Little's Law and Related Results [*]

Ronald W. Wolff
Department of Industrial Engineering and Operations Research
University of California at Berkeley

January 29, 2011

Abstract

For queues and similar systems, Little's Law, often written $L = \lambda W$, states that the time average of the number of customers in system is the product of the arrival rate and the customer average of their waiting times. We treat *sample-path* and *stationary* versions of this result and extensions, including $H = \lambda G$, and a distributional Little's Law. We illustrate how these results are applied and treat the connection between these results and a rate conservation law.

Keywords: time average, customer average, sample path, $L = \lambda W$, $H = \lambda G$, RCL

Queueing theory is about the analysis of mathematical models where entities called *customers* arrive at some facility called the *system*, spend *waiting time* in system, and then depart.

In its early history, methods of analysis were tailored to individual models and varied widely. There were no theorems of any consequence that held across the board from model to model. This changed in 1961 with "$L = \lambda W$" [12] by John Little, often called "Little's Law" today, and we will sometimes abbreviate as "LL". It, together with extensions, is of fundamental importance for both the foundations of queueing theory and its applications. The literature on this topic since then is enormous. It is impossible to review all of that here. Instead, our goal is to explain LL and important extensions, demonstrate their usefulness, show the connection with some related topics, and provide a literature guide sufficient for further investigation.

LL states that the time average of the number of customers in system is the product of the arrival rate and the customer average of the waiting times.

This article is organized as follows. We present an intuitive explanation of LL and illustrate applications in Section 1. We prove the sample-path version of Little's Law in Section 2, treat the stationary version in Section 3, and present the extension of LL to $H = \lambda G$ in Section 4. We relate these results to a rate conservation law in Section 5, present a distributional Little's Law in Section 6, and briefly review the literature and other related results in Section 7.

[*] To appear in *Wiley Encyclopedia of Operations Research and Management Science*, J.J. Cochran Ed., Vol. 4, pp. 2828–2841, John Wiley & Sons, New York, 2011.

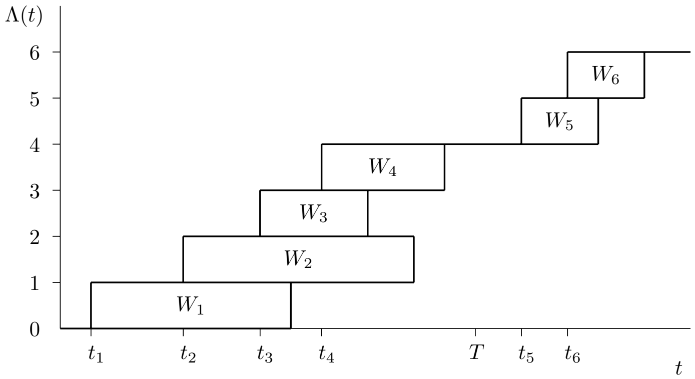

Figure 1: Waiting Time Data

1 Notation, Explanation, Applications

For $i \geq 1$, let customer C_i arrive at time t_i, have waiting W_i, and depart at $t_i + W_i$, where $t_0 \equiv 0 \leq t_1 \leq t_2 \leq \ldots$. For any $t \geq 0$, C_i is *in the system* at t if $t_i \leq t < t_i + W_i$. Define the corresponding indicator, $I_i(t) = 1$ if $t_i \leq t < t_i + W_i$, and $I_i(t) = 0$ otherwise, where $\int_0^\infty I_i(t)\, dt = W_i$. Let $N(t) = \sum_{i=1}^\infty I_i(t)$, the *number of customers in system* at time t, and $\Lambda(t) = \max\{i : t_i \leq t\}$, the number of arrivals by time t. With $\Lambda(t)$, we rewrite $N(t) = \sum_{i=1}^{\Lambda(t)} I_i(t)$.

From customer data $\{t_i\}$ and $\{W_i\}$ for all i, we can construct $\{N(t)\}$ for all t. We plot $\{\Lambda(t)\}$ as a step function and show some customer waiting-time data in Figure 1, and the constructed $\{N(t)\}$ in Figure 2.

In Figure 1, C_3 arrives after but departs before C_2. There is no way to deduce this from Figure 2, which tells us only how many customers are in the system, not *which* customers are there. It has less information than Figure 1.

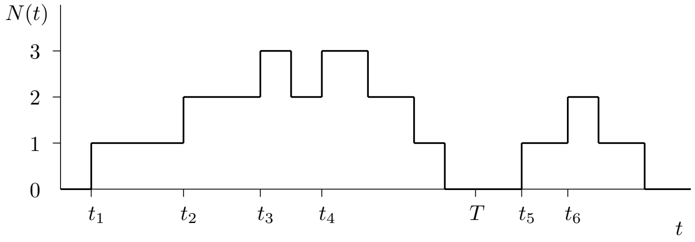

Figure 2: Number of Customers in System Data

While waiting times are lengths of time, they are also displayed as rectangles of height one in Figure 2. Thus *length* W_i is also *area* W_i. For each t, $N(t)$ is simply the number of waiting time rectangles that contain point t.

In a (stochastic) queueing *model*, t_i, W_i, and $N(t)$ are random variables; $\{t_i, i \geq 1\}$, $\{W_i, i \geq 1\}$, and $\{N(t), t \geq 0\}$ are stochastic processes. These random quantities are defined on some sample space. At ω, some arbitrary point in this space, $t_i(\omega)$, $W_i(\omega)$, and $N(t,\omega)$ are numbers, and the collection of stochastic processes is called a *sample path* or *realization.* A simulation of a queueing model generates a sample path. We may also observe a real system as it evolves over time without having a formal model, and view the observed evolution of arrivals, waiting times, and number in system as a sample path.

Usually, assumptions are made such that *stationary versions* of these processes exist, where in particular for these versions, W_i has the same distribution for all i and $N(t)$ has the same distribution for all t. Let W and N be corresponding random variables that have these distributions.

There are two versions of Little's Law. The *sample-path* version relates sample-path averages at some ω in the sample space; the *stationary* version relates the expected values of W and N. The former is easy to understand and so intuitively appealing that to some, it may not require a proof (we think it does). It also applies to data collected on real systems. The latter is a common way to find these quantities, and is an important tool for the analysis of queueing models. We now present the sample-path version, but omit writing ω explicitly; e.g., we write W_i rather than $W_i(\omega)$.

