Notes on Little's Law: $L = \lambda W$

IEOR 4615: Service Engineering

Professor Whitt, January 27, 2015

1 Introduction

Little's law or $L = \lambda W$ is a **conservation law** that provides important insight into queueing systems. The relation $L = \lambda W$ can be quickly stated:

> The average number of customers waiting in line (or items in a system), L, is equal to the arrival rate (or throughput), λ, multiplied by the average waiting time (time spent in system) per customer, W.

If we know any two of these quantities, then we necessarily know all three. The easily understood reason is reviewed in §2.3.

Even though the basic relation is remarkably simple, and can be easily understood, the full implications for applications can be quite surprising. That is illustrated by the examples done in class, which mostly follow Avishai Mandelbaum's notes [9]. An early paper illustrating possible applications is [10].

For stochastic queueing models where λ is known, the relation $L = \lambda W$ yields a relation between the two steady-state mean values. Given λ, we know both L and W whenever one has been calculated. For example, many common queueing models are birth-and-death processes, as in §6.3 of Ross [11], where λ is part of the model data and L can be readily computed; we can apply Little's law to obtain $W = L/\lambda$; e.g., see §§8.2-8.3 in Ross [11].

As reviewed in [14, 15], Little's law can be expressed in terms of steady-state distributions in stochastic models or as limits of averages as the sample size increases. A version of the result involving limits of averages is given here in Theorem 2.1 in §2.2 and proved in §3.1. The classic paper is Little [7], but the version here follows Stidham [13]. See El-Taha and Stidham [1] and Little [8] for more discussion.

The main idea is surprisingly simple, but a careful proof is surprisingly difficult. There are also rather deep implications of the key relationships that go beyond the basic relation. For example, there is the central limit theorem (CLT) version of $L = \lambda W$ [2]; it is basically an extension of the main idea, once properly understood. This is reviewed in [14]. No doubt there remains more interesting things to discover. See [17] for a short recent addition to the theory.

However, most applications involve measurements over finite time intervals. It is important to recognize that Little's law typically does *not* apply exactly with such **finite measurements**. We discuss how the finite averages are related in §2.1 and §2.3, following §2 of Kim and Whitt [5]; see Little [8] and Mandelbaum [9] for further discussion. For data collected over finite intervals, it is natural to take a **statistical approach**, e.g., and estimate confidence intervals. We discuss the statistical approach in §4, pointing toward [5, 6] for further discussion.

Organization. Here is how the rest of these notes are organized: We tell the main story in §2. In §2.1 we carefully define the finite averages. In §2.2 we state a version of the limit theorem for averages. In §2.3 we carefully examine how the finite averages are related. We provide technical details in §3. In §3.1 we give a detailed proof of the limit theorem in §2.2. That proof relies on two basic technical propositions in §3.2. We conclude with a brief discussion of the statistical approach in §4.

2 The Main Story

2.1 The Performance Functions and Their Averages

We initially consider a finite time interval $[0, t]$. Consistent with most applications, we assume that the system was in operation in the past, prior to time 0, and that it will remain in operation after time t. We will use standard queueing terminology, referring to the items being counted as customers. We focus on the customers that are in the system at some time during the interval $[0, t]$. Let these customers be indexed in order of their arrival time, which could be prior to time 0 if the system is not initially empty (with some arbitrary method to break ties, if any).

For customer k, let A_k be the arrival time, D_k the departure time and $W_k \equiv D_k - A_k$ the waiting time (time in system), where $-\infty < A_k < D_k < \infty$, $[0, t] \cap [A_k, D_k] \neq \emptyset$ and $\equiv$ denotes "equality by definition." Let $R(0)$ count the customers that arrived before time 0 that remain in the system at time 0; let $A(t)$ count the total number of new arrivals in the interval $[0, t]$; and let $L(t)$ be the number of customers in the system at time t. Thus, $A(t) = \max\{k \geq 0 : A_k \leq t\} - R(0)$, $t \geq 0$, and $L(0) = R(0) + A(0)$, where $A(0)$ is the number of new arrivals at time 0, if any. We will carefully distinguish between $R(0)$ and $L(0)$, but the common case is to have $A(0) = 0$ and $L(0) = R(0)$.

