Ch. 6[H] 317

Note Title 2015-03-19

Little's Law

Theorem 6.1 (Little's Law for Open Systems)

For any ergodic open system, we have that

$$E[N] = \lambda E[T], \text{ where}$$

$E[N]$ is the expected # jobs in the system

λ is the avg. arrival rate

$E[T]$ is the mean time a job spends in the system

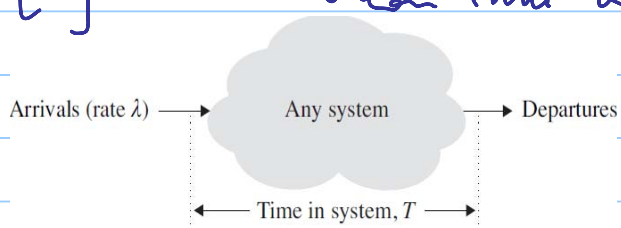
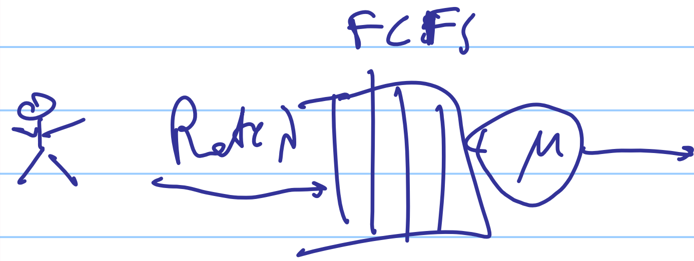


Fig. 6.1 [H]

Intuition (6.2 [H])



$$E[T] = \frac{1}{\lambda} E[N]$$

time to process one job

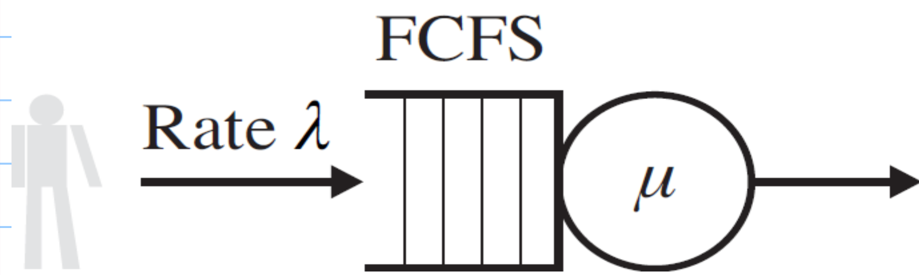


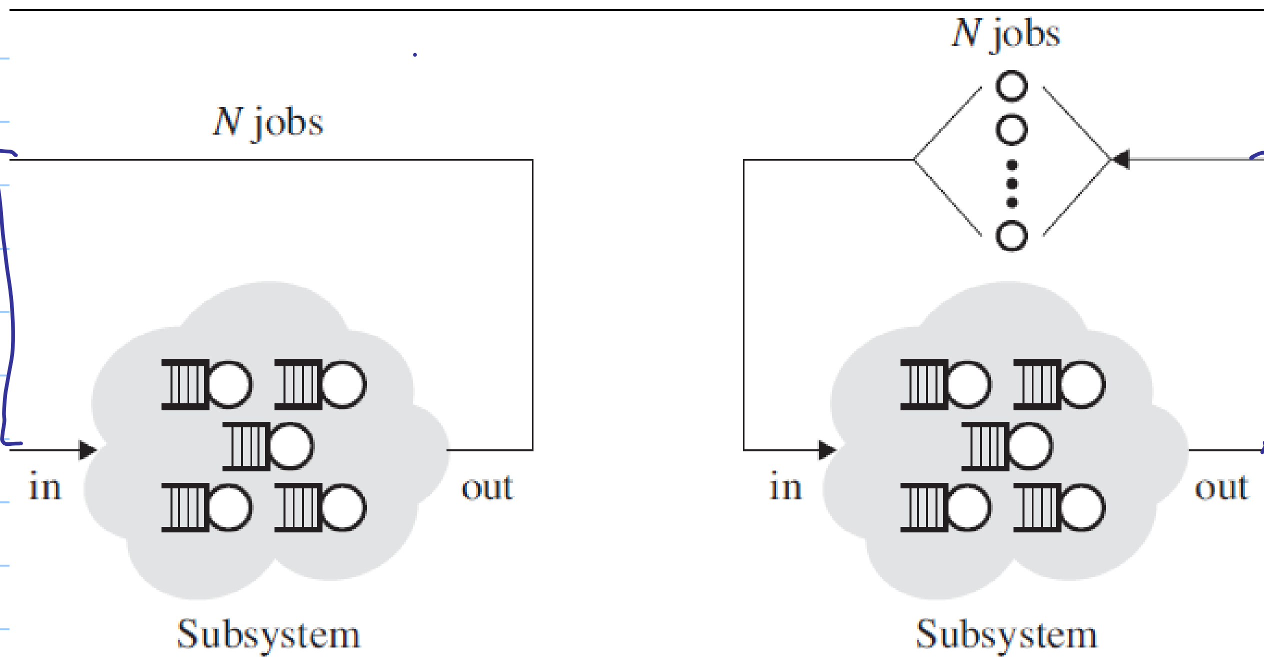
Fig. 6.2 [H]

Little's Law for Closed Systems (6.3 [H])

Theorem 6.2 (Little's Law for Closed Systems) *Given any ergodic closed system,*

$$N = X \cdot \mathbf{E}[T],$$

where N is a constant equal to the multiprogramming level, X is the throughput (i.e., the rate of completions for the system), and $\mathbf{E}[T]$ is the mean time jobs spend in the system.



