Queueing Syst (2014) 76:267–281
DOI 10.1007/s11134-013-9364-8

Little's law when the average waiting time is infinite

Ronald W. Wolff · Yi-Ching Yao

Received: 6 May 2012 / Revised: 3 May 2013 / Published online: 8 June 2013

Abstract One version of Little's law, written as $L = \lambda w$, is a relation between averages along a sample path. There are two others in a stochastic setting; they readily extend to the case where the average waiting time w is infinite. We investigate conditions for the sample-path version of this case to hold. Published proofs assume (our) Eq. (3) holds. It is only sufficient. We present examples of what may happen when (3) does not hold, including one that may be new where w is infinite and L is finite. We obtain a sufficient condition called "weakly FIFO" that is weaker than (3), and through truncation, a necessary and sufficient condition. We show that (3) is sufficient but not necessary for the departure rate to be equal to the arrival rate.

Keywords Sample path · FIFO · Truncation · Departure rate · Stability

Mathematics Subject Classification 60G17 · 60G52 · 60K25 · 68M20 · 90B22

1 Introduction

It is often written, for example, in [12] and [14], that there are two versions of Little's law, **SP**, the purely *sample-path* (deterministic) result as in Stidham [11], and **D**, a relation between the means of stationary *distributions*, as in Franken et al. [4].

However, the oldest (general) formulation, that by Little in [5], as perfected by Brumelle in [2], is neither of these. We call this version a *hybrid* **H**: It is a sample-path

R. W. Wolff (✉)
Department of Industrial Engineering and Operations Research, University of California,
Berkeley, CA 94720, USA
e-mail: wolff@ieor.berkeley.edu

Y.-C. Yao
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, ROC

result in a stochastic setting; see also [7] and page 249 of [12]. Versions **D** and **H** have the same starting point: The joint sequence of inter-arrival times and waiting times is stationary, where the waiting times have an arbitrary joint distribution.

Whatever the version, we write Little's law as "$L = \lambda w$," where λ is the arrival rate, $0 < \lambda < \infty$, and for **SP**, L is the time average of the number of customers in system and w is the customer average of their waiting times. This departs from the usual notation, where W is the customer-average waiting time. We prefer to define W as a stationary waiting time with mean $w = E(W)$. Thus for version **D**, $L = \lambda E(W)$, where L is the mean of the stationary number-of-customers-in-system process. We will often abbreviate Little's law as LL.

The proof of LL via **H** is easier than via **D**, but **D** obtains much more, the construction of the stationary number-of-customers-in-system process. Both readily extend to the case where $w = \infty$.

For finite w, the proof via **SP** is easy to understand and uses more elementary tools. It holds for an arbitrary point (*sample path*) in the sample space, or for sequences of numbers without defining a sample space. We do not have to be concerned about whether stationary versions of processes exist.

For **SP** when $w = \infty$, it is usually assumed that (3) below still holds. This makes the proof easy, but it does not hold in some elementary situations.

We investigate conditions for Little's law to hold on sample paths, with primary consideration to the case where $0 < \lambda < \infty$ and $w = \infty$.

In Sect. 2, we review how (3) is used to prove LL, and through a series of examples, show what can happen when it does not hold. This includes an example where $0 < \lambda < \infty$, $L < \infty$, and $w = \infty$.

In Sect. 3, we show that LL holds under a "weakly FIFO" condition that is weaker than (3), and like (3), it is sufficient but not necessary. In particular, FIFO is sufficient, and is different (neither weaker nor stronger) from (3). We show that (3) is sufficient but not necessary for the departure rate to be equal to the arrival rate. We relate this result to the literature at the end of this section in Remark 4, and in Sect. 6.

In Sect. 4, we use truncation to obtain a necessary and sufficient condition for LL to hold. We also obtain a stronger sufficient condition. While not necessary, it is applied in the stochastic setting of **H** to give what may be the most elementary proof of LL when $w = \infty$.

Section 5 is devoted primarily to establishing that for a stochastic model where $L = \lambda w = \infty$, (3) does not hold. This example is important because, unlike our deterministic examples, the waiting times are stable. See Sect. 5.1 for the meaning of terms "stable" and "unstable," and further analysis of this model. We also point out what are probably inadvertent errors in several places in a well-known publication regarding the necessity of (3) for LL when $w = \infty$.

In Sect. 6, we briefly discuss the two ways infinite w may arise in queueing models and some of the implications of each of them.

2 Version SP when $w = \infty$

In what follows, the quantities we define are for some arbitrary sample path in a sample space, or are sequences of numbers without defining a sample space.

For $i \geq 1$, let customer C_i arrive at time t_i, have waiting time W_i, and depart at $d_i = t_i + W_i$, where $t_0 \equiv 0 \leq t_1 \leq t_2 \leq \cdots$, and $t_i \to \infty$ as $i \to \infty$. For any $t \geq 0$, C_i is *in the system* at t if $t_i \leq t < d_i$. Define the corresponding indicator, $I_i(t) = 1$ if $t_i \leq t < d_i$, and $I_i(t) = 0$ otherwise, where $\int_0^\infty I_i(t)\,\mathrm{d}t = W_i$. 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 $N(t) = \sum_{i=1}^{\Lambda(t)} I_i(t)$. Let $D(t) = \#\{i : d_i \leq t\}$, the number of departures by time t.