The respective averages over the time interval $[0, t]$ are

$$\bar{\lambda}(t) \equiv t^{-1}A(t), \quad \bar{L}(t) \equiv t^{-1}\int_0^t L(s)\,ds, \quad \bar{W}(t) \equiv (1/A(t))\sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \tag{1}$$

where $0/0 \equiv 0$ for $\bar{W}(t)$. The first two are time averages, while the last, $\bar{W}(t)$, is a customer average, but over all arrivals during the interval $[0, t]$.

We will focus on these averages over $[0, t]$ in (1), but we could equally well consider the averages associated with the first n arrivals. To do so, let T_n be the arrival epoch of the n^{th} new arrival, i.e., $T_n \equiv A_{n+R(0)}$, $n \geq 0$,

$$\bar{\lambda}_n \equiv n/T_n, \quad \bar{L}_n \equiv (1/T_n)\int_0^{T_n} L(s)\,ds, \quad \bar{W}_n \equiv n^{-1}\sum_{k=R(0)+1}^{R(0)+n} W_k. \tag{2}$$

As in (1), the first two averages in (2) are time averages, but over the time interval $[0, T_n]$, while the last, $\bar{W}_n$, is a customer average over the first n (new) arrivals. If there is only a single arrival at time T_n, then the averages in (2) can be expressed directly in terms of the averages in (1): $\bar{\lambda}_n = \bar{\lambda}(T_n)$, $\bar{L}_n = \bar{L}(T_n)$ and $\bar{W}_n = \bar{W}(T_n)$, so that conclusions for (1) yield analogs for (2).

2.2 The Relation Among Limits of the Averages

There are various statements of $L = \lambda W$ depending on what is assumed. Here is a fairly general statement, which assumes existence of only two limits. For practical purposes, all the limits may be assumed to exist; then the main conclusion is the relation among the limits: $L = \lambda W$. Even assuming that *all* limits exist, establishing the relation is still somewhat challenging.

Theorem 2.1 ($L = \lambda W$, *Little's law*) *If* $\bar{\lambda}(t) \to \lambda$ *as* $t \to \infty$ *and* $\bar{W}_n \to W$ *as* $n \to \infty$ *for* $\bar{\lambda}(t)$ *in* (1) *and* $\bar{W}_n$ *in* (2), *where* $0 < \lambda < \infty$ *and* $W < \infty$, *then*

$$(\bar{L}(t), \bar{\lambda}(t), \bar{W}(t)) \to (L, \lambda, W) \quad \textit{as} \quad t \to \infty \tag{3}$$

and

$$(\bar{L}_n, \bar{\lambda}_n, \bar{W}_n) \to (L, \lambda, W) \quad as \quad n \to \infty, \tag{4}$$

where

$$L = \lambda W. \tag{5}$$

We next examine the relation among the finite averages, from which the main content of Theorem 2.1 becomes evident. However, we also give a detailed proof of Theorem 2.1 in §3.1. The proof draws on two basic technical propositions stated and proved in §3.2.

2.3 How the Finite Averages in (1) Are Related

Figures 1 and 2 below show how the three averages in (1) are related. These averages are related via $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$ if the system starts and ends empty, i.e., if $R(0) = L(t) = 0$, as we show in Theorem 2.2 below. However, more generally, these averages are not simply related. To illustrate, in Figures 1 and 2 a bar of height 1 is included for each of the customers in the system at some time during $[0, t]$ with the bar extending from the customer's arrival time to its departure time. (In this example the customers do not depart in the same order they arrived.) Thus the width of the bar is the customer's waiting time. For $0 \le s \le t$, the number of bars above any time s is $L(s)$.