Closed systems; batch

interactive

$E[T] = E[R] + E[Z]$

$E[T]$: time in system

$E[R]$: response time

$E[Z]$: think time

6.4 [H] Proof of Little's Law for Open Systems

using time averages

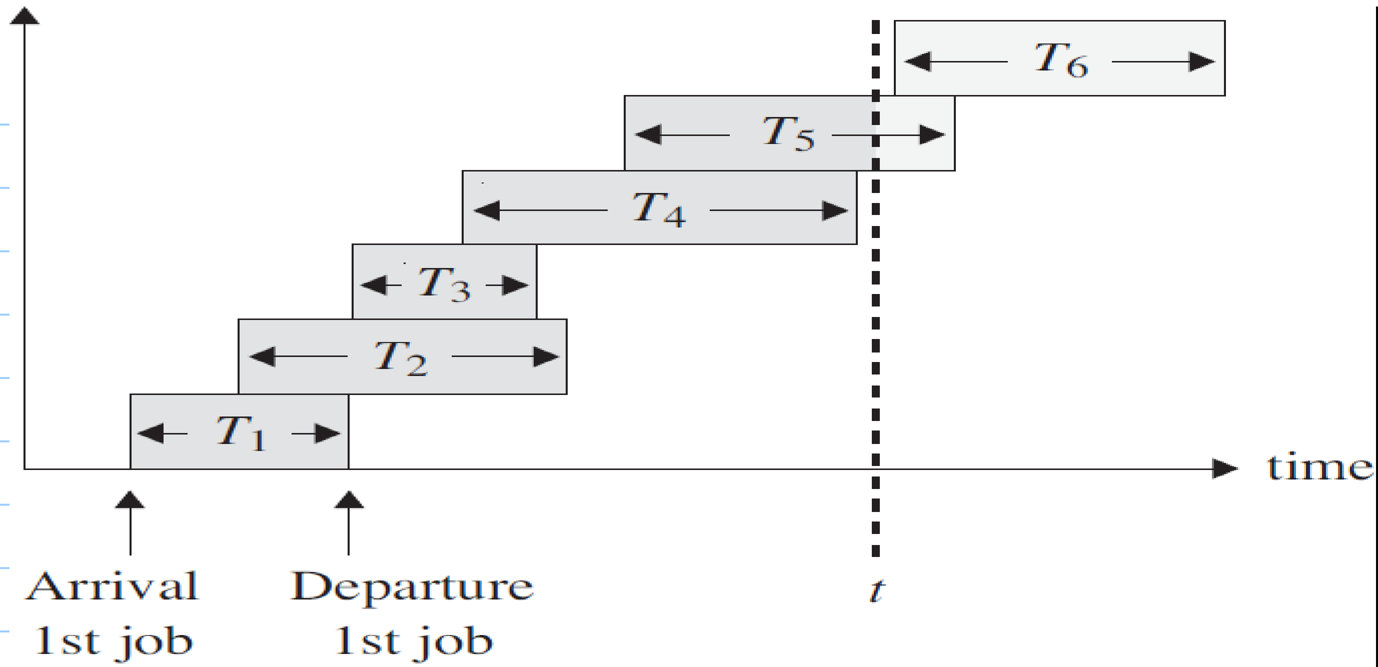
Theorem 6.3 (Little's Law for Open Systems Restated) *Given any system where $\overline{N}^{\text{Time Avg}}$, $\overline{T}^{\text{Time Avg}}$, λ, and X exist and where $\lambda = X$, then*

$$\overline{N}^{\text{Time Avg}} = \lambda \cdot \overline{T}^{\text{Time Avg}}.$$

$\lambda = \lim_{t\to\infty} \frac{A(t)}{t}$ and $X = \lim_{t\to\infty} \frac{C(t)}{t}$. $\lambda = X$, if jobs are not dropped, except in special cases.

arrival rate throughput

Ergodicity implies the assumptions of Thm. 6.3.



