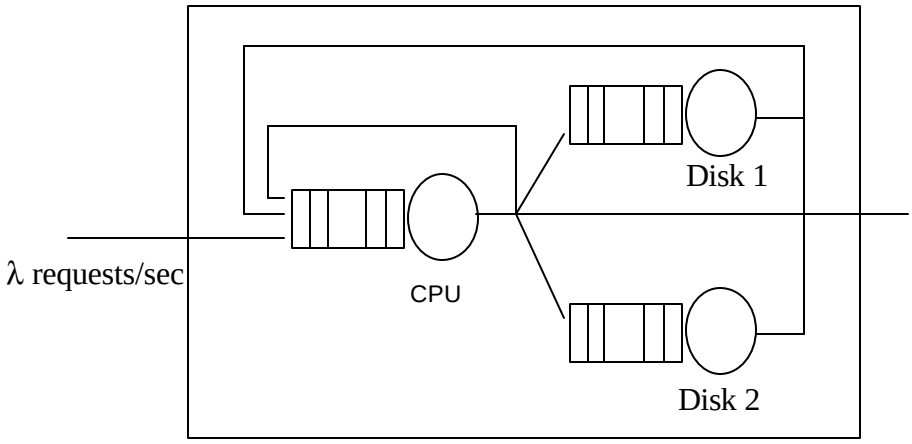
Performance Modeling – Part II
Queuing Networks

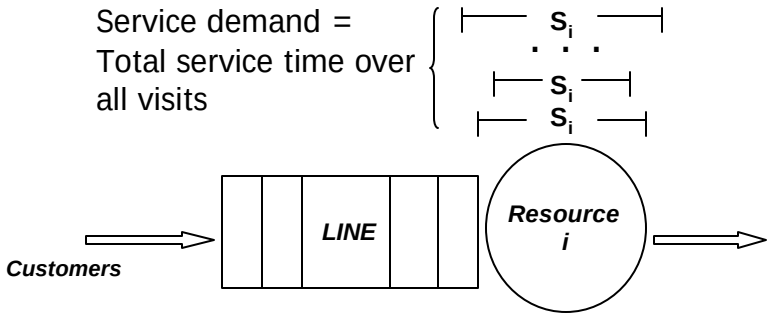
Prof. Daniel A. Menascé
Dept. of Computer Science
George Mason University

A Computer System as a Network of Queues



computer system

Service Demand (D_i)



