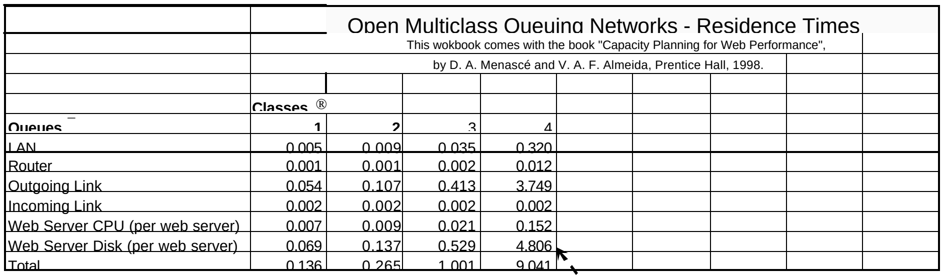
Open Multiclass Queuing Networks - Queue Lengths

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®				
Queues –	1	2	3	4	Total
LAN	0.010	0.028	0.030	0.019	0.087
Router	0.001	0.002	0.001	0.001	0.005
Outgoing Link	0.113	0.322	0.347	0.225	1.007
Incoming Link	0.003	0.005	0.001	0.000	0.009
Web Server CPU (per web server)	0.014	0.026	0.018	0.009	0.067
Web Server Disk (per web server)	0.144	0.412	0.444	0.288	1.289

A Web Server Example (cont'd)



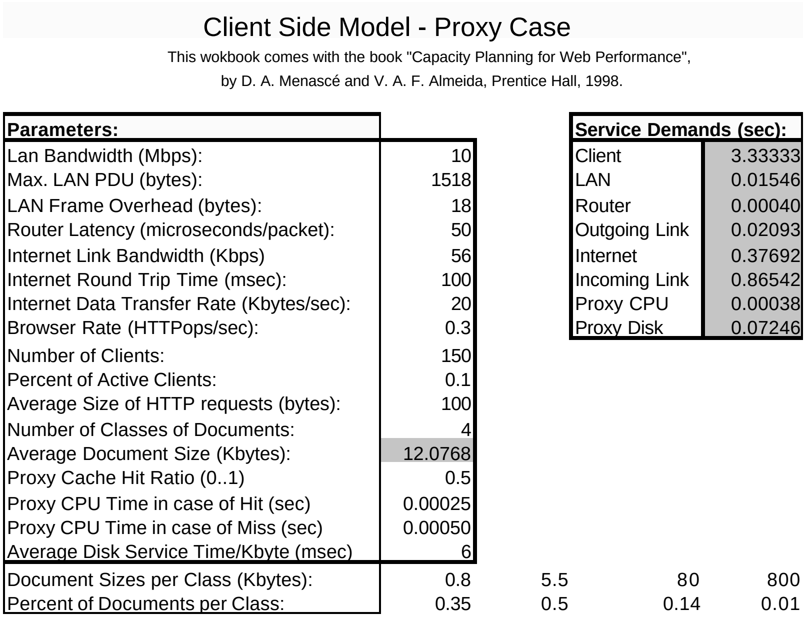
Open Multiclass Queuing Networks - Residence Times

This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®			
Queues ¯	1	2	3	4
LAN	0.005	0.009	0.035	0.320
Router	0.001	0.001	0.002	0.012
Outgoing Link	0.054	0.107	0.413	3.749
Incoming Link	0.002	0.002	0.002	0.002
Web Server CPU (per web server)	0.007	0.009	0.021	0.152
Web Server Disk (per web server)	0.069	0.137	0.529	4.806
Total	0.136	0.265	1.001	9.041

major contribution to response time



Client Side Model - Proxy Case

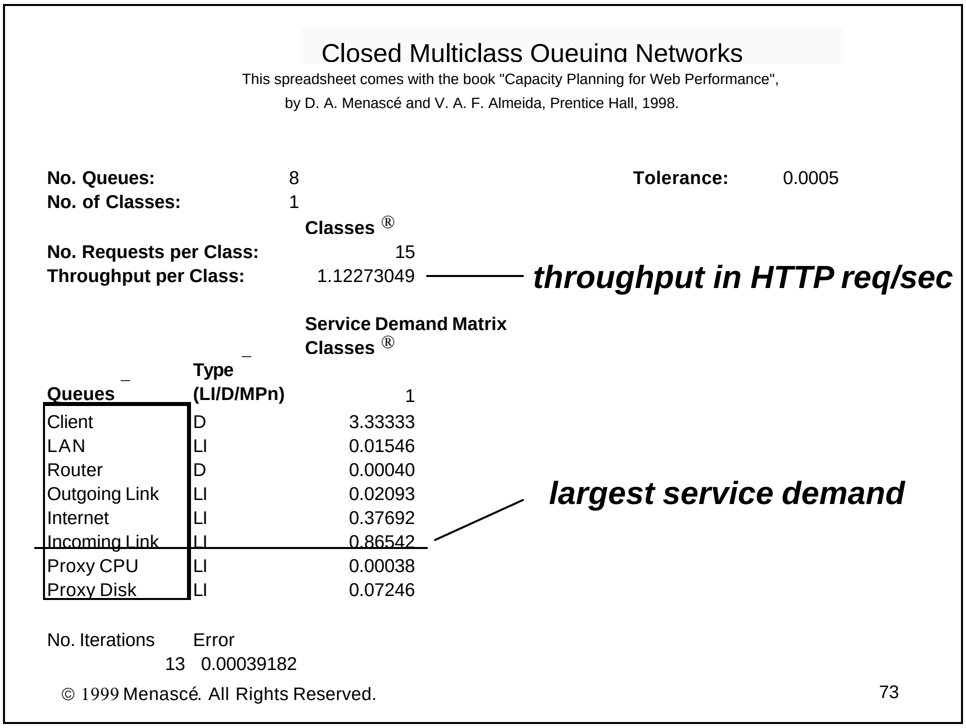
This wokbook comes with the book "Capacity Planning for Web Performance",