Q: area in dark part of rectangles

Graph of arrivals in an open system (Fig. 6.5 [H])

$$\sum_{i \in C(t)} T_i \leq Q \leq \sum_{i \in Q(t)} T_i$$

completed by time t — arrived by time t

$$Q = \int_0^t N(s)\,ds$$ ("vertical view"; sum # jobs in system at any moment in time)

So:
$$\frac{\sum_{i \in C(t)} T_i}{t} \leq \frac{\int_0^t N(s)\,ds}{t} \leq \frac{\sum_{i \in Q(t)} T_i}{t}$$, or equivalently

$$\frac{\sum_{i \in C(t)} T_i}{C(t)} \cdot \frac{C(t)}{t} \leq \frac{\int_0^t N(s)\,ds}{t} \leq \frac{\sum_{i \in Q(t)} T_i}{Q(t)} \cdot \frac{Q(t)}{t}.$$

Taking limits as $t \to \infty$,

$$\lim_{t\to\infty} \frac{\sum_{i \in C(t)} T_i}{C(t)} \cdot \lim_{t\to\infty} \frac{C(t)}{t} \leq \bar{N}^{\text{Time Avg}} \leq \lim_{t\to\infty} \frac{\sum_{i \in Q(t)} T_i}{Q(t)} \cdot \lim_{t\to\infty} \frac{Q(t)}{t}$$

$$\underbrace{\bar{T}^{\text{Time Avg}}}_{\substack{\text{avg time} \\ \text{in system} \\ \text{(viewing} \\ \text{completions)}}} \cdot \underbrace{X}_{\substack{\text{completion} \\ \text{rate} \\ \text{(Throughput)}}} \leq \bar{N}^{\text{Time Avg}} \leq \underbrace{\bar{T}^{\text{Time Avg}}}_{\substack{\text{avg time} \\ \text{in system} \\ \text{(viewing arrivals)}}} \cdot \underbrace{\lambda}_{\substack{\text{arrival} \\ \text{rate}}}$$

Since $\lambda = X$, $\quad \bar{N}^{\text{Time Avg}} = \lambda \bar{T}^{\text{Time Avg}}$

Corollary 6.4 (Little's Law for Time in Queue) *Given any system where* $\overline{N_Q}^{\text{Time Avg}}$, $\overline{T_Q}^{\text{Time Avg}}$, λ, *and* X *exist and where* $\lambda = X$, *then*

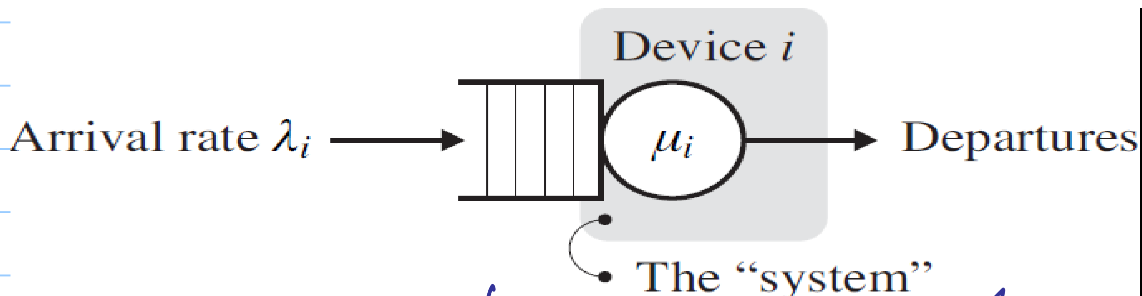
$$\overline{N_Q}^{\text{Time Avg}} = \lambda \cdot \overline{T_Q}^{\text{Time Avg}},$$

where N_Q *represents the number of jobs in queue in the system and* T_Q *represents the time jobs spend in queues.*

The same kind of "geometric" proof can be carried out, except that now the "rectangles" $T_Q(i)$ represent time in queue for job i, and they can be broken up as jobs leave a queue and enter a processor.

Corollary 6.5 (Utilization Law) *Consider a single device i with average arrival rate λ_i jobs/sec and average service rate μ_i jobs/sec, where $\lambda_i < \mu_i$. Let ρ_i denote the long-run fraction of time that the device is busy. Then*

$$\rho_i = \frac{\lambda_i}{\mu_i}.$$



(queue not included in the "system")

ρ_i is called the device utilization or device load (for device i)

The expected number of jobs in the system is

$1 \times P\{\text{system is busy}\} + 0 \times \{\text{system is idle}\} = 1 \times \rho_i + 0 \times (1-\rho_i) = \rho_i$.

So, applying Little's Law, we have;

ρ_i = Expected number of jobs in the system =

= (arrival rate in the system) $\times$ (mean time in the system) =

$= \lambda_i \cdot E[\text{service time at device } i] = \lambda_i \cdot \frac{1}{\mu_i}$.

