Operational Laws

Raj Jain
Washington University in Saint Louis
Jain@eecs.berkeley.edu or Jain@wustl.edu

A Mini-Course offered at UC Berkeley, Sept-Oct 2012

These slides and audio/video recordings are available on-line at:

http://amplab.cs.berkeley.edu/courses/queue

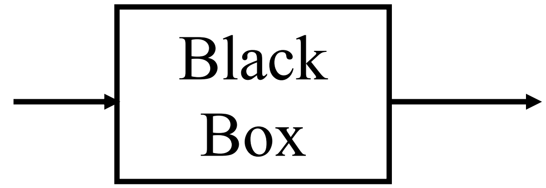
and http://www.cse.wustl.edu/~jain/queue

- What is an Operational Law?
  1. Utilization Law
  2. Forced Flow Law
  3. Little's Law
  4. General Response Time Law
  5. Interactive Response Time Law
  6. Bottleneck Analysis

Operational Laws

- Relationships that do not require any assumptions about the distribution of service times or inter-arrival times.
- Identified originally by Buzen (1976) and later extended by Denning and Buzen (1978).
- **Operational** $\Rightarrow$ Directly measured.
- **Operationally testable assumptions** $\Rightarrow$ assumptions that can be verified by measurements.
  - For example, whether number of arrivals is equal to the number of completions?
  - This assumption, called job flow balance, is operationally testable.
  - A set of observed service times is or is not a sequence of independent random variables is not is not operationally testable.

Operational Quantities



- Quantities that can be directly measured during a finite observation period.
- T = Observation interval — A_i = Number of arrivals
- C_i = Number of completions — B_i = Busy time B_i

$$\text{Arrival Rate } \lambda_i = \frac{\text{Number of arrivals}}{\text{Time}} = \frac{A_i}{T}$$

$$\text{Throughput } X_i = \frac{\text{Number of completions}}{\text{Time}} = \frac{C_i}{T}$$

$$\text{Utilization } U_i = \frac{\text{Busy Time}}{\text{Total Time}} = \frac{B_i}{T}$$

$$\text{Mean service time } S_i = \frac{\text{Total time Served}}{\text{Number served}} = \frac{B_i}{C_i}$$

Utilization Law

$$
\begin{aligned}
\text{Utilization } U_i &= \frac{\text{Busy Time}}{\text{Total Time}} = \frac{B_i}{T} \\
&= \frac{C_i}{T} \times \frac{B_i}{C_i} = \frac{\text{Completions}}{\text{Time}} \times \frac{\text{Busy Time}}{\text{Completions}} \\
&= \text{Throughput} \times \text{Mean Service Time} = X_i S_i
\end{aligned}
$$

- This is one of the operational laws
- Operational laws are similar to the elementary laws of motion For example,

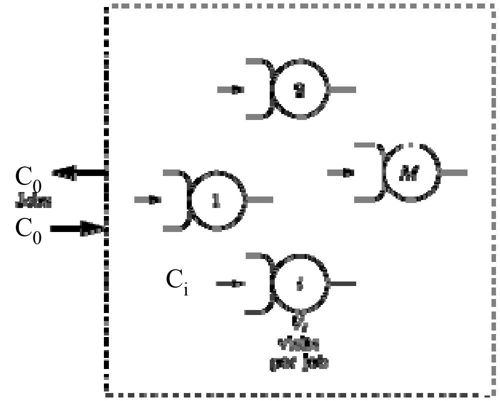
$$d = \frac{1}{2} a t^2$$

- Notice that distance d, acceleration a, and time t are **operational quantities**. No need to consider them as expected values of random variables or to assume a distribution.

Example 33.1

- Consider a network gateway at which the packets arrive at a rate of *125* packets per second and the gateway takes an average of two milliseconds to forward them.
- Throughput X_i = Exit rate = Arrival rate = 125 packets/second
- Service time S_i = 0.002 second
- Utilization $U_i = X_i S_i = 125 \times 0.002 = 0.25 = 25\%$
- This result is valid for any arrival or service process. Even if inter-arrival times and service times to are not IID random variables with exponential distribution.

Forced Flow Law



- Relates the system throughput to individual device throughputs.
- In an open model,
  System throughput = # of jobs leaving the system per unit time
- In a closed model, System throughput = # of jobs traversing OUT to IN link per unit time.
- If observation period T is such that $A_i = C_i$
  $\Rightarrow$ Device satisfies the assumption of *job flow balance*.
- Each job makes V_i requests for i^{th} device in the system
- $C_i = C_0 V_i$ or $V_i = C_i/C_0$ V_i is called visit ratio
- $$\text{System throughput } X = \frac{\text{Jobs completed}}{\text{Total time}} = \frac{C_0}{T}$$

Forced Flow Law (Cont)

- Throughput of i^{th} device:

$$\text{Device Throughput } X_i = \frac{C_i}{T} = \frac{C_i}{C_0} \times \frac{C_0}{T}$$

- In other words:

$$X_i = XV_i$$

- This is the **forced flow law**.

Bottleneck Device

- Combining the forced flow law and the utilization law, we get:

$$\text{Utilization of } i^{\text{th}} \text{ device } U_i = X_i S_i = X V_i S_i$$

$$U_i = X D_i$$

- Here $D_i = V_i S_i$ is the total service demand on the device for all visits of a job.
- The device with the highest D_i has the highest utilization and is the **bottleneck device**.

