Performance Modeling – Single Queues

CS 700

1

Acknowledgement

These slides are based on presentations created and copyrighted by Prof. Daniel Menasce (GMU)

2

Purpose of Models

- Provide a way to derive performance metrics from model parameters.
- Examples of performance metrics:
  - Response time
  - Throughput
  - Availability
- Types of parameters:
  - Workload intensity (e.g., arrival rates)
  - Service demands.

3

Type of Models

- Simulation: mimic flow of transactions through a system.
  - Distribution-driven
  - Trace-driven
- Analytic: set of formulas or computational algorithms.
  - Exact
  - Approximate
- Hybrid

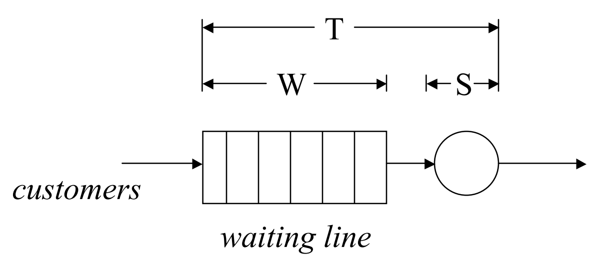
4

When to Use?

- Use Exact Analytic Models Whenever Possible.
- Use Approximate Analytic Models:
  - For first-cut analysis
  - If validated by simulation
  - To reduce combinations of input parameters to simulation models.
- Use Simulation:
  - If there is no tractable analytic model.

5

Single Queue



$$T = W + S$$

6

Background: Stochastic Processes

- A stochastic process is a family of random variables $\{X(t) | t \in T\}$, defined on a given probability space, indexed by the parameter t, where t varies over the index set T
  - The values assumed by the random variable X(t) are called states
    - If state space is discrete, then the stochastic process is a discrete-state process, often referred to as a chain, otherwise it is a continuous-state process
  - If the index set is discrete, the process is called a discrete parameter process, otherwise it is a continuous parameter process

7

Stochastic processes cont'd

- Consider a single-server queue. We can identify several stochastic processes
  - N_k - number of customers in the system at the time of departure of the kth customer.
    - $\{N_k | k = 1,2,\ldots\}$ is a discrete parameter, discrete-state process
  - X(t) - number of customers in the system at time t
    - $\{X(t) | 0 < t < \infty\}$ is a continuous parameter, discrete state process
  - W_k - time the kth customer has to wait to receive service
    - $\{W_k | k = 1, 2,\ldots\}$ is a discrete parameter, continuous state process
  - Y(t) - cumulative service requirement of all jobs in the system at time t
    - $\{Y(t) | 0 < t < \infty\}$ is a continuous parameter, continuous state process

8

Stochastic processes - some types

- Markov process/chain -- if the future states of a process are independent of the past and depend only on the current state, the process is called a Markov process
- Birth-death processes -- discrete state Markov processes in which transitions are restricted to neighboring states only
- Poisson process -- if the inter-arrival times at a queue are IID (independent and identically distributed) and exponentially distributed, the arrival process is called a Poisson process
  - This is because the number of arrivals over a given interval of time will have a Poisson distribution

9

Example of An Analytic Model: M/G/1 Queue

- Single server.
- Arrival process is Poisson (inter-arrival times are exponentially distributed).
- Service time is arbitrarily distributed.

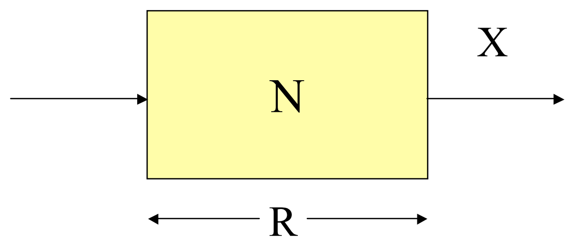
$$T = E[S] + \frac{\lambda\, E[S^2]}{2(1-\rho)} = E[S] + \frac{\rho E[S](1+C_s^2)}{2(1-\rho)}$$

Where

$$\rho = \lambda E[S] < 1$$

10

Little's Law



The average number of customers in a "black box" is equal to the average time spent in the box multiplied by the throughput of the box.

$$N = R \times X$$

11

Little's Law Example I

- An NFS server was monitored during 30 min and the number of I/O operations performed during this period was found to be 32,400. The average number of active requests (N_{req}) was 9.
- What was the average response time per NFS request at the server?

12

Little's Law Example I

- An NFS server was monitored during 30 min and the number of I/O operations performed during this period was found to be 32,400. The average number of active requests (N_{req}) was 9.
- What was the average response time per NFS request at the server?