by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

Parameters:	
Lan Bandwidth (Mbps):	10
Max. LAN PDU (bytes):	1518
LAN Frame Overhead (bytes):	18
Router Latency (microseconds/packet):	50
Internet Link Bandwidth (Kbps)	56
Internet Round Trip Time (msec):	100
Internet Data Transfer Rate (Kbytes/sec):	20
Browser Rate (HTTPops/sec):	0.3
Number of Clients:	150
Percent of Active Clients:	0.1
Average Size of HTTP requests (bytes):	100
Number of Classes of Documents:	4
Average Document Size (Kbytes):	12.0768
Proxy Cache Hit Ratio (0..1)	0.5
Proxy CPU Time in case of Hit (sec)	0.00025
Proxy CPU Time in case of Miss (sec)	0.00050
Average Disk Service Time/Kbyte (msec)	6

Service Demands (sec):	
Client	3.33333
LAN	0.01546
Router	0.00040
Outgoing Link	0.02093
Internet	0.37692
Incoming Link	0.86542
Proxy CPU	0.00038
Proxy Disk	0.07246

Document Sizes per Class (Kbytes):	0.8	5.5	80	800
Percent of Documents per Class:	0.35	0.5	0.14	0.01



Closed Multiclass Queuing Networks

This spreadsheet comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

No. Queues: 8 **Tolerance:** 0.0005
No. of Classes: 1

	Classes ®
No. Requests per Class:	15
Throughput per Class:	1.12273049

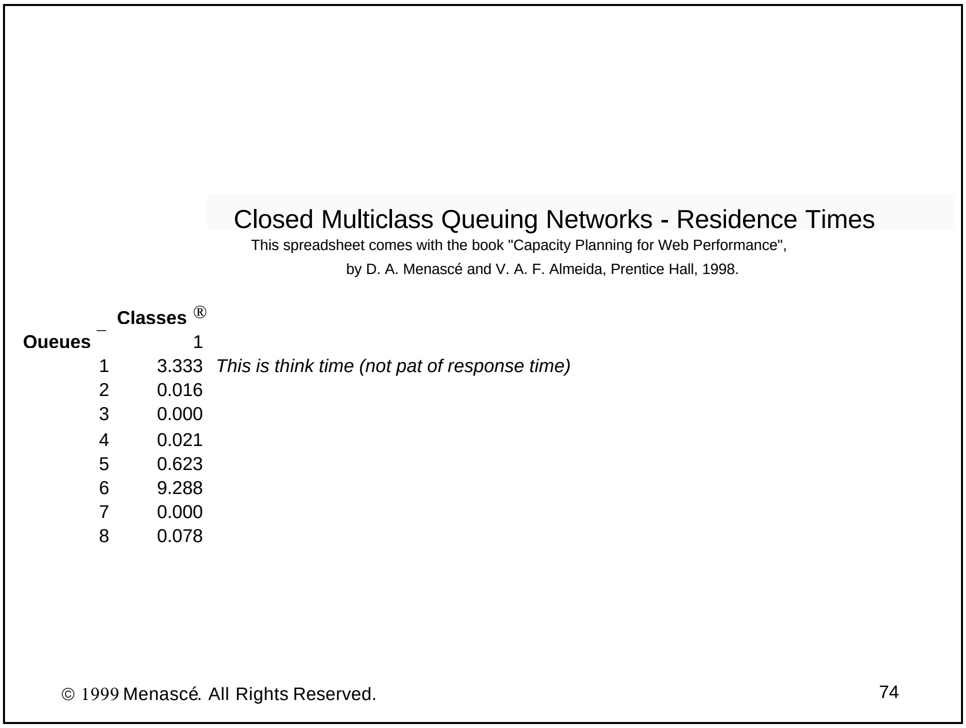
throughput in HTTP req/sec

Service Demand Matrix
Classes ®

Queues ¯	**Type ¯ (LI/D/MPn)**	1
Client	D	3.33333
LAN	LI	0.01546
Router	D	0.00040
Outgoing Link	LI	0.02093
Internet	LI	0.37692
Incoming Link	LI	0.86542
Proxy CPU	LI	0.00038
Proxy Disk	LI	0.07246

largest service demand

No. Iterations	Error
13	0.00039182



Closed Multiclass Queuing Networks - Residence Times

This spreadsheet comes with the book "Capacity Planning for Web Performance",
by D. A. Menascé and V. A. F. Almeida, Prentice Hall, 1998.