Rectangular area W_i is the contribution to the area under $\{N(t)\}$ that is made by customer C_i. Now consider the area under $\{N(t)\}$ on some interval $(0,T)$, where as in Figures 1 and 2, T has the property that $N(T) = 0$. For any such T, this area may be written either as an integral or a sum,

$$\int_0^T N(t)\,dt = \int_0^T \sum_{i=1}^{\Lambda(T)} I_i(t)\,dt = \sum \int = \sum_{i=1}^{\Lambda(T)} W_i, \tag{1}$$

where in our figures, $\Lambda(T) = 4$. For any T where $N(T) > 0$, (1) does not hold because a portion of at least one rectangle in the sum contributes to the area under $\{N(t)\}$ *after* T. In this case, sum $-$ integral $\equiv$ error ≥ 0. We define long-run averages as limits, when they exist, and name them.

$$L = \lim_{T\to\infty} \frac{1}{T}\int_0^T N(t)\,dt, \quad w = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} W_i, \quad \lambda = \lim_{t\to\infty} \frac{\Lambda(t)}{t}. \tag{2}$$

L is *the average number of customers in the system,*
w is *the average waiting time* (of customers in the system), and
λ is *the arrival rate.*

L and w are the two most common *performance measures* for a queue. We present an intuitive explanation below and a proof in Section 2 of this fundamental relation between them (sample-path version):

Theorem 1 (Little's Law, LL, or $L = \lambda w$). *If limits λ and w exist and are finite, L exists and is finite, where*

$$L = \lambda w. \tag{3}$$

The basic reason for (3) is (1). For any T (where now $N(T) \geq 0$) we write

$$\frac{\int_0^T N(t)dt}{T} = \frac{\sum_{i=1}^{\Lambda(T)} W_i}{T} - \frac{\text{error}}{T} = (\frac{\Lambda(T)}{T})(\frac{\sum_{i=1}^{\Lambda(T)} W_i}{\Lambda(T)}) - \frac{\text{error}}{T}.$$

As T gets large, the first quantity in parenthesis on the right approaches λ, the second approaches w, and the product approaches λw. Hence Theorem 1 holds if the term on the far right approaches 0 as $T \to \infty$. The error term itself need not get small, but only grow more slowly than T.

Note that (3) holds for some unspecified ω. If limits w and λ hold as constants on a set of ω with probability 1 (w.p.1), (3) holds as a constant w.p.1.

Usually, "$L = \lambda w$" is written "$L = \lambda W$". We prefer to use W and N as defined above, in order to write the stationary version as $E(N) = \lambda E(W)$; see Section 3. The notation "L" has a long history, meaning the average length of a waiting *line*. We now present an application, based on personal experience.

Example 1 (Wine Cellars). There are about 3000 bottles in my friend Loy's wine cellar ($L = 3000$). He doesn't keep records of when each bottle was bought and consumed. He wondered "about how old, on average, are these wines when they are drunk?" He estimates that (with some help) he consumes about 300 bottles per year ($\lambda = 300$). The answer to his question is $w = L/\lambda = 10$ years. Not quite true. They are on average 10 years older than when he bought them.

When designing my wine cellar, I had to decide its size. I estimated that on average, wines would be held 5 years when they are drunk ($w = 5$ years), and I would drink about 200 bottles per year ($\lambda = 200$). Thus $L = \lambda w = 1000$ bottles, and the capacity of the wine cellar should be somewhat larger to account for fluctuations about the average. (Fluctuations are much less in wine cellars than in a typical queueing model.) It turned out that the wine cellar was too small. What went wrong? Not LL. I had underestimated λ!

Viewing the system as a "black box", we don't have to know anything about what goes on inside the box; LL holds. In this example, it is natural to think of a collection of wine bottles as *inventory*, and in fact, queueing theory is closely related to what is called *inventory theory.* These theories differ in how we model what goes on inside the box, what aspects of the model may be subject to some control, and what decisions are being contemplated.

In most queueing models, the arrival process is treated as given. At a wine shop, arrivals are scheduled. At a wine cellar, both arrivals and departures are scheduled. We now describe some queueing models of what is inside the box.

In elementary situations, each arrival at a facility has a *service time* to be performed by a server, where the facility has one or more servers. Arrivals finding all servers busy serving other customers wait in a queue. While in the system, a customer may spend some time in queue, followed by a service time.

The time that a customer spends in queue before service begins we call *delay in queue* or just *delay* (sometimes called *waiting time in queue*). Thus a customer's waiting time is the *sum* of the corresponding delay and service time.

Similarly, at any time there will be a *queue length* (*number of customers in queue*) and a *number of customers in service*, where their sum is the *waiting-line length* (*number of customers in system*).

In Figure 3, we show a "snapshot" of a system where customers are represented as circles and servers as square boxes. We have five customers in queue, two servers, two customers in service, and a total of seven customers in system.

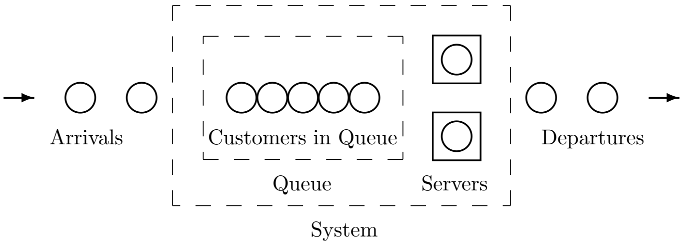

Figure 3: Snapshot of a Queueing System

Let C_i have delay D_i and service time S_i, $i \geq 1$, where $W_i = D_i + S_i$. Similarly, let $N_q(t)$ and $N_s(t)$ be the respective (constructed) number of customers in queue and in service at time t, $t \geq 0$, where $N(t) = N_q(t) + N_s(t)$.