"black box" = NFS server

$$X_{server} = 32{,}400 / 1{,}800 = 18 \text{ requests/sec}$$

$$R_{req} = N_{req} / X_{server} = 9 / 18 = 0.5 \text{ sec}$$

13

Little's Law Example II

- A large portal service offers free email service. The number of registered users is two million and 30% of them send send mail through the portal during the peak hour. Each mail takes 5.0 sec on average to be processed and delivered to the destination mailbox. During the busy period, each user sends 3.5 mail messages on average. The log file indicates that the average size of an e-mail message is 7,120 bytes.
- What should be the capacity of the spool for outgoing mails during the peak period?

14

Little's Law Example II

- A large portal service offers free email service. The number of registered users is two million and 30% of them send send mail through the portal during the peak hour. Each mail takes 5.0 sec on average to be processed and delivered to the destination mailbox. During the busy period, each user sends 3.5 mail messages on average. The log file indicates that the average size of an e-mail message is 7,120 bytes.
- What should be the capacity of the spool for outgoing mails during the peak period?

AvgNumberOfMails = Throughput x ResponseTime
= (2,000,000 x 0.30 x 3.5 x 5.0) / 3,600 =
2,916.7 mails

AvgSpoolFile = 2,916.7 x 7,120 bytes = 19.8 MBytes

15

Little's Law Example III

- A Web-based brokerage company runs a three-tiered site. The site is used by 1.1 million customers. During the peak hour, 20,000 users are logged in simultaneously. The e-commerce site processes 3.6 million business functions per hour in a peak-load hour.
- What is the average response time of an e-commerce function during the peak hour?

16

Little's Law Example III

- A Web-based brokerage company runs a three-tiered site. The site is used by 1.1 million customers. During the peak hour, 20,000 users are logged in simultaneously. The e-commerce site processes 3.6 million business functions per hour on a peak-load hour.
- What is the average response time of an e-commerce function during the peak hour?

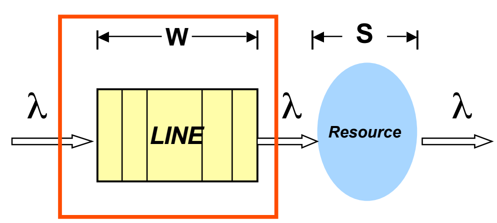
Black box = E-commerce site

AverageResponseTime = AvgNumberOfUsers / SiteThroughput

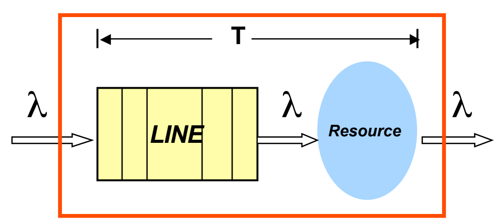
= 20,000 / (3,600,000 / 3,600) = 20 sec

Using Little's Law in the M/G/1 Queue

$$E[N_q] = \frac{\rho^2(1+C_s^2)}{2(1-\rho)}$$



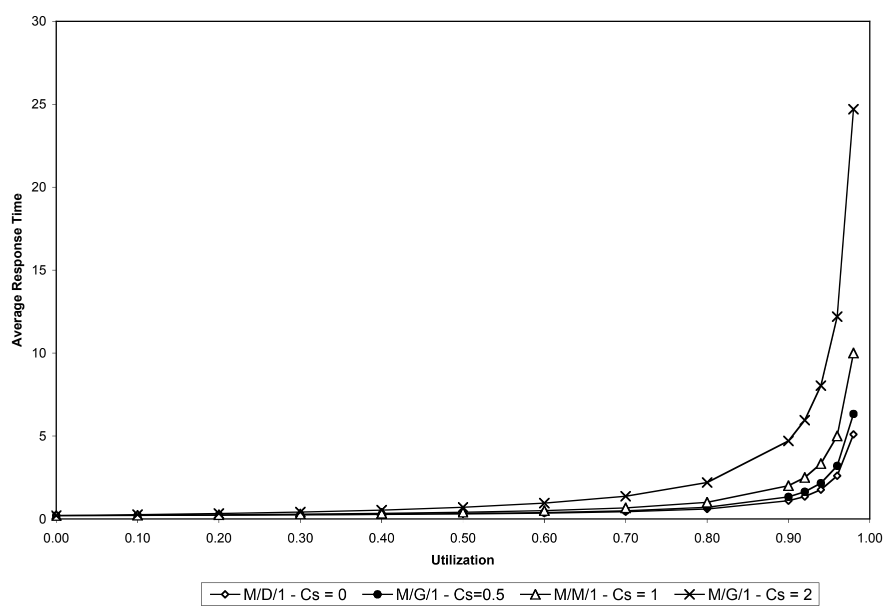
$$E[N] = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)}$$



Exercise

- Plot the response time for M/G/1 as a function of ρ for M/M/1, M/D/1, and distributions with coefficient of variation equal to _ and 2. Assume that E[S] = 0.2. Vary λ accordingly.
- What conclusions do you take from looking at the graphs?