Example 33.2

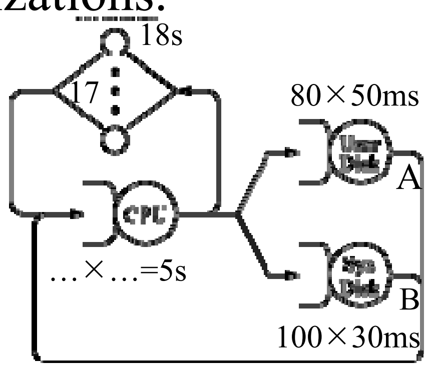
- In a timesharing system, accounting log data produced the following profile for user programs.
  - Each program requires five seconds of CPU time, makes 80 I/O requests to the disk A and 100 I/O requests to disk B.
  - Average think-time of the users was 18 seconds.
  - From the device specifications, it was determined that disk A takes 50 milliseconds to satisfy an I/O request and the disk B takes 30 milliseconds per request.
  - With 17 active terminals, disk A throughput was observed to be 15.70 I/O requests per second.
- We want to find the system throughput and device utilizations.

$D_{CPU} = 5$ seconds $\quad V_A = 80,$

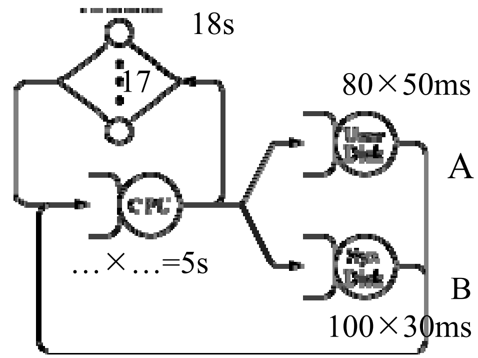
$V_B = 100, \quad Z = 18$ seconds,

$S_A = 0.050$ seconds, $\quad S_B = 0.030$ seconds,

$N = 17,$ and $\quad X_A = 15.70$ jobs/second



Example 33.2 (Cont)



$D_{CPU} = 5$ seconds, $V_A = 80$, $V_B = 100$, $Z = 18$ seconds, $S_A = 0.050$ seconds, $S_B = 0.030$ seconds, $N = 17$, and $X_A = 15.70$ jobs/second

- Since the jobs must visit the CPU before going to the disks or terminals, the CPU visit ratio is: $V_{CPU} = V_A + V_B + 1 = 181$

$$\begin{aligned} D_{CPU} &= 5 \text{ seconds} \\ D_A &= S_A V_A = 0.050 \times 80 = 4 \text{ seconds} \\ D_B &= S_B V_B = 0.030 \times 100 = 3 \text{ seconds} \end{aligned}$$

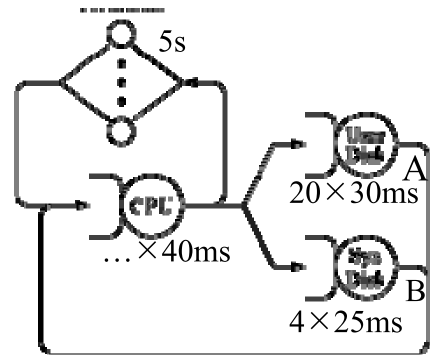
- Using the forced flow law, the throughputs are:

$$\begin{aligned} X &= \frac{X_A}{V_A} = \frac{15.70}{80} = 0.1963 \text{ jobs/second} \\ X_{CPU} &= X V_{CPU} = 0.1963 \times 181 \\ &= 35.48 \text{ requests/second} \\ X_B &= X V_B = 0.1963 \times 100 \\ &= 19.6 \text{ requests/second} \end{aligned}$$

- Using the utilization law, the device utilizations are:

$$\begin{aligned} U_{CPU} &= X D_{CPU} = 0.1963 \times 5 = 98\% \\ U_A &= X D_A = 0.1963 \times 4 = 78.4\% \\ U_B &= X D_B = 0.1963 \times 3 = 58.8\% \end{aligned}$$

Quiz 33A



- The visit ratios and service time per visit for a system are as shown:
- For each device what is the total service demand:
  - CPU: V_i = ________, S_i= ________, D_i= ________
  - Disk A: V_i = ________, S_i= ________, D_i= ________
  - Disk B: V_i = ________, S_i= ________, D_i= ________
  - Terminals: V_i = ________, S_i= ________, D_i= ________
- If disk A utilization is 50%, what's the utilization of CPU and Disk B?
  - $X_A = U_A/D_A$ =________
  - $U_{CPU} = X\,D_{CPU}$ = ________
  - $U_B = X\,D_B$ = ________
- What is the bottleneck device? ________

Key: $U_i=X_iS_i=XD_i$, $D_i=S_iV_i$, $X= X_i/V_i$

Transition Probabilities

- p_{ij} = Probability of a job moving to j^{th} queue after service completion at i^{th} queue
- Visit ratios and transition probabilities are equivalent in the sense that given one we can always find the other.
- In a system with job flow balance: $C_j = \sum_{i=0}^{M} C_i p_{ij}$

$i = 0 \Rightarrow$ visits to the outside link

- p_{i0} = Probability of a job exiting from the system after completion of service at i^{th} device
- Dividing by C_0 we get:

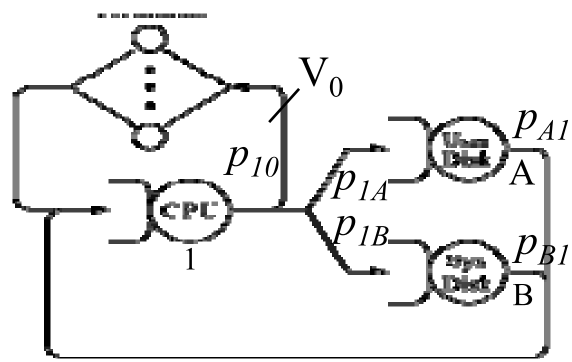
$$V_j = \sum_{i=0}^{M} V_i p_{ij}$$

Transition Probabilities (Cont)