To better communicate what is going on visually, we have ordered the customers in a special way. In Figures 1 and 2, the customers that arrive before time 0 but are still there at time 0 are placed first, starting at the bottom and proceeding upwards. These customers are ordered according to the arrival time, so the customers that arrived before time 0 appear at the bottom. One of these customers also departs after time t. The customers that arrived before time 0 and are still in the system at time 0 contribute to the regions A, B and C in Figure 2.

After the customers that arrived before time 0, we place the customers that arrive after time 0 and depart before time t, in order of arrival; they constitute region D in Figure 2. Finally, we place the customers that arrive after time 0 but depart after time t. These customers are ordered according to their arrival time as well; they constitute regions E and F in Figure 2. Three extra horizontal lines are included, along with the vertical lines at times 0 and t, to separate the regions. The arrival numbers are indicated along the vertical y axis. The condition $R(0) = L(t) = 0$ arises in Figure 2 as the special case in which all regions except region D are empty.

The averages can be expressed in terms of the two *cumulative processes*,

$$C_L(t) \equiv \int_0^t L(s)\, ds \quad \text{and} \quad C_W(t) \equiv \sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \quad t \ge 0. \tag{6}$$

The difference between these two cumulative processes can be expressed in terms of the process $T_W^{(r)}(t)$, recording the *total residual waiting time* of all customers in the system at time t, i.e.,

$$T_W^{(r)}(t) \equiv \sum_{k=1}^{L(t)} W_k^{r,t}, \tag{7}$$

where $W_k^{r,t}$ is the remaining waiting time at time t for customer k in the system at time t (with index k assigned at time t among those remaining). The averages in (1) are the *time average* $\bar{L}(t) \equiv t^{-1}C_L(t)$ and the *customer average* $\bar{W}(t) \equiv C_W(t)/A(t)$. For a region A in Figure 2,

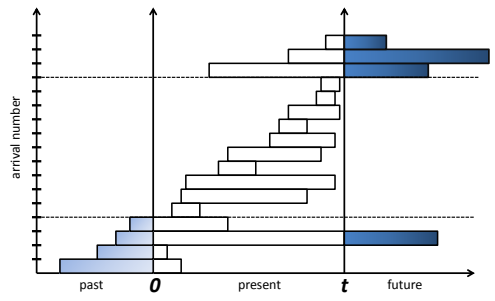

Figure 1: The total work in the system during the interval $[0, t]$ with edge effects: including arrivals before time 0 and departures after time t.

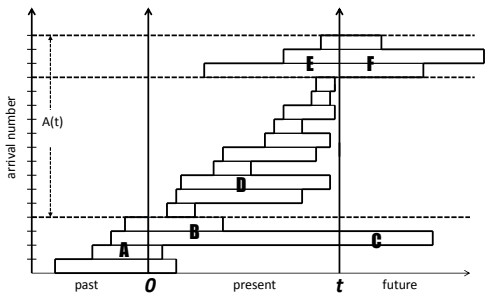

Figure 2: Six regions: waiting times (i) of customers that both arrive and depart inside $[0, t]$ (D), (ii) of arrivals before time 0 $(A \cup B \cup C)$ and (iii) of departures after time t $(C \cup E \cup F)$.

let $|A|$ be the area of A. In general, the cumulative processes can be expressed in terms of the regions in Figure 2 as $C_L(t) = |B \cup D \cup E|$ and $C_W(t) = |D \cup E \cup F|$, while $T_W^{(r)}(0) = |B \cup C|$ and $T_W^{(r)}(t) = |C \cup F|$, so that

$$C_L(t) - C_W(t) = |B| - |F| = |B \cup C| - |F \cup C| = T_W^{(r)}(0) - T_W^{(r)}(t). \tag{8}$$