We define long-run averages as limits, when they exist, and name them.

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

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.* Unless the contrary is stated, we assume $0 < \lambda < \infty$.

Customer data t_i and W_i for all i determine $N(t)$ for all t, where for any time $T > 0$, the inequalities

$$\sum_{\{i:d_i\leq T\}} W_i \leq \int_0^T N(t)\,\mathrm{d}t \leq \sum_{i=1}^{\Lambda(T)} W_i \tag{2}$$

are easily obtained, as in [11,14], page 162 of [3], and page 287 of [13].

We now sketch what is essentially the same proof of LL in all of these references for the case where w is finite and $0 < \lambda < \infty$.

Divide (2) 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 $\liminf \geq \lambda w$. Toward that end, it is easily shown that finite w implies

$$\lim_{i\to\infty} W_i/i = 0, \tag{3}$$

and that (3), together with finite w and λ, implies

$$\lim_{i\to\infty} W_i/t_i = 0. \tag{4}$$

(When $0 < \lambda < \infty$, (3) and (4) are equivalent, whether or not w is finite.)

The remainder of the proof uses (4) to show that for any $\epsilon > 0$, the left-hand expression in (2), when divided by T, has $\liminf \geq \lambda w/(1+\epsilon)$, and because ϵ is arbitrary, has $\liminf \geq \lambda w$. Thus we have

$$\liminf_{T\to\infty} \frac{1}{T} \int_0^T N(t)\,\mathrm{d}t \geq \lambda w, \tag{5}$$

which completes the proof. The key property is (3).

It is easy to extend LL to the case where $w = \infty$ if we assume that (3) still holds, which is usually what is done, e.g., on page 241 of [12].

To illustrate, consider a $GI/G/1$ first-in-first-out (FIFO) queue with average service time $E(S)$, where $\lambda E(S) < 1$, and variance $V(S) = \infty$. We have $w = \infty$, and it is easy to show that (3) holds.

However, this is not always true, as in an ∞-server queue with i.i.d. service times S_i, 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 (3) holds w.p.1.; for discussion and a proof, see Sect. 5. As we show in Example 6, LL holds for this model even when w is infinite. Thus we have an example of stable waiting times where, when $w = \infty$, LL holds but (3) does not.

On the other hand, when $0 < \lambda < \infty$, the second inequality in (2) gives

$$L = \infty \implies w = \infty. \tag{6}$$

It was first noted by Brumelle [2], more explicitly by Stidham [11], and many times since that there is an asymmetry to LL. When $0 < \lambda < \infty$, finite w implies finite L, but not the converse. When L is finite, it has long been known that w may not exist, and we show in Example 4 below that w may be infinite. Thus infinite L implies infinite w, but not the converse.

The following four sample-path examples show what can happen when (3) does not hold. The first one appears on page 289 of [13].

In all of these examples, waiting times are either "long" or "short" (equal to zero, but this is not essential), where the arrival rate of customers with long service times is zero. In all but Example 3, we introduce some additional notation: Let the jth customer with a long waiting time arrive at a_j, depart at b_j, and have waiting time $W_{a_j} = b_j - a_j,\ j \geq 1$.

Example 1 Let $t_i = i$ for $i \geq 1$ (so $\lambda = 1$). Let $a_j = 2^{j-1}$, $b_j = a_{j+1}$, and $W_{a_j} = b_j - a_j = 2^{j-1}$, $j \geq 1$. For any arrival time $t_i = i$, $i \notin \{a_1, a_2, \ldots\}$, departure time $d_i = i$ and $W_i = 0$ (short waiting times). For customers with long waiting times, one departs just as the next one arrives. $N(t) = 1$ for $t \geq 1$, and $L = 1$. It is easy to show that the sequence $\{\overline{W}_n = \sum_{i=1}^{n} W_i/n\}$ fluctuates between one and two as $n \to \infty$. Limit w does not exist.

Example 2 This is the same as Example 1, except that we make the long waiting times longer. Let $a_j = 2^{j-1}$ (the same as in Example 1) and $b_j = a_{2j},\ j \geq 1$. We still have $\lambda = 1$. It is easy to see that $N(t)$ increases without bound and $L = \infty$. From (6), $w = \infty$.

Thus, while (3) is sufficient, it is not necessary for $w = L = \infty$.

So far, we have examples with $0 < \lambda < \infty$ where L is finite and limit w does not exist and where both L and w are infinite. As we show next, it is easy to construct examples where, when $\lambda = 0$, we have finite L and infinite w.

Example 3 Let $\{W_i\}$ be any increasing sequence with limit ∞, hence $w = \infty$. Let $t_1 = 1$, and define

$$t_{i+1} = t_i + W_i, \quad i \geq 1.$$

As in Example 1, $N(t) = 1,\ t \geq 1$, and $L = 1$, but now, $\lambda = 0$. Compared with Example 1, we have deleted the arrival times and waiting times of those customers who had short waiting times.

A sample-path example in [11] has finite L, but neither λ nor w exist. Our next example, which is of much greater interest and requires more ingenuity, has

$$0 < \lambda < \infty,\ L < \infty, \quad \text{and} \quad w = \infty. \tag{7}$$

We are not aware of a previously published example of (7).