Let the customer average of $\{D_i\}$ and $\{S_i\}$ be d and $E(S)$ respectively, where the notation "$E(S)$" is often used because in many applications, the service times S_i are independent and identically distributed (i.i.d.) random variables, and their customer average is their expected value. Let the time average of $\{N_q(t)\}$ and $\{N_s(t)\}$ be respectively Q and L_s. From LL, we immediately have

$$Q = \lambda d \quad \text{and} \quad L_s = \lambda E(S). \tag{4}$$

To obtain (4), we didn't mention the order of service of customers in queue. In Figure 3, it would appear to be first-in-first-out (FIFO), also called first-come-first-served (FCFS); FIFO is easier to pronounce. Some alternatives are last-in-first-out (LIFO) and shortest job (service time) first (SJF). SJF usually (but not necessarily) reduces both Q and d. (4) holds in each case. We don't have to know what is going on inside the customers-in-queue box.

Is λ in the expression for Q in (4) correct? Arrivals finding an idle server don't join the queue. To investigate, let β be the (long run) fraction of arrivals who find all servers busy and join the queue, so $\beta\lambda$ is the arrival rate of these customers. Let d_β be the average delay of these customers. From LL, $Q = \beta\lambda d_\beta$. The average delay of customers who don't join the queue is 0, and averaging over both groups, $d = \beta d_\beta + (1-\beta)0 = \beta d_\beta$, and indeed, $Q = \lambda d$ is correct. For LL, L, λ, and w must be averages over the same stream of customers.

The second equation in (4) also deserves a closer look. It is the average number of customers in service, assuming all are served. This will be true when the number of servers exceeds $\lambda E(S)$, which is also the average number of busy servers. When there is one server, it is the fraction of time that server is busy.

Little's Law may be applied in several ways. If either L or w is known or can be found by an independent analysis (and we know λ), then we know the other. In other applications, this equation, together with other information, determines both quantities. In a typical queueing *model*, λ is known (is a parameter). In a real system, it is not known; sometimes it is estimated from estimates of L and w. We now illustrate a different application.

Example 2 (Order of Service). At one time at banks and post offices in the USA, a separate queue formed at each teller (server). These days, a single queue feeds all tellers. (We exclude from consideration special-purpose tellers.) How does changing from multiple queues (MQ) to a single queue (SQ) affect customers? In particular, what is the effect on *average* waiting time?

To consider this question, assume arrival times t_i and service times S_i are fixed. We are changing the *order* of service. (Should there be a positive queue behind one teller when another teller becomes free, we assume that a customer will shift immediately to the idle teller.) With SQ, customers are served FIFO. With MQ, service times *in the order served* are a rearrangement of the S_i.

Suppose customer service times are i.i.d., independent of all else. Service times in the order served will be i.i.d., and the *distribution* of the number of customers in system will not change. That is, $\{N(t), t \geq 0\}$ under MQ is essentially the same process as under SQ. (The number in system processes are said to be *stochastically equivalent.*) Hence time-average L has not changed. As λ has not changed,

$$w = L/\lambda \textit{ has not changed.}$$

This does not mean that customers are indifferent to the change. The waiting time *distribution* will change. In fact, from the FIFO property of SQ, it is easily shown that the *variance* of *delay* under SQ is smaller than under MQ, and also under any alternative of this nature, such as an SQ that operates LIFO.

Some queueing models are much more complex. A farmers' market may be viewed as a *queueing network*, where each arriving customer is served at some sequence of stands (stations), and then departs. LL applies to the arrival and departure of customers at the market, and also to the arrival and departure of customers at individual stations. This is an *open network.* In a *closed network*, customers move from station to station, but there are no arrivals to or departures from the system. For closed networks, LL holds at individual stations.

2 Sample-Path Proof of Little's Law

Our intuitive explanation for LL in Section 1 is the basis for a formal proof:

Proof. Motivated by Figures 1 and 2, we obtain the inequalities in (5), where the sum on the left is over those waiting-time rectangles that have ended by T.

$$\sum_{\{i:t_i+W_i\leq T\}} W_i \leq \int_0^T N(t)\,dt \leq \sum_{i=1}^{\Lambda(T)} W_i. \tag{5}$$

Now divide (5) by T and let $T \to \infty$. The right-hand expression has limit λw, which implies the left-hand expression has $\limsup \leq \lambda w$. LL follows if it has limit λw. Toward that end, we need two preliminary results. First, write

$$\frac{W_n}{n} = \frac{1}{n}\sum_{i=1}^{n} W_i - (\frac{n-1}{n})(\frac{1}{n-1})\sum_{i=1}^{n-1} W_i.$$

For finite w, this expression has limit $w - w = 0$ as $n \to \infty$. That is,

$$\lim_{n\to\infty} W_n/n = 0. \tag{6}$$

Finite λ implies $t_n \to \infty$ and hence $\Lambda(t_n)/t_n \to \lambda$ as $n \to \infty$. If the arrival times are distinct, $\Lambda(t_n) = n$; otherwise, $\Lambda(t_n) \geq n$, and we write

$$\frac{W_n}{t_n} = \frac{W_n}{n}\frac{n}{t_n} \leq \frac{W_n}{n}\frac{\Lambda(t_n)}{t_n}.$$

For finite λ and w and from (6), we have

$$\lim_{n\to\infty} W_n/t_n = 0. \tag{7}$$

For any $\epsilon > 0$, (7) implies that for some finite m and all $i > m$, $W_i < t_i\epsilon$ and $t_i + W_i < t_i(1+\epsilon)$. Thus $t_i \leq T/(1+\epsilon) \implies t_i + W_i \leq T$ for every $i > m$, which gives this *lower bound* on the left-hand expression in (5):

$$\sum_{i=1}^{\Lambda(\frac{T}{1+\epsilon})} W_i - \sum_{i=1}^{m} W_i \leq \sum_{\{i: t_i + W_i \leq T\}} W_i. \tag{8}$$

Now divide (8) by T and let $T \to \infty$, noting that the second term on the left is a constant. The left-hand side has limit $\lambda w/(1+\epsilon)$, which implies the right-hand side has $\liminf \geq \lambda w/(1+\epsilon)$. Because ϵ can be arbitrarily small, this implies the $\liminf \geq \lambda w$. Combining with $\limsup$, the limit is λw, and we have (3). □

Some early proofs (but after [12]) assumed the system would empty from time to time. Little did not require that, nor do we, aside from our convention that the system is empty initially (which can be dispensed with). To illustrate, suppose C_i arrives at $t_i = 2i$, with service time $S_i = 3$, $i \geq 1$, at a 2-server system. It is easy to see that (draw the figure!) even though the system never empties, the limits are $\lambda = 1/2$, $w = 3$, and $L = \lambda w = 1.5$.