S_i : service time

D_i : service demand

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

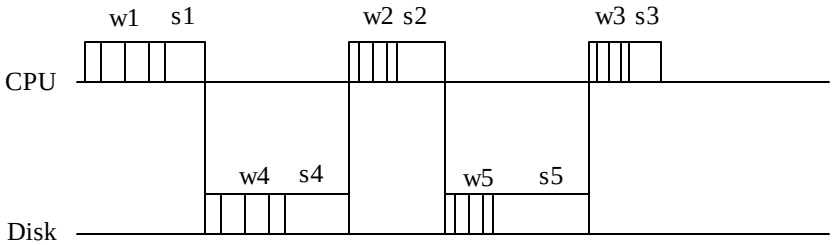
Queuing Basic Concepts

- 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*

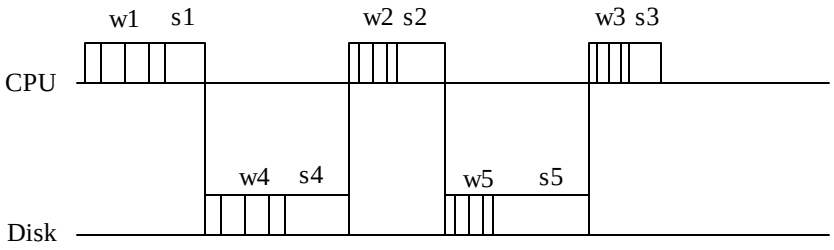
Queuing Time



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

Waiting time Service time

Service Demand



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

Waiting time Service time

Basic Queuing Concepts

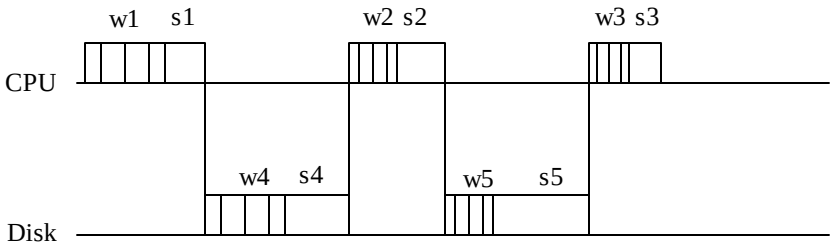
- Service Demand (D_i) is the sum of all service times for a request at resource *i*

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

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

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

Residence Time

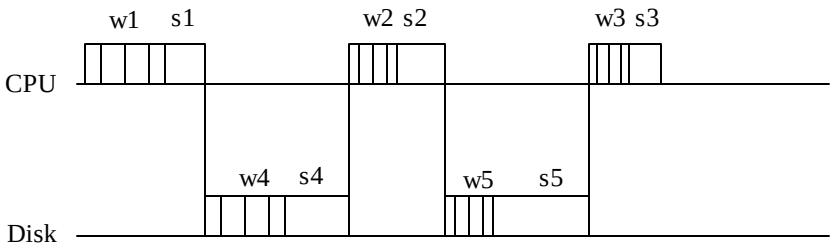


Residence time at the CPU = w1 + s1 + w2 + s2 + w3 + s3
Residence time at the disk = w4 + s4 + w5 + s5

Waiting time Service time

9

Response Time



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

Waiting time Service time

10

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.

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

11

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*,
- $D_i = V_i \times S_i$

12

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.

Basic Performance Laws

Utilization Law

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

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

Example of Utilization Law: iostat in Unix

r/s	w/s	Kr/s	Kw/s	svc_t_(msec)
0.8	7.4	6.2	131.2	136.7
0.2	4.4	1.6	113.6	61
1	14.8	8	438.4	61.3
13	1.2	128	134.4	16.8
0.2	0	1.6	0	12.4
0	0.2	0	25.6	40.9
0	0	0	0	0
0	4	0	28.6	116
0	0	0	0	0
0	0	0	0	0
3	0	24	0	11.4
0	0.6	0	35.2	35.2
0	0	0	0	0
0	0.2	0	1.6	17.3
1.30	2.34	12.10	64.90	36.36

$X_{disk} = 1.3 + 2.34 = 3.64$ IOs/sec

$U_{disk} = X_{disk} \times S_{disk} = 3.64 \times 0.03636 = 13.24\%$

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?

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?

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

17

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.*

$$X_i = V_i * X_o$$

18

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?

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?

$$X_{server} = 7{,}200 / 3{,}600 = 2 \text{ tps}$$

$$X_{disk} = V_{disk} * X_{server} = 4.5 * 2 = 9 \text{ tps}$$

$$U_{disk} = X_{disk} * S_{disk} = 9 * 0.02 = 0.18 = 18\%$$

Basic Performance Results

Service Demand Law

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

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

Example of Service Demand Law: vmstat

in	sy	cs	us	sy	idle
119	65	24	1	0	99
296	2491	289	13	6	81
260	5586	213	44	7	49
326	2822	474	21	7	72
352	1913	271	13	4	83
304	2058	280	17	5	78
275	3072	506	21	7	72
322	3340	417	18	8	74
301	2000	201	9	3	87
261	1952	282	10	4	86
251	1870	220	9	4	87
412	4646	763	33	12	54
					76.83

Interval:
12*5sec= 60 sec

Number of Requests:
20

$$U_{cpu} = 1 - 0.7683 = 0.232 = 23.2\%$$

$$X_0 = 20 / 60 = 0.333 \text{ requests/sec}$$

$$D_{cpu} = \frac{U_{cpu}}{X_0} = 0.232 / 0.333 = 0.695 \text{ sec}$$

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?