Example 4 Let $t_i = i,\ i \geq 1$; $\lambda = 1$. We construct sequences $\{a_j\}$ and $\{b_j\}$, with $a_1 = 1,\ a_2 = 2$, and integers $2 < b_1 < b_2$. Let $\{r_j\},\ j \geq 1$, be an increasing sequence of positive integers with limit ∞. For $j = 1, 2, \ldots$, define

$$b_{j+2} = b_j + b_{j+1} r_j,$$

so that $\{b_j\}$ is strictly increasing. Let $a_{j+2} = b_j$ for $j = 1, 2, \ldots$, so that $\{a_j\}$ is strictly increasing. For $t_i = i,\ i \notin \{a_1, a_2, \ldots\},\ d_i = i$ and $W_i = 0$.

Because customers with long waiting times depart FIFO and $a_{j+2} = b_j,\ N(t) = 2$ for every $t > a_2$, which implies $L = 2$. We now show that $w = \infty$. For n in the range $a_{j+2} = b_j \leq n < b_{j+1},\ j \geq 1$,

$$\sum_{i=1}^{n} W_i \geq W_{a_{j+2}} = b_{j+2} - a_{j+2} = b_{j+2} - b_j = b_{j+1} r_j, \quad \text{so that}$$

$$\frac{\sum\limits_{i=1}^{n} W_i}{n} > \frac{b_{j+1} r_j}{b_{j+1}} = r_j, \quad \text{for} \quad n \in [b_j, b_{j+1}).$$

Because $r_j \to \infty$ as $n \to \infty$, this gives

$$w = \lim_{n \to \infty} \frac{\sum_{i=1}^{n} W_i}{n} = \infty.$$

3 A weaker sufficient condition

As defined in (1), λ is the limit of $\Lambda(t)/t$ as $t \to \infty$, which is easily shown to be equivalent to $t_i/i \to 1/\lambda$ as $i \to \infty$. Arrival i departs at $d_i = t_i + W_i$. If customers depart FIFO, d_i is the ith (ordered) departure time. If we also have (3), $W_i/i \to 0$ as $i \to \infty$, it is immediate that $d_i/i \to 1/\lambda$ as $i \to \infty$, which is

$$\text{departure rate } = \text{arrival rate}. \tag{8}$$

Actually, it is immediate that (slightly) more is true, which we state as

Lemma 1 *For $0 < \lambda \le \infty$, and when we have FIFO, any two of*

$$(a)\ arrival\ rate = \lambda, \quad (b)\ departure\ rate = \lambda, \quad (c)\ W_i/i \to 0 \ \ as\ i \to \infty$$

implies the third.

It turns out that (3) is sufficient for (8) without requiring FIFO; see Sect. 3.1. Nevertheless, (3) imposes what we call a *weakly FIFO* condition on departure order that we now define (FIFO and (3) are special cases).

Definition (*Weakly FIFO*). For given $0 \le$ (integer) $B < \infty$ and $0 < \epsilon \le 1$, a sequence of arrival and departure times $\{(t_i, d_i), i \ge 1\}$ is called (B, ϵ)-FIFO if $d_j \le d_k$ for all j and k satisfying $B < j < \epsilon k$.

Remark 1 (0, 1)-FIFO is the same as FIFO. If $\{(t_i, d_i)\}$ is (B, ϵ)-FIFO, then it is (B', ϵ')-FIFO provided that $B' \ge B$ and $\epsilon' \le \epsilon$.

Remark 2 If $\{(t_i, d_i)\}$ is (B, ϵ)-FIFO and $0 \le \lambda < \infty$, $\{(t_i, d_i)\}$ is $(0, \epsilon')$-FIFO for sufficiently small $\epsilon' > 0$.

Remark 3 If $0 < \lambda < \infty$ and (3) holds for $\{(t_i, d_i)\}$, then it is (B, ϵ)-FIFO, where ϵ can be set to any positive value < 1, together with a suitably chosen (possibly large) B. Conversely, a (B, ϵ)-FIFO sequence need not in general satisfy (3). In fact, Example 7 in Sect. 5 has a FIFO sequence for which $L = \lambda w = \infty$ and $W_i/i \to 1$.

Next we show that weakly FIFO is sufficient for $L = \infty$. It is not necessary; for example, it can be shown that weakly FIFO does not hold in Example 2.

Theorem 1 *If $0 < \lambda < \infty$ and $w = \infty$ for $\{(t_i, d_i)\}$, then $L = \lambda w = \infty$ holds provided that $\{(t_i, d_i)\}$ is (B, ϵ)-FIFO for $0 \le B < \infty$ and $0 < \epsilon \le 1$.*

Proof We claim that for all $T > 0$,

$$\int_0^T N(t)\,\mathrm{d}t \ge \min\left\{\frac{T}{4}\Big[\Lambda(3T/4) - \Lambda(T/2)\Big], \sum_{\{i:B<i<\epsilon I(T/2)\}} W_i\right\}, \tag{9}$$

where $I(t) = \inf\{i : t_i \ge t\}$. Assume the claim holds.
As $T \to \infty$, $\frac{1}{T}[\Lambda(3T/4) - \Lambda(T/2)] \to \lambda/4 > 0$, $\frac{\text{first term of the min}}{T} \to \infty$, and

$$\frac{1}{T}\sum_{\{i:B<i<\epsilon I(T/2)\}} W_i = \frac{\epsilon I(T/2)}{T}\frac{1}{\epsilon I(T/2)}\sum_{\{i:B<i<\epsilon I(T/2)\}} W_i \to (\epsilon\lambda/2)w = \infty.$$

It follows that $\frac{1}{T}\int_0^T N(t)\,\mathrm{d}t \to \infty = L$.