This relation for $C_L(t)$ is easy to see if we let ν be the total number of arrivals and departures in the interval $[0, t]$, τ_k be the k^{th} ordered time point among all the arrival times and departure times in $[0, t]$, with ties indexed arbitrarily and consistently, $\tau_0 \equiv 0$ and $\tau_{\nu+1} = t$. Then

$$C_L(t) \equiv \int_0^t L(s)\, ds = \sum_{j=1}^{\nu+1} \int_{\tau_{j-1}}^{\tau_j} L(s)\, ds = \sum_{j=1}^{\nu+1} L(\tau_{j-1})(\tau_j - \tau_{j-1}) = |B \cup D \cup E|,$$

where the last relation holds because $L(\tau_{j-1})$ is the number of single-customer unit-height bars above the interval $[\tau_{j-1}, \tau_j]$. Since $C_L(t) = C_W(t) = |D|$ if $R(0) = L(t) = 0$, we necessarily have the following well known result, appearing as Theorem I of [4].

Theorem 2.2 (*traditional finite-time Little's law*) *If* $R(0) = L(t) = 0$, *then* $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$.

Proof. Under the condition, $\bar{L}(t) \equiv \frac{C_L(t)}{t} = \frac{C_W(t)}{t} = \left(\frac{A(t)}{t}\right)\left(\frac{C_W(t)}{A(t)}\right) \equiv \bar{\lambda}(t)\bar{W}(t)$. ∎

On the other hand, for the common case in which there are customers in the system during $[0, t]$ that arrived before time 0 and/or depart after time t, as in Figures 1 and 2, there is no simple relation between these cumulative processes and the associated averages, because of the interval edge effects. Nevertheless, the analysis above exposes the relationship that does hold. This material is taken from §2 of [5]. Variants of these relations are needed to establish sample-path limits in Little law theory; e.g., see Theorem 1 of [2]. A variant appears on p. 17.4 of [9], who credits it to his student Abir Koren and emphasizes its importance for looking at data.

In the following theorem we relate the direct finite averages in (1) to the following indirect estimators, which we might use to estimate one from the others:

$$\bar{L}_{W,\lambda}(t) \equiv \bar{\lambda}(t)\bar{W}(t), \quad \bar{\lambda}_{L,W}(t) \equiv \frac{\bar{L}(t)}{\bar{W}(t)} \quad \text{and} \quad \bar{W}_{L,\lambda}(t) \equiv \frac{\bar{L}(t)}{\bar{\lambda}(t)}. \tag{9}$$

The following result is a consequence of the reasoning above.

Theorem 2.3 (*extended finite-time Little's law*) *The averages in* (1) *and* (9) *are related by*

$$\begin{aligned}
\Delta_L(t) &\equiv \bar{L}_{W,\lambda}(t) - \bar{L}(t) = \frac{|F| - |B|}{t} = \frac{T_W^{(r)}(t) - T_W^{(r)}(0)}{t}, \\
\Delta_W(t) &\equiv \bar{W}_{L,\lambda}(t) - \bar{W}(t) = \frac{|B| - |F|}{A(t)} = -\frac{\Delta_L(t)}{\bar{\lambda}(t)} = \frac{T_W^{(r)}(0) - T_W^{(r)}(t)}{A(t)}, \\
\Delta_\lambda(t) &\equiv \bar{\lambda}_{L,W}(t) - \bar{\lambda}(t) = \left(\frac{|B| - |F|}{|D| + |E| + |F|}\right)\bar{\lambda}(t) = -\frac{\Delta_L(t)}{\bar{W}(t)},
\end{aligned} \tag{10}$$

where $|B|$ *is the area of the region* B *in Figure* 2 *and* $T_W^{(r)}(t)$ *is defined in* (7).

3 Technical Details

This section is devoted to a proof of Theorem 2.1. We give the proof in §3.1 and then prove two basic technical propositions used in that proof in §3.2.

3.1 Proof of Theorem 2.1

From the previous section, the main conclusion of Theorem 2.1, the relation $L = \lambda W$ in (5), should be evident, for the most part. As the sample size grows by letting $t \to \infty$ in (1) or $n \to \infty$ in (2), then the edge effects identified in Theorem 2.3 should remain stable while the cumulative processes grow. However, a good general mathematical result is not entirely straightforward. We present a variant of the proof by Stidham [13]. Our proof relies heavily on two basic technical lemmas stated and proved in §3.2.

