Mean Value Analysis and Related Techniques

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

Audio/Video recordings of this lecture are available at:

http://www.cse.wustl.edu/~jain/cse567-15/

Overview

1. Analysis of Open Queueing Networks
2. Mean-Value Analysis
3. Approximate MVA
4. Balanced Job Bounds

Analysis of Open Queueing Networks

- Used to represent transaction processing systems, such as airline reservation systems, or banking systems.
- Transaction arrival rate is not dependent on the load on the computer system.
- Arrivals are modeled as a Poisson process with a mean arrival rate λ.
- Exact analysis of such systems
- **Assumption**: All devices in the system can be modeled as either **fixed-capacity service centers** (single server with exponentially distributed service time) or **delay centers** (infinite servers with exponentially distributed service time).

Analysis of Open Queueing Networks

- For all fixed capacity service centers in an open queueing network, the response time is:

$$R_i = S_i\,(1+Q_i)$$

- On arrival at the i^{th} device, the job sees Q_i jobs ahead (including the one in service) and expects to wait $Q_i\,S_\text{i}$ seconds. Including the service to itself, the job should expect a total response time of $S_i(1+Q_i)$.
- **Assumption**: Service is memory-less (not operationally testable) $\Rightarrow$ Not an operational law
- Without the memory-less assumption, we would also need to know the time that the job currently in service has already consumed.

Mean Performance

- Assuming *job flow balance*, the throughput of the system is equal to the arrival rate:

$$X = \lambda$$

- The throughput of i^{th} device, using the *forced flow law* is:

$$X_i = X\,V_i$$

- The utilization of the i^{th} device, using the *utilization law* is:

$$U_i = X_i\,S_i = X\,V_i\,S_i = \lambda\,D_i$$

- The queue length of i^{th} device, using Little's law is:

$$Q_i = X_i\,R_i = X_i\,S_i(1+Q_i) = U_i(1+Q_i)$$

Or $Q_i = U_i/(1-U_i)$

- Notice that the above equation for Q_i is identical to the equation for M/M/1 queues.

Mean Performance

- The device response times are:

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

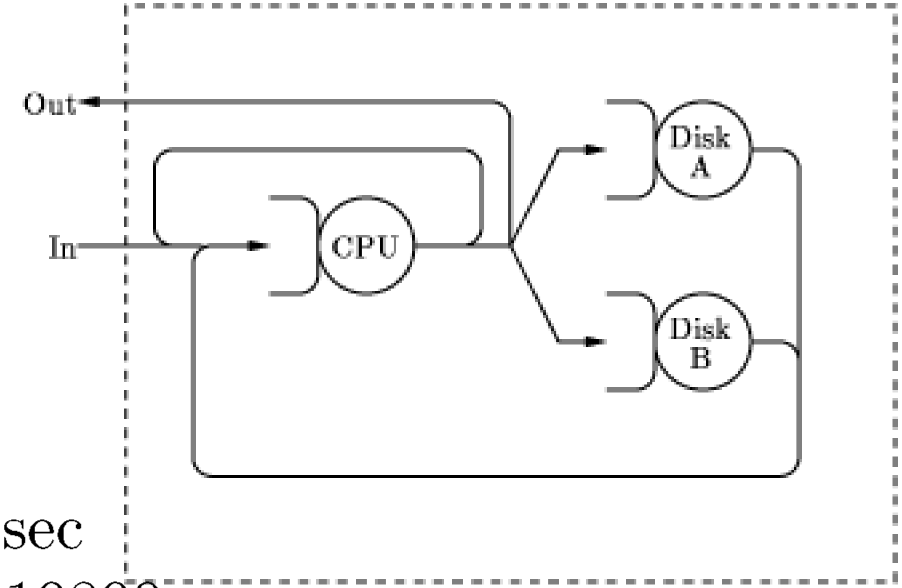
- In delay centers, there are infinite servers and, therefore:

$$\begin{aligned} R_i &= S_i \\ Q_i &= R_i X_i = S_i X V_i = X D_i = U_i \end{aligned}$$

- Notice that the utilization of the delay center represents the mean number of jobs receiving service and does not need to be less than one.