It remains to prove (9). Consider the following two cases.

Case (i): $d_i > T$ for all i such that $\frac{T}{2} \leq t_i < T$. We have

$$\int_0^T N(t)\,\mathrm{d}t \geq \int_{T/2}^T N(t)\,\mathrm{d}t \geq \sum_{\{i:\frac{T}{2}<t_i\leq\frac{3}{4}T\}} (T-t_i)$$
$$\geq \frac{T}{4}\#\left\{i : \frac{T}{2} < t_i \leq \frac{3}{4}T\right\} = \frac{T}{4}[\Lambda(3T/4) - \Lambda(T/2)],$$

so that (9) holds.

Case (ii): $d_{i_0} \leq T$ for some i_0 such that $\frac{T}{2} \leq t_{i_0} < T$. By the (B,ϵ)-FIFO condition, we have $d_i \leq d_{i_0}$ for all i in the range $B < i < \epsilon i_0$. Since $i_0 \geq I(T/2)$,

$$\int_0^T N(t)\,\mathrm{d}t \geq \sum_{\{i:B<i<\epsilon i_0\}} W_i \geq \sum_{\{i:B<i<\epsilon I(T/2)\}} W_i,$$

so that (9) holds. This completes the proof. □

FIFO is an important special case because it is often assumed in models of queues and frequently occurs in the real world. For example, suppose customers "take a number" on arrival, so that they depart FIFO from the queue. Theorem 1 applies to the entire system when there is one server, and to the queue when there is one or more servers, without regard to whether (3) holds.

Related but different is the use of FIFO in proving LL on pages 163–164 of [3]. Theorem 6.2 assumes (8), but not (3), and allows $0 \leq \lambda \leq \infty$. Corollary 6.3 requires finite w.

3.1 Constructing FIFO departure processes

When the actual departure process is not FIFO, we will see that it is useful to construct another departure process that has this property. One example occurs on page 167 of [3] for what we call the *hat* departure process: Let the mth departure time $\hat{d}_m = \max\{d_1, \ldots, d_m\}$, $m \geq 1$, and $\hat{D}(t) = \#\{m : \hat{d}_m \leq t\}$, $t \geq 0$, which clearly is FIFO. Lemma 6.9 on page 169 of [3] states that for $0 < \lambda \leq \infty$, any two of: arrival rate $= \lambda$, $\lim_{t\to\infty} \hat{D}(t)/t = \lambda$, and (3) implies the third.

For an alternative, let $W'_m = \max\{W_1, \ldots, W_m\}$, $m \geq 1$. Define the *prime* departure process $d'_m = t_m + W'_m$, $m \geq 1$, which is the mth ordered departure (the FIFO property), and $D'(t) = \#\{m : d'_m \leq t\}$, $t \geq 0$. Note that the hat departure process is identical to the original when the latter is FIFO, whereas this is not the case for prime.

Applying Lemma 1 to the prime departure process, noting that $W'_m/m \to 0$ is equivalent to (3) (see Remark 4), we have

Lemma 2 *For* $0 < \lambda \leq \infty$, *any two of: arrival rate* $= \lambda$, $\lim_{t\to\infty} D'(t)/t = \lambda$, *and (3) implies the third.*

From $d'_m \geq d_m \geq t_m$ for all m, $D'(t) \leq D(t) \leq \Lambda(t)$ for all t. From these inequalities, the following theorem is an immediate consequence of Lemma 2.

Theorem 2 *For arrival rate λ, where $0 < \lambda \leq \infty$, (3) is sufficient for (8).*

Of course, the hat process and Lemma 6.9 can be used instead of Lemma 2 to complete the proof of Theorem 2. It is such an obvious consequence of either lemma that an explicit statement of this result is hard to find. We do not claim it to be new. Also note that Lemma 6.9 is an easy consequence of the inequalities

$$t_m \leq \hat{d}_m \leq t_m + W'_m = d'_m, \text{ and } W_m \leq \hat{W}_m \equiv \hat{d}_m - t_m \leq W'_m, \quad m \geq 1,$$

where we also have that $\hat{W}_m/m \to 0$ is equivalent to (3).

What may have been missed is that (3) is not necessary for (8). For example, suppose $0 < \lambda < \infty$, where, in the terminology used in Examples 1–4, the long-run fraction of arrivals that have short ($W_i = 0$) waiting times is one. (8) holds, regardless of the size of the longs. If $W_i = i$ for an infinite subsequence of longs, (3) does not hold.

Remark 4 See Lemma 2.10 on page 45 of [3] for an easy proof that (i) $W_m/m \to c$, as $m \to \infty$, is equivalent to (ii) $W'_m/m \to c$, as $m \to \infty$, where $c = 0$. Their proof is correct for $c = 0$, but their claim that this result also holds for $c > 0$ is not true. While (i) implies (ii) for this case, the converse does not hold. Here is a counterexample. Let $W_i = i$ for even i, $W_i = 0$ for odd i. Limit (ii) holds, where $c = 1$. Limit (i) does not exist.