19



20

M/G/1, M/M/1, and M/D/1

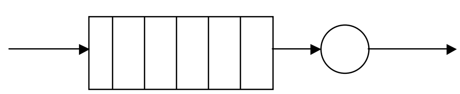
M/G/1: $$W = \frac{\rho E[S](1 + C_S^2)}{2(1-\rho)}$$

M/D/1: $$W = \frac{\rho E[S]}{2(1-\rho)}$$

M/M/1: $$W = \frac{\rho E[S]}{(1-\rho)}$$

21

G/G/1 Queue



$$\rho = \lambda E[S] < 1$$

$$p_0 = 1 - \rho$$

22

An Approximation for G/G/1

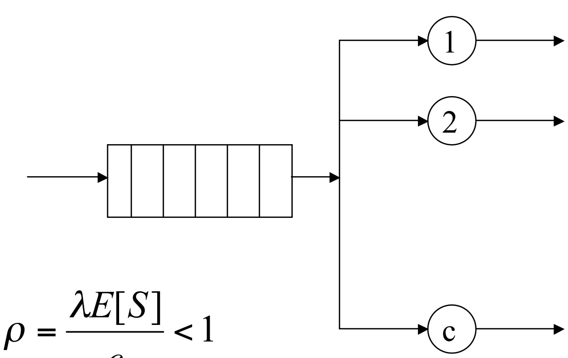
$$W \approx \frac{C_a^2 + \rho^2 C_s^2}{1 + \rho^2 C_s^2} \times \frac{\rho(1 + C_s^2)}{2(1-\rho)/E[S]}$$

C_a^2 : coefficient of variation of the interarrival time.

Approximation is exact for M/G/1, good for G/M/1, and fair for G/G/1.
The approximation improves as ρ increases.

23

G/G/c Queue



$$\rho = \frac{\lambda E[S]}{c} < 1$$

24

An Approximation for G/G/c

$$W \approx \frac{C(\rho,c)}{c(1-\rho)/E[S]} \times \frac{C_a^2 + C_s^2}{2}$$

where $C(\rho,c) = \dfrac{(c\rho)^c / c!}{(1-\rho)\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!}}$ is Erlang's C formula.

Approximation is exact for M/M/c.
The error increases with C_a and C_s.

25

The M/M/c Queue

$$W = \frac{C(\rho,c)}{c(1-\rho)/E[S]}$$

where $C(\rho,c) = \dfrac{(c\rho)^c / c!}{(1-\rho)\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!}}$ is Erlang's C formula.

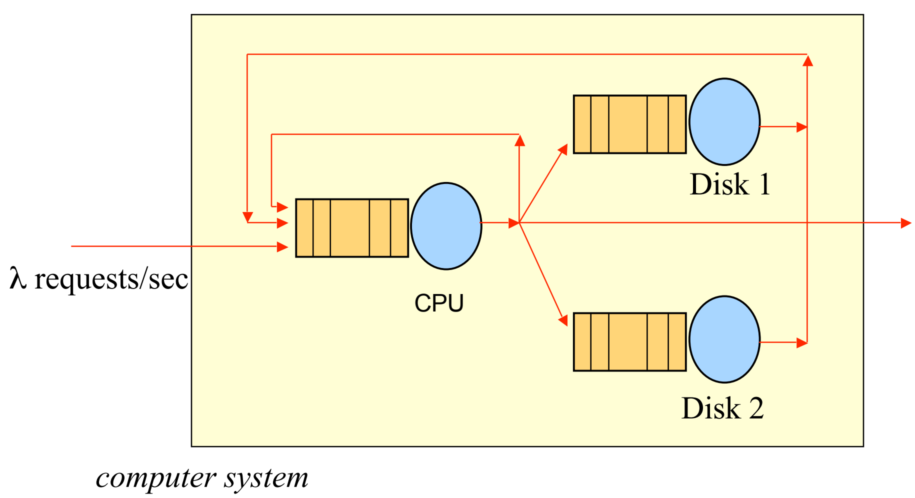
26

Performance Modeling – Queuing Networks

CS 700

27

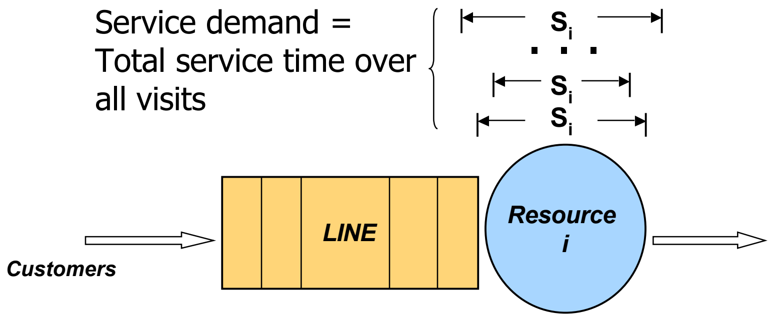
A Computer System as a Network of Queues



28

Service Demand (D_i)

