Little's Law (1961)

Theorem 6.1 (Little's Law for Open Systems)

For any ergodic open system, we have that

\[ E[N] = \lambda E[T], \]

where

- \( E[N] \) is the expected # jobs in the system
- \( \lambda \) is the avg. arrival rate
- \( E[T] \) is the mean time a job spends in the system

Fig. 6.1 [HT]
Theorem (6.2 [H] [H])

FCFS

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

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,

\[ N = X \cdot E[T], \]

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

Note that \( N \) is often set because of considerations such as memory space, in batch systems (cf. Section 2.6.2 [H]). In interactive (terminal-driven) systems, \( N \) is the number of terminals (cf. Section 2.6.1 [H]).

See figures below.
Closed systems; beta

\[ E[T] = E[R] + E[?] \]

- \( E[T] \): time in system
- \( E[R] \): response time
- \( E[?] \): think time
6.4 [t7] Proof of Little's Law for Open Systems

\textbf{Theorem 6.3 (Little's Law for Open Systems Restated)} \quad \text{Given any system where} \quad \lambda, \ X \ \text{exist and where} \ \lambda = X, \ \text{then}

\[ \frac{N}{T_{\text{Time Avg}}} = \lambda \cdot \frac{X}{T_{\text{Time Avg}}} \]

\[ \lim_{t \to \infty} \frac{A(t)}{t} \quad \text{and} \quad \lim_{t \to \infty} \frac{C(t)}{t} \]

\( \lambda = X \), if jobs are not dropped, except in special cases.

Ergodicity implies the assumptions of Thm. 6.3.
Graph of arrivals in an open system (Fig. 6.5 [H])
\[ \sum_{i \in C(t)} T_i \leq \Omega \leq \sum_{i \in A(t)} T_i \]

- \( \sum_{i \in C(t)} T_i \): jobs completed by time \( t \)
- \( \sum_{i \in A(t)} T_i \): jobs arrived by time \( t \)

\[ \Omega = \int_0^t N(s) \, ds \quad \text{("vertical view": sum # jobs in system at any moment in time)} \]

So:
\[ \sum_{i \in C(t)} T_i \leq \int_0^t N(s) \, ds \leq \sum_{i \in A(t)} T_i \]

or equivalently:
\[ \sum_{i \in C(t)} \frac{T_i}{C(t)} \leq \frac{\int_0^t N(s) \, ds}{t} \leq \sum_{i \in A(t)} \frac{T_i}{\Omega(t)} \cdot \frac{\Omega(t)}{t} \]
Taking limits as \( t \to \infty \),

\[
\lim_{t \to \infty} \frac{\sum_{i=0}^{\infty} Ti}{\mathcal{C}(t)} \leq \frac{\mathcal{N}}{\mathcal{T} \text{ Time Avg}} \leq \lim_{t \to \infty} \frac{\sum_{i=0}^{\infty} Q(t)}{\mathcal{Q}(t)} \leq \lim_{t \to \infty} \mathcal{Q}(t)
\]

- \( \mathcal{F} \text{ Time Avg} \)
- \( \mathcal{N} \text{ Time Avg} \)
- \( \mathcal{T} \text{ Time Avg} \)
- \( \mathcal{d} \text{ completion in system rate} \)
- \( \mathcal{d} \text{ completion (throughput)} \)
- \( \mathcal{d} \text{ arrival on system rate} \)

Since \( d = X \),

\[
\frac{\mathcal{N}}{\mathcal{T} \text{ Time Avg}} = 1
\]
Corollary 6.4 (Little’s Law for Time in Queue) Given any system where $N_Q^{\text{Time Avg}}$, $T_Q^{\text{Time Avg}}$, $\lambda$, and $X$ exist and where $\lambda = X$, then

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

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 with average arrival rate $\lambda_i$ jobs/sec and average service rate $\mu_i$ jobs/sec, where $\lambda_i < \mu_i$. Let $\rho_i$ denote the long-run fraction of time that the device is busy. Then

$$\rho_i = \frac{\lambda_i}{\mu_i}.$$
The expected number of jobs in the system is

1 \times P_{\text{system is busy}} + 0 \times P_{\text{system is idle}} = 1 \times P_{\text{system is busy}} = P_i.