Example 34.1

- File server consisting of a CPU and two disks, A and B.
- With 6 clients systems:



Observation interval = 3600 sec

Number of client requests = 10800

CPU busy time = 1728 sec

Disk A busy time = 1512 sec

Disk B busy time = 2592 sec

Number of visits (I/O requests) to Disk A = 75600

Number of visits (I/O requests) to Disk B = 86400

Example 34.1 (Cont)

$$
\begin{aligned}
X &= \text{Throughput} = 10800/3600 \\
  &= \text{3 client requests per sec} \\
V_A &= 75600/10800 = \text{7 visits per client request to disk A} \\
V_B &= 86400/10800 = \text{8 visits per client request to disk B} \\
V_{CPU} &= 1+7+8 \\
  &= \text{16 visits per client requests to CPU} \\
D_{CPU} &= 1728/10800 = \text{0.16 sec of CPU time per client request} \\
D_A &= 1512/10800 \\
  &= \text{0.14 sec of disk A time per client request} \\
D_B &= 2592/10800 \\
  &= \text{0.24 sec of disk B time per client request} \\
S_{CPU} &= 0.16/16 = \text{0.01 sec per visit to CPU} \\
S_A &= 0.14/7 = \text{0.02 sec per visit to disk A} \\
S_B &= 0.24/8 = \text{0.03 sec per visit to disk B}
\end{aligned}
$$

Example 34.1 (Cont)

- Device utilizations using the utilization law are:

$$\begin{aligned} U_{CPU} &= XD_{CPU} = 3 \times 0.16 = 0.48 \\ U_A &= XD_A = 3 \times 0.14 = 0.42 \\ U_B &= XD_B = 3 \times 0.24 = 0.72 \end{aligned}$$

Example 34.1 (Cont)

- The device response times using Equation 34.2 are:

$$\begin{aligned}
R_{CPU} &= S_{CPU}/(1-U_{CPU}) \\
&= 0.01/(1-0.48) = 0.0192 \text{ sec} \\
R_A &= S_A/(1-U_A) = 0.02/(1-0.42) = 0.0345 \text{ sec} \\
R_B &= S_B/(1-U_B) = 0.03/(1-0.72) = 0.107 \text{ sec}
\end{aligned}$$

- Server response time:

$$\begin{aligned}
R &= \sum V_i R_i \\
&= 16 \times 0.0192 + 7 \times 0.0345 + 8 \times 0.107 \\
&= 1.406 \text{ sec}
\end{aligned}$$

- We can quantify the impact of the following changes:

Example 34.1 (Cont)

- Q: What if we increase the number of clients to 8?
  ⇒ Request arrival rate will go up by a factor of 8/6.

$$
\begin{aligned}
X &= 4 \text{ requests/sec} \\
U_{CPU} &= XD_{CPU} = 4 \times 0.16 = 0.64 \\
U_A &= XD_A = 4 \times 0.14 = 0.56 \\
U_B &= XD_B = 4 \times 0.24 = 0.96 \\
R_{CPU} &= S_{CPU}/(1 - U_{CPU}) \\
&= 0.01/(1 - 0.64) = 0.0278 \text{ sec} \\
R_A &= S_A/(1 - U_A) = 0.02/(1 - 0.56) = 0.0455 \text{ sec} \\
R_B &= S_B/(1 - U_B) = 0.03/(1 - 0.96) = 0.75 \text{ sec} \\
R &= 16 \times 0.0278 + 7 \times 0.0455 + 8 \times 0.75 = 6.76 \text{ sec}
\end{aligned}
$$

- Conclusion: Server response time will degrade by a factor of 6.76/1.406= 4.8

Example 34.1 (Cont)

- Q: What if we use a cache for disk B with a hit rate of 50%, although it increases the CPU overhead by 30% and the disk B service time (per I/O) by 10%.
- A:

$$\begin{aligned} V_B &= 0.5 \times 8 = 4 \\ S_{CPU} &= 1.3 \times 0.01 = 0.013 \Rightarrow D_{CPU} = 0.208 \text{ sec} \\ S_B &= 1.1 \times 0.03 = 0.033 \Rightarrow D_B = 4 \times 0.033 = 0.132 \text{ sec} \end{aligned}$$