Now let $\Lambda^d(t)$ be the number of departures by t. When (6) holds, we have (7). By the same method, it is easy to show that

$$\lim_{t\to\infty} \Lambda^d(t)/t = \lambda, \tag{9}$$

that is, the departure rate is equal to the arrival rate. (We don't require $w < \infty$; see Section 2.1.)

2.1 Technical Considerations; the Case $w = \infty$

In Figures 1 and 2, $\{W_i\}$ determines $\{N(t)\}$, but not the converse. It also turns out that when finite limits L and $\lambda > 0$ exist, w may not exist; for examples, see [20] and p. 289 of [24]. In both cases, the waiting time average oscillates, and limit (6) does not exist.

Suppose $0 < \lambda < \infty$ and w exist, with $w = \infty$. It would be nice to have $L = \infty$ as well, and this is known to be true for a variety of queueing models.

To deal with this question, suppose (6) still holds, which gives (7) and (8). Now divide (8) by T. For $w = \infty$, the left-hand side $\to \infty$ as $T \to \infty$. We have: When $0 < \lambda < \infty$, $w = \infty$, and (6) holds, L is well defined, where

$$L = \infty.$$

For a single-server FIFO queue with finite average service time $E(S)$, where $\rho \equiv \lambda E(S) < 1$, it is easy to show that (6) holds, even when $w = \infty$. However, this is not always true. For example, consider an ∞-server queue with i.i.d. service times, so the $W_i = S_i$ are i.i.d. For i.i.d. W_i, $w < \infty$ w.p.1 if and only if (6) holds w.p.1; see p. 239 of [13]. The point is not whether $L = \infty$ in this example, but rather to question when it is valid to use (6) to prove it.

3 Stationary Version of Little's Law

Queueing models are rarely so simple that either L or w can be found directly as sample-path averages. Instead, we find the distributions of N or W, or enough about them to find their means. This is usually done by a *stationary analysis*.

A substantial portion of the queueing literature is about continuous-time Markov-chain models, sometimes with complex structure. For a queueing network, we let $N_k(t)$ be the number of customers at station k at time t, and $N(t)$ be their sum over k. We replace exponential service with *phase-type* service, and so on. For positive-recurrent chains, stationary N exists, with distribution determined by the balance equations, and $L = E(N)$ w.p.1, even when infinite. In these models, $\lambda > 0$ (in open networks), and $w = L/\lambda$. Sometimes sophisticated techniques are required to determine positive recurrence and to solve for stationary distributions, or their means, or approximations of them, but we have changed the playing field from finding sample-path averages directly.

Sometimes, we find properties of W directly, and from them, properties of N. For a single-server queue with *renewal arrivals* (i.i.d. inter-arrival times) and i.i.d. service times (the $GI/G/1$ queue), FIFO delays D_i satisfy the recursion

$$D_{i+1} = \max\{D_i + S_i - T_i, 0\}, \; i \geq 1, \tag{10}$$

where *inter-arrival time* $T_i = t_{i+1} - t_i$. Assume $0 < \lambda E(S) < 1$. By letting D_i and D_{i+1} in (10) have the same distribution (a stationary analysis), we find approximations and bounds for the mean and other properties of D, W, and N.

3.1 Little's Law via Stationary Marked Point Processes

A (one-dimensional) *marked point process* (mpp) is a sequence $\{t_i, k_i\}$, $i \geq 1$, where $0 \leq t_1 \leq t_2 \leq \dots$ are the points, and the k_i are the marks. For queueing theory applications, the t_i are arrival times, so $\{t_i\}$ (the point process) is an arrival process, and for each i, k_i is something associated with arrival i; in a single-server queue, for example, it would be the service time of arrival i.

A mpp is called *stationary* (smpp) if $t_1 > 0$ and $\{\Lambda(t)\}$ has stationary increments, which means that for every t, the distribution of $\Lambda(t+h) - \Lambda(h)$ is independent of h. For a smpp, it is easily shown that for all t, $E\{\Lambda(t)\} = \lambda t$, where $\lambda > 0$ is the arrival rate. For renewal arrivals, where $T_i = t_{i+1} - t_i$, $i \geq 1$, have distribution function F and mean $1/\lambda$, we get the corresponding smpp by letting t_1 have *equilibrium* density function $f_e(u) = \lambda[1 - F(u)]$, $u \geq 0$.

$\{\Lambda(t)\}$ is *time* stationary. For some $\{\Lambda(t)\}$, we would like to determine the distribution of *event* (or *customer*) stationary $\{T_i\}$, and the converse. For renewal arrivals, this is easy, but in general, where the T_i are dependent, this is not the case. One of the major achievements of point-process theory was to *invert* (determine one from the other) the distributions of $\{\Lambda(t)\}$ and $\{T_i\}$ (the latter is called the *Palm* distribution), where the one we start with is stationary and ergodic. Originally, this was for *simple* arrival processes, where this means $T_i > 0$. Later, this was extended to allow $P(T_i = 0) > 0$ (batch arrivals).

We also have marks. Starting with stationary $\{T_i, W_i\}$, the corresponding $\{\Lambda(t), N(t)\}$ was constructed in [6] (and in earlier work referenced there), where $\{N(t)\}$ is stationary, and the stationary-version of Little's Law,

$$E(N) = \lambda E(W), \tag{11}$$

was shown, even when infinite, without appealing to the sample-path version.

Describing how the inversion or this construction is done is beyond our scope. In addition to [6], we cite [4] for point-process theory, [18] for an empirical-inversion approach, [15] for more on Palm theory, and all four for an historical account and references.

(11) appears to be very general. For example, we can view waiting times in some model as though they were generated by an ∞-server queue, and set $W_i = S_i$, where the S_i are stationary and ergodic. However, this is somewhat misleading because a queueing *model* usually would consist of an arrival process, service requirements, and details about the service facility. Stationary $\{W_i\}$ has to be *constructed*. This is easy to do for a single-server queue, but for, say, a network of multi-server stations, we don't know how to do this, or even whether such a $\{W_i\}$ exists, without having some structure, such as regeneration points (in a general sense), or a Markov process that is Harris recurrent.