- Since each visit to the outside link is defined as the completion of the job, we have: $V_0 = 1$
- These are called **visit ratio equations**
- In central server models, after completion of service at every queue, the jobs always move back to the CPU queue:

$$p_{i1} = 1 \quad \forall i \neq 1$$

$$p_{ij} = 0 \quad \forall i, j \neq 1$$

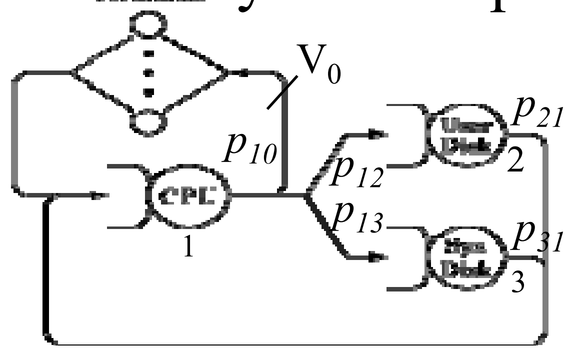


Transition Probabilities (Cont)

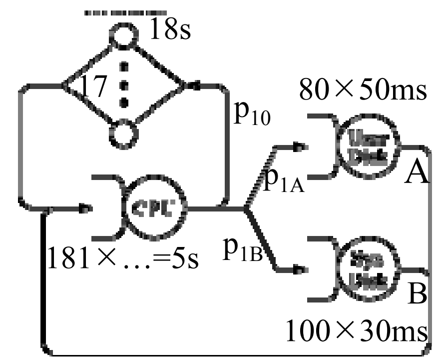
- The above probabilities apply to exit and entrances from the system (i=0), also. Therefore, the visit ratio equations become:

$$
\begin{aligned}
1 &= V_1 p_{10} \quad \Rightarrow \quad V_1 = \frac{1}{p_{10}} \\
V_1 &= 1 + V_2 + V_3 + \cdots + V_M \\
V_j &= V_1 p_{1j} = \frac{p_{1j}}{p_{10}} \quad j = 2, 3, \ldots, M
\end{aligned}
$$

- Thus, we can find the visit ratios by dividing the probability p_{1j} of moving to j^{th} queue from CPU by the exit probability p_{10}.



Example 33.3



- Consider the queueing network:
- The visit ratios are V_A=80, V_B=100, and V_{CPU}=181.
- After completion of service at the CPU the probabilities of the job moving to disk A, disk B, or terminals are 80/181, 100/181, and 1/181, respectively. Thus, the transition probabilities are p_{1A}=0.4420, p_{1B}=0.5525, and p_{10}=0.005525.
- Given the transition probabilities, we can find the visit ratios by dividing these probabilities by the exit probability (0.005525):

$$V_A = \frac{p_{1A}}{p_{10}} = \frac{0.4420}{0.005525} = 80$$

$$V_B = \frac{p_{1B}}{p_{10}} = \frac{0.5525}{0.005525} = 100$$

$$V_{CPU} = 1 + V_A + V_B = 1 + 80 + 100 = 181$$

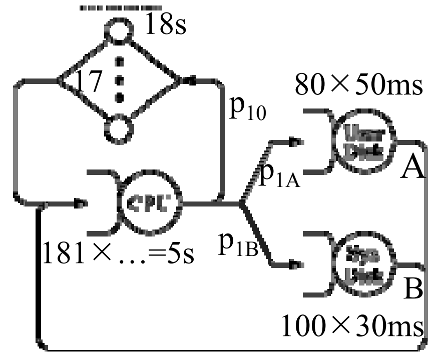
Little's Law

$$
\begin{aligned}
&\text{Mean number in the device} \\
&= \text{Arrival rate} \times \text{Mean time in the device} \\
&Q_i = \lambda_i R_i
\end{aligned}
$$

- If the job flow is balanced, the arrival rate is equal to the throughput and we can write:

$$Q_i = X_i R_i$$

Example 33.4



- The average queue length in the computer system of Example 33.2 was observed to be: 8.88, 3.19, and 1.40 jobs at the CPU, disk A, and disk B, respectively. What were the response times of these devices?
- In Example 33.2, the device throughputs were determined to be: $X_{CPU} = 35.48,\ X_A = 15.70,\ \text{and } X_B = 19.6$
- The new information given in this example is:

$$Q_{CPU} = 8.88,\ Q_A = 3.19,\ \text{and } Q_B = 1.40$$

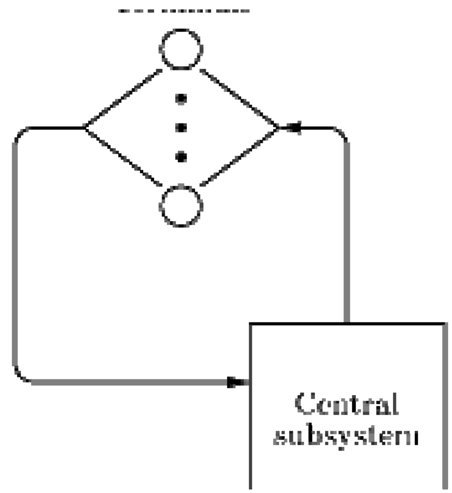
- Using Little's law, the device response times are:

$$R_{CPU} = Q_{CPU}/X_{CPU} = 8.88/35.48 = 0.250 \text{ seconds}$$

$$R_A = Q_A/X_A = 3.19/15.70 = 0.203 \text{ seconds}$$

$$R_B = Q_B/X_B = 1.40/19.6 = 0.071 \text{ seconds}$$

General Response Time Law



- There is one terminal per user and the rest of the system is shared by all users.
- Applying Little's law to the central subsystem:

$$Q = XR$$

- Here,
- Q = Total number of jobs in the system
- R = system response time
- X = system throughput

$$Q = Q_1 + Q_2 + \cdots + Q_M$$

$$XR = X_1R_1 + X_2R_2 + \cdots + X_MR_M$$

General Response Time Law (Cont)

$$XR = X_1R_1 + X_2R_2 + \cdots + X_MR_M$$

- Dividing both sides by X and using forced flow law:

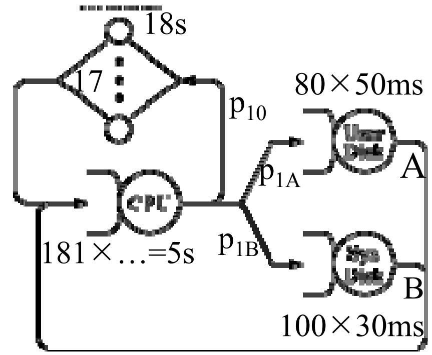
$$R = V_1R_1 + V_2R_2 + \cdots + V_MR_M$$

- or,

$$R = \sum_{i=1}^{M} R_iV_i$$

- This is called the **general response time law**.

Example 33.5



- Let us compute the response time for the timesharing system of Example 33.4
- For this system:

$$V_{CPU} = 181,\ V_A = 80,\ \text{and}\ V_B = 100$$

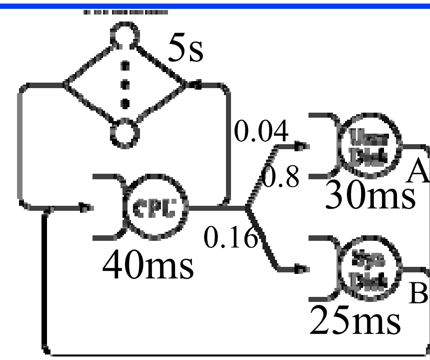
$$R_{CPU} = 0.250,\ R_A = 0.203,\ \text{and}\ R_B = 0.071$$

- The system response time is:

$$\begin{aligned} R &= R_{CPU}V_{CPU} + R_A V_A + R_B V_B \\ &= 0.250 \times 181 + 0.203 \times 80 + 0.071 \times 100 \\ &= 68.6 \end{aligned}$$

- The system response time is 68.6 seconds.

Quiz 33B



- The transition probabilities of jobs exiting CPU and device service times are as shown.
- Find the visit ratios:
  - $V_A = p_{1A}/p_{10}$ = __________
  - $V_B = p_{1B}/p_{10}$ = __________
  - $V_{CPU} = 1 + V_A + V_B$ = __________
- The queue lengths at CPU, disk A, and disk B was observed to be 6, 3, and 1, respectively. The system throughput is 1 jobs/sec. What is the system response time?
  - $R_{CPU} = Q_{CPU}/X_{CPU} = Q_{CPU}/(XV_{CPU})$ = __________
    $R_A = Q_A/(X_A)$ = __________
  - $R_B = Q_B/(X_B)$ = __________
  - $R = R_{CPU}V_{CPU} + R_AV_A + R_BV_B$ = __________
  - Check: $Q = X R$ __________

Key: $U_i = X_iS_i = XD_i$, $D_i = S_iV_i$, $X = X_i/V_i$, $Q_i = X_iR_i$, $R = \sum_{i=1}^{M} R_iV_i$

Interactive Response Time Law

- If Z = think-time, R = Response time
  - The total cycle time of requests is $R+Z$
  - Each user generates about $T/(R+Z)$ requests in T
- If there are N users:

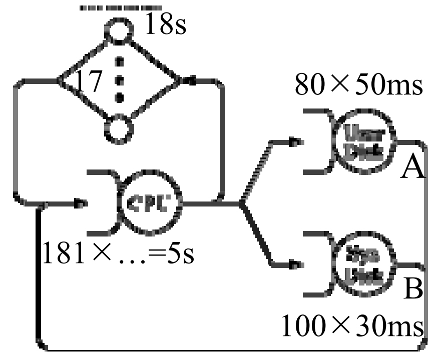
$$\begin{aligned} \text{System throughput } X &= \text{Total \# of requests/Total time} \\ &= N(T/(R+Z))/T \\ &= N/(R+Z) \end{aligned}$$

or

$$R = (N/X) - Z$$

- This is the interactive response time law

Example 33.6



- For the timesharing system of Example 33.2:

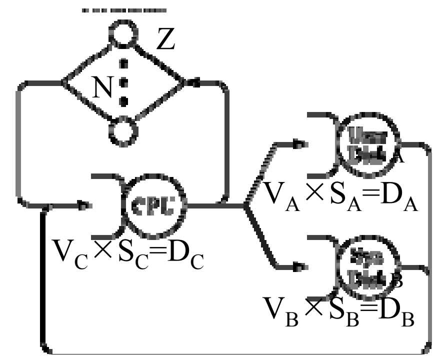
  $$X = 0.1963, N = 17, \text{ and } Z = 18$$

  The response time can be calculated as follows:

  $$R = \frac{N}{X} - Z = \frac{17}{0.1963} - 18 = 86.6 - 18 = 68.6 \text{ seconds}$$

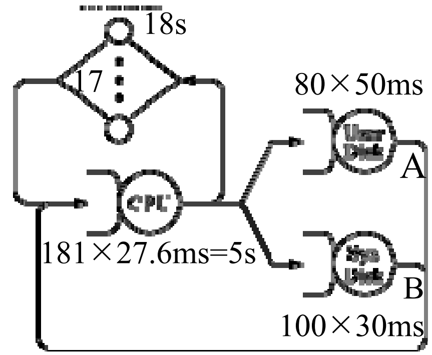
- This is the same as that obtained earlier in Example 33.5.