Example 34.1 (Cont)

- The analysis of the changed systems is as follows:

$$\begin{aligned}
U_{CPU} &= XD_{CPU} = 3 \times 0.208 = 0.624 \\
U_A &= XD_A = 3 \times 0.14 = 0.42 \\
U_B &= XD_B = 3 \times 0.132 = 0.396 \\
R_{CPU} &= S_{CPU}/(1 - U_{CPU}) = 0.013/(1 - 0.624) \\
&= 0.0346 \text{ sec} \\
R_A &= S_A/(1 - U_A) = 0.02/(1 - 0.42) = 0.0345 \text{ sec} \\
R_B &= S_B/(1 - U_B) = 0.033/(1 - 0.396) = 0.0546 \text{ sec} \\
R &= 16 \times 0.0346 + 7 \times 0.0345 + 4 \times 0.0546 = 1.013 \text{ sec}
\end{aligned}$$

- Thus, if we use a cache for Disk B, the server response time will improve by *(1.406-1.013)/1.406 = 28%.*

Example 34.1 (Cont)

- Q: What if we have a lower cost server with only one disk (disk A) and direct all I/O requests to it?

$$
\begin{aligned}
V_B &= 0 \\
V_A &= 7+8=15 \\
D_{CPU} &= 0.16 \text{ sec (as before)} \\
D_A &= 15 \times 0.02 = 0.3 \text{ sec} \\
U_{CPU} &= XD_{CPU} = 3 \times 0.16 = 0.48 \\
U_A &= XD_A = 3 \times 0.3 = 0.90 \\
R_{CPU} &= S_{CPU}/(1-U_{CPU}) = 0.01/(1-0.48) = 0.0192 \text{ sec} \\
R_A &= S_A/(1-U_A) = 0.02/(1-0.90) = 0.2 \text{ sec} \\
R &= 16 \times 0.0192 + 15 \times 0.2 = 3.31 \text{ sec}
\end{aligned}
$$

- A: the server response time will degrade by a factor of 3.31/1.406 = 2.35

MVA (Cont)

- Mean-value analysis (**MVA**) allows solving closed queueing networks in a manner similar to that used for open queueing networks
- It gives the mean performance. The variance computation is not possible using this technique.
- Initially limit to fixed-capacity service centers. Delay centers are considered later. Load-dependent service centers are also considered later.
- Given a closed queueing network with N jobs:

$$R_i(N) = S_i\,(1+Q_i(N-1))$$

- Here, $Q_i(N-1)$ is the mean queue length at i^{th} device with $N-1$ jobs in the network.
- It assumes that the service is memoryless

MVA (Cont)

- Since the performance with no users (N=0) can be easily computed, performance for any number of users can be computed iteratively.
- Given the response times at individual devices, the system response time using the general response time law is:

$$R(N) = \sum_{i=1}^{M} V_i R_i(N)$$

- The system throughput using the interactive response time law is:

$$X(N) = \frac{N}{R(N) + Z}$$

MVA (Cont)

- The device throughputs measured in terms of jobs per second are:

$$X_i(N) = X(N)\, V_i$$

- The device queue lengths with N jobs in the network using Little's law are:

$$Q_i(N) = X_i(N)\, R_i(N) = X(N)\, V_i\, R_i(N)$$

- Response time equation for delay centers is simply:

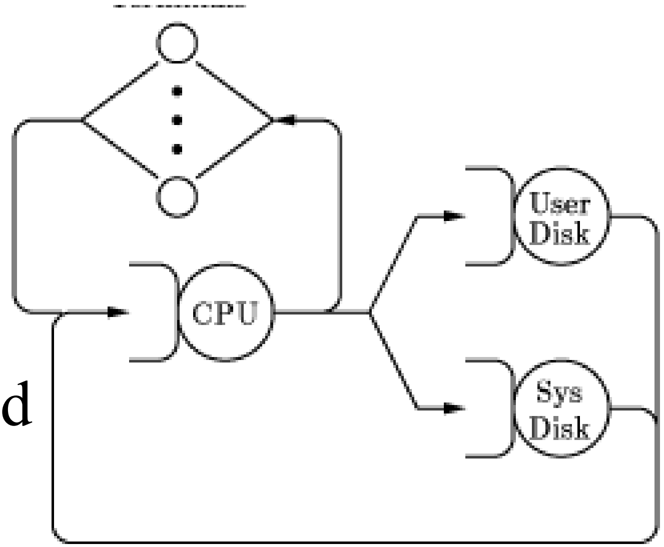
$$R_i(N) = S_i$$

- Earlier equations for device throughputs and queue lengths apply to delay centers as well.

$$Q_i(0) = 0$$

Example 34.2

- Consider a timesharing system
- Each user request makes ten I/O requests to disk A, and five I/O requests to disk B.
- The service times per visit to disk A and disk B are 300 and 200 milliseconds, respectively.
- Each request takes two seconds of CPU time and the user think time is four seconds.



$$S_A = 0.3,\ V_A = 10 \Rightarrow D_A = 3$$
$$S_B = 0.2,\ V_B = 5 \Rightarrow D_B = 1$$
$$D_{CPU} = 2,\ V_{CPU} = V_A + V_B + 1 = 16 \Rightarrow S_{CPU} = 0.125$$
$$Z = 4,\ \text{and}\ N = 20$$

Example 34.2 (Cont)

- Initialization:
  - Number of users: $N=0$
  - Device queue lengths: $Q_{CPU}=0$, $Q_A=0$, $Q_B = 0$
- Iteration 1:
  - Number of users: N=1
  - Device response times:

$$R_{CPU} = S_{CPU}(1 + Q_{CPU}) = 0.125(1+0) = 0.125$$
$$R_A = S_A(1+Q_A) = 0.3(1+0) = 0.3$$
$$R_B = S_B(1+Q_B) = 0.2(1+0) = 0.2$$

  - System Response time:

$$\begin{aligned} R &= R_{CPU}V_{CPU} + R_AV_A + R_BV_B \\ &= 0.125 \times 16 + 0.3 \times 10 + 0.2 \times 5 = 6 \end{aligned}$$

Example 34.2 (Cont)

- System Throughput: *X=N/(R+Z)=1/(6+4)=0.1*
- Device queue lengths:

$$Q_{CPU} = XR_{CPU}V_{CPU} = 0.1 \times 0.125 \times 16 = 0.2$$
$$Q_A = XR_AV_A = 0.1 \times 0.3 \times 10 = 0.3$$
$$Q_B = XR_BV_B = 0.1 \times 0.2 \times 5 = 0.1$$

- Iteration 2:
  - Number of users: *N=2*
  - Device response times:

$$R_{CPU} = S_{CPU}(1 + Q_{CPU}) = 0.125(1 + 0.2) = 0.15$$
$$R_A = S_A(1 + Q_A) = 0.3(1 + 0.3) = 0.39$$
$$R_B = S_B(1 + Q_B) = 0.2(1 + 0.1) = 0.22$$

Example 34.2 (Cont)

- System Response time:

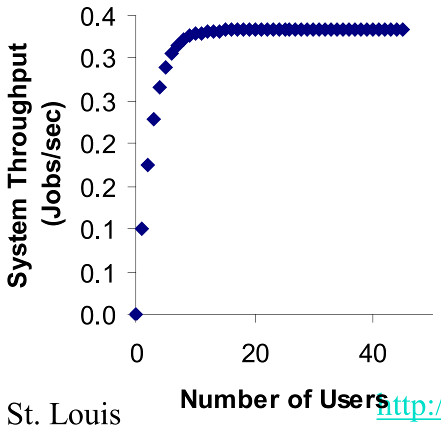
$$\begin{aligned} R &= R_{CPU}V_{CPU} + R_A V_A + R_B V_B \\ &= 0.15 \times 16 + 0.39 \times 10 + 0.22 \times 5 = 7.4 \end{aligned}$$