We now present a deterministic example where customer- and time-stationary processes are easily constructed independently.

Example 3. Consider a single-server queue with $t_1 = 1$, subsequent inter-arrival times that alternate, $1, 4, 1, 4, \dots$ (so $\lambda = 1/2.5 = 0.4$), and constant service times $S_i = 2$. We plot $\{N(t)\}$ in Figure 4.

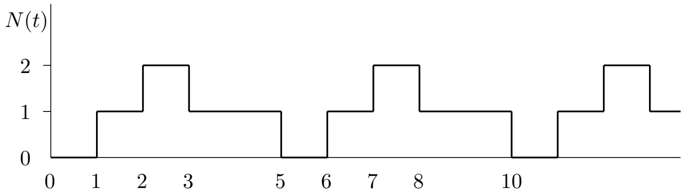

Figure 4: Constructing Stationary Versions

When $T_i = 1$, C_i finds the system empty, and $W_i = 2$. Similarly, when $T_i = 4$, $W_i = 3$. These events alternate. We have jointly-stationary $\{T_i, W_i\}$, where events $\{T_i = 1, W_i = 2\}$ and $\{T_i = 4, W_i = 3\}$ each has probability 1/2.

It is easy to see (as fractions of time) that stationary N has distribution

$$P(N = 0) = P(N = 2) = 0.2, \text{ and } P(N = 1) = 0.6, \tag{12}$$

with $E(N) = 1.0$. From the distribution of $\{T_i, W_i\}$, $E(W) = 2.5$, and (of course!) $E(N) = \lambda E(W) = (0.4)(2.5)$.

Constructing $\{\Lambda(t)\}$ with stationary increments: As the arrival process is deterministic, its stationary version is completely determined by where the initial point $t_0 = 0$ falls. We call inter-arrival intervals either *long* (length 4) or *short* (length 1). If t_0 falls in a long, the time until the first arrival is uniformly distributed on $(0, 4)$, with subsequent inter-arrival times $1, 4, 1, 4, \ldots$. If it falls in a short, the time until the first arrival is uniformly distributed on $(0, 1)$, with subsequent inter-arrival times $4, 1, 4, 1, \ldots$. The probability that t_0 falls in a long is 0.8, the fraction of the real line that is covered by long intervals.

Selecting $t_0 = 0$ in this manner, we get waiting-time sequence $2, 3, 2, 3, \ldots$ with probability 0.8 (t_0 falls in a long), and $3, 2, 3, 2, \ldots$ otherwise. $\{W_i\}$ is *not* stationary. Except in special cases such as renewal arrivals, it is not possible for $\{N(t)\}$ and $\{W_i\}$ to be stationary simultaneously on the same sample space.

4 Extension to $H = \lambda G$; Work

For Little's Law, the C_i (and unspecified system behavior) generate $\{W_i\}$ and $\{N(t)\}$, where C_i contributes W_i to the area under $\{N(t)\}$. We now present an extension, where the C_i may generate other discrete- and continuous-time processes that are related in the same way. We consider only the sample-path version, at sample point ω, and again omit writing ω explicitly.

For each C_i, we associate function $f_i(t)$, $t \geq 0$, where $\int_0^\infty |f_i(t)|\, dt < \infty$, and for some finite $l_i > 0$, $f_i(t) = 0$ for $t \notin [t_i, t_i + l_i)$. Now define

$$G_i = \int_0^\infty f_i(t)\, dt, \; i \geq 1, \quad \text{and} \quad H(t) = \sum_{i=1}^\infty f_i(t), \; t \geq 0.$$

When $f_i(t) \geq 0$, G_i is the area under $\{H(t)\}$ contributed by C_i. Often, $l_i = W_i$. We define customer and time averages, and λ as defined before:

$$\overline{G} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} G_i, \quad \text{and} \quad \overline{H} = \lim_{T\to\infty} \frac{1}{T} \int_0^T H(t)\, dt.$$

The following is an important extension of LL:

Theorem 2 ($\overline{H} = \lambda\overline{G}$). *If limits λ and $\overline{G}$ exist and are finite, and technical condition (TC): $l_i/t_i \to 0$ as $i \to \infty$, $\overline{H}$ exists, where*

$$\overline{H} = \lambda\overline{G}. \tag{13}$$

In the literature, the usual notation is "$H = \lambda G$", but we prefer to reserve H and G for the stationary version. For LL, $l_i = W_i$, and TC is implied by $w < \infty$. Without TC, some of the area contributed by C_i may occur so far in the future, that, as i increases, some of it is "lost" in the limit, where, possibly,

$$\overline{H} < \lambda\overline{G},$$

or $\overline{H}$ may not exist. However, TC is not necessary. For a necessary and sufficient condition, see [5], p. 178. The easily-proven version here is widely applicable.

The same issue arises with LL when we have *multiple visits*; that is, the same customer visits the system many times, and we define W_i as the sum of C_i's waiting times on all visits. Let C_i depart (for the last time) at $t_i + l_i$. If w is finite but TC does not hold, it is possible to have

$$L < \lambda w.$$

So-called *counterexamples* to Little's Law have been constructed in this manner.

We sketch a proof of $\overline{H} = \lambda\overline{G}$, which is almost identical to that for LL. We assume $f_i(t) \geq 0$. When this is not the case, split the functions into their positive and negative parts, $f_i^+(t) = \max\{f_i(t), 0\}$ and $f_i^-(t) = -\min\{f_i(t), 0\}$, $t \geq 0$. Go through the steps below for the positive parts and the negative parts separately, and combine the results.

When $f_i(t) \geq 0$, the bounds on the integral below are immediate:

$$\sum_{\{i:t_i+l_i\leq T\}} G_i \leq \int_0^T H(t)\, dt \leq \sum_{i=1}^{\Lambda(T)} G_i. \tag{14}$$

Now divide (14) by T and let $T \to \infty$. The right-hand expression has limit $\lambda\overline{G}$, and the left-hand expression has $\limsup \leq \lambda\overline{G}$. To complete the proof, we obtain a lower bound on the left-hand expression in (14). We already have $l_n/t_n \to 0$, and following the argument for (7), we get $l_n/t_n \to 0$, as $n \to \infty$. Following the argument for (8), we have the lower bound

$$\sum_{i=1}^{\Lambda(\frac{T}{1+\epsilon})} G_i - \sum_{i=1}^{m} G_i \leq \sum_{\{i:t_i+l_i\leq T\}} G_i.$$

The remaining steps are the same as those for (3), and we have proven (13).