Review of Operational Laws



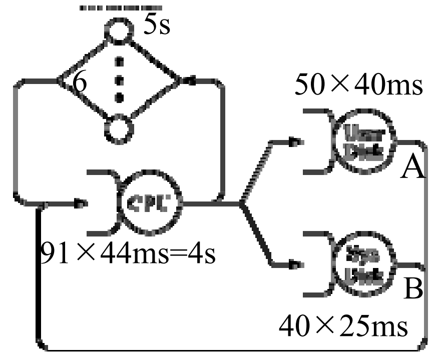
- **Operational quantities**:
  Can be measured by operations personnel
  - V_i = # of visits per job to device *i*
  - S_i = Service time per job at device *i*
  - D_i = Total service demands per job at device *i* = S_iV_i
  - X_i = Throughput of device i
  - X = Throughput of the system
  - Z = User think time
  - N = Number of users in a time shared system
- **Operational assumptions**: That can be easily validated.
  # Input = # output (**flow balance**) can be validated
  Distributions and independence can not be validated.
- **Operational Laws**: Relationships between operational quantities
  These apply regardless of distribution, burstiness, arrival patterns.
  The only assumption is flow balance.
  1. Utilization Law: $U = X_iS_i = XD_i$
  2. Forced Flow Law: $X_i = XV_i$
  3. Little's Law: $Q_i = X_i R_i$
  4. General Response Time Law: $R = \Sigma R_iV_i$
  5. Interactive Response Time Law: $R = N/X - Z$

Example



- **Operational quantities**:
  Can be measured by operations personnel
  V_i = # of visits per job to device i = *181, 80, 100*
  S_i = Service time per job at device i = *27.6ms, 50ms, 30ms*
  D_i = Total service demands per job at device $i = S_iV_i$ = *5s, 4s, 3 s*
  Z = User think time = 18s
  N = Number of users in a time shared system = 12
- **Operational Laws**: Given U_A= 75%, Q_A=2.41, Q_B=1.21, Q_C=5
  1. Utilization Law: $U=X_iS_i = XD_i$
     $X = U_A/D_A = 0.75/4 = 0.188$ jobs/s
     $U_C = X \times D_C = 0.188 \times 5 = 0.939$
     $U_B = X \times D_B = 0.188 \times 3 = 0.563$
  2. Forced Flow Law: $X_i = XV_i$
     $X_A = X \times 80 = 0.188 \times 80 = 15$ *jobs/s*
     $X_B = X \times 100 = 0.188 \times 100 = 18.8$ *jobs/s*
     $X_C = X \times 181 = 0.188 \times 181 = 34$ *jobs/s*
  3. Little's Law: $Q_i = X_i R_i$
     $R_A=Q_A/X_A=2.41/15=0.161$, $R_B=1.21/18.8=0.064$, $R_C=5/34=0.147$
  4. General Response Time Law: $R=\Sigma R_iV_i = 0.161\times80+0.064\times100+0.147\times181$
     $=45.89$s
  5. Interactive Response Time Law: $R = N/X - Z = 12/0.188-18 = 45.83$s

Quiz 33C



- **Operational quantities**:
  Can be measured by operations personnel_
  V_i = # of visits per job to device i = *91, 50, 40*
  S_i = Service time per job at device i = *0.044s, 0.040s, 0.025s*
  Z = User think time = 5s N = Number of users = 6
- **Operational Laws**: Given U_A= 48%, R_A=0.0705s, R_B=0.0323s, R_C=0.1668s
  1. D_i = Total service demands per job at device $i = S_i V_i$
     $D_C = S_C V_C$ = ______ × ___ = ___, D_A= ______ × _____= ___, D_B= ______ × ___= ___
  2. Utilization Law: $U = X_i S_i = XD_i$
     $X = U_A/D_A$ = ______ / ______ = ______ jobs/s
     $U_C = X\ D_C$ = ______ × ______ = ______
     $U_B = X\ D_B$ = ______ × ______ = ______
  3. Forced Flow Law: $X_i = XV_i$
     $X_A = X\ V_A$ = ______ × ______ = ______*jobs/s*
     $X_B = X\ V_B$ = ______ × ______ = ______*jobs/s*
     $X_C = X\ V_C$ = ______ × ______ = ______*jobs/s*
  4. Little's Law: $Q_i = X_i R_i$
     Q_A= _____ × _____ = _____, Q_B= _____ × _____ = _____, Q_C= _____ × _____ = _____
  5. General Response Time Law: $R = \Sigma R_i V_i$
     = _____ × _____ + _____ × _____ + _____ × ______ = _____ s
  6. Interactive Response Time Law: $R = N/X - Z$ = ______ / ______ - ______ = ______ s

Bottleneck Analysis

- From forced flow law:

$$U_i \propto D_i$$

- The device with the highest total service demand D_i has the highest utilization and is called the bottleneck device.
- Note: Delay centers can have utilizations more than one without any stability problems. Therefore, delay centers cannot be a bottleneck device.
- Only queueing centers used in computing D_{max}.
- The bottleneck device is the key limiting factor in achieving higher throughput.

Bottleneck Analysis (Cont)

- Improving the bottleneck device will provide the highest payoff in terms of system throughput.
- Improving other devices will have little effect on the system performance.
- Identifying the bottleneck device should be the first step in any performance improvement project.