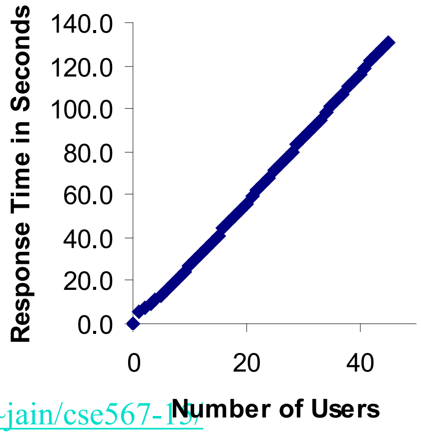
- System Throughput: X=N/(R+Z)=2/(7.4+4)=0.175
- Device queue lengths:

$$Q_{CPU} = X R_{CPU} V_{CPU} = 0.175 \times 0.15 \times 16 = 0.421$$
$$Q_A = X R_A V_A = 0.175 \times 0.39 \times 10 = 0.684$$
$$Q_B = X R_B V_B = 0.175 \times 0.22 \times 5 = 0.193$$

MVA Results for Example 34.2

Iteration	Response Time				System	Queue Lengths		
#	CPU	Disk A	Disk B	System	Throughput	CPU	Disk A	Disk B
1	0.125	0.300	0.200	6.000	0.100	0.200	0.300	0.100
2	0.150	0.390	0.220	7.400	0.175	0.421	0.684	0.193
3	0.178	0.505	0.239	9.088	0.229	0.651	1.158	0.273
4	0.206	0.647	0.255	11.051	0.266	0.878	1.721	0.338
5	0.235	0.816	0.268	13.256	0.290	1.088	2.365	0.388
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
17	0.370	3.962	0.300	47.045	0.333	1.974	13.195	0.499
18	0.372	4.259	0.300	50.032	0.333	1.981	14.187	0.499
19	0.373	4.556	0.300	53.022	0.333	1.987	15.181	0.500
20	0.373	4.854	0.300	56.016	0.333	1.991	16.177	0.500





Box 34.2: MVA Algorithms

Inputs:

N = number of users
Z = think time
M = number of devices
S_i = service time/visit to ith device
V_i = number of visits to ith device

Outputs:

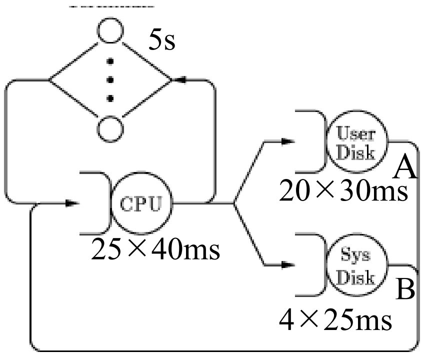
X = system throughput
Q_i = average # of jobs at ith device
R_i = response time of ith device
R = system response time
U_i = utilization of the ith device

```
Initialization: FOR $i = 1$ TO $M$ DO $Q_i = 0$

Iterations:
FOR $n = 1$ TO $N$ DO
BEGIN
    FOR $i = 1$ TO $M$ DO $R_i = \begin{cases} S_i(1+Q_i) & \text{Fixed capacity} \\ S_i & \text{Delay centers} \end{cases}$
    $R = \sum_{i=1}^{M} R_i V_i$
    $X = \frac{n}{Z+R}$
    FOR $i = 1$ TO $M$ DO $Q_i = X V_i R_i$
END
Device throughputs: $X_i = X V_i$
Device utilizations: $U_i = X S_i V_i$
```

Homework 34A: MVA

Part 1: Fill in the rows for N=0 and N=1 only.



$$R_i = S_i(1 + Q_i) \qquad R = \sum_{i=1}^{M} R_i V_i$$

$$X = \frac{N}{Z+R} \qquad Q_i = X V_i R_i$$

V_i	25	20	4					Z=5				
S_i	0.04	0.03	0.025									
N	R_C	R_A	R_B	$V_C R_C$	$V_A R_A$	$V_B R_B$	R	R+Z	X	Q_C	Q_A	Q_B
0	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _
1	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _
2	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _	_____ _

Part 2: Fill in the row for N=2.