The Utilization Law is also written

$$\rho_i = \lambda_i E[S_i] = X_i E[S_i].$$

6.5 [H] Proof of Little's Law for Closed Systems

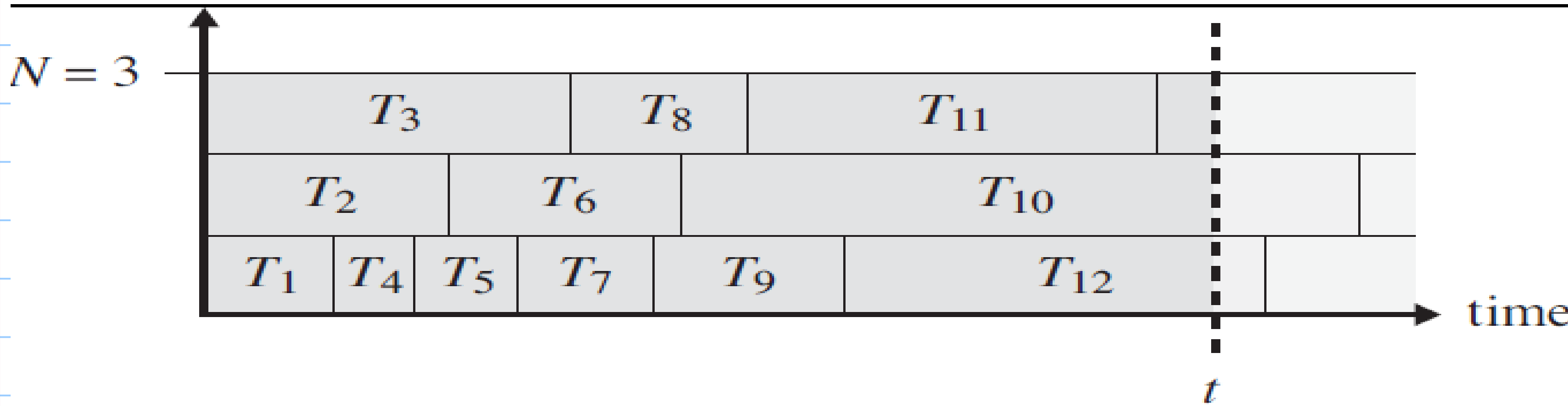
Theorem 6.6 (Little's Law for Closed Systems Restated) *Given any closed system (either interactive or batch) with multiprogramming level N and given that $\overline{T}^{\text{Time Avg}}$ and X exist and that $\lambda = X$, then*

$$N = X \cdot \overline{T}^{\text{Time Avg}}.$$

$X = \lim_{t \to \infty} \frac{C(t)}{t}$, where $C(t)$ is the number of system completions by time t

$\lambda = \lim_{t \to \infty} \frac{Q(t)}{t}$, where $Q(t)$ is the number of jobs generated by time t.

(Note: not the number of arrivals.)



N = 3
T3
T8
T11
T2
T6
T10
T1
T4
T5
T7
T9
T12
time
t

6.6 [H] Generalized Little's Law

Little's Law has been generalized to higher moments, e.g., $E[N^2]$, $E[T^2]$, but only under restrictive conditions, such as a system with a single FCFS queue.

6.7 [H] Examples applying Little's Law

Example 1 (Closed Interactive System)

What is the throughput, X, of the system?

$$N = X \cdot E[T] = X \cdot (E[Z] + E[R])$$

$$\Rightarrow X = \frac{N}{E[Z] + E[R]} = \frac{10}{5+15} = \frac{1}{2} \frac{\text{jobs}}{\text{sec}}$$

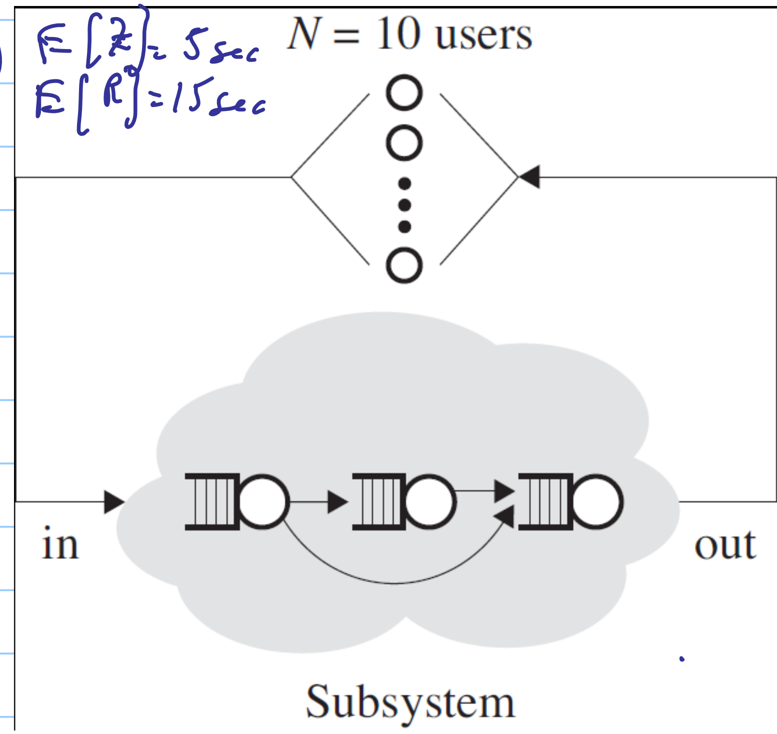


Fig. 6.8 [H]

Response Time Law for Closed Systems:

$$E[R] = \frac{N}{X} - E[Z]$$

$L \quad \frac{10}{1/2} - 5 = 20 - 5 = 15$

Example 2: A more complex interactive system

$X_{disk3} = 40 \frac{requests}{sec}$

$E[S_{disk3}] = 0.0225$ sec

$E[N_{disk3}] = 4$ jobs

What is the utilization of disk 3?