	Classes ®	
Oueues ¯	1	
1	3.333	*This is think time (not pat of response time)*
2	0.016	
3	0.000	
4	0.021	
5	0.623	
6	9.288	
7	0.000	
8	0.078	

Intranet Example
Increasing the Bandwidth of the link to the ISP

Link Bandwidth (Kbps)	Throughput (HTTP req/sec)	Resp. Time (sec)	Bottleneck
56	1.125	9.996	Link to ISP
128	2.393	2.935	ISP+Internet+External Web Server
256	3.475	0.984	ISP+Internet+External Web Server
1500	3.883	0.530	ISP+Internet+External Web Server

Outline

❑Component-level Models

❑Computing Service Demands

❑Open Queuing Models

❑Closed Queuing Models

❑Examples

❑Models of E-commerce Servers

Incorporating New Phenomena in the Workload Characterization

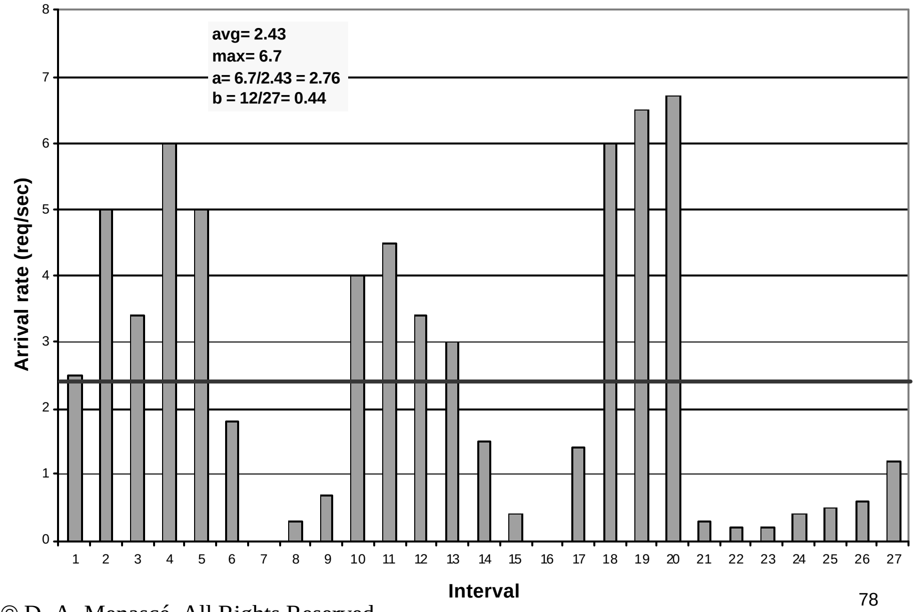
Burstiness Modeling

❑ 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.

77

Burstiness Parameters *a* and *b*



avg= 2.43
max= 6.7
a= 6.7/2.43 = 2.76
b = 12/27= 0.44

78

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 = (number of epochs for which $\lambda_k > \lambda$) / n
- above-average arrival rate, $\lambda^+ = Arr^+ / (b*\tau)$
- $a = \lambda^+ / \lambda = Arr^+ / (b*L)$

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?

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

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

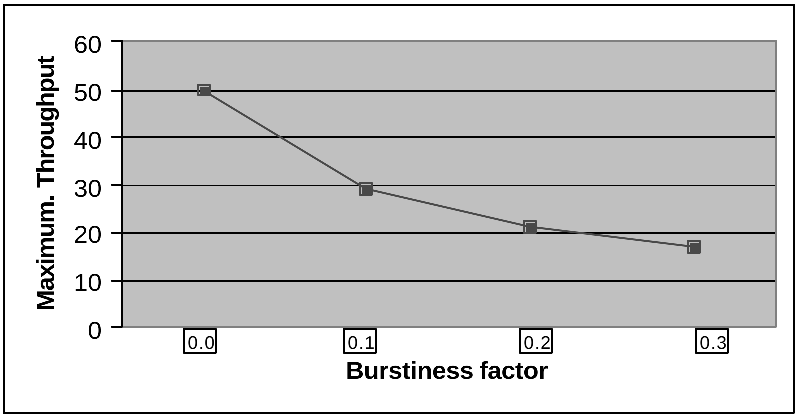
The Impact of Burstiness: an example

- ❑ 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
  - α = (0.18/0.903 - 0.06/0.323)/(0.273-0.182) = 0.149
  - the adjustment factor is: $\alpha \times b$ = 0.149 × 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)



Summary

- Component-level models are used to represent the various components of a networked system.
- Parameters for component-level models include the service demands on system resources, i.e. total time spent by a request receiving service from a resource.

Summary (cont'd)

- ❑ Waiting times, response times, throughputs can be computed using
  - ▪ open models (e.g., web servers)
  - ▪ closed models (e.g., intranets)
- ❑ In modeling e-commerce servers, need to consider software as well as hardware contention.