MVA Assumptions

- MVA is applicable only if the network is a product form network.
- This means that the network should satisfy the conditions of job flow balance, one step behavior, and device homogeneity.
- Also assumes that all service centers are either fixed-capacity service centers or delay centers.
- In both cases, we assumed exponentially distributed service times.

Homework 34B

For a timesharing system with two disks (user and system), the probabilities for jobs completing the service at the CPU were found to be 0.**75** to disk A, **0.15** to disk B, and **0.1** to the terminals. The user think time was measured to be 5 seconds, the disk service times are 30 milliseconds and 25 milliseconds, while the average service time per visit to the CPU was 40 milliseconds.

Using the queueing network model shown in Figure 32.8:

- Use MVA to compute system throughput and response time for N=1,…,5 interactive users

Balanced Job Bounds

- A system without a bottleneck device is called a balanced system.
- Balanced system has a better performance than a similar unbalanced system
  ⇒ Allows getting two sided bounds on performance
- An unbalanced system's performance can always be improved by replacing the bottleneck device with a faster device.
- Balanced System: Total service time demands on all devices are equal.

Balanced Job Bounds (Cont)

- Thus, the response time and throughput of a time-sharing system can be bounded as follows:

$$\max\left\{ND_{max}-Z, D+(N-1)D_{avg}\frac{D}{D+Z}\right\} \leq R(N) \leq D+(N-1)D_{max}\frac{(N-1)D}{(N-1)D+Z}$$

$$\frac{N}{Z+D+(N-1)D_{max}\frac{(N-1)D}{(N-1)D+Z}} \leq X(N) \leq \min\left\{\frac{1}{D_{max}}, \frac{N}{Z+D+(N-1)D_{avg}\frac{D}{D+Z}}\right\}$$

- Here, $D_{avg}=D/M$ is the average service demand per device.
- These equations are known as **balanced job bounds**.
- These bounds are very tight in that the upper and lower bound are very close to each other and to the actual performance.
- For batch systems, the bounds can be obtained by substituting $Z=0$

Balanced Job Bounds (Cont)

- Assumption: All service centers except terminals are fixed-capacity service centers.
- Terminals are represented by delay centers. No other delay centers are allowed because the presence of delay centers invalidates several arguments related to D_{max} and D_{avg}.

Derivation of Balanced Job Bounds

Steps:

1. Derive an expression for the throughput and response time of a balanced system.
2. Given an unbalanced system, construct a corresponding `best case' balanced system such that the number of devices is the same and the sum of demands is identical in the balanced and unbalanced systems. This produces the upper bounds on the performance.
3. Construct a corresponding `worst case' balanced system such that each device has a demand equal to the bottleneck and the number of devices is adjusted to make the sum of demands identical in the balanced and unbalanced systems. This produces the lower bounds on performance.