Service demand =
Total service time over
all visits



S_i : service time
D_i : service demand

29

Service Demand Example

Database transactions use two disks. The service times at each of the disks for each I/O carried out by a single transaction are

	Service Time (msec)	
I/O	**Disk 1**	**Disk 2**
1	12	12
2	20	15
3	15	14
4	18	-
	65	41

30

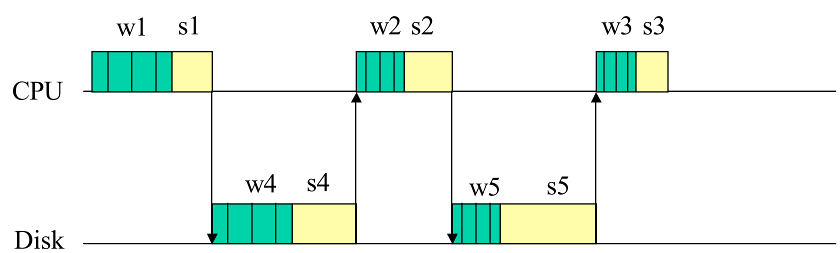
Queuing Basic Concepts

Total time spent by a request during the j^{th} visit to a resource *i*:

- Service time (S_i^j): period of time a request is receiving service from resource *i*, such as CPU or disk.
- Waiting time (W_i^j): the time spent by a request waiting access to resource *i*

31

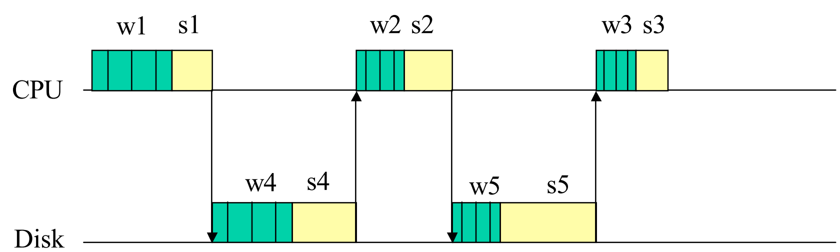
Queuing Time



Queuing time at the CPU = w1 + w2 + w3
Queuing time at the disk = w4 + w5

32

Service Demand



Service demand at the CPU = s1 + s2 + s3
Service demand at the disk = s4 + s5

Waiting time   Service time

33

Basic Queuing Concepts

Service Demand (D_i) is the sum of all service times for a request at resource *i*

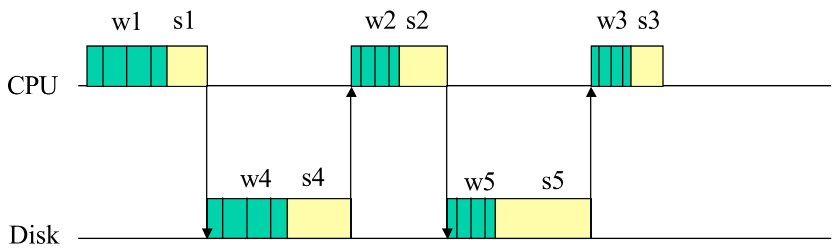
$$D_{scpu} = S^1_{scpu} + S^2_{scpu}$$

Queuing Time (Q_i) is the sum of all waiting times for a request at resource *i*

$$Q_{scpu} = W^1_{scpu} + W^2_{scpu}$$

34

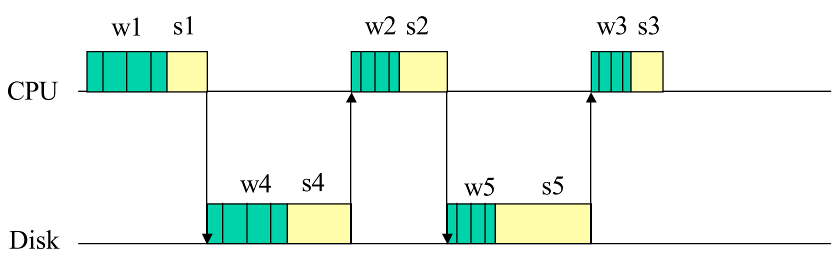
Residence Time



Residence time at the CPU = w1 + s1 + w2 + s2 + w3 + s3
Residence time at the disk = w4 + s4 + w5 + s5

35

Response Time



Response time = Residence time at the CPU + Residence time at the disk

36

Basic Queuing Concepts

Residence Time (R'_i) at resource *i* is the sum of service demand plus queuing time.

$$R'_i = Q_i + D_i$$

Response time (R_r) of a request *r* is the sum of that request's residence time at all resources.

$$R_{server} = R'_{cpu} + R'_{disk}$$

37

Notation

V_i: average number of visits to queue *i* by a request;

$S_{i:}$ average service time of a request at queue *i* per visit to the resource;

λ_i average arrival rate of requests to queue *i*

D_i service demand of a request at queue *i*,

$$D_i = V_i \times S_i$$

38

More Notation

N_i: average number of requests at queue *i*, waiting or receiving service from the resource

X_i: average throughput of queue *i*, i.e. average number of requests that complete from queue *i* per unit of time

X_o: average system throughput, defined as the number of requests that complete per unit of time.

