

Ch.33 Operational Laws

- Stochastic Analysis : Markovian Queueing Network
 - Assumptions
 - The system is modeled by stationary stochastic processes
 - Jobs are stochastically independent
 - Job steps from device to device follow a Markov chain
 - The system is in stochastic equilibrium
 - The service time at each device conforms to an exponential distribution
 - The system is ergodic (long-term time averages converges to those at stochastic equilibrium)
 - Problems with the assumptions
 - Difficult to understand - “stationary” “equilibrium” ...
 - They cannot be proved to hold by observing in a finite time period
 - Most of them are incorrect
 - parameters change over time
 - jobs are dependent
 - device to device transitions do not follow Markov chain
 - service distributions are seldom exponential...

Operational Analysis

- Operational Principles (assumptions)
 - All quantities are precisely measurable (and directly testable)
 - The system is flow balanced (arrivals = departures)
 - The devices are homogeneous (routing/service time are not dependent on other devices)
- Operational Laws
 - Simple relationships that do not assume about the distribution of service times or interarrival times
 - Derives results using high-school algebra (“ $f=ma$ ” style)
 - Called “laws” since they hold without any assumptions

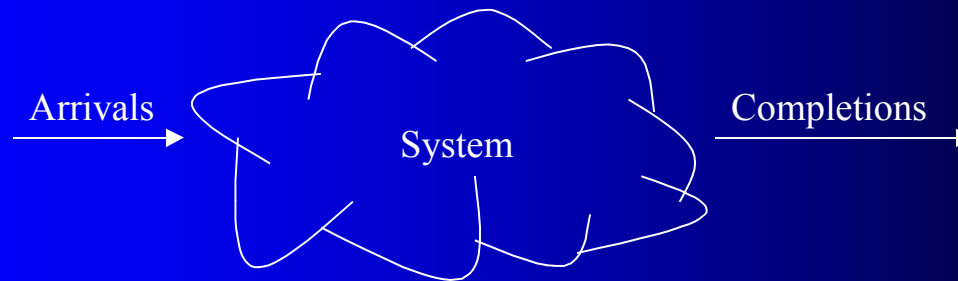
Fundamental Laws

- Basic Quantities (Measured Q)

- T : Observation time interval
- A : # Arrivals during T
- C : # Completions during T
- B : Busy time during T

(Derived Quantities)

λ : Arrival rate = A/T
X : Completion rate = C/T
(Throughput)
U : Utilization = B/T
S : Mean service time = B/C



Utilization Law

- Utilization Law :

$$U=XS$$

- For a Single Resource Subsystem
- $U_i = B_i/T = C_i/T \times B_i/C_i = X_i S_i$
- Example: 2 cars per minute, each requires 0.2 minutes
=> 0.4 minutes' service is required per minute = 40% utilization
- “Job Flow Balance” assumption: $A=C$
=> $\lambda=X$ and hence, $U = XS = \lambda S$

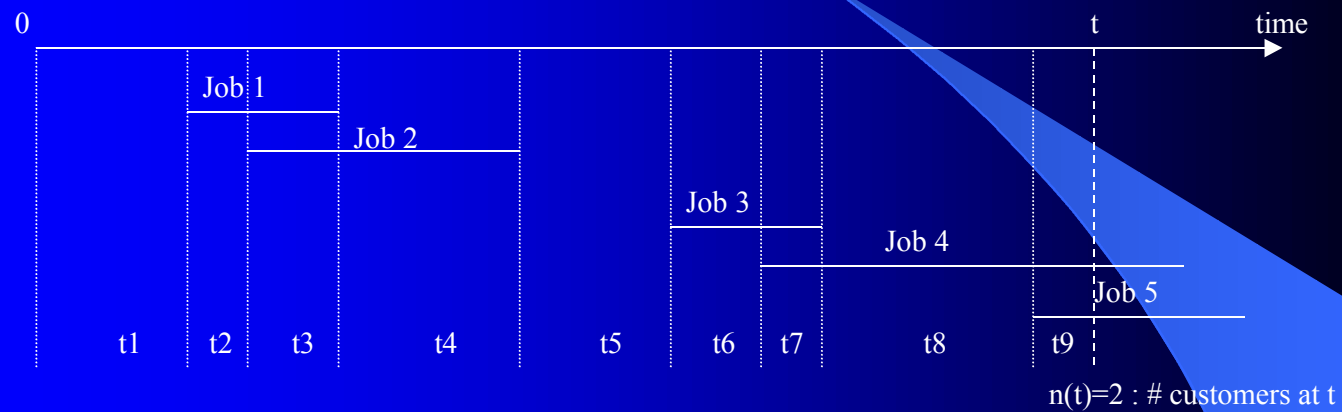
Utilization Law (cont'd)

- Question : A desktop PC is connected to a 10Mbps LAN and is used for video conferencing. The software package sends 10 frames/sec and uses an encoding scheme that requires an average of 39321 bits/frame. How much is the LAN connection utilized ?