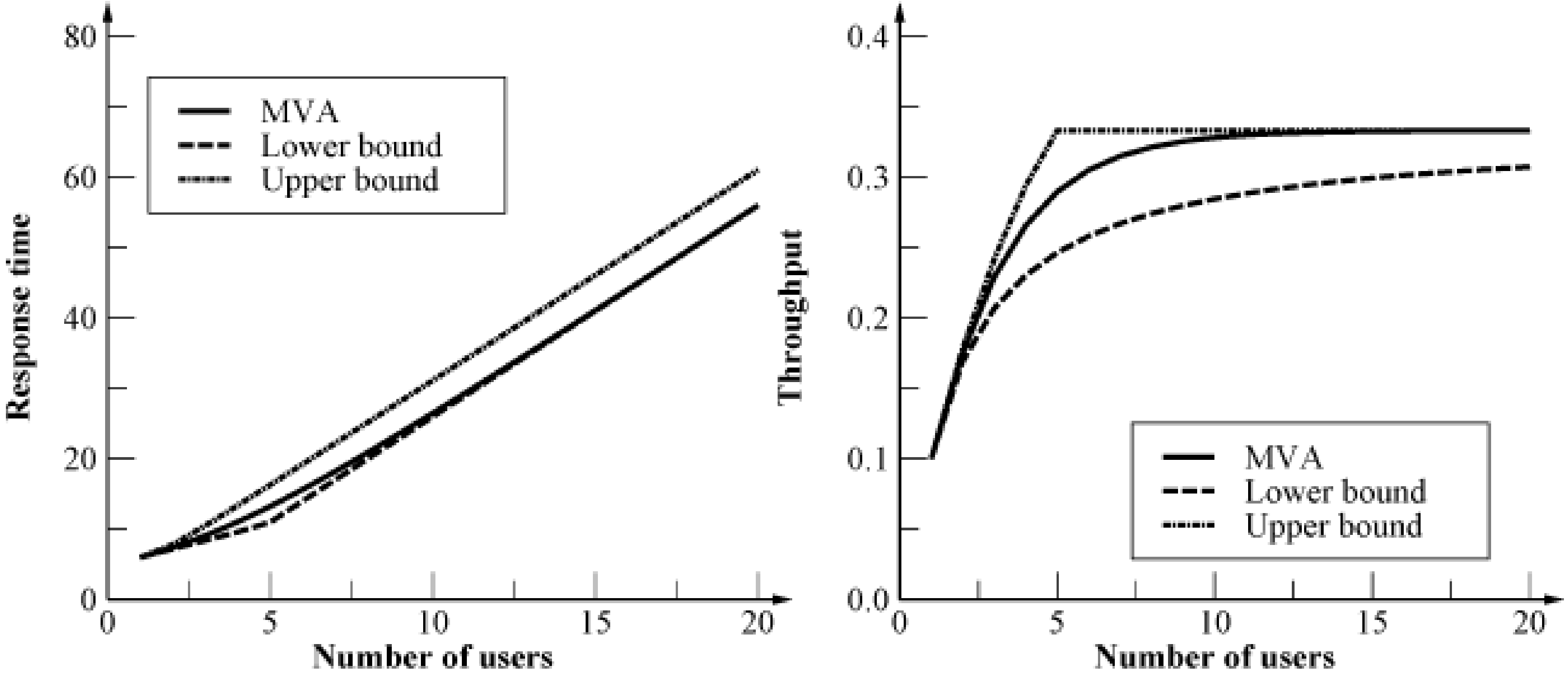
Example 34.4

- For the timesharing system of Example 34.1
  - $D_{CPU} = 2\ ,\ \ D_A = 3\ ,\ \ D_B = 1\ ,\ \ Z = 4$
  - $D = D_{CPU} + D_A + D_B = 2+3+1 = 6$
  - $D_{avg} = D/3 = 2$
  - $D_{max} = D_A = 3$
- The balanced job bounds are:

$$\frac{N}{4+6+(N-1)3\frac{6(N-1)}{6(N-1)+4}} \leq X(N) \leq \min\left\{\frac{1}{3}, \frac{N}{4+6+(N-1)2\frac{6}{6+4}}\right\}$$

$$\max\left\{3N-4, 6+(N-1)2\frac{6}{6+4}\right\} \leq R(N) \leq 6+(N-1)3\frac{6(N-1)}{6(N-1)+4}$$

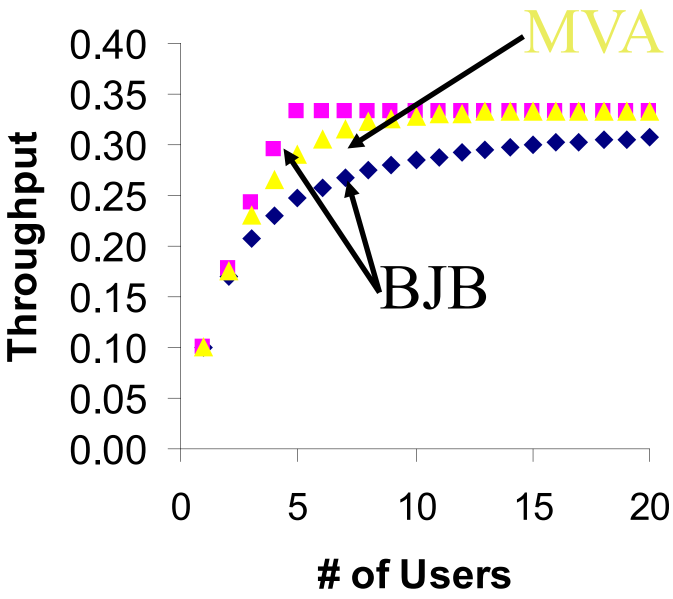
Example 34.4 (Cont)

