1 Notes on Little's Law ($l = \lambda w$)

We consider here a famous and very useful law in queueing theory called Little's Law, also known as $l = \lambda w$, which asserts that *the time average number of customers in a queueing system, l, is equal to the rate at which customers arrive and enter the system, λ, $\times$ the average sojourn time of a customer, w.* For example, in a four-year college, in which (on average) 5000 first-year students enter per year, the average number of students present at this college is given by $5000 \times 4 = 20,000$.[1] Our presentation is based on a sample-path analysis and the reader should not assume apriori that any specific stochastic assumptions are in force. Imagine instead that a sample path is being studied of some stochastic queueing process.

1.1 Little's Law

We consider a queueing "system" in which customers arrive from the outside, spend some time in the system and then depart. C_n denotes the n^{th} customer, and this customer arrives and *enters* the system at time t_n. The point process $\{t_n : n \geq 1\}$ is assumed an increasing (to ∞) sequence of non-negative numbers with counting process $\{N(t) : t \geq 0\}$; $N(t) = \max\{n : t_n \leq t\}$ ($= 0$ if there are no arrivals by time t), the number of arrivals during $(0, t]$. Upon entering the system, C_n spends $W_n \geq 0$ units of time inside the system (C_n's *sojourn time*) and then departs the system at time $t_n^d = t_n + W_n$. Note that the departure times are not necessarily ordered, which means that we do not require that customers depart in the same order that they arrived (think of a supermarket). $\{N^d(t) : t \geq 0\}$ denotes the counting process for the departure times $\{t_n^d\}$; $N^d(t) =$ the number of customers who have departed by time t; note that $N^d(t) \leq N(t),\ t \geq 0$.

A customer C_n is in the system at time t if and only if $t_n \leq t < t_n^d = t_n + W_n$, and we define $L(t)$, the total number of customers in the system at time t, by

$$L(t) = \sum_{n=1}^{\infty} I\{t_n \leq t < t_n^d\} \tag{1}$$

$$= \sum_{\{n : t_n \leq t\}} I\{W_n > t - t_n\} \tag{2}$$

$$= \sum_{n=1}^{N(t)} I\{W_n > t - t_n\}. \tag{3}$$

Define (when the limits exist)

$$\lambda \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{N(t)}{t}, \text{ \textit{the arrival rate into the system,}} \tag{4}$$

$$w \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} W_j, \text{ \textit{average sojourn time,}} \tag{5}$$

[1] Little's Law is named after John D.C. Little, who was the first to prove a version of it, in 1961. Little's original framework was stochastic however. In 1974 S. Stidham proved a sample-path version which is what we present here.

(6) $$l \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{1}{t}\int_0^t L(s)ds, \text{ time average number in system.}$$

Theorem 1.1 **(** $l = \lambda w$**)** *If both λ and w exist and are finite, then l exists and $l = \lambda w$.*

$L = \lambda W$ is one of the most general and versatile laws in queueing theory, and, if used in clever ways, can lead to remarkably simple derivations. The trick is to choose what the "system" is, and what the arrivals to this system are. For example, given a complicated network of queues, the "system" can be the waiting area of an isolated node of interest, or it can be one (or all together) of the service areas, etc.

The area under the path of $L(s)$ from 0 to t, $\int_0^t L(s)ds$, is simply the sum of whole and partial sojourn times (e.g., rectangles of height 1 and lengths W_j). If the system is empty at time t, then the area is exactly $W_1 + \cdots + W_{N(t)}$; otherwise some partial pieces must be considered. The following inequality is easily derived:

(7) $$\sum_{\{j:t_j^d\le t\}} W_j \le \int_0^t L(s)ds \le \sum_{\{j:t_j\le t\}} W_j = \sum_{j=1}^{N(t)} W_j.$$

To see this:

$$\begin{aligned}
&(8) & \int_0^t L(s)ds &= \int_0^t \{\sum_{\{j:t_j\le s\le t\}} I\{W_j > s - t_j\}\}ds \\
&(9) & &= \sum_{\{j:t_j\le t\}} \int_{t_j}^t I\{W_j > s - t_j\}ds \\
&(10) & &= \sum_{\{j:t_j\le t\}} \min\{W_j, t - t_j\}.
\end{aligned}$$