They used the incorrect version of Lemma 2.10 ($c > 0$ when $\lambda < \infty$) in their proof of Lemma 6.9. Because the t_i are nondecreasing, however, it is enough to use only the $c = 0$ case, as we have done. Thus Lemma 6.9 is correct, as we have shown. These errors seem to have originated in [9], where the first error occurs in Eq. (12) on page 153, and the second follows two lines later.

4 A necessary and sufficient condition

We now derive a necessary and sufficient condition for $L = \infty$ via truncation, where, for $n = 1, 2, \ldots$, and $k > 0$ (k may be a function of n), we define

$$\overline{W}_n(k) = \frac{1}{n}\sum_{i=1}^{n} \min(W_i, k). \tag{10}$$

We also replace the second inequality in (2) by the more precise

$$\int_0^T N(t)\,\mathrm{d}t = \sum_{i=1}^{\Lambda(T)} \min(W_i, T - t_i). \tag{11}$$

Theorem 3 *When* $0 < \lambda < \infty$,

$$\lim_{n\to\infty} \overline{W}_n(n) = \infty \tag{12}$$

is necessary and sufficient for $L = \lambda w = \infty$.

Proof To show that (12) implies $L = \infty$, fix $0 < c < 1/(1+\lambda) < 1$. Since

$$\lim_{T\to\infty} [\Lambda(cT) + t_{\Lambda(cT)}]/T = c\lambda + c < 1,$$

$$\Lambda(cT) < T - t_{\Lambda(cT)} \text{ for all sufficiently large } T. \tag{13}$$

From (11),

$$\begin{aligned}\frac{1}{T}\int_0^T N(t)\,\mathrm{d}t &\geq \frac{1}{T}\sum_{i=1}^{\Lambda(cT)} \min(W_i, T - t_i)\\ &\geq \frac{1}{T}\sum_{i=1}^{\Lambda(cT)} \min(W_i, T - t_{\Lambda(cT)}), \text{ and from (13) for large } T,\\ &\geq \frac{1}{T}\sum_{i=1}^{\Lambda(cT)} \min(W_i, \Lambda(cT)) = \frac{\Lambda(cT)}{T}\overline{W}_{\Lambda(cT)}(\Lambda(cT)).\end{aligned}$$

As $T \to \infty$, the last expression $\to \infty$, which implies $L = \infty$.

To show that $L = \infty$ implies (12), fix $0 < c' < \min(\lambda, 1)$. Since

$$\begin{aligned}&\lim_{n\to\infty} t_{[c'n]}/n = c'/\lambda < 1,\\ &t_{[c'n]} < n \quad \text{for all sufficiently large } n,\end{aligned} \tag{14}$$

where $[x] \equiv$ greatest integer $\leq x$. From (10), (14), and noting that $[c'n] < n$,

$$\begin{aligned}\overline{W}_n(n) &\geq \frac{1}{n}\sum_{i=1}^{[c'n]} \min(W_i, t_{[c'n]}) \quad \text{(for large } n\text{)}\\ &\geq \frac{1}{n}\sum_{i=1}^{[c'n]} \min(W_i, t_{[c'n]} - t_i) = \frac{1}{n}\int_0^{t_{[c'n]}} N(t)\,\mathrm{d}t\\ &= \frac{t_{[c'n]}}{n}\frac{1}{t_{[c'n]}}\int_0^{t_{[c'n]}} N(t)\,\mathrm{d}t.\end{aligned}$$

As $n \to \infty$, the last expression $\to (\frac{c'}{\lambda})L = \infty$, which implies (12) and completes the proof. □

We may replace (12) by $\lim_{n\to\infty}\overline{W}_n(rn) = \infty$ with $0 < r < \infty$. In fact, it is easily shown that $\lim_{n\to\infty}\overline{W}_n(r'n) = \infty$ for some $0 < r' < \infty$ implies that $\lim_{n\to\infty}\overline{W}_n(rn) = \infty$ for all $0 < r < \infty$.

A key feature of (12) is that the length of truncated intervals increases with n as we take the limit. We now describe a more conventional approach, which is to fix truncation length, k, let $n \to \infty$, and then let $k \to \infty$. As $n \to \infty$ in (10),

$$\liminf_{n\to\infty}\overline{W}_n(k) < \infty$$

exists (the limit may not). It is easy to see that

$$\lim_{k\to\infty}\liminf_{n\to\infty}\overline{W}_n(k) = \infty \tag{15}$$

implies (12). Hence we have

Corollary *When $0 < \lambda < \infty$, (15) is a sufficient condition for $L = \infty$.*

As we show in Example 5, however, (15) is not necessary. We now investigate how (12) and (15) apply to earlier examples.

Example 5 First consider Example 2. For fixed $k > 0$, any $n \geq 1$, and W_i for $i = 1, 2, \ldots, n$, define $\#_0(n)$ and $\#_n(n)$ as the number of the W_i that, respectively, are > 0 and $\geq n$, where $\#_0(n) \geq \#_n(n)$. It is easy to show that $\#_0(n)/n \to 0$ as $n \to \infty$. From this, $\overline{W}_n(k) \leq \#_0(n)k/n \to 0$ as $n \to \infty$. Hence (15) does not hold, showing, as noted above, that it is not necessary for $L = \infty$.