$\rho_{disk3} = X_{disk3} \cdot E[S_{disk3}] =$

$= 40 \cdot 0.0225 = 90\%$

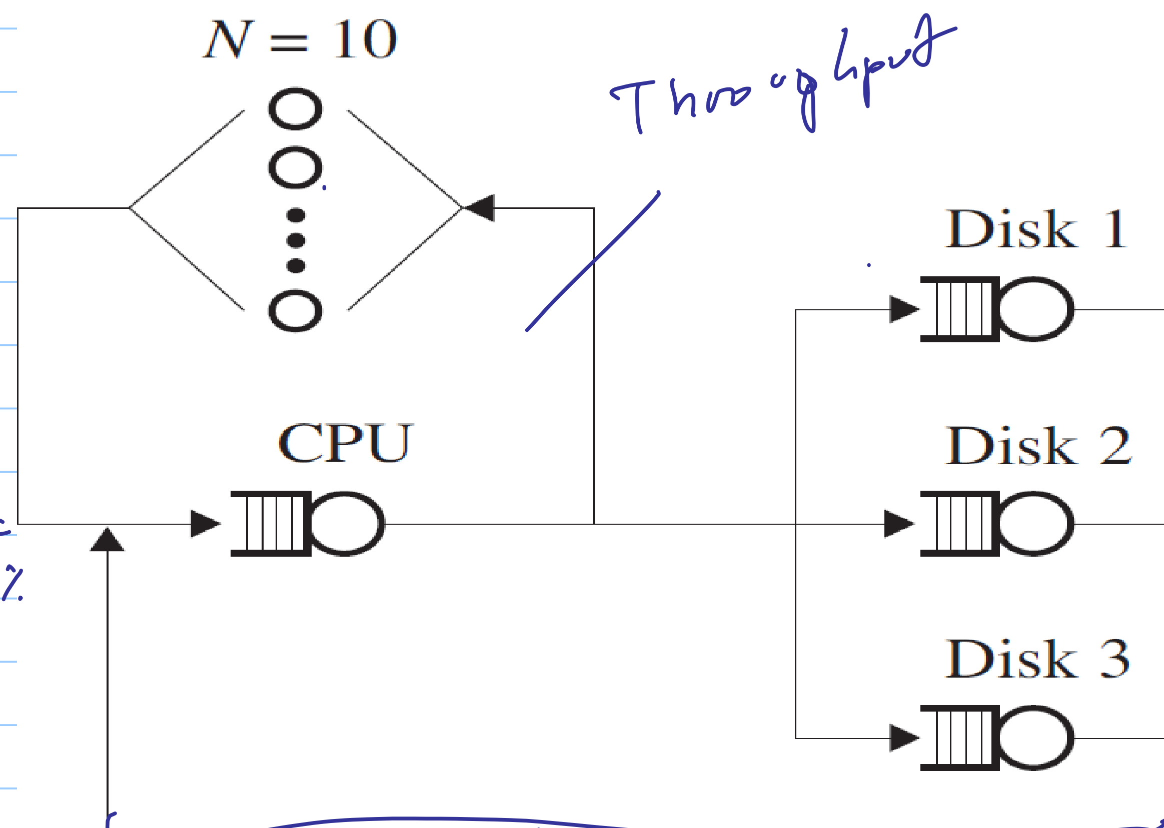


Fig. 6.9 [H]

What is the mean time spent queueing at disk 3?

$T_{disk\,3}$ is the time spent queueing plus serving at disk 3

T_Q^{disk3} is " " " in the queue at disk 3.

$$E[T_{disk3}] = \frac{E[N_{disk3}]}{X_{disk3}} = \frac{4}{40} = 0.1 \text{ sec}$$

$$E[T_Q^{disk3}] = E[T_{disk3}] - E[S_{disk3}] = 0.1 - 0.0225 = 0.0775 \text{ sec}$$

Find the number of requests queued at disk 3.

$$E[N_Q^{disk3}] = E[N_{disk3}] - E[\text{Number served at disk 3}] =$$

$$= 4 - \rho_{disk3} = 4 - 0.9 = 3.1$$

Alternatively, use Little's Law on the queue at Disk3:

$$E[N_Q^{disk3}] = E[T_Q^{disk3}] \cdot X_{disk3} = 0.075 \times 40 = 3.1$$

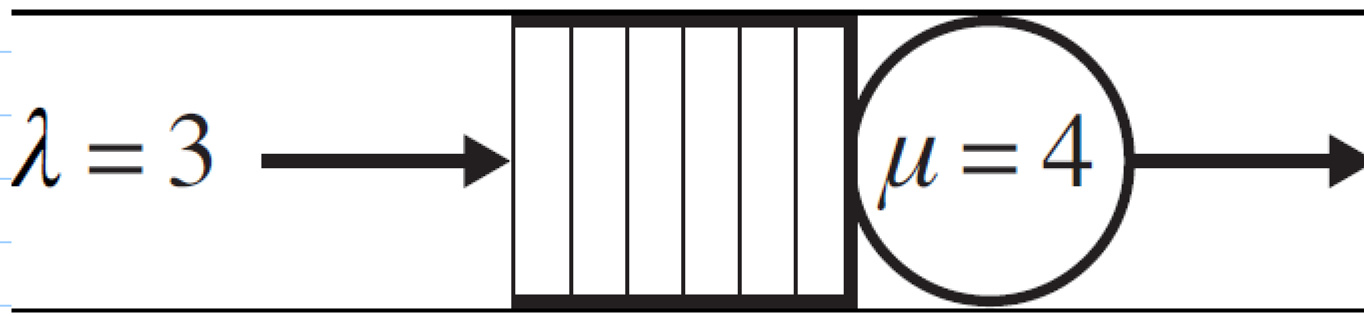
What is the system throughput?