Since $\min\{W_j, t - t_j\} \le W_j$, the upper bound in (7) is immediate. For the lower bound

$$\begin{aligned}
&(11) & \sum_{\{j:t_j\le t\}} \min\{W_j, t - t_j\} &= \sum_{\{j:t_j+W_j\le t\}} W_j + \sum_{\{j:t_j\le t,\ t_j+W_j>t\}} t - t_j \\
&(12) & &\ge \sum_{\{j:t_j+W_j\le t\}} W_j = \sum_{\{j:t_j^d\le t\}} W_j.
\end{aligned}$$

Dividing the upper bound by t, and re-writing $1/t = (N(t)/t)(1/N(t))$, we obtain

$$(\frac{N(t)}{t})\ \frac{1}{N(t)}\sum_{j=1}^{N(t)} W_j.$$

Taking the limit as $t\longrightarrow\infty$ yields λw, due to the assumed existence of the two limts in (4) and (5) for λ and w (and their assumed finiteness). Thus the proof of $L = \lambda w$ can be completed by showing that the lower bound in (7) when divided by t converges to λw as well, that is, we must show that

(13) $$\lim_{t\to\infty} \frac{1}{t}\sum_{\{j:t_j^d\le t\}} W_j = \lambda w.$$

Lemma 1.1 *If λ and w exists and are finite, then*

$$\lim_{n\to\infty} \frac{W_n}{n} = 0, \tag{14}$$

$$\lim_{n\to\infty} \frac{W_n}{t_n} = 0. \tag{15}$$

Proof :

$$\frac{W_n}{n} = \frac{1}{n}\sum_{j=1}^{n} W_j - \frac{1}{n}\sum_{j=1}^{n-1} W_j \tag{16}$$

$$= \frac{1}{n}\sum_{j=1}^{n} W_j - (\frac{n-1}{n})(\frac{1}{n-1})\sum_{j=1}^{n-1} W_j \tag{17}$$

$$\to w - w = 0, \tag{18}$$

by (5) and finiteness of w. (14) is thus proved.

From (4) it follows that $M(t_n)/t_n \to \lambda$ because it is assumed that $t_n \to \infty$. Assuming that the arrival times are strictly increasing yields $M(t_n) = n$ and thus that

$$\frac{n}{t_n} = \frac{M(t_n)}{t_n} \to \lambda.$$

If the arrival times are not strictly increasing (so-called *batch* arrivals), then

$$\frac{n}{t_n} \leq \frac{M(t_n)}{t_n} \to \lambda.$$

Thus in either case, from (14)

$$\begin{aligned} \frac{W_n}{t_n} &= \frac{W_n}{n}\,\frac{n}{t_n} \\ &\leq \frac{W_n}{n}\frac{M(t_n)}{t_n} \\ &\to 0\ \lambda = 0, \end{aligned}$$

because λ is assumed finite. (15) is thus proved. ■

We are now prepared to finish the proof of $L = \lambda w$:
Proof :[$l = \lambda w$] To prove (13) it suffices to prove

$$\underline{\lim}_{t\to\infty} \frac{1}{t} \sum_{\{j:t_j^d \leq t\}} W_j \geq \lambda w, \tag{19}$$

because we already established λw as an upper bound.

To this end, choose any $\epsilon > 0$ no matter how small. From Lemma 1.1 there exists an integer n_0 such that $W_j \leq \epsilon t_j,\ j \geq n_0$, and thus that $t_j^d = t_j + W_j \leq (1+\epsilon)t_j,\ j \geq n_0$.