Service Demand Law: example

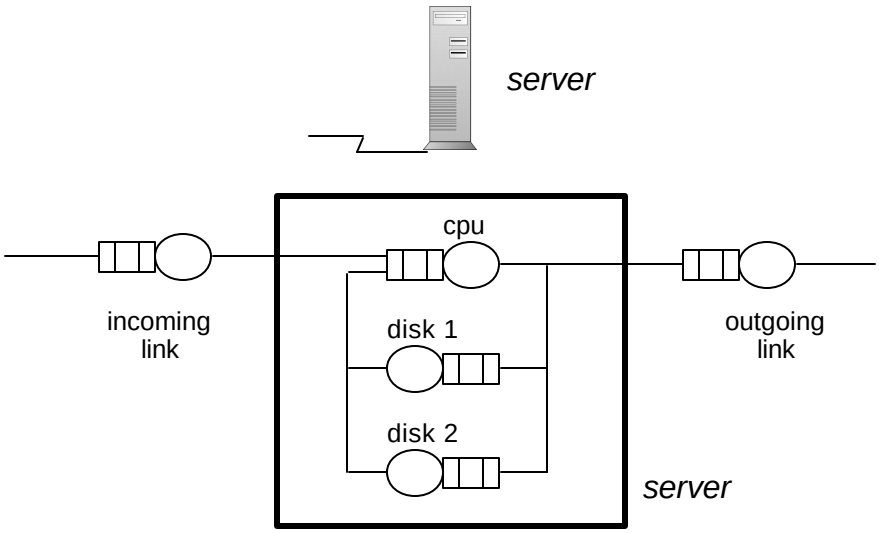
- A Web server was monitored for 10 minutes. It was observed that the CPU was 90% busy during the monitoring period. The number of HTTP requests counted in the log was 30,000.
- What is the CPU service demand of an HTTP request?

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

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

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

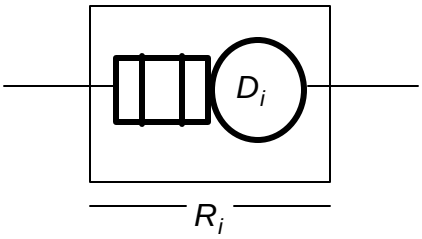
An Open Queuing Model Example



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.

Open QN Models
Computing Residence Times



Service demand at resource i

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

Utilization of resource i (U_i)

Derivation of Residence Time

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

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

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

$$R_i = S_i + SX_i R_i = S_i + U_i R_i$$

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

multiplying both sides by V_i:

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

Open Model Equations

$$U_i = \lambda \times D_i$$

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

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

Bound on Throughput

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

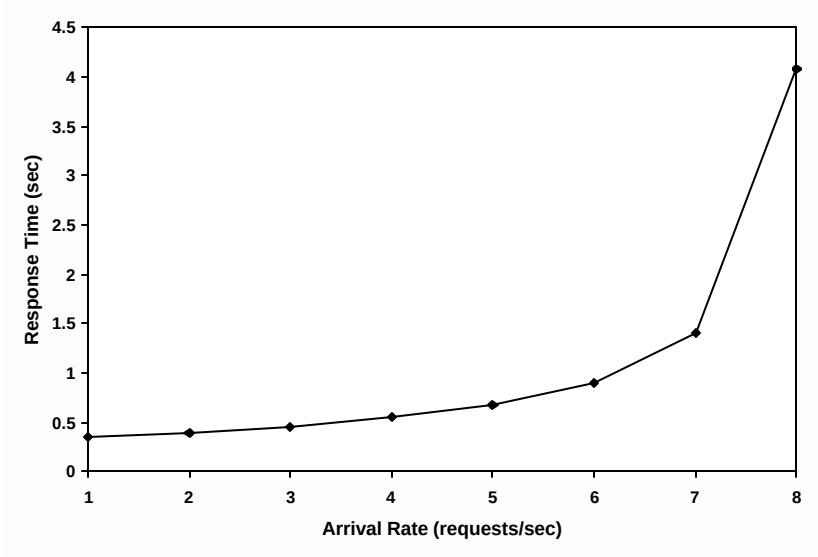
  (Hint: the utilization cannot exceed 100%)

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.