$$\boxed{\begin{array}{l} E[N_{not\,thinking}] = 7.5 \\ N = 6 \\ E[Z] = 5 \end{array}}$$

$$\left\{\begin{array}{l} X = \dfrac{N}{E[R] + E[Z]} = \dfrac{10}{E[R] + 5} \\[2ex] E[R] = \dfrac{E[N_{not\text{-}thinking}]}{X} = \dfrac{7.5}{X} \end{array}\right.$$

$$\Rightarrow X = .5, \quad E[R] = 15$$

Example 3: A finite buffer



7 jobs in system:
6 in queue, 1 served

Fig 6.10 [H]

$\lambda \neq X$, so Little's Law does not apply to the finite buffer system.
However, the rate of jobs that get through is $\lambda(1 - P\{7 \text{ jobs in the system}\})$; this is the <u>effective arrival rate</u>. Little's Law applies with the effective arrival rate:

$$E[N] = \lambda(1 - P\{7 \text{ jobs in the system}\}) \cdot E[T].$$

6.8 [H] More operational laws: the forced flow law

$$X_i = E[V_i] \cdot X$$

X is the system throughput
X_i is the device throughput
V_i is the number of visits to device i per job.

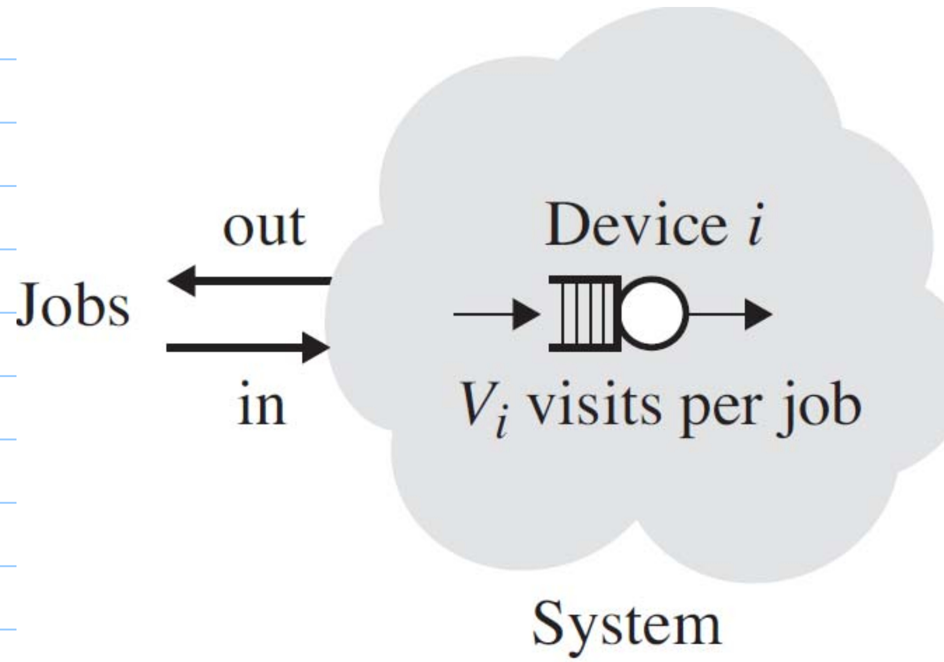


Fig. 6.11 [H]