So, applying Little's Law, we have:

\[ E_i = E_i = \text{Expected number of jobs in the system} = \]

\[ = (\text{arrival rate in the system}) \times (\text{mean time in the system}) = \]

\[ = \lambda_i \times E[s_i] = \lambda_i \times \frac{1}{\mu_i}. \]

The Utilization Law is also written

\[ E_i = d_i \times E[s_i] = x_i \times E[s_i]. \]
6.5 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 $\bar{T}_{\text{Time Avg}}$, and $X$ exist and that $\lambda = X$, then

$$N = X \cdot \bar{T}_{\text{Time Avg}}.$$ 

$$X = \lim_{t \to \infty} \frac{C(t)}{t},$$ 
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.)
\[ \sum_{i \in \mathcal{I}(t)} T_i \leq N \cdot t \leq \sum_{i \in \mathcal{A}(t)} T_i \]

\[ \sum_{i \in \mathcal{I}(t)} T_i \leq N \leq \sum_{i \in \mathcal{A}(t)} T_i \]
\[
\sum_{i \in C(t)} C(t) \leq N \leq \sum_{i \in \sigma(t)} \alpha(t) \quad t \in T.
\]

\[
\lim_{t \to \infty} \sum_{i \in C(t)} \frac{C(t)}{t} \leq N \leq \sum_{i \in \sigma(t)} \lim_{t \to \infty} \frac{\alpha(t)}{t} \quad t \in T.
\]

\[
T \leq N \leq T + 1
\]

\[
N = X \cdot \text{Trim}\text{Ang}
\]
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?

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

$\Rightarrow X = \frac{N}{E[T] + E[R]} = \frac{10}{5+15} = \frac{1}{2} \text{ (bbs/sec)}$

Response Time Law for Closed Systems:

$E[R] = \frac{N}{X} \cdot E[T]$
Example 2: A more complex interactive system

$X_{disk3} = \mu_0 \text{ request }_3 \text{ sec}$

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

$E[N_{disk3}] = \text{ # jobs}$

What is the utilization of Disk 3?

$E_{disk3} = X_{disk3} \cdot E[S_{disk3}] = 0.0 \times 0.0225 = 0.0%$

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

$N = 10$

Throughput

Fig. 6.9 [H]
What is the mean time spent querying at disk 3?

$T_{\text{disk }3}$ is the time spent querying plus serving at disk 3.

$T_{\text{disk }3} = T_{\text{qs}} + \ldots + \text{time in queue at disk }3$

$E[T_{\text{disk }3}] = \frac{E[N_{\text{disk }3}]}{4\sigma} = \frac{4}{40} = 0.1 \text{ sec}$

$E[T_{\text{qs}}] = E[T_{\text{disk }3}] - E[S_{\text{disk }3}] = 0.1 - 0.025 = 0.075 \text{ sec}$

Find the number of requests queued at disk 3.

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

$= 4 - E[\text{disk }3] = 4 - 0.9 = 3.1$
Alternatively, use Little's Law on the queue at Disk 3:

\[ E[N_{\text{disk}3}] = E[T_{\text{disk}3}] \times 40 = 0.075 \times 40 = 3.1 \]

What is the system throughput?

\[ X = \frac{N}{E[R] + E[Q]} = \frac{10}{E[R] + 5} \]

\[ E[R] = \frac{E[N_{\text{not-throttled}}]}{X} = \frac{7.5}{X} \]

\[ \Rightarrow X = \frac{5}{X}, \quad E[R] = 15 \]
Example 3: A finite buffer

\[ \lambda = 3 \quad \text{jobs in system;} \quad \mu = 4 \quad \text{in queue, 1 served} \]

\[ E[N] = \lambda (1 - P \{ \text{7 jobs in the system} \}) \cdot E[T]. \]
6.8 [HF] 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 [HF]**
Example of Forced Flow Law

\[ C_a = C_{cpu} \cdot \frac{80}{181} \]

\[ C_b = C_{cpu} \cdot \frac{100}{181} \]

\[ C_c = C_{cpu} \cdot \frac{1}{181} \]

\[ C_{cpu} = C_a + C_b + C_c \quad \text{So,} \]