There is a technical complication, which requires care. The result does not require that the customers depart in the same order that they arrive. That is common in multi-server queues. Since the customers need not depart in the same order they arrive, D_k is the departure time of the k^{th} arrival, but not necessarily the k^{th} departure overall. To work with the departure times, let

$$D_k^{\uparrow} \equiv \max_{1 \le j \le k} D_j. \tag{11}$$

Let $D^{\uparrow}(t)$ be the associated counting processes, defined by

$$D^{\uparrow}(t) \equiv \max\{k \ge 1 : D_k^{\uparrow} \le t\} - R(0), \quad t \ge 0, \tag{12}$$

with $D^{\uparrow}(t) \equiv 0$ if $D_1^{\uparrow} > t$. Note that $D^{\uparrow}(t)$ is defined in terms of $D_k^{\uparrow}$ just like $A(t)$ is defined in terms of A_k in §2.1. In contrast, let $D(t)$ be the number of departures by time t. Note that $D(t) = R(0) + A(t) - L(t)$, but $D(t)$ is *not* defined in terms of D_k as in (12).

Reasoning as in the previous section, we start with the key relation

$$\sum_{j=1}^{D^{\uparrow}(t)} W_k - \sum_{j=1}^{R(0)} W_k \le \int_0^t Q(s)\,ds \le \sum_{j=R(0)+1}^{R(0)+A(t)} W_k + \sum_{j=1}^{R(0)} W_k, \quad t \ge 0. \tag{13}$$

We now elaborate on (13). First, by (11) and (12), the first $D^{\uparrow}(t)$ arrivals will all have departed by time t. Hence that initial sum of waiting times is dominated by the integral of the queue length, except it may have extra elapsed waiting times of arrivals that came before time 0. The term subtracted on the left is at least that large. Hence we have the first inequality. Second, the term on the right contains the sum of all the waiting times, corresponding to all six regions in Figure 2. Hence we have the second inequality.

We next observe that the initial effect due to the $R(0)$ initial customers in the system at time 0 cannot affect the limit for the averages; after dividing by t, this constant quantity will become asymptotically negligible as $t \to \infty$. Hence, without loss of generality, we omit these terms in the outer terms and assume that $R(0) = 0$.

We then observe that the assumed convergence $\bar{\lambda}(t) \to \lambda$ as $t \to \infty$ for $\bar{\lambda}(t)$ in (1) is equivalent to the convergence, $\bar{\lambda}_n \to 1/\lambda$ as $n \to \infty$ for $\bar{\lambda}_n$ in (2) (see Proposition 2 in §3.2), so that both hold. We then observe that the assumed convergence $\bar{W}_n \to W$ as $n \to \infty$ for $\bar{W}_n$ in (2) implies that $n^{-1}W_n \to 0$ as $n \to \infty$. Since $W_k = D_k - A_k$, $k \geq 1$, we necessarily also have the convergence

$$\bar{\delta}_n \equiv n^{-1}D_n \to 1/\lambda \quad \text{as} \quad n \to \infty. \tag{14}$$

However, by Proposition 1 in §3.2, the convergence in (14) implies the convergence

$$\bar{\delta}_n^{\uparrow} \equiv n^{-1}D_n^{\uparrow} \to 1/\lambda \quad \text{as} \quad n \to \infty. \tag{15}$$

That in turn, by Proposition 2 again, implies the convergence

$$\bar{\delta}^{\uparrow}(t) \equiv t^{-1}D^{\uparrow}(t) \to \lambda \quad \text{as} \quad n \to \infty. \tag{16}$$