When $\lambda > 0$ and $\overline{G} = \infty$, and we still have TC, $\overline{H} = \infty$.

4.1 Work in System

After Little's Law, the most important applications of (13) involve various representations of process $\{V(t)\}$, where $V(t)$, called *work in system* or just *work* at time $t \geq 0$, is *the sum of the remaining service times of all customers in system at time* t.

With the notation in Section 1, suppose we have a *single-server* queue with arrival times t_i and service times S_i. $V(t)$ jumps by S_i at t_i, and where positive, decreases with slope -1 between jumps, because the remaining service time of any customer in service decreases at that rate. A typical sample path of $V(t)$ is shown as the heavy horizontal (where $V(t) = 0$) or slanted lines in Figure 5.

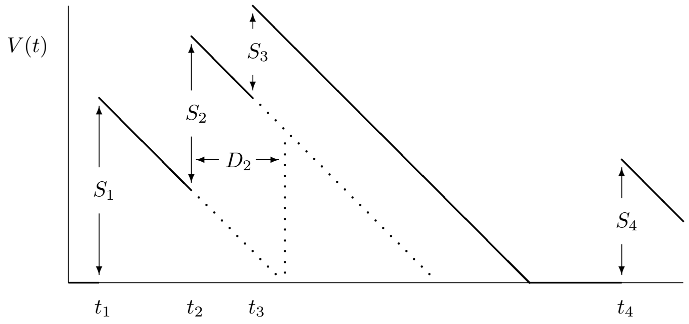

Figure 5: Work in System for a Single-Server Queue

When customers are served FIFO, the diagonal dotted lines show when the first two customers finish service. The vertical dotted line shows when C_2 enters service, and the time between t_2 and that line is D_2. Similarly, D_3 (not shown) is easy to represent. The contribution of customer C_i to the area under $\{V(t)\}$ is a trapezoid that has a rectangular piece with area D_iS_i, and a triangular piece with area $S_i^2/2$. So for this example, $G_i = D_iS_i + S_i^2/2$, and from (13)

$$E(V) = \lambda \left[\overline{DS} + \overline{S^2}/2\right]. \tag{15}$$

To be consistent with our notation, "$E(V)$" should be "$\overline{V}$". From the stationary version, however, it is also the mean of a stationary V. The *form* of (15) does not require FIFO order of service, but only that customers are served to completion, without interruption. For example, it holds under LIFO and SJF. [While each customer still generates a trapezoid, the D_i would change. If we had LIFO in Figure 5, D_3 would be the time between t_3 and the vertical dotted line.] Note that TC is the same as (6) here, and is satisfied when $w < \infty$.

When we make additional assumptions, (15) can be written in different ways. For example, when service times are i.i.d. and independent of the arrival process, the D_i and S_i are independent and $\overline{DS} = dE(S)$; (15) becomes

$$E(V) = \rho d + \lambda E(S^2)/2, \tag{16}$$

where $\rho = \lambda E(S)$. This includes FIFO and LIFO, but *not* SJF. When we also have Poisson arrivals (now an $M/G/1$ queue) and FIFO, $d = E(V)$ (from PASTA, see e.g. [23]), and from (16), we get

$$d = \lambda E(S^2)/2(1-\rho), \tag{17}$$

the average delay in queue for this model. (As in Example 2, d is the same under LIFO; only the *argument* for (17) requires FIFO). When arrivals are not Poisson, and we can show either $d > E(V)$ or $d < E(V)$, then from (16), we get a *bound* on d.

(15) holds for multi-server queues, but the sample paths of $\{V(t)\}$ would change. When service times are i.i.d., Brumelle [2] used these ideas to obtain an important lower bound on d_c, the average delay in queue for a c-server system, in terms of the average delay for a corresponding fast single server; that is, a server with service times S_i/c.

Sometimes a queue operates under rules that give preferential treatment to some customers or classes of customers over others. This may permit interrupting service on one customer in order to serve another. Suppose we have *work conservation*, which means that the total time in service of every customer is rule invariant. For a *single-server* queue, this implies that the sample paths of $\{V(t)\}$, as in Figure 5, are the same for all rules, as is $E(V)$. This fact plays an important role in the analysis of these rules. For many examples, see [24].

5 A Rate Conservation Law (RCL)

As with Little's Law and its extensions, the Rate Conservation Law (RCL) by Miyazawa is of fundamental importance in queueing theory. We cite [14], a review article; his publications in this area go back to 1983. The theory of *level crossings* (e.g., see [1]) is a special case.

As conceived by Miyazawa, RCL is a relation between the expected values of time-stationary and arrival-stationary quantities, as is the case for Little's Law in (11). Sigman [17] showed that the sample-path version of RCL is equivalent to the corresponding version of $\overline{H} = \lambda\overline{G}$.

However, the starting points for these results (geometrically-appealing areas in one case, jumps and derivatives in the other) are quite different. In an application, one is often much easier and more intuitive to use than the other.

In this brief introduction, we present only the sample-path version of RCL.

Consider a sample path $\{Y(t,\omega),\ t \geq 0\}$ of some stochastic process at sample point ω. As before, we drop ω, writing $Y(t) = Y(t,\omega)$, and let $\mathbf{Y} = \{Y(t)\}$. We assume $\mathbf{Y}$ is right-differentiable for all t, and define

$$Y'(t) = \lim_{h\downarrow 0}\,[Y(t+h) - Y(t)]/h,$$

where we assume that $\mathbf{Y}'$ is right-continuous and has left-hand limits. Now $\mathbf{Y}$ may have discontinuities. Let "events" occur at points in an underlying point process (events could be arrivals), $0 < t_1 < t_2 < \cdots$, where $r(t)$ is the number of points that occur by t, such that each discontinuity of $\mathbf{Y}$ occurs at one of these points. (We don't require $\mathbf{Y}$ to be discontinuous at every such point.) Define

$$-J_i = Y(t_i+) - Y(t_i-), \text{ the size of the } i^{\text{th}} \text{ jump.}$$

Then we have

$$\int_0^t Y'(u)\,du = Y(t) - Y(0) + \sum_{i=1}^{r(t)} J_i, \tag{18}$$

which simply means that the *change* in $\mathbf{Y}$ between 0 and t is the sum of the change while continuous plus the sum of the jumps.