Little's Law

- Little's Law : $N = XR$
- For an Entire System
- The most important relationship in queueing analysis
- Utilization Law is a special case of Little's Law
- “Total time in system, W (job-seconds)”
 - $W = RC$: Average response time \times # Jobs
 - $W = NT$: Average # jobs \times Observation Time $\Rightarrow N = RC/T = RX$

Derivation of Little's Law



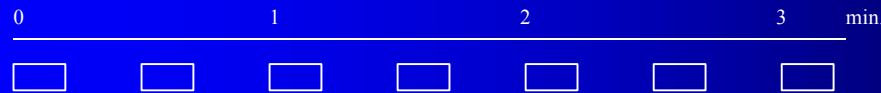
- Total response time during t, W (job-seconds)
 - $W = \text{length}(\text{Job1}) + \text{length}(\text{Job2}) + \text{length}(\text{Job3}) + \text{part of length}(\text{Job4 \& 5})$
 $= 0 \cdot t1 + 1 \cdot t2 + 2 \cdot t3 + 1 \cdot t4 + 0 \cdot t5 + 1 \cdot t6 + 2 \cdot t7 + 1 \cdot t8 + 2 \cdot t9$
 $= \int_0^t n(\tau) d\tau$

Derivation of Little's Law (cont'd)

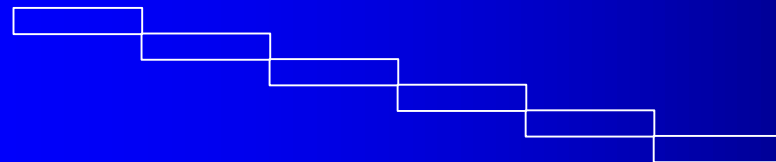
- Average wait time per customer, R
 - $R = \{\text{len}(\text{Job1}) + \text{len}(\text{Job2}) + \text{len}(\text{Job3}) + \text{part of len}(\text{Job4 \& 5})\} / \#\text{jobs}$
 $= W/C$
- Average number of customers in system, N
 - $N = 0 \cdot t_1/t + 1 \cdot t_2/t + 2 \cdot t_3/t + 1 \cdot t_4/t + 0 \cdot t_5/t + 1 \cdot t_6/t + 2 \cdot t_7/t + 1 \cdot t_8/t + 2 \cdot t_9/t$
 $= (\int_0^t n(\tau) d\tau) / t$
 $= W/t$
- Little's Law : $N = XR$ \Rightarrow $N = X(R+Z)$, Z is think time
 $R = N/X - Z$: Response Time Law
 - $N = W/t$
 $= R \cdot C / t = R \cdot X$

Utilization Law vs Little's Law

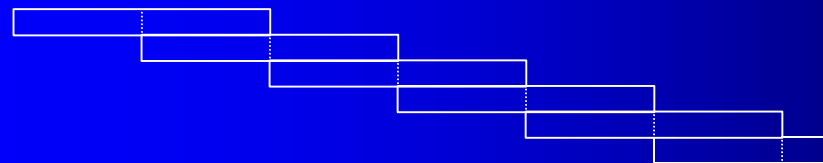
- Utilization Law : $U = XS$ (of a server) : special case of Little's Law
- Little's Law : $N = XR$ (of an entire system)
- Example: 2 cars arrive per minute (λ) = 2 cars depart per minute (X)



Each requires 0.2 min.
 $U = XS = 2 \times 0.2 = 40\%$ utilization
 $N = XR = 0.4$ customer in the system



Each requires 0.5 min.
 $U = XS = 2 \times 0.5 = 100\%$ utilization
 $N = XR = 1$ customer in the system



Each requires 1 min.
 $U = XS = 2 \times 1 = 200\%$ utilization ???
 $N = XR = 2$ customers in the system
 (ex: pipeline of servers)

Little's Law (cont'd)

- Question : “Symmany” server receives 100,000 messages per second on the average, each requires an average of 2.5 seconds to deliver to the user. How many messages the Symmany server must buffer on the average ?