Hence, when we rewrite (13) without the terms involving $R(0)$ as

$$\left(\frac{D^{\uparrow}(t)}{t}\right)\left(\frac{\sum_{j=1}^{D^{\uparrow}(t)} W_k}{D^{\uparrow}(t)}\right) \leq \frac{\int_0^t Q(s)\,ds}{t} \leq \left(\frac{A(t)}{t}\right)\left(\frac{\sum_{j=1}^{A(t)} W_k}{A(t)}\right), \quad t \geq 0, \tag{17}$$

we see that the left and right sides both converge to λW as $t \to \infty$. Hence, by a "sandwiching" argument (see part (i) of Proposition 2 for more details about the argument), we necessarily have the convergence

$$\bar{L}(t) \equiv \frac{\int_0^t Q(s)\,ds}{t} \to \lambda W \quad \text{as} \quad t \to \infty. \tag{18}$$

Along the way, we have shown that

$$(\bar{\lambda}_n, \bar{\delta}_n, \bar{\delta}_n^{\uparrow}) \to (1/\lambda, 1/\lambda, 1/\lambda) \quad \text{as} \quad n \to \infty \tag{19}$$

and

$$(\bar{\lambda}(t), \bar{\delta}(t), \bar{\delta}^{\uparrow}(t)) \to (\lambda, \lambda, \lambda) \quad \text{as} \quad t \to \infty. \tag{20}$$

We also easily get $\bar{W}(t) \to W$ as $t \to \infty$ and $\bar{L}_n \to L$ as $n \to \infty$ from the other established results. Hence, the proof is complete. ∎

3.2 Two Supporting Basic Technical Propositions

The proof above depends critically upon two basic technical propositions involving the preservation of convergence of sequences of real numbers under mappings. Everything in this section depends on the mathematical notion of convergence; see Chapter 3 of Rudin [12] for further discussion.

Definition 3.1 (*convergence of a sequence of real numbers*) *For any sequence of real numbers* $\{x_n : n \geq 1\}$, *we say that* x_n *converges to a limit* x *as* $n \to \infty$, *and write* $\lim_{n\to\infty} x_n = x$ *or* $x_n \to x$ *as* $n \to \infty$, *if for all* $\epsilon > 0$ *there exists a positive integer* $n_0 \equiv n_0(\epsilon)$ *depending on* ϵ *such that*

$$|x_n - x| < \epsilon \quad \textit{for all} \quad n \geq n_0.$$

For any sequence of real numbers $\{x_n : n \geq 1\}$, let $\{x_n^{\uparrow} : n \geq 1\}$ be the associated sequence of successive maxima, defined by

$$x_n^{\uparrow} \equiv \max\{x_k : 1 \leq k \leq n\}, \quad n \geq 1. \tag{21}$$

The first proposition says that convergence is "preserved" under the maximum function in (21).

Proposition 1 (*preservation of convergence under a maximum*) *If*

$$\lim_{n\to\infty} \frac{x_n}{n} = x \geq 0, \tag{22}$$

then

$$\lim_{n\to\infty} \frac{x_n^{\uparrow}}{n} = x \geq 0. \tag{23}$$

Proof. This is a variant of Proposition 3.3.1 in [16]. It suffices to prove, under the condition, that for any $\epsilon > 0$ there exists $n_1 \equiv n_1(\epsilon)$ such that $|(x_n^{\uparrow}/n) - x| < \epsilon$ for all $n \geq n_1$. For $\epsilon > 0$ given, the condition implies that there exists $n_0 \equiv n_0(\epsilon)$ such that

$$(x - \epsilon)n \leq x_n \leq (x + \epsilon)n \quad \text{for all} \quad n \geq n_0. \tag{24}$$

Hence,

$$(x - \epsilon)n \leq x_n^{\uparrow} \leq x_{n_0}^{\uparrow} \vee (x + \epsilon)n \quad \text{for all} \quad n \geq n_0, \tag{25}$$

where $a \vee b \equiv \max\{a, b\}$. Now choose $n_1 \geq n_0$ such that $x_{n_0}^{\uparrow} \leq n_1\epsilon$, which implies that $x_{n_0}^{\uparrow} \leq n\epsilon$ for all $n \geq n_1$. Since $x \geq 0$, we can combine this last step with the relations (24) and (25) to get