To show that (12) holds, consider $n \in [a_j, a_{j+1})$.

For j odd, $j > 3$, and $s = (j+3)/2, (j+5)/2, \ldots, j$,

$$W_{a_s} = b_s - a_s = 2^{2s-1} - 2^{s-1} > 2^{j+2} - 2^{j-1} > 2^j > n.$$

For j even, $j > 2$, and $s = (j+2)/2, (j+4)/2, \ldots, j$,

$$W_{a_s} > 2^{j+1} - 2^{j-1} > 2^j > n.$$

In both cases, $\#_n(n) \geq (j-1)/2$ for $n \in [a_j, a_{j+1}),\ j \geq 4$. Hence $\#_n(n) \to \infty$ as $n \to \infty$. From $\overline{W}_n(n) \geq \#_n(n)$, (12) holds!

Now consider Example 4, using the same notation. By the same analysis, (15) does not hold. For $n \in [b_j, b_{j+1})$, the *sum* of the waiting times of customers with arrival times $a_j, a_{j-2}, \ldots$ is at most b_j, and the sum of the waiting times of customers with arrival times $a_{j-1}, a_{j-3}, \ldots$ is at most b_{j-1}. Among the long waiting times in $\{W_1, \ldots, W_n\}$, only those with arrival times a_{j+1} and a_{j+2} are not included in one of these sums. It follows that $\overline{W}_n(n)$ is bounded above by 4; (12) does not hold.

Although only a sufficient condition, (15) has a natural connection to the $w = \infty$ case in a stochastic setting, where it holds, and gives an easy proof.

Example 6 Now consider the stochastic setting in **H** where the W_i are stationary and ergodic, with $w = \infty$. The truncated quantities $W_i(k) = \min(W_i, k)$ are also stationary and ergodic, but with finite mean

$$E[W_i(k)] = \int_0^k P(W_i > u)\,\mathrm{d}u,$$

where $\lim_{k\to\infty} E[W_i(k)] = \infty$. Thus the lim inf in (15) is equal to $E[W_i(k)]$, and (15) holds! Now apply the sample-path proof of LL for finite w (sketched above) to the truncated quantities, and bring in the stochastic setting at the end, to conclude that $w = \infty \Longrightarrow L = \infty$, but this is "only" w.p.1. Note that this includes the special case of i.i.d. W_i, where (3) does not hold when $w = \infty$ (see Theorem 4 in Sect. 5).

In Example 6, the same argument goes through when the W_i are stationary but not ergodic, where for invariant sigma field $\mathcal{F}$, $E(W \mid \mathcal{F}) = \infty$, and

$$E[W_i(k) \mid \mathcal{F}] = \int_0^k P(W_i > u \mid \mathcal{F})\,\mathrm{d}u.$$

5 An earlier stochastic example

In Sect. 2, we asserted that for i.i.d. W_i,

$$w < \infty \text{ w.p.1 if and only if (3) holds w.p.1,} \tag{16}$$

and showed in Example 6 that $L = \infty$ w.p.1. when $w = \infty$ w.p.1. (16) was presented as a counterexample to the necessity of (3) when $w = \infty$ on p. 298 of [13] without a reference, and in [14] with reference to Loève [8]. However, this reference is problematic because of possible confusion of edition number and difficulty of finding the result. Here is a closely related result:

For i.i.d. W_i, where *i.o.* denotes *infinitely often*,

$$w < \infty \text{ w.p.1 if and only if } P(W_i > i \; i.o.) = 0. \tag{17}$$

(17) appears as Problem 10 on page 44 of [1]. Breiman also references Loève, with similar difficulty of verification.

While proofs of (16) and (17) appear elsewhere, we prove them here because we need (16) and because the proofs are easy and short.

Theorem 4 *(16) and (17) hold.*

Proof For (17), let W be a generic waiting time, where $w = \int_0^\infty P(W > u)\,\mathrm{d}u$.

$$w < \infty \text{ w.p.1 if and only if } \sum_{i=1}^{\infty} P(W > i) < \infty, \text{ which is equivalent to}$$
$$\sum_{i=1}^{\infty} P(W_i > i) < \infty,$$

because the W_i are identically distributed. Replacing probabilities by indicators, and taking expected value outside the sum, $w < \infty$ w.p.1 if and only if

$$E\left[\sum_{i=1}^{\infty} I\{W_i > i\}\right] < \infty.$$

From the Borel–Cantelli lemma, we have (17) because events $\{W_i > i\}$ are independent.

When $w < \infty$ w.p.1, the strong law of large numbers readily gives that (3) holds w.p.1. That is, $w < \infty$ is sufficient (as it is on sample paths). To show that it is necessary, suppose $w = \infty$. From (17) and by Kolmogorov's 0-1 law, $P(W_i > i \; i.o.) = 1$. For any sample path where $W_i > i \; i.o.$, (3) does not hold. Hence (3) holds w.p.0 for this case, and we have (16). □

Thus (16), important because it is for a conventional queueing model with stable waiting times, shows that $L = \lambda w = \infty$ may hold when (3) does not. Example 2 is not a conventional model. As we now illustrate, there are much simpler deterministic examples that are conventional, but they are unstable.