Little's Law (cont'd)

- Question : Suppose a server in Webcrawler currently has an average of 250 TCP connections open. Assuming that all current users are connected by 28.8 kbps modems, how many TCP connections would be open if everyone switched to ISDN (128 kbps) ?

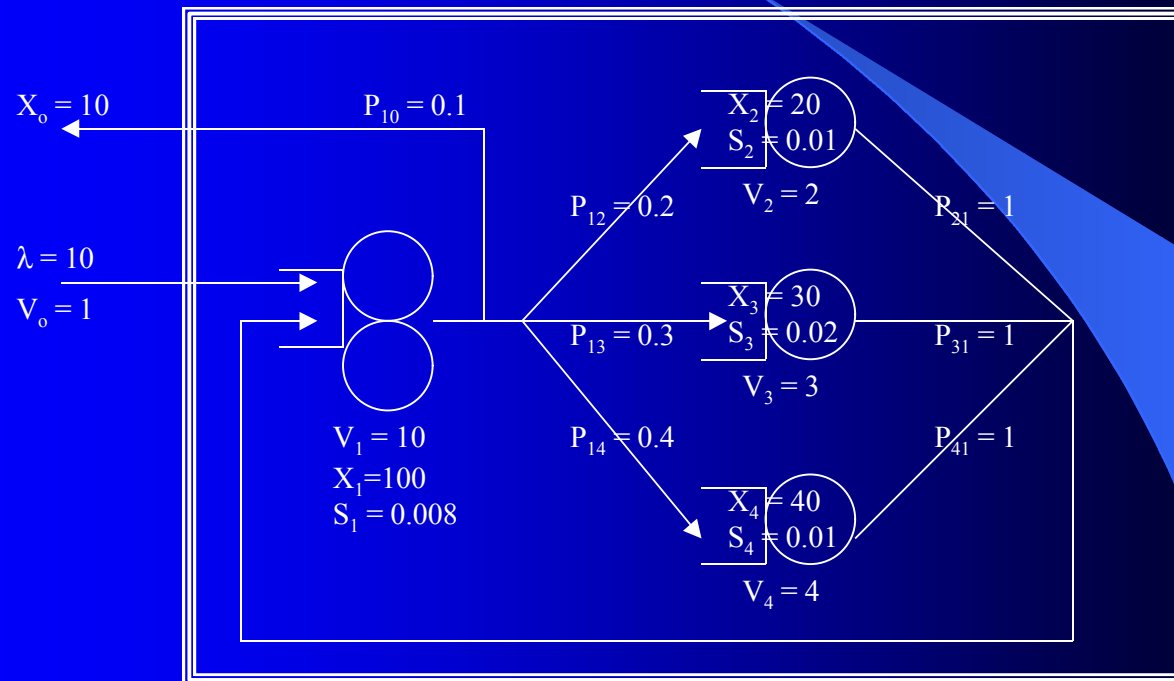
Forced Flow Law

- Relationship between an individual resource and the entire system
- Per system quantities
 - C_0 : # completions from the system
- Per device quantities
 - C_i : # Completions from device i
 - V_i : # Visits (each job makes V_i visits for the device i) = C_i / C_0
- Forced Flow Law : $X_i = X V_i$
 - $X_i = C_i/T = C_0/T \times C_i/C_0 = X V_i$

Forced Flow Law (cont'd)

- Per device quantities
 - B_i : Busy time of device i
 - S_i : Service (demand) per visit (per job) = B_i / C_i
 - D_i : (Service) Demand with V_i visits (per job) = $V_i S_i$
- Per system quantities
 - D : (Service) Demand of system = $\sum D_i$
- Utilization Law : $U_i = X_i S_i = X D_i$
 - $U_i = X_i S_i = X \cdot V_i S_i = X D_i$
 - Utilization Law with system throughput

Forced Flow Law (cont'd)



* Confirm the Forced Flow Law : $X_i = \lambda V_i$
 Check the Utilization Law : $U_i = X_i S_i = \lambda D_i$

Response Time Law

- General Response Time Law

- Total time = Sum of times at all servers
- $N = XR = N_1 + N_2 + \dots + N_M$
 $= X_1R_1 + X_2R_2 + \dots + X_MR_M$
 $\Rightarrow R = V_1R_1 + V_2R_2 + \dots + V_MR_M$

- Interactive Response Time Law

- $N = XR$
 $\Rightarrow N = X(R+Z)$ or
 $R = N/X - Z$
 $X = N/(R+Z)$
- Z : Think time, which is generated by a delay center
(an infinite server, $M/M/\infty$, where response time = service time)

Response Time Law (cont'd)