39

Basic Performance Laws

Utilization Law

The utilization (U_i) of resource *i* is the fraction of time that the resource is busy.

$$U_i = X_i * S_i = \lambda_i * S_i$$

40

Utilization Law: example

- A network segment transmits 1,000 packets/sec. Each packet has an average transmission time equal to 0.15 msec.
- What is the utilization of the LAN segment?

41

Utilization Law: example

- A network segment transmits 1,000 packets/sec. Each packet has an average transmission time equal to 0.15 msec.
- What is the utilization of the LAN segment?

$$U_{LAN} = X_{LAN} * S_{LAN} = 1{,}000 * 0.00015 = 0.15 = 15\%$$

42

Basic Performance Results

Forced Flow Law

By definition of the average number of visits V_i, each completing request has to pass V_i times, on the average, by queue *i*. So, if X_o requests complete per unit of time, V_i*X_o requests will visit queue *i*.

$$X_i = V_i * X_o$$

43

Forced Flow Law: example I

- Database transactions perform an average of 4.5 I/O operations on the database server. During a one-hour monitoring period, 7,200 transactions were executed.
- What is the average throughput of the disk?
- If each I/O takes 20 msec on the average, what is the disk utilization?

44

Forced Flow Law: example I

- Database transactions perform an average of 4.5 I/O operations on the database server. During a one-hour monitoring period, 7,200 transactions were executed.
- What is the average throughput of the disk?
- If each I/O takes 20 msec on the average, what is the disk utilization?

$$X_{server} = 7{,}200 / 3{,}600 = 2 \text{ tps}$$
$$X_{disk} = V_{disk} * X_{server} = 4.5 * 2 = 9 \text{ tps}$$
$$U_{disk} = X_{disk} * S_{disk} = 9 * 0.02 = 0.18 = 18\%$$

45

Basic Performance Results

Service Demand Law

The service demand D_i is related to the system throughput and utilization by the following:

$$D_i = V_i * S_i = (X_i/X_o)(U_i/X_i) = U_i / X_o$$

46

Service Demand Law: example

- A Web server was monitored for 10 minutes. It was observed that the CPU was 90% busy during the monitoring period. The number of HTTP requests counted in the log was 30,000.
- What is the CPU service demand of an HTTP request?

47

Service Demand Law: example

- A Web server was monitored for 10 minutes. It was observed that the CPU was 90% busy during the monitoring period. The number of HTTP requests counted in the log was 30,000.
- What is the CPU service demand of an HTTP request?

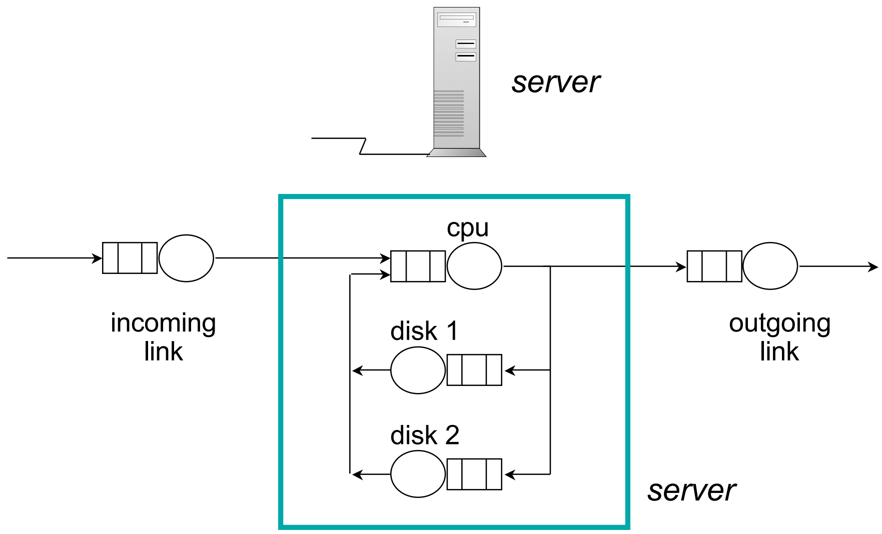
$$U_{cpu} = 90\%$$

$$X_{server} = 30{,}000 / (10*60) = 50 \text{ requests/sec}$$

$$D_{cpu} = V_{cpu} * S_{cpu} = U_{cpu} / X_{server} = 0.90 / 50 = 0.018 \text{ sec}$$

48

An Open Queuing Model Example



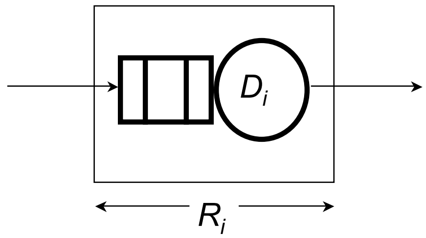
49

Open QN Models

- The number of requests in the system is not bounded.
- Input parameters: arrival rate of requests and service demands.
- Output metrics: response time, queue lengths, and utilizations.