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

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

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

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

\[ \Rightarrow E[V_{cpu}] = 181, E[V_a] = 80, E[V_b] = 100. \]

Fig. 6.12 - Calculating the visit ratios.
6. 9c(H) Combining 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 \( (E[\ell_{disk}] = 0.3) \)

0.025 sec avg. service time per visit to that disk \( (E[S_{disk}] = 0.025) \)

What is the mean response time \( (E[R] = E[T] - E[Z]) \)?

The response time law for closed system states:
\[ E[C] = \frac{N}{X} - E[Z] \]  \[ N = 25, \ E[Z] = 18 \]

The Forced Flow lower state:

\[ X_i = E[V_i] \cdot X \Rightarrow X = \frac{X_{disk}}{E[V_{disk}]} \]

\[ E[V_{disk}] = 20 \]

The Utilization lower state:

\[ E_i = \frac{d_i}{\mu_i}, \ \text{or (p. 101)} \]

\[ E_i = X_i \cdot E[Z_i], \ \text{i.e.} \]

\[ X_{disk} = \frac{E_{disk}}{E[S_{disk}]} = \frac{3}{0.25} = 12 \]

Working backwards
Hander example (Zorowska et al.)

\( N = 23 \)
\( E[2] = 21 \text{ sec} \)
\( X = 0.45 \text{ interactions per second} \)
\( E[N \text{ getting memory}] = 11.65 \)
\( E[V_{\text{CPU}}] = 3 \text{ visits to CPU per interaction} \)
\( E[S_{\text{CPU}}] = 0.21 \text{ sec} \)

System throughput seen here

Jobs sometimes are swapped out of memory

Central subsystem

Disk a

CPU

Disk b

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

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

By the Response Time Law,

\[ E[\text{Response Time}] = \frac{N}{K} \]

\[ E[\text{Time to get memory}] = E[\text{Number Getting Memory}] \times \frac{21}{0.45} = 47.78\text{ sec} \]

By Little's law for closed systems \( (\bar{N} = \frac{X}{F}) \)

\[ E[\text{Time to get memory}] = \frac{E[\text{Number Getting Memory}]}{X} \times \frac{21}{0.45} = 25.88\text{ sec} \]
What is the CPU utilization? By the utilization law (version of p. 101):

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

\[ = X \cdot E[V_{CPU}] \cdot E[S_{CPU}] = 0.45 \cdot 3 \cdot 0.21 \approx 0.28 \]
6.10 Define Device demands

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

$$D_i = \sum_{j=1}^{V_i} S_{ij}$$

where $S_{ij}$ is the time required by the $j$-th visit of a job to device $i$.

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

To compute $E[D_i]$:

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

\[ \rho_i = X \cdot E[D_i] \]

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

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

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 \( S = \sum_{i=1}^{N} X_i \), \( N \perp X_i \perp X_i \)

\[
E[S] = E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i|N=n] \cdot P_{N=n} = (N + X_i) = \ldots
\]
\[= \sum \mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] P\{N=n\} = (X_i \sim X \in N) = \sum \mathbb{E}[nX] \frac{P\{N=n\}}{n} = \sum \mathbb{E}[X] P\{N=n\} = \mathbb{E}[X] \sum \frac{P\{N=n\}}{n} = \mathbb{E}[X] \mathbb{E}[N] \]

Example:

\[X = 3 \text{ in } \text{sec} \quad \mathbb{E}[V_{disk}] = 10 \quad \mathbb{E}[S_{disk}] = 0.01 \text{ sec} \]

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

\[E_{disk} = X \cdot \mathbb{E}[O_{disk}] = 3 \times 0.1 = 0.3 \]