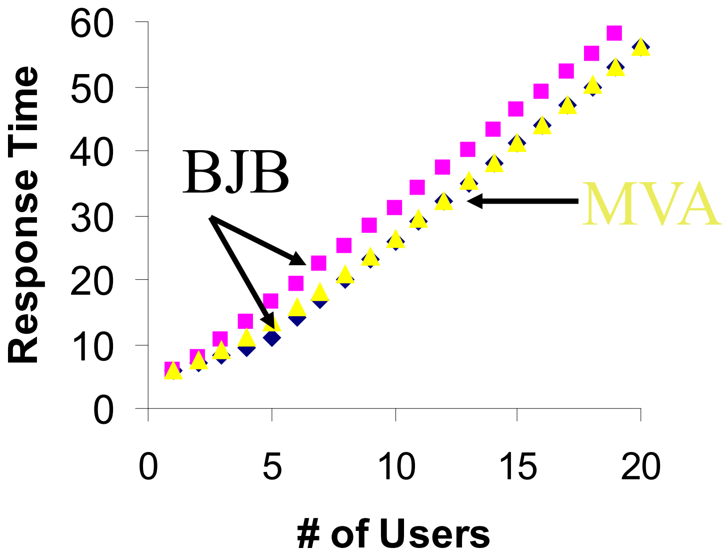


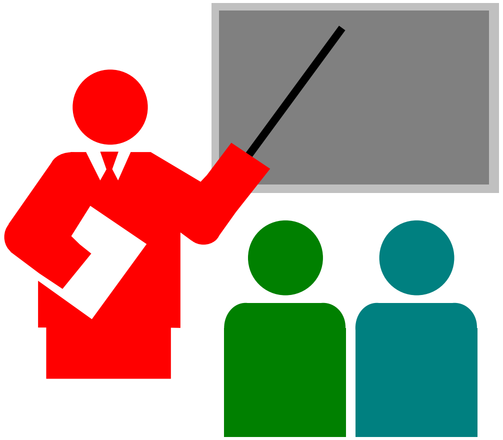
Example 34.4 (Cont)

N	Response Time			Throughput		
	Lower		Upper	Lower		Upper
#	BJB	MVA	BJB	BJB	MVA	BJB
1	6.000	6.000	6.000	0.100	0.100	0.100
2	7.200	7.400	7.800	0.169	0.175	0.179
3	8.400	9.088	10.500	0.207	0.229	0.242
4	9.600	11.051	13.364	0.230	0.266	0.294
5	11.000	13.256	16.286	0.246	0.290	0.333
⋮	⋮	⋮	⋮	⋮	⋮	⋮
15	41.000	41.089	46.091	0.299	0.333	0.333
16	44.000	44.064	49.085	0.301	0.333	0.333
17	47.000	47.045	52.080	0.303	0.333	0.333
18	50.000	50.032	55.075	0.305	0.333	0.333
19	53.000	53.022	58.071	0.306	0.333	0.333
20	56.000	56.016	61.068	0.307	0.333	0.333

Example 34.4 (Cont)







Summary

1. Open queueing networks of M/M/1 or M/M/∞ can be analyzed exactly
2. MVA allows exact analysis of closed queueing networks. Given performance of N-1 users, get performance for N users.
3. Balanced Job bounds: A balanced system with $D_i = D_{avg}$ will have better performance and an unbalanced system with some devices at D_{max} and others at 0 will have worse performance.

Homework 34C

For the data of homework 34A:

- Write the expressions for balanced job bounds on the system throughput and response time of the system and compute the bounds for up to N=30 users.