50

Open QN Models
Computing Residence Times



Service demand at resource i

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

Utilization of resource i (U_i)

51

Derivation of Residence Time

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

$$\overline{n}_i^A = \overline{n}_i \text{ for open systems}$$

$$\overline{n}_i = X_i R_i \text{ from Little's Law}$$

$$R_i = S_i + S X_i R_i = S_i + U_i R_i$$

$$\Rightarrow R_i = \frac{S_i}{1 - U_i}$$

multiplying both sides by V_i :

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

52

Open Model Equations

$$U_i = \lambda \times D_i$$

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

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

53

Bound on Throughput

Give an expression for the maximum throughput of a computer system as a function of the service demands $D_1, \ldots, D_K$

(Hint: the utilization cannot exceed 100%)

54

Equations for Open Multiple Class QN Models

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

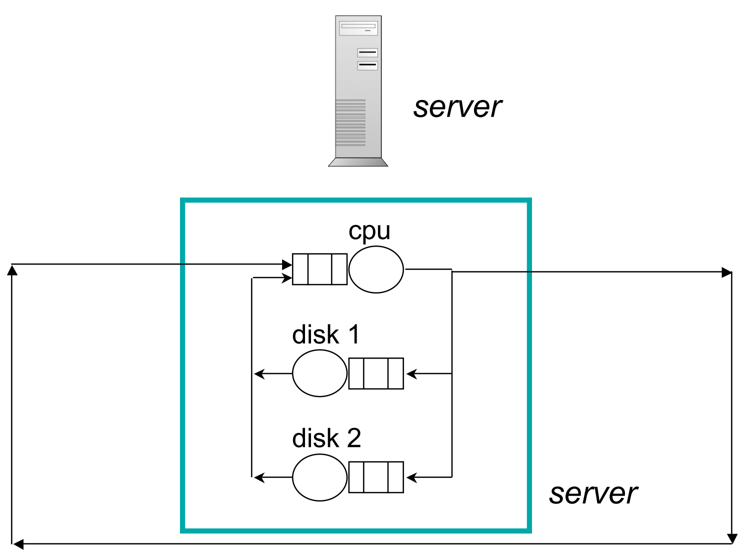
$$U_i = \sum_{r=1}^{R} U_{i,r}$$

$$R'_{i,r} = \frac{D_{i,r}}{1 - U_i}$$

$$R_{o,r} = \sum_{i=1}^{K} R'_{i,r}$$

55

A Closed Queuing Model Example



56

Closed QN Models

- The number of requests in the system is constant: a completing request is immediately replaced by a new request.
- Input parameters: number of requests in the system and service demands.
- Output metrics: throughput, response time, queue lengths, and utilizations.
- Solution technique: Mean Value Analysis (MVA)

57

Closed QN Model
MVA Equations

Residence Time Equation:

my total waiting time at resource i

$$R_i'(n) = \underbrace{D_i}_{\text{my total service time}} + \overbrace{D_i \times \underbrace{\bar{n}_i(n-1)}_{\text{avg. number of requests at resource } i \text{ found upon my arrival}}}^{\text{my total waiting time at resource } i}$$

my total service time

avg. number of requests at resource i found upon my arrival

58

Closed QN Model: MVA Equations

Residence Time Equation:

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

59

Closed QN Model: MVA Equations

Throughput Equation. Using Little's Law:

throughput

$$n = X_o(n) \times R_o(n)$$

total response time

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

60

Closed QN Model: MVA Equations

Throughput Equation:

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

61

Closed QN Model: MVA Equations

Queue Length Equations. Applying Little's Law and the Forced Flow Law to the resource i.

$$\overline{n}_i(n) = X_o(n) \times R_i^{'}(n)$$

62

MVA Equations

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

$$X_o(n) = \frac{n}{\sum_{i=1}^{K} R_i^{'}(n)}$$

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

63

Solving the Model

$$R_{cpu}^{'}(1) = D_{cpu} \times \left[1 + \bar{n}_{cpu}(0)\right] = D_{cpu}$$

$$R_{disk}^{'}(1) = D_{disk} \times \left[1 + \bar{n}_{disk}(0)\right] = D_{disk}$$