$$(x - \epsilon)n \leq x_n^{\uparrow} \leq \epsilon n \vee (x + \epsilon)n = (x + \epsilon)n \quad \text{for all} \quad n \geq n_1, \tag{26}$$

as required. ∎

For any sequence of *nondecreasing nonnegative* real numbers $\{x_n : n \geq 1\}$, let $\{c(t) : t \geq 0\}\}$ be the associated counting function, defined by

$$c(t) \equiv \max\{k \geq 0 : x_k \leq t\}, \quad t \geq 0, \tag{27}$$

where $x_0 \equiv 0$. Recall the basic inverse relation for renewal counting processes (or any counting process) given in (7.20) on p. 423 of Ross [11]. We use that relation here:

Lemma 3.1 (*basic inverse relation*) *For any sequence of nondecreasing nonnegative numbers* $\{x_n : n \geq 1\}$,

$$x_n \leq t \quad \textit{if and only if} \quad c(t) > n. \tag{28}$$

The second proposition says that convergence is "preserved" under the inverse function in (27). In fact, there is an equivalence of convergence.

Proposition 2 (*preservation of convergence under inverse*) *For any sequence of nondecreasing nonnegative numbers* $\{x_n : n \geq 1\}$*, there is convergence*

$$\lim_{n\to\infty} \frac{x_n}{n} = x > 0, \tag{29}$$

if and only if there is convergence

$$\lim_{t\to\infty} \frac{c(t)}{t} = 1/x > 0. \tag{30}$$

Proof. This is implied by Theorem 3.4.1 and Corollary 3.4.1 in [16]. We do the two directions in turn.

(i). One direction: (29) implies (30). This direction is a variant of the standard proof of the strong law of large numbers (SLLN) for renewal processes, Proposition 7.1 on p. 728 of Ross [11]. Here is a direct proof: First, observe that the definition of $c(t)$ in (27) implies that

$$x_{c(t)} \leq t < x_{c(t)+1} \quad \text{for all} \quad t > 0. \tag{31}$$

so that, after dividing through by $c(t)$,

$$\frac{x_{c(t)}}{c(t)} \leq \frac{t}{c(t)} < \frac{x_{c(t)+1}}{c(t)} = \left(\frac{x_{c(t)+1}}{c(t)+1}\right)\left(\frac{c(t)+1}{c(t)}\right) \quad \text{for all} \quad t > 0. \tag{32}$$

Since the limit in (29) is assumed to hold, necessarily $c(t) \to \infty$ as $t \to \infty$. Hence $(c(t) + 1)/c(t) \to 1$ as $t \to \infty$ and, by virtue of (29),

$$\frac{x_{c(t)}}{c(t)} \to x \quad \text{and} \quad \frac{x_{c(t)+1}}{c(t)+1} \to x \quad \text{as} \quad t \to \infty. \tag{33}$$

Hence, the lower and upper bounds on $t/c(t)$ in (32) both converge to x as $t \to \infty$. By this "sandwich" argument, we deduce that

$$\lim_{t\to\infty} \frac{t}{c(t)} = x. \tag{34}$$

That can be further justified by applying the limit inferior and limit supremum. From (33), we get

$$x = \lim_{t\to\infty} \frac{x_{c(t)}}{c(t)} \leq \liminf_{t\to\infty} \frac{t}{c(t)} \leq \limsup_{t\to\infty} \frac{t}{c(t)} \leq \lim_{t\to\infty} \frac{x_{c(t)+1}}{c(t)+1} = x \tag{35}$$

and

$$x \leq \liminf_{t\to\infty} \frac{t}{c(t)} \leq \limsup_{t\to\infty} \frac{t}{c(t)} \leq x, \tag{36}$$

which implies (34). In turn (34) implies that

$$\lim_{t\to\infty} \frac{c(t)}{t} = \frac{1}{x}, \tag{37}$$

which completes the proof in one direction: (29) implies (30).