31

Response vs. Arrival Rate



32

Equations for Open Multiple Class QN Models

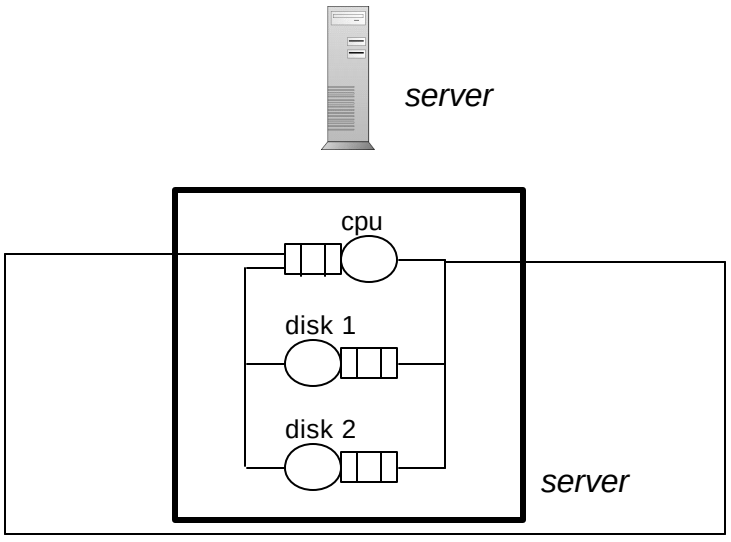
$$U_{i,r} = l \times D_{i,r}$$

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

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

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

A Closed Queuing Model Example

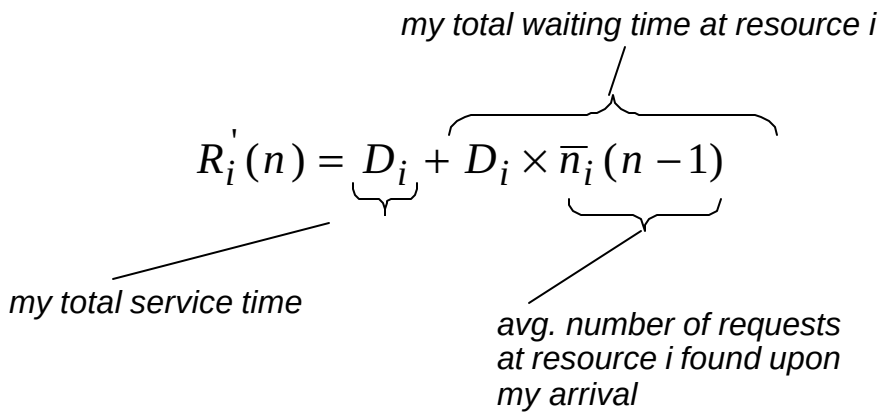


Closed QN Models

- 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)

Closed QN Model MVA Equations

- Residence Time Equation:



Closed QN Model
MVA Equations

- Residence Time Equation:

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

Closed QN Model
MVA Equations

- Throughput Equation. Using Little's Law:

throughput

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

total response time

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

Closed QN Model
MVA Equations

- Throughput Equation:

$$X_o(n) = \frac{n}{R_0(n)} = \frac{n}{\sum_{i=1}^{K} R_i'(n)}$$

Closed QN Model
MVA Equations

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

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

MVA Equations

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

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

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

Solving the Model

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

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

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

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

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

Solving the Model

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

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

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

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

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

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%?