Example 7 Consider a single-server FIFO queue with $t_i = i$ and service times $S_i = 2,\ i \geq 1$. It is easy to see that $W_i = 1 + i$ and $N(t)$ both increase without bound. We have $\lambda = 1$, both L and w are infinite, and (3) does not hold. Incidentally, (15) holds here.

While writing this paper, we did a literature search to be sure that we were not "rediscovering the wheel." Probably the most complete treatment of LL is in the El-Taha and Stidham book, [3], which we refer to several times. We searched this book for necessary and sufficient conditions for LL and also for the more general $H = \lambda G$, and were surprised, and worried at first, at what we found.

Theorem 6.17 on page 179 claims that a necessary and sufficient condition for $H = \lambda G$ (including the case $G = \infty$) is that an error term converges to zero (Condition R). In the following Example 6.2, they conclude that LL holds when λw is well defined, where the convergence of the error term turns out to be (3). Of course, this is incorrect. Condition R (and (3)) is not necessary when G (w) is infinite. The result is correct if G is restricted to $0 \leq G < \infty$. This is an unfortunate oversight.

Also unfortunate are similar statements in at least two other places. The last sentence on page 160 states in words the incorrect Theorem 6.17. Independent of that, the first sentence in Sect. 6.3 on page 170 states the same thing about LL, and refers incorrectly to an earlier result.

5.1 Stability properties of our infinite server model

The stability of queues is a research area where there is usually a stochastic framework, and the definition of a stable queue depends on this framework. See Appendix C, stability in stochastic models, in [3] for a treatment of the question: what is a stable queue?

From this, we come away with the fact that "stable queue" has no generally accepted definition that covers all situations. Instead, it is useful to think of a *queueing model*, which is a set of assumptions that determine how a queue evolves over time. These assumptions determine how certain stochastic processes evolve, of which the most prominent (usually) are waiting-time sequence $\{W_i\}$ and number-of-customers-in-system process $\{N(t)\}$.

We define a stochastic process to be stable if it converges (W_i as $i \to \infty$ or $N(t)$ as $t \to \infty$) to a proper limiting distribution, independent of initial conditions. These are rough definitions that leave out certain possibilities, such as periodic behavior. We define a *queueing model* to be stable if *both* $\{W_i\}$ and $\{N(t)\}$ are stable.

Similarly, we define a stochastic process to be *unstable* if it grows without bound to ∞ (we are dealing here with non-negative quantities that can grow in only one direction), usually w.p.1. We define a *queueing model* to be unstable if *both* $\{W_i\}$ and $\{N(t)\}$ are unstable.

As an example, consider an $M/M/1$ queue with arrival rate λ and service rate μ, with $\rho = \lambda/\mu$. $\{N(t)\}$ is an irreducible Markov chain that is positive recurrent, null recurrent, or transient if (respectively) $\rho < 1,\ \rho = 1$, or $\rho > 1$. The positive recurrent case is stable; the transient case is unstable. In the null case, the chain returns over and over to an empty system, but it is not stable. If we define unstable to be growth to ∞ w.p.1, it is not unstable. The waiting-time sequence has the identical properties for each range of ρ.

In our sample-path examples, these definitions are not helpful, with the exception of Example 7. For our infinite-server model with i.i.d. waiting times (ISM for short), however, our definitions apply, where (obviously) $\{W_i\}$ is stable. What about $\{N(t)\}$?

It is difficult to answer this question without defining the arrival process. It will be helpful to consider this special case: Arrivals are Poisson at rate λ, so that the ISM is an $M/G/\infty$ queue. Our analysis of this model is a refinement and extension of Example 2–6 on page 75 of [13].

Let G be the service (waiting time) distribution. Extend the time horizon, $-\infty < t < \infty$. Define s-customers to be those with waiting time $\leq s$, and for every t, let $N_s(t)$ be the number of s-customers in the system at time t. By subscript s, we denote quantities $\lambda_s,\ L_s$, and w_s for the system consisting only of s-customers.

To be in the system at time t, an s-customer must have arrived in the interval $(t - s, t]$, where the number of s-customers that arrive in this interval is Poisson with mean $\lambda s G(s)$. An s-customer that arrives at time u in this interval will be in the system at time t if its waiting time exceeds $t - u$, an event with probability $[G(s) - G(t - u)]/G(s)$. Given the number of s-customer arrivals in this interval, their (unordered) arrival times are i.i.d. uniform with density $1/s$. Hence the probability that any one of them is in the system at time $t,\ p_t$, is

$$p_t = [sG(s)]^{-1} \int_{t-s}^{t} \big[G(s) - G(t-u)\big]\,\mathrm{d}u = [sG(s)]^{-1} \int_{0}^{s} [G(s) - G(u)]\,\mathrm{d}u.$$

We have what is referred to in [13] as a “random partition” of the s-arrivals in the interval $(t-s, t]$, and it follows that $N_s(t)$ has a Poisson distribution with mean, L_s in our notation above,

$$L_s = \lambda s G(s) p_t = \lambda \int_{0}^{s} [G(s) - G(u)]\,\mathrm{d}u. \tag{18}$$

Noting that $\int_0^s [1 - G(u)/G(s)]\,\mathrm{d}u = w_s$, the RHS of (18) is equal to $\lambda_s w_s$.