$$X_o(1) = \frac{1}{R_o(1)} = \frac{1}{R_{cpu}^{'}(1) + R_{disk}^{'}(1)}$$

$$\bar{n}_{cpu}(1) = X_o(1) \times R_{cpu}^{'}(1)$$

$$\bar{n}_{disk}(1) = X_o(1) \times R_{disk}^{'}(1)$$

64

Solving the Model

$$R'_{cpu}(2) = D_{cpu} \times \left[1 + \bar{n}_{cpu}(1)\right]$$

$$R'_{disk}(2) = D_{disk} \times \left[1 + \bar{n}_{disk}(1)\right]$$

$$X_o(2) = \frac{2}{R_o(2)} = \frac{2}{R'_{cpu}(2) + R'_{disk}(2)}$$

$$\bar{n}_{cpu}(2) = X_o(2) \times R'_{cpu}(2)$$

$$\bar{n}_{disk}(2) = X_o(2) \times R'_{disk}(2)$$

65

Closed QN Example

An online transaction processing system has one CPU and one disk. Transactions use an average of 18 msec of CPU time and do 3.5 I/Os on average. Each I/O takes 8 msec on average.

1. Compute the service demands at the CPU and disk.
2. Compute the maximum throughput.
3. Plot the system response time and the throughput as function of the number of concurrent requests in execution.
4. What would you do to improve the maximum throughput by 30%?

66

Open QN Example

An online transaction processing system has one CPU and one disk. Transactions use an average of 18 msec of CPU time and do 3.5 I/Os on average. Each I/O takes 8 msec on average.

1. Compute the service demands at the CPU and disk.
2. Compute the maximum throughput.
3. Plot the system response time as function of the arrival rate of requests.

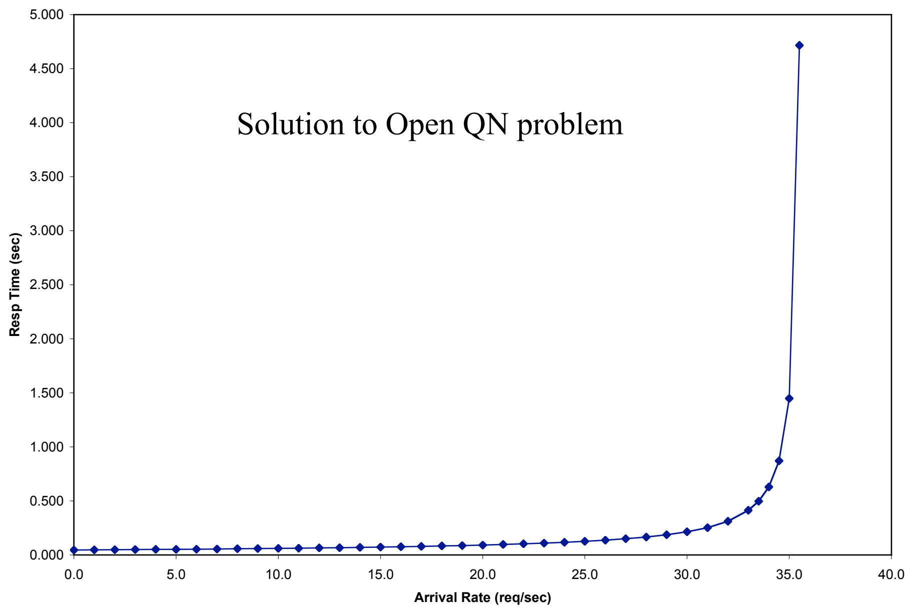
67

Dcpu	0.018	sec
Ddisk	0.028	sec
Max Throughput	35.71429	req/sec

Lambda	Ucpu	Udisk	R'cpu (msec)	R'disk (msec)	Resp Time (msec)
0.0	0.000	0.000	0.018	0.028	0.046
1.0	0.018	0.028	0.018	0.029	0.047
2.0	0.036	0.056	0.019	0.030	0.048
3.0	0.054	0.084	0.019	0.031	0.050
4.0	0.072	0.112	0.019	0.032	0.051
5.0	0.090	0.140	0.020	0.033	0.052
6.0	0.108	0.168	0.020	0.034	0.054
7.0	0.126	0.196	0.021	0.035	0.055
8.0	0.144	0.224	0.021	0.036	0.057
9.0	0.162	0.252	0.021	0.037	0.059
10.0	0.180	0.280	0.022	0.039	0.061
11.0	0.198	0.308	0.022	0.040	0.063
12.0	0.216	0.336	0.023	0.042	0.065
13.0	0.234	0.364	0.023	0.044	0.068
14.0	0.252	0.392	0.024	0.046	0.070
15.0	0.270	0.420	0.025	0.048	0.073
16.0	0.288	0.448	0.025	0.051	0.076
17.0	0.306	0.476	0.026	0.053	0.079
18.0	0.324	0.504	0.027	0.056	0.083
19.0	0.342	0.532	0.027	0.060	0.087
20.0	0.360	0.560	0.028	0.064	0.092
21.0	0.378	0.588	0.029	0.068	0.097
22.0	0.396	0.616	0.030	0.073	0.103
23.0	0.414	0.644	0.031	0.079	0.109
24.0	0.432	0.672	0.032	0.085	0.117
25.0	0.450	0.700	0.033	0.093	0.126
26.0	0.468	0.728	0.034	0.103	0.137
27.0	0.486	0.756	0.035	0.115	0.150
28.0	0.504	0.784	0.036	0.130	0.166
29.0	0.522	0.812	0.038	0.149	0.187
30.0	0.540	0.840	0.039	0.175	0.214
31.0	0.558	0.868	0.041	0.212	0.253
32.0	0.576	0.896	0.042	0.269	0.312
33.0	0.594	0.924	0.044	0.368	0.413
33.5	0.603	0.938	0.045	0.452	0.497
34.0	0.612	0.952	0.046	0.583	0.630
34.5	0.621	0.966	0.047	0.824	0.871
35.0	0.630	0.980	0.049	1.400	1.449
35.5	0.639	0.994	0.050	4.667	4.717