Example of Forced Flow Law

$C_a = C_{cpu} \cdot 80/181$

$C_b = C_{cpu} \cdot 100/181$

$C = C_{cpu} \cdot 1/181$

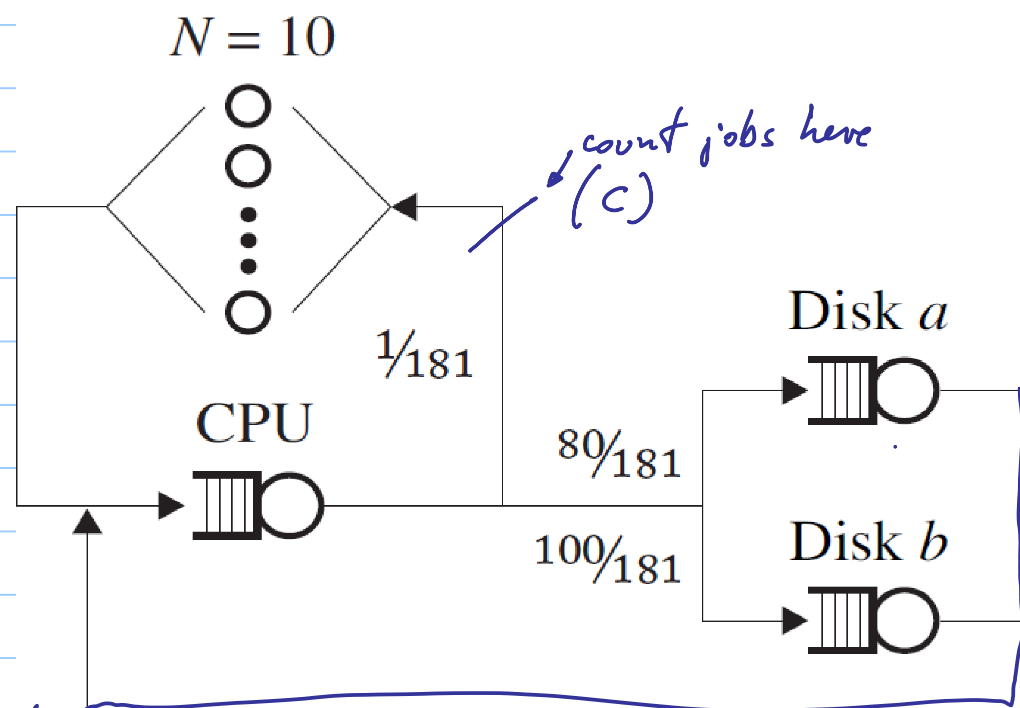
$C_{cpu} = C_a + C_b + C$. So,

$E[V_a] = E[V_{cpu}] \cdot 80/181$

$E[V_b] = E[V_{cpu}] \cdot 100/181$

$1 = E[V_{cpu}] \cdot \frac{1}{181}$

$E[V_{cpu}] = E[V_a] + E[V_b] + 1$



$\Rightarrow E[V_{cpu}] = 181,\ E[V_a] = 80,\ E[V_b] = 100$. // Fig. 6.12 [H] Calculating the visit ratios

6.9 [H] Computing operational laws

Simple Example

$N = 25$ (25 terminals), 18 sec avg think time ($E[Z] = 18$)

20 visits per interaction on avg. to a specific disk ($E[V_{disk}] = 20$)
30% utilization of that disk ($\rho_{disk} = 0.3$)
0.025 sec avg. service time per visit to that disk ($E[S_{disk}] = .025$)
What is the mean response time ($E[R] = E[T] - E[Z]$)?

The Response Time Law for Closed System states:

$E[R] = \frac{N}{X} - E[Z] = \quad N = 25, \; E[Z] = 18. \quad X?$

$= \frac{25}{0.6} - 18 \simeq 41.7 - 18 = 23.7$ sec

The Forced Flow Law states

$X_i = E[V_i] \cdot X \Rightarrow X = \frac{X_{disk}}{E[V_{disk}]} \quad E[V_{disk}] = 20. \quad X_{disk}?$

$= \frac{12}{20} = 0.6 \; \frac{\text{interactions}}{\text{sec}}$

The Utilization Law states

$\rho_i = \frac{\lambda_i}{\mu_i}$, or (p. 101), $\rho_i = X_i \, E[S_i]$, i.e. $X_{disk} = \frac{\rho_{disk}}{E[S_{disk}]} = \frac{.3}{.025} = 12 \; \frac{\text{requests}}{\text{sec}}$

Working backwards

Harder example
(Lazowska et al.)