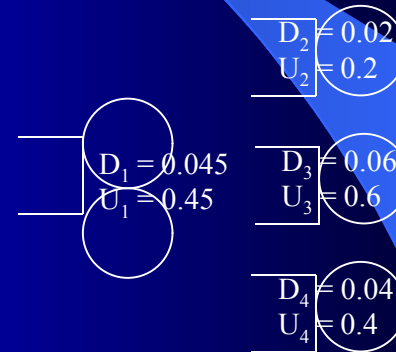
- Question: A file server handles 80 requests per second from 50 workstations, each tries to access the file server after an average of 0.5 second's computation. What is the file server response time ?
What if the throughput is 10% higher (88 request/sec) ?

* Bounds on Performance

- Why study on performance bounds ?
 - Provides valuable insights into primary factors affecting the performance
 - Can be computed quickly
 - Asymptotic Bounds : wider application, simpler
 - Balanced System Bounds : tighter bounds
- Low/High performance bounds
 - Bounds on system throughput X
 - For a closed queueing model
 - As a function of the number of customers in the system (N)

Bottleneck Analysis

- Find the bottleneck center
 - We want to increase X to 20
 $\Rightarrow U_1 = 90\%$, $U_2 = 40\%$, $U_4 = 80\%$
But $U_3 = 120\%$ (???)
 - Device 3 is the bottleneck center
which has D_{\max}
 - X is bounded by $1/D_{\max}$
 $= 1/0.06 = 16.7$



System Throughput $X=10$

Asymptotic Bounds

- Bottleneck Analysis

- $U_i = X D_i$
 $\Rightarrow U_{\max} = X D_{\max} \leq 1$ (bottleneck device)
 $\Rightarrow X \leq 1/D_{\max}$

- Minimum Response Time

- $R(N) = N/X - Z \geq ND_{\max} - Z$
- $R(1) = D_1 + D_2 + \dots + D_M = D$
(when there is only one customer)
 $\Rightarrow R(N) \geq D$
(when there are $N \geq 1$ customers)

- Maximum Response Time

- $R_{\max} = \text{Max queueing delay} + \text{Service}$
 $= (N-1)D + D = ND$

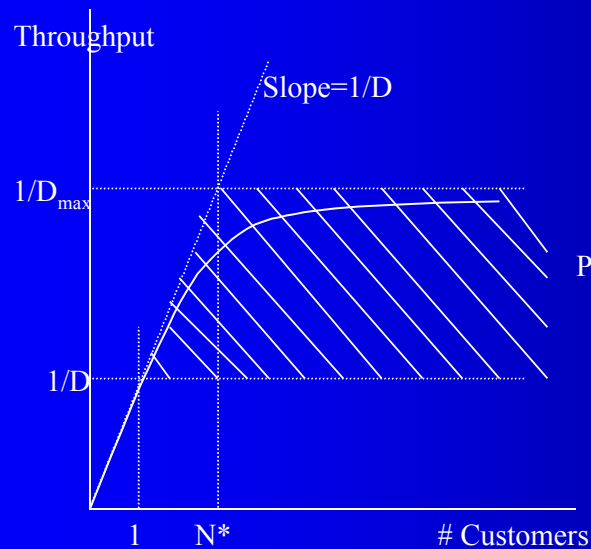
$$R(N) \geq \max \{D, ND_{\max} - Z\}$$
$$R(N) \leq ND$$

Since $X = N/(R+Z)$ & $X \leq 1/D_{\max}$,

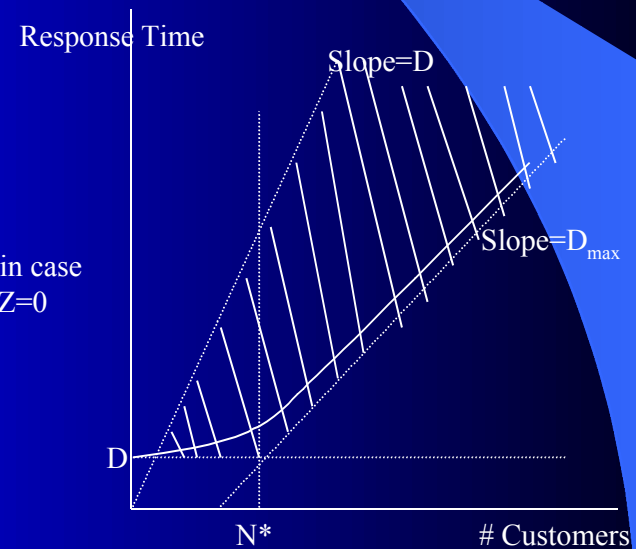
$$X(N) \leq \min \{1/D_{\max}, N/(D+Z)\}$$
$$N/(ND+Z) \leq X(N)$$