Now let $s \to \infty$. When w is finite, $N(t)$ has a Poisson distribution with mean λw. The process $\{N(t)\}$ is stationary and hence stable.

When w is infinite, we have very different behavior. The limit of (18) is ∞. Because, at any point in the sample space, $N_s(t)$ can only increase as s increases, $N(t) = \infty$ w.p.1.

Thus process $\{N(t)\}$ is unstable, even though process $\{W_i\}$ is stable. The queueing model? We could say it is not stable, which is not to say it is unstable. In our experience, processes $\{N(t)\}$ and $\{W_i\}$ behaved alike; either both were stable or both were unstable. We do not have a name for what happens here.

We believe that this peculiar limiting behavior will hold for much more general arrival processes, but it will be difficult to carry out the analysis, provided that the arrival process is independent of the waiting times. When they are dependent, there may be other possibilities. For example, if the arrival process is defined as in Example 3, and the W_i are i.i.d., $L = 1$, but now we have $\lambda = 0$.

6 Concluding remarks

Some of our examples are not conventional queueing models, but this should not be regarded as an objection. The beauty and simplicity of Little’s law is that it makes no assumptions about model structure. This also accounts for the ubiquity of its applications; see [6] and the references therein.

In an earlier version of this paper, we commented that (3) is sufficient for (8) without requiring FIFO, but gave no reference. A referee requested that we supply one. We thought this was well settled and intended to comply, but an investigation revealed that there are surprising gaps in the literature. We are aware of a proof by Sigman [10], using a different approach, but it is not published in a refereed journal or a book, and does not cover the case $\lambda = \infty$. It is not archival and may be deleted at any time.

We find the approach in [3] and [9] more intuitive as well as more general. For any FIFO departure process, such as hat or prime, where the arrival rate $\lambda > 0$ exists, it is immediate from the elementary Lemma 1 that (3) is necessary and sufficient for (8). Without FIFO, we then showed that (3) is sufficient but not necessary. Note that we used the correct part of Lemma 2.10 in [3] only for waiting times; the arrival process plays no role.

In a typical queueing model, arriving customers at a facility have service requirements, whereas the facility has constraints on its ability to meet these requirements. For a single-server queue, the constraint is that, at most, one customer is served at a time (an infinite-server queue removes such a constraint). Sometimes, the facility cannot keep up with the service requirements, for example, the arrival rate at a single-server queue is greater than the service rate. These queues are unstable. Both L and w are infinite, but there is no reason to expect (3) to hold. This is the situation in Example 7.

Stable queues may also have infinite w. This is true, for example, in a $GI/G/1$ FIFO queue with $\lambda/(\text{service rate}) < 1$ and infinite service time *variance*. Infinite w may occur because some customers are delayed by others, even though, in this case, the system empties from time to time. An important component of waiting time in this example is the remaining service time of the customer in service, if any, found by an arrival. It has an infinite mean, and is the reason w is infinite.

Acknowledgements We are grateful to Karl Sigman for many comments on this manuscript, and in particular, for contributing the proof of (17), and to Rhonda Righter, whose comments helped us improve the exposition.

References

1. Breiman, L.: Probability. Addison Wesley, Reading (1968)
2. Brumelle, S.: On the relation between customer and time averages in queues. J. Appl. Probab. **8**, 508–520 (1971)
3. El-Taha, M., Stidham Jr, S.: Sample-Path Analysis of Queueing Systems. Kluwer Academic Publishers, Boston (1999)
4. Franken, P., König, D., Arndt, U., Schmidt, V.: Queues and Point Processes. Wiley, New York (1982)
5. Little, J.D.C.: A proof of the queuing formula: $L = \lambda W$. Oper. Res. **9**, 383–387 (1961)
6. Little, J.D.C.: Little's law as viewed on its 50th anniversary. Oper. Res. **59**, 536–549 (2011)
7. Little, J.D.C., Wolff, R.W.: The "Flaw" in Little (1961), its Identification and Fixes. Online commentary on the INFORMS web site. http://orforum.blog.informs.org/ (2011)
8. Loève, M.: Probability Theory, 3rd edn. Van Nostrand, Princeton (1963)
9. Serfozo, R.F.: Little laws for utility processes and waiting times in queues. Queueing Syst. **17**, 137–181 (1994)
10. Sigman, K.: Notes on Little's law. www.columbia.edu/ks20/stochastic-I/stochastic-I-LL.pdf (2009)
11. Stidham Jr, S.: A last word on $L = \lambda W$. Oper. Res. **22**, 417–421 (1974)
12. Whitt, W.: A review of $L = \lambda W$ and extensions. Queueing Syst. **9**, 235–268 (1991)
13. Wolff, R.W.: Stochastic Modeling and the Theory of Queues. Prentice-Hall, Englewood Cliffs (1989)
14. Wolff, R.W.: Little's law and related results. In: Cochran, J.J. (ed.) Wiley Encyclopedia of Operations Research and Management Science, vol. 4, pp. 2828–2841. Wiley, Hoboken (2011)