(ii). The other direction: (30) implies (29). Now suppose that (30) holds. Thus, for $\epsilon > 0$ specified, there exists t_0 such that

$$t(x^{-1} - \epsilon) \leq c(t) \leq t(x^{-1} + \epsilon) < t(x^{-1} + 2\epsilon) \quad \text{for all} \quad t \geq t_0. \tag{38}$$

We now use the floor and ceiling functions: $\lfloor x \rfloor$ is the greatest integer less than or equal to x, while $\lceil x \rceil$ is the least integer greater than or equal to x. Using these functions, we have

$$n_1(t) \equiv \lfloor t(x^{-1} - \epsilon) \rfloor \le c(t) < \lceil t(x^{-1} + 2\epsilon) \rceil \equiv n_2(t) \quad \text{for all} \quad t \ge t_0. \tag{39}$$

Then Lemma 3.1 implies that

$$x_{n_1(t)} \le t < x_{n_2(t)} \quad \text{for all} \quad t \ge t_0. \tag{40}$$

Now let $t_1(n)$ and $t_2(n)$ be functions of n defined by

$$t_1(n) \equiv \frac{n}{x^{-1} - \epsilon} \quad \text{and} \quad t_2(n) \equiv \frac{n}{x^{-1} + 2\epsilon} \tag{41}$$

and observe that

$$n_1(t_1(n)) = n_2(t_2(n)) = n \quad \text{for all} \quad n. \tag{42}$$

Hence, for all

$$n \ge n_0 \equiv \lceil t_0(x^{-1} + 2\epsilon) \rceil, \tag{43}$$

we have

$$t_1(n_0) = \frac{\lceil t_0(x^{-1} + 2\epsilon) \rceil}{x^{-1} - \epsilon} > \frac{\lceil t_0(x^{-1} + 2\epsilon) \rceil}{x^{-1} + 2\epsilon} = t_2(n_0) \ge t_0 \tag{44}$$

and, by (40),

$$t_2(n) \le x_{n_2(t_2(n))} = x_n = x_{n_1(t_1(n))} \le t_1(n) \tag{45}$$

or, equivalently,

$$\frac{n}{x^{-1} + 2\epsilon} \le x_n \le \frac{n}{x^{-1} - \epsilon}, \tag{46}$$

so that

$$\frac{1}{x^{-1} + 2\epsilon} \le \frac{x_n}{n} \le \frac{1}{x^{-1} - \epsilon} \tag{47}$$

or

$$x - \frac{2\epsilon x}{1 + 2\epsilon x} \le \frac{x_n}{n} \le x + \frac{\epsilon}{1 - \epsilon x} \tag{48}$$

From (48), we see that, for any $x > 0$ and target error bound δ for $|(x_n/n) - x|$, we can choose $\epsilon \equiv \epsilon(x, \delta)$ suitably small to achieve it. Then, with ϵ specified, we let $t_0 \equiv t_0(\epsilon)$ be as needed to obtain (38) and then we choose $n_0 \equiv n_0(t_0)$ as in (43). That completes the proof. ∎

4 A Statistical Approach

We advocate taking a statistical approach with data over a finite time interval. Thus, in a stationary setting, we regard the finite averages as realizations of random estimators of underlying unknown "true" values L, λ and W. We suggest estimating confidence intervals, just as in steady-state simulation, and discuss how to do so in [5].

In nonstationary settings, we can also use a statistical approach. Since the parameters L, λ and W are no longer defined, we regard the finite averages as estimators of their expected values. Then we may rely on samples from multiple days to provide a basis for estimating these expected values. Again we can estimate confidence intervals. Without stationarity, it is important to consider bias. In [5, 6] we suggest refined estimators to reduce the bias.

Here is the essence of a typical application: We start with the observation of $L(s)$, the number of items in the system at time s, for $0 \le s \le t$. From that sample path, we can directly observe the arrivals (jumps up) and departures (jumps down). Hence, we can easily

estimate the arrival rate λ and the average number in system L. However, based only on the available information, we typically cannot determine the time each item spends in the system, because the items need not depart in the same order that they arrived. Nevertheless, we can