Asymptotic Bounds (cont'd)

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



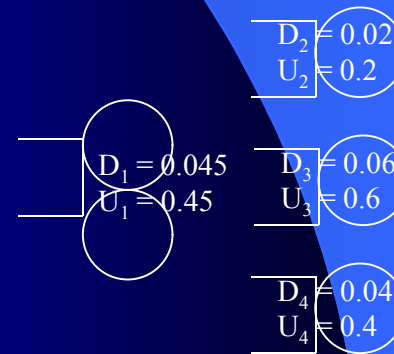
Plot in case
of $Z=0$



Asymptotic Bounds (cont'd)

- If $Z=0$,
 - $1/D \leq X(N) \leq \min\{1/D_{\max}, N/D\}$
 - $\max\{D, ND_{\max}\} \leq R(N) \leq ND$

- Example
 - $D = 0.045+0.02+0.06+0.04 = 0.165, D_{\max} = 0.06$
 - $1/D \leq X(N) \leq \min\{1/D_{\max}, N/D\}$
 $\Rightarrow 1/0.165 \leq X(N) \leq \min\{1/0.06, N/0.165\}$
 $\Rightarrow 6.06 \leq X(N) \leq \min\{16.67, 6.06N\}$
 - $N^* = D/D_{\max} = 0.165/0.06 = 2.75$
 \Rightarrow Beyond that, linear thruput increase stops
 and queueing must occur



System Throughput $X=10$

Asymptotic Bounds (cont'd)

- Question: We measure the followings on a simplified interactive system
 - $T = 900$ seconds : observation interval
 - $B1 = 400$ seconds : CPU busy
 - $B2 = 100$ seconds : slow disk busy
 - $B3 = 600$ seconds : fast disk busy
 - $C = 200$ jobs : completed jobs
 - $C2 = 2000$: slow disk operations
 - $C3 = 20000$: fast disk operations
 - $Z = 15$ seconds : think time
- What will be the effect if we (a) replace CPU with one that is twice as fast
(b) shift some files from the fast disk to slow one, balancing their demands
(c) add a second fast disk to handle the load of the busier existing disk

Balanced Job Bounds

- Asymptotic Bounds are one-sided
 - Good (tight) best-case performance bound but not a good worst-case performance bound
 - Not a tight upper bound for $R(N)$, not a tight lower bound for $X(N)$
$$\max \{D, ND_{\max}\} \leq R(N) \leq ND$$
$$1/D \leq X(N) \leq \min \{1/D_{\max}, N/D\}$$
- Balanced system
 - Balanced system : system without a bottleneck device
 - Balanced system provides good bound for both sides
 - Best case balanced system: Total D is equally distributed to all devices
 - Worst case balanced system: D_{\max} is distributed to some devices

Balanced Job Bounds (cont'd)

- Best case is that there's no bottleneck center
 - Service centers has equal service demand
or the system is balanced : $D_1 = D_2 = D_3 = \dots = D_M = D/M$
- Worst case is that total demand is distributed to minimal # centers
 - D/D_{\max} centers with D_{\max} , other centers = 0



Best & Worst Balanced System

- In the balanced system, we have tighter bounds
 - $U_i(N) = N/(N+M-1)$: we'll see later
 - $X(N) = U_i/D_i = N/\{(N+M-1)D_i\}$
 - $N/\{(N+M-1)D_{\max}\} \leq X(N) \leq N/\{(N+M-1)D_{\min}\}$

- Best-case balanced system ($D_1 = D_2 = D_3 = \dots = D_M = D/M$)

$$X(N) \leq N/\{(N+M-1)D_{\text{avg}}\} = N/\{(N-1)D_{\text{avg}}+D\}$$

- Worst-case balanced system (D/D_{\max} centers with D_{\max} , others = 0)

$$N/\{(N+M-1)D_{\max}\} = N/\{(N+D/D_{\max}-1)D_{\max}\} = N/\{(N-1)D_{\max}+D\} \leq X(N)$$

$$N/\{(N-1)D_{\max}+D\} \leq X(N) \leq \min\{1/D_{\max}, N/\{(N-1)D_{\text{avg}}+D\}\}$$

$$\max\{ND_{\max}, D+(N-1)D_{\text{avg}}\} \leq R(N) \leq D+(N-1)D_{\max}$$