Define the limits, when they exist: event rate λ, event average $\overline{J}$, and time average $\overline{Y'}$. Now divide (18) by t and let $t \to \infty$. The following has been shown:

Theorem 3 (Sample-Path RCL). *When λ and $\overline{J}$ are finite, and $Y(t)/t \to 0$ as $t \to \infty$,*

$$\overline{Y'} = \lambda \overline{J}. \tag{19}$$

(The stationary-version notation is to write $\overline{Y'} = E(Y')$ and $\overline{J} = E(J)$.)

It is easier to obtain (15) via (13) (it is immediate), rather than via RCL. On the other hand, for finding an expression for $P(V > v)$, for stationary work in system V, RCL is the better tool. See p. 664 of [17].

6 A Distributional Little's Law

The stationary version of Little's Law raises this question: Is there a similar relation between the distributions of N and W, a *distributional* Little's Law?

In Example 2, we compared a bank under different rules of operation. Under both rules, the distribution of N is the same, but the distribution of W is different. There is no hope for a distributional law in a general setting. Nevertheless, we derive a law in the restricted setting of Haji and Newell, [7]. For this result, further restricted to Poisson arrivals, with applications, see [10].

Let $t_0 = 0$ be a random time point, where $N(0) = N$ is stationary, and $\{\Lambda(t),\ -\infty < t < \infty\}$, has stationary increments. Number customers backward in time, where C_i occurs at $-t_i$, $0 \le t_1 \le t_2 \le \cdots$. So C_1 is the most recent arrival. We make the following assumptions:
(a) Customers depart FIFO *from the system*,
(b)$\{W_i\}$ is stationary, and for every i,
(c) W_i is independent of the arrival process after C_i arrives; in particular, it is independent of t_i, which is "how long ago" that customer arrived.

Thus, customer i is in the system at time $t_0 = 0$ if and only if $\{W_i \ge t_i\}$, and because of (a), these are the same events:

$$\{N \ge i\} = \{W_i \ge t_i\},\ i \ge 1,$$

where (c) W_i and t_i are independent, and (b) $P(N \ge i) = P(W \ge t_i)$, where W is an arbitrary stationary waiting time, independent of *every* t_i. The event

$\{W \geq t_i\}$ means that at least i arrivals occur in interval $[-W, 0]$, where W is independent of $\{\Lambda(t)\}$, and the probability of this event does not depend on the interval location. We have:

Theorem 4 (A Distributional Little's Law, (DLL)). *Under (a) - (c),*

$$N \text{ has the same distribution as } \Lambda(W), \tag{20}$$

where $\{\Lambda(t)\}$ and W are independent.

We now introduce some new quantities to compare this result with another. For a queueing system, let N_a and N_d be, respectively, the stationary number of customers in system *found by an arrival*, and *left behind by a departure*. It is easily shown that, in general, N_a and N_d have the *same* distribution. On the other hand, it is usually the case that N_a and N have *different* distributions and means. For simplicity, assume inter-arrival times $T_i > 0$ (no batches). Let $\{\Lambda_c(t)\}$ be the *customer-stationary* arrival process, which begins with an arrival at $t_0 = 0$, and $\Lambda_c(t)$ is the number of arrivals in interval $(0, t]$. When we have (a), N_d is the number of arrivals during the departing customer's waiting time, and when we also have renewal arrivals (Poisson arrivals is a special case),

$$N_d \text{ has the same distribution as } \Lambda_c(W), \tag{21}$$

where $\{\Lambda_c(t)\}$ and W are independent. While (21) is similar to (20), it is not as useful. Usually, $E\{\Lambda_c(t)\} \neq \lambda t$ and $E(N_d) = E\{\Lambda_c(W)\} \neq \lambda E(W)$.

To illustrate, consider a single-server queue, where $t_i = 10i$ and $S_i = 9$ for all i (a $D/D/1$ queue). Stationary waiting time $W_i = 9$, a constant. As fractions of time, it is easy to see that N is either 0 or 1, with distribution

$$P(N = 0) = 0.1 \quad \text{and} \quad P(N = 1) = 0.9. \tag{22}$$

We have $\lambda = 0.1$ and $E(N) = \lambda E(W)$, but $N_a = N_d = \Lambda_c(W) = 0$. For $\{\Lambda(t)\}$, t_1 (backward in time) is uniformly distributed on $(0, 10)$. C_1 is in the system at $t_0 = 0$ if and only if $\{t_1 \leq 9\}$, and $\Lambda(W) = 1$. $\Lambda(W)$ has distribution (22); DLL holds. This is a special case of a FIFO $GI/G/1$ queue, where DLL also holds.

Confusion between $\{\Lambda(t)\}$ and $\{\Lambda_c(t)\}$ may be partly responsible for the following *incorrect* "intuitive" and "elementary" explanation of the stationary version of Little's Law that has appeared several times in the literature:
(i) $E(N_a) = E(N_d)$ and (ii) $E(N_d) = \lambda E(W)$.

While (i) is true, there are several problems with this argument. Implicit in (ii) is FIFO, as in (a) in Theorem 4. That is merely a restriction. More serious is that (ii) is often false, as in the elementary example above. Also serious is the implicit equating of N_a and N. We are left with

$$L = E(N) \neq E(N_a) = E(N_d) \neq \lambda E(W). \tag{23}$$

For the $M/G/1$ queue, both inequalities in (23) are equalities, the first from PASTA (N and N_a have the same distribution); for the second, a Poisson process has stationary and independent increments [$\{\Lambda(t)\}$ and $\{\Lambda_c(t)\}$ are equivalent]. In this case, the explanation is correct, but it is neither intuitive nor elementary.