Asymptotic Bounds

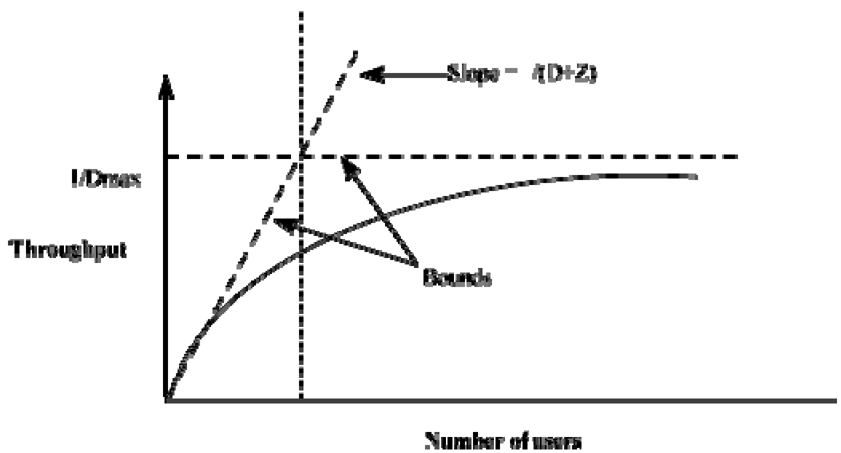
- Throughput and response times of the system are bound as follows:

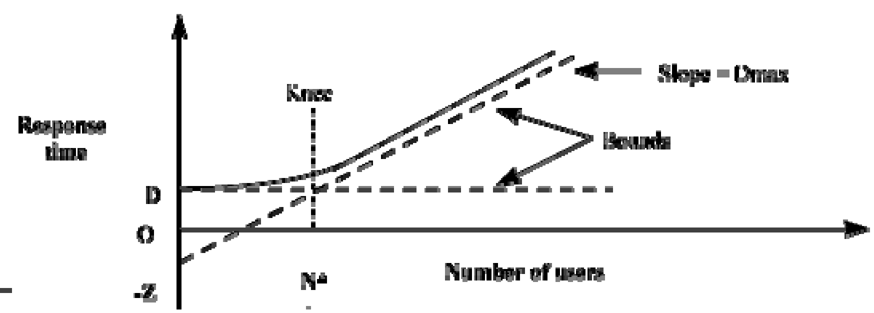
$$X(N) \leq min\{\frac{1}{D_{max}}, \frac{N}{D+Z}\}$$

and

$$R(N) \geq max\{D, ND_{max} - Z\}$$

- Here, $D = \sum D_i$ is the sum of total service demands on all devices except terminals.





Asymptotic Bounds: Proof

- The asymptotic bounds are based on the following observations:
  1. The utilization of any device cannot exceed one. This puts a limit on the maximum obtainable throughput.
  2. The response time of the system with N users cannot be less than a system with just one user. This puts a limit on the minimum response time.
  3. The interactive response time formula can be used to convert the bound on throughput to that on response time and vice versa.

Proof (Cont)

1. For the bottleneck device b we have:

$$U_b = XD_{max}$$

Since U_b cannot be more than one, we have:

$$XD_{max} \leq 1$$

$$X \leq \frac{1}{D_{max}}$$

2. With just one job in the system, there is no queueing and the system response time is simply the sum of the service demands:

$$R(1) = D_1 + D_2 + \cdots + D_M = D$$

With more than one user there may be some queueing and so the response time will be higher. That is:

$$R(N) \geq D$$

Proof (Cont)

3. Applying the interactive response time law to the bounds:

$$R = (N/X) - Z$$

$$R(N) = \frac{N}{X(N)} - Z \geq N D_{max} - Z$$

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

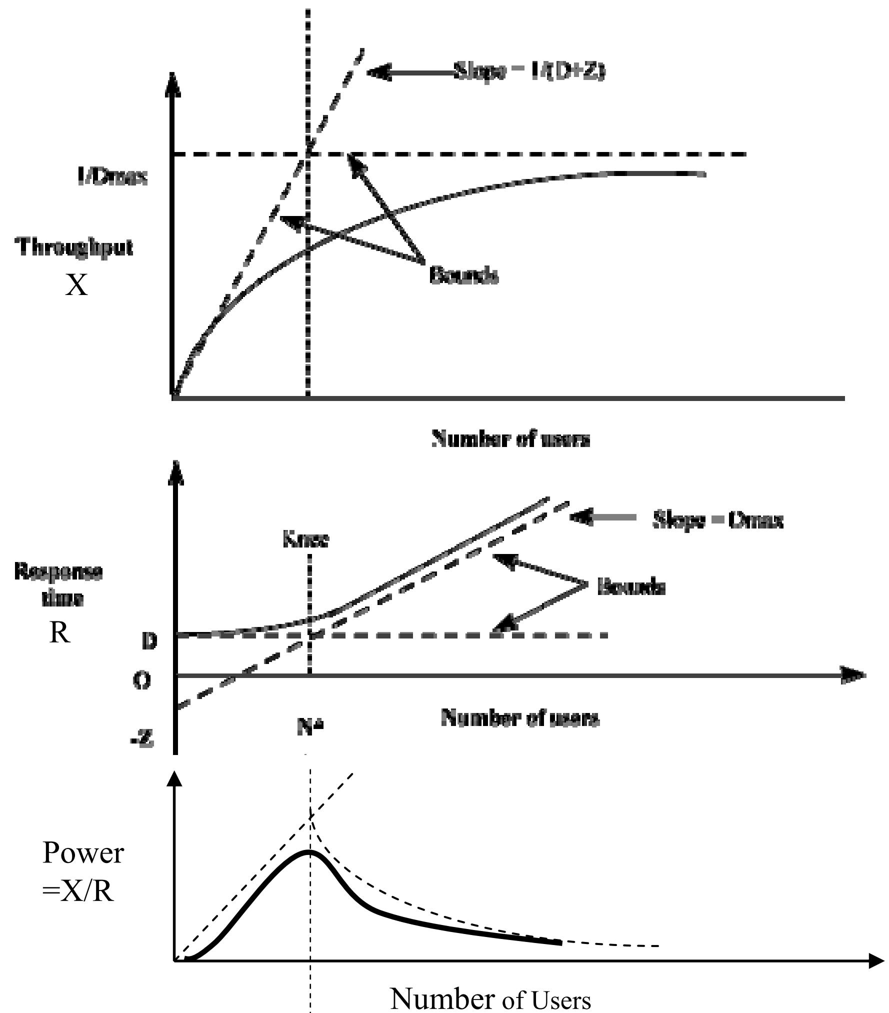
Optimal Operating Point

- The number of jobs N^* at the knee is given by:

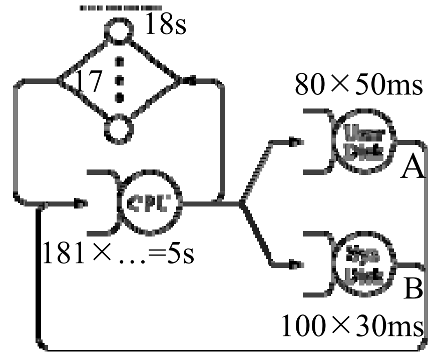
$$D = N^* D_{max} - Z$$

$$N^* = \frac{D + Z}{D_{max}}$$

- If the number of jobs is more than N^*, then we can say with certainty that there is queueing somewhere in the system.
- The asymptotic bounds can be easily explained to people who do not have any background in queueing theory or performance analysis.
- Control Strategy:
  Increase N iff dP/dN is positive



Example 33.7



- For the timesharing system of Example 33.2:

$$D_{CPU} = 5, D_A = 4, D_B = 3, Z = 18$$

$$D = D_{CPU} + D_A + D_B = 5 + 4 + 3 = 12$$

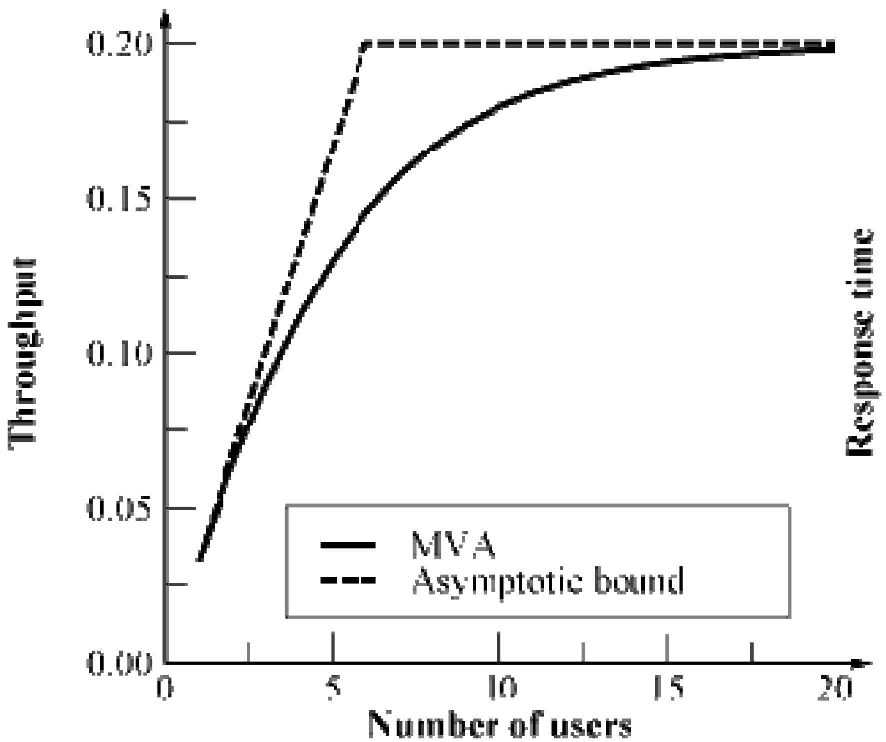
$$D_{max} = D_{CPU} = 5$$

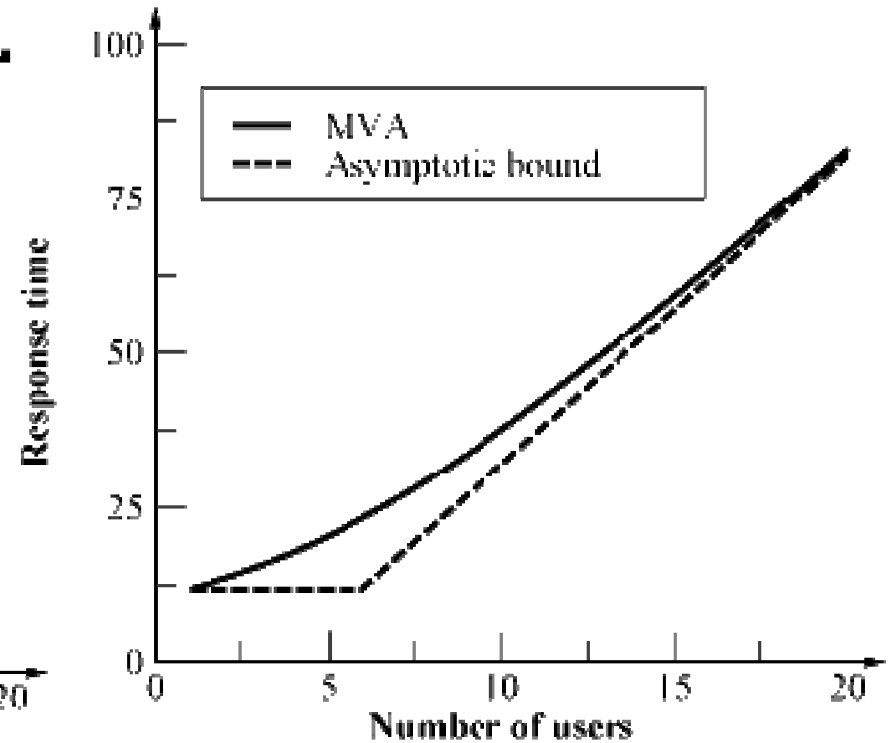
- The asymptotic bounds are:

$$X(N) \leq \min\left\{\frac{N}{D+Z}, \frac{1}{D_{max}}\right\} = \min\left\{\frac{N}{30}, \frac{1}{5}\right\}$$

$$R(N) \geq \max\{D, ND_{max} - Z\} = \max\{12, 5N - 18\}$$

Example 33.7: Asymptotic Bounds



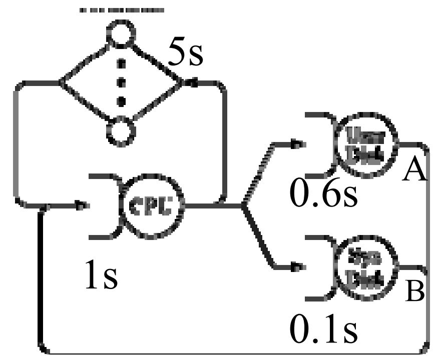


- The knee occurs at:

$$12 = 5N^* - 18$$

$$N^* = \frac{12 + 18}{5} = \frac{30}{5} = 6$$

Quiz 33D



- The total demands on various devices are as shown.
- What is the minimum response time?
  $R = D = D_{CPU} + D_A + D_B =$ __________
- What is the bottleneck device? __________
- What is the maximum possible utilization of disk B?
  $U_B =$ __________
- What is the maximum possible throughput? $X =$ __________
- What is the upper bound on throughput with N users?

  ____

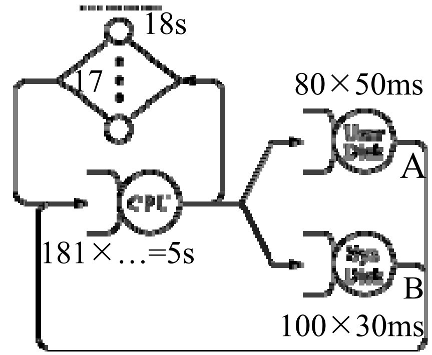
- What is the lower bound on response time with N users?

  ____

- What is the knee capacity of this system? ____________________

Key: $R \geq \max\{D, ND_{max} - Z\}$, $X \leq \min\{1/D_{max}, N/(D+Z)\}$

Example 33.8

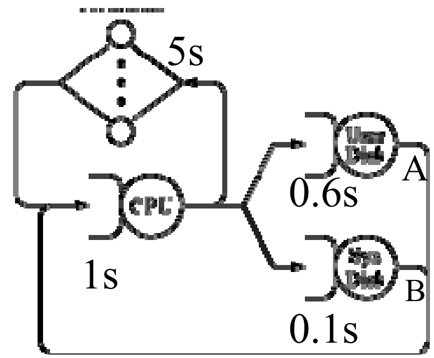


- How many terminals can be supported on the timesharing system of Example 33.2 if the response time has to be kept below 100 seconds?
- Using the asymptotic bounds on the response time we get:

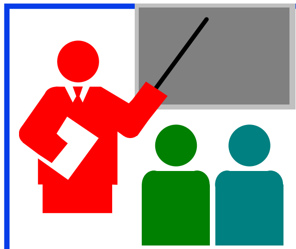
$$R(N) \geq \max\{12, 5N - 18\}$$

- The response time will be more than 100, if: $5N - 18 \geq 100$
- That is, if: $N \geq 23.6$ the response time is bound to be more than 100. Thus, the system cannot support more than 23 users if a response time of less than 100 is required.

Quiz 33E



- For this system, which device would be the bottleneck if:
- The CPU is replaced by another unit that is twice as fast? ______
- Disk A is replaced by another unit that is twice as slow? ______
- Disk B is replaced by another unit that is twice as slow? ______
- The memory size is reduced so that the jobs make 25 times more visits to disk B due to increased page faults? ______



Summary

Utilization Law:	U_i	$=$	$X_i S_i = X D_i$
Forced Flow Law:	X_i	$=$	$X V_i$
Little's Law:	Q_i	$=$	$X_i R_i$
General Response Time Law:	R	$=$	$\sum_{i=1}^{M} R_i V_i$
Interactive Response Time Law:	R	$=$	$\frac{N}{X} - Z$
Asymptotic Bounds:	R	$\geq$	$max\{D, ND_{max} - Z\}$
	X	$\leq$	$min\{1/D_{max}, N/(D+Z)\}$

- Symbols:

D	$=$	Sum of service demands on all devices $= \sum_i D_i$
D_i	$=$	Total service demand per job for ith device $= S_i V_i$
D_{max}	$=$	Service demand on the bottleneck device $= \max_i\{D_i\}$
N	$=$	Number of jobs in the system
Q_i	$=$	Number in the ith device
R	$=$	System response time
R_i	$=$	Response time per visit to the ith device
S_i	$=$	Service time per visit to the ith device
U_i	$=$	Utilization of ith device
V_i	$=$	Number of visits per job to the ith device
X	$=$	System throughput
X_i	$=$	Throughput of the ith device
Z	$=$	Think time

Homework 33

- Draw a diagram showing the flow of jobs in your system including waiting for disk I/O and network I/O.