Thus

$$\{j : t_j^d \le t\} \supset \{j : j \ge n_0,\ (1+\epsilon)t_j \le t\} = \{j : j \ge n_0,\ t_j \le \frac{t}{1+\epsilon}\},$$

from which it follows that

$$\sum_{\{j:t_j^d \le t\}} W_j \ge \sum_{j=n_0}^{N(\frac{t}{1+\epsilon})} W_j.$$

The rhs of the above can be re-written as

$$\sum_{j=1}^{N(\frac{t}{1+\epsilon})} W_j - \sum_{j=1}^{n_0-1} W_j.$$

Dividing the first piece by t and letting $t \to \infty$ yields $\lambda w/(1+\epsilon)$ by the same argument used on the upper bound in (7). The second piece is a constant hence when divided by t, tends to 0. Thus we conclude that for any $\epsilon > 0$,

$$\underline{\lim}_{t\to\infty} \frac{1}{t} \sum_{\{j:t_j^d \le t\}} W_j \ge \lambda w/(1+\epsilon).$$

Since $\epsilon > 0$ was chosen arbitrary, we conclude that (19) holds.

■

A consequence of the proof of Theorem 1.1 ($l = \lambda w$) is

Proposition 1.1 *If λ exists and is finite, and if $W_n/n \to 0$, then*

$$\lim_{t\to\infty} \frac{N^d(t)}{t} = \lambda,$$

the departure rate exists and equals the arrival rate λ: Departure rate = arrival rate.

Proof :

Since $N^d(t) \le N(t)$, $\overline{\lim}_{t\to\infty} \frac{N^d(t)}{t} \le \overline{\lim}_{t\to\infty} \frac{N(t)}{t} = \lambda$; an upper bound is established. Now for a lower bound: (15) followed from (14) only (a condition that is weaker than assuming w exists and is finite); hence as in the proof of $l = \lambda w$, for every $\epsilon > 0$ there exists an integer n_0 such that $N^d(t) \ge N(t/(1+\epsilon)) - n_0$, yielding

$$\underline{\lim}_{t\to\infty} \frac{N^d(t)}{t} \ge \lambda.$$

■

1.2 Applications of $l = \lambda w$

1. $Q = \lambda d$: If we let the "system" be the queue area (where customers wait before entering service), then average sojourn time is average delay in queue, d, l becomes average number waiting in queue, Q, and $l = \lambda w$ takes on the form $Q = \lambda d$.

2. *Infinite server queue:* For any infinite server queue with arrival rate $\lambda < \infty$ and average service time $1/\mu < \infty$, l exists and $l = \rho = \lambda/\mu$, because $w = 1/\mu$ here: $W_n = S_n$.

3. *Proportion of time the server is busy in a single-server queue:* Customers arrive to the queue at rate $\lambda < \infty$ and have average service time $1/\mu < \infty$. Let λ_s denote the rate at which customers *enter* service. Letting the "system" be the server, and letting $L_s(t)$ denote the number of customers in service at time t, with time-average l_s, we conclude that $l_s = \lambda_s(1/\mu)$, because $W_n = S_n$ here. It can be proved that $\lambda_s = \lambda$ when $\rho < 1$ and $\lambda_s = \mu$ when $\rho \geq 1$. Thus $l_s = \rho$ if $\rho < 1$; $l_s = 1$ if $\rho \geq 1$. But since $L_s(t) = 1$ if the server is busy at time t, and $L_s(t) = 0$ if the server is idle at time t, we conclude (from the fact that l_s is a time average) that l_s is the long-run proportion of time the server is busy:

   For any single-server queue, with arrival rate λ and mean service time $1/\mu$, the long-run proportion of time the server is busy exists and is equal to $\min\{\rho, 1\}$.

4. *FIFO M/M/1 queue with $\rho < 1$:* We earlier solved for the balance equations of number in system; $P_n = (1-\rho)\rho^n,\ n \geq 0$. Thus $l = \sum_n P_n = \rho/(1-\rho)$. $l = \lambda w$ implies that $w = l/\lambda$ yielding the following expression for average sojourn time:

$$w = \frac{\frac{1}{\mu}}{1-\rho}.$$

   This has a nice interpretation: From PASTA, the average number of customers found in the system by an arrival is the same as the time average number in system, l. Moreover, by the memoryless property of the service times, the customer in service has a remaining service as good as new. Thus (adding in their own service time), such an arriving customer must wait, on average, for the completion of $1 + l$ iid service times each of mean $1/\mu$. Thus $w = (1/\mu)(1+l) = \frac{\frac{1}{\mu}}{1-\rho}$.