Now return to DLL. When does it hold? Condition (a) holds for single-server queues when customers in queue are served FIFO. It holds for a tandem (series) arrangement of single-server queues. It holds for multi-server queues under FIFO and constant service. Looking only at the queue, there is a corresponding relation between the number of customers in queue and the delay in queue. It holds for multi-server queues under FIFO, with a general service distribution. There are other examples for priority classes in what are called priority queues.

Condition (c) appears to require renewal arrivals, which includes the Poisson process, but not (say) the typical arrival process at the second station of a tandem queue. There is an exception to this requirement when there are an infinite number of servers. DLL does not hold in Example 3.

The DLL is very useful when it applies, but unlike the ubiquitous LL, the conditions for it to hold are rather restrictive.

7 Brief History and Literature Review

Little's Law was believed to be true long before 1961. On p. 75 of [16], Morse noted that it holds for every model he was aware of, and challenged someone to either come up with a general proof or a counterexample. Unfortunately, this pre-1961 status inhibited the use of LL as a tool in analysis of queueing models.

We have emphasized the sample-path version because it is the easiest to understand and prove, and it holds in situations when there may be no stationary version. Viewed this way, it is a conservation law (of area). Furthermore, it is rare in an application to use the full stationary version in (11). Instead, a stationary analysis is performed to determine properties of *one* of the stationary distributions sufficient to find, approximate, or make other statements about its mean. The sample-path version then gives us the other.

At first glance, Little's formulation seems to be sample-path, but it is not, as it relies on stochastic properties to obtain limits. There also is a flaw in the formulation, as noted in [3] and elsewhere. In 1967, Jewell [9] proved LL for (classically) regenerative processes where the system empties from time to time, and a new cycle begins. Thus over a cycle, (1) is exact! (Shortly thereafter, several authors either proved LL or gave intuitive explanations for systems that empty periodically.) What Jewell showed may be viewed as a sample-path result and may have provided intuition for what followed. However, his conditions are very restrictive. Stidham has the first sample-path proof, [19], and in 1974 [20], a proof similar to the one here.

What is now usually called $H = \lambda G$ has a similar history. Brumelle [3] proved it in 1971 in a stochastic setting similar to Little. He also obtained the important equation (15). Heyman and Stidham [8] has a sample-path proof in 1980 similar to the one presented here. Brumelle's result is not the equivalent stationary version (11) in Section 3. See the references there.

There are too many related results to discuss in the space we have. Here are three. For $H = \lambda G$, a customer's contribution is an integral, so it accumulates gradually. This has been extended to allow "lumps" at certain times. Often a model is simulated to estimate (say) Q or d, and direct (straightforward) estimators of each, $\hat{Q}$ or $\hat{d}$, are available. From LL, an *indirect* estimator of (say) Q is $\lambda\hat{d}$. The statistical efficiency of indirect estimators was investigated in [11], and by others later. Finally, there are also central-limit-theorem versions of these results.

For more on these and other results, many other references, and discussion of the literature, see [5, 21, 22].

References

[1] Brill, P, and Posner, M. 1977. Level Crossings in Point Processes Applied to Queues: Single-Server Case, *Oper. Res.*, **25**: 662–674.

[2] Brumelle, S. 1971. Some Inequalities for Parallel-Server Queues, *Oper. Res.*, **19**: 402–413.

[3] Brumelle, S. 1971. On the Relation Between Customer and Time Averages in Queues, *J. Appl. Prob.*, **8**: 508-520.

[4] Daley, DJ, and Vere-Jones, D. 1988. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York.

[5] El-Taha, M, and Stidham Jr. S. 1999. *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publishers, Boston.

[6] Franken, P, König, D, Arndt, U, and Schmidt, V. 1982. *Queues and Point Processes*. John Wiley & Sons, New York.

[7] Haji, R, and Newell, GF. 1971. A Relation between Stationary Queue and Waiting Time Distributions. *J. Appl. Prob.,* **8**: 617–620.

[8] Heyman, DP, and Stidham Jr., S. 1980. The Relation Between Customer and Time Averages in Queues. *Oper. Res.*, **28**: 983–994.

[9] Jewell, WS. 1967. A simple Proof of $L = \lambda W$. *Oper. Res.*, **15**: 1109–1116.

[10] Keilson, J, and Servi, LD. 1988. A Distributional Form of Little's Law. *Oper. Res. Let.*, **7**: 223–227.

[11] Law, AM. 1975. Efficient Estimators for Simulated Queueing Systems. *Mgmt. Sci.*, **22**: 30–41.

[12] Little, JDC. 1961. A Proof of the Queuing Formula: $L = \lambda W$, *Oper. Res.*, **9**: 383–387.

[13] Loève, M. 1963. *Probability Theory*, 3rd Ed. Van Nostrand, Princeton, N.J.

[14] Miyazawa, M. 1994. Rate Conservation Laws: A Survey. *Queueing Systems: Theory and Applications,* **15**: 1–58.

[15] Miyazawa, M, Nieuwenhuis, G, and Sigman, K. 2001. Palm Theory for Random Time Changes. *J. Appl. Math. and Stoc. Anal.*, **14**: 55–74.

[16] Morse, PM. 1958. *Queues, Inventories and Maintenance*. John Wiley & Sons, New York.

[17] Sigman, K. 1991. A Note on a Sample-Path Rate Conservation Law and its Relationship with $H = \lambda G$. *Adv. Appl. Prob.,* **23**: 662–665.

[18] Sigman, K. 1995. *Stationary Marked Point Processes*, Chapman & Hall, New York.

[19] Stidham Jr., S. 1972. $L = \lambda W$: A Discounted Analogue and a New Proof. *Oper. Res.*, **20**: 1115–1126.

[20] Stidham Jr., S. 1974. A Last Word on $L = \lambda W$. *Oper. Res.*, **22**: 417–421.

[21] Stidham Jr., S, and El-Taha, M. 1995. Sample-Path Techniques in Queueing Theory. Pp 119–166 in Dshalalow, JH, ed. *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton.

[22] Whitt, W. 1991. A Review of $L = \lambda W$ and Extensions, *Queueing Systems*, **9**: 235–268.

[23] Wolff, RW. 1982. Poisson Arrivals See Time Averages, *Oper. Res.*, **30**: 223–231.

[24] Wolff, RW. 1989. *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, NJ.