$N = 23$

$E[Z] = 21$ sec

$X = 0.45$ $\frac{\text{interactions}}{\text{second}}$

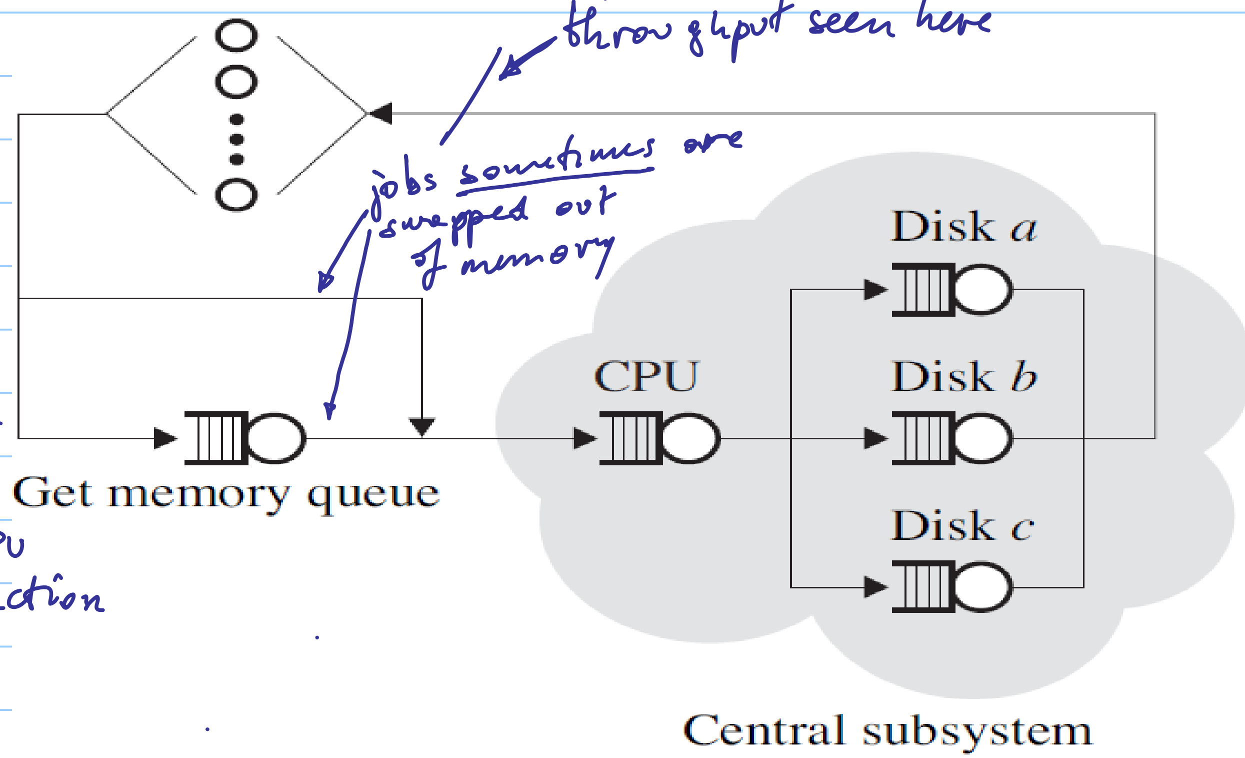
$E[N_{\text{getting memory}}] = 11.65$

$E[V_{CPU}] = 3$ visits to CPU per interaction

$E[S_{CPU}] = 0.21$ sec

System throughput seen here

jobs sometimes are swapped out of memory



What is the average amount of time that elapses between getting a memory partition and completing the interaction?

$E[\text{Time in Central Subsystem}] = E[\text{Response Time}] - E[\text{Time to get memory}]$

4.23 30.11 25.88

By the Response Time Law,

$$E[\text{Response Time}] = \frac{N}{X} - E[Z] = \frac{23}{0.45} - 21 \cong 51.11 - 21 = \underline{30.11 \text{ sec}}$$

By Little's Law for closed systems ($\bar{N} = X\,\bar{T}$);

$$\Rightarrow E[\text{Time to get memory}] = \frac{E[\text{Number Getting Memory}]}{X} = \frac{11.65}{0.45} \cong 25.88 \text{ sec}$$

i) What is the CPU utilization? By the utilization law (version of p. 101):

$$\rho_{CPU} = X_{CPU} \cdot E[S_{CPU}] = (\text{Forced Flow Law}) =$$

$$= X \cdot E[V_{CPU}] \cdot E[S_{CPU}] = 0.45 \cdot 3 \cdot 0.21 \cong 0.28$$

6.10 [H] Device demands

Define D_i as the total demand of one job to device i:

$D_i = \sum_{j=1}^{V_i} S_i^{(j)}$, where $S_i^{(j)}$ is the time required by the j-th visit of a job to device i.

$E[D_i] = E[V_i] \cdot E[S_i]$ if V_i and $S_i^{(j)}$ are independent. (See below)

To compute $E[D_i]$:

$$E[D_i] = \frac{B_i}{C} = \frac{\text{total busy time of device } i \text{ (for a long time } t)}{\text{number of system completions on time } t}.$$

utilization law (3rd version p. 101)

$$e_i = X_i \cdot E[S_i] = X \cdot E[V_i] \cdot E[S_i] = X \cdot E[D_i]$$

forced flow law

(assumption of) independence of V_i and S_i

$$\boxed{\rho_i = X \cdot \mathbf{E}[D_i]}$$

The Bottleneck Law

If $D_i = \sum_{j=1}^{V_i} S_i^{(j)}$ and V_i and the $S_i^{(j)}$ are independent, then

service demand at device i

$$E[D_i] = E[V_i] \cdot E[S_i].$$

visits to device i

The independence assumption may be rephrased as: The number of visits a job makes to a device does not affect (and is not affected by) its service demand at the device.

We show a generic version of the equality in red:

Let $S = \sum_{i=1}^{N} X_i$, $N \perp X_i \ \forall i$

$$E[S] = E\left[\sum_{i=1}^{N} X_i\right] = \sum_n E\left[\sum_{i=1}^{N} X_i \,\middle|\, N=n\right] \cdot P\{N=n\} = (N \perp X_i) =$$

$$= \sum_n E\left[\sum_{i=1}^{n} X_i\right] P\{N=n\} = \quad (X_i \sim X \ \forall i) =$$

$$= \sum_n E[nX] \cdot P\{N=n\} = \sum_n n E[X] P\{N=n\} = E[X] \sum_n n P\{N=n\} =$$

$$= E[X] \, E[N]$$

Example

$X = 3 \frac{\text{jobs}}{\text{sec}}$, $\quad E[V_{disk}] = 10$, $\quad E[S_{disk}] = 0.01 \text{sec}$

$$E[D_{disk}] = E[V_{disk}] \cdot E[S_{disk}] = 10 \times 0.01 = 0.1 \text{ sec}$$

$$e_{disk} = X \cdot E[D_{disk}] = 3 \times 0.1 = 0.3$$