Solution to Open QN problem

68



69

Closed QN Example

An online transaction processing system has one CPU and one disk. Transactions use an average of 18 msec of CPU time and do 3.5 I/Os on average. Each I/O takes 8 msec on average.

1. Compute the service demands at the CPU and disk.
2. Compute the maximum throughput.
3. Plot the system response time and the throughput as function of the number of concurrent requests in execution.
4. What would you do to improve the maximum throughput by 30%?

70

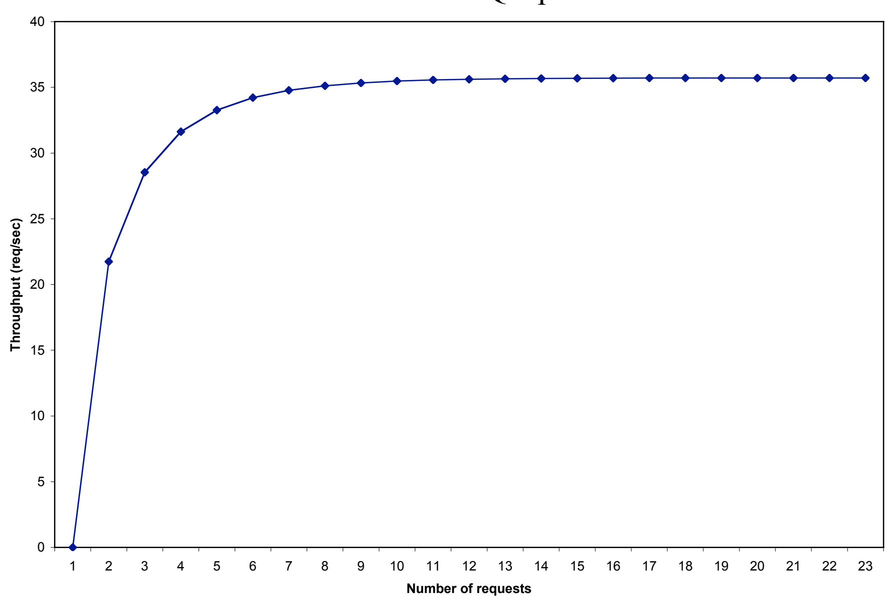
Solution to Closed QN problem

Dcpu 0.018 sec
Ddisk 0.028 sec
Max Throughput 35.71429 req/sec

n	R'cpu	R'disk	Ro	Xo	ncpu	ndisk
0				0	0	0
1	0.02	0.03	0.05	21.74	0.39	0.61
2	0.03	0.05	0.07	28.54	0.71	1.29
3	0.03	0.06	0.09	31.63	0.98	2.02
4	0.04	0.08	0.12	33.27	1.18	2.82
5	0.04	0.11	0.15	34.21	1.34	3.66
6	0.04	0.13	0.17	34.77	1.47	4.53
7	0.04	0.15	0.20	35.12	1.56	5.44
8	0.05	0.18	0.23	35.34	1.63	6.37
9	0.05	0.21	0.25	35.47	1.68	7.32
10	0.05	0.23	0.28	35.56	1.71	8.29
11	0.05	0.26	0.31	35.61	1.74	9.26
12	0.05	0.29	0.34	35.65	1.76	10.24
13	0.05	0.31	0.36	35.67	1.77	11.23
14	0.05	0.34	0.39	35.69	1.78	12.22
15	0.05	0.37	0.42	35.70	1.79	13.21
16	0.05	0.40	0.45	35.70	1.79	14.21
17	0.05	0.43	0.48	35.71	1.79	15.21
18	0.05	0.45	0.50	35.71	1.80	16.20
19	0.05	0.48	0.53	35.71	1.80	17.20
20	0.05	0.51	0.56	35.71	1.80	18.20
21	0.05	0.54	0.59	35.71	1.80	19.20
22	0.05	0.57	0.62	35.71	1.80	20.20

71

Solution to Closed QN problem



72

Additional Reading

- Two columns from "Programming Pearls" by Jon Bentley on "Back of the envelope" calculations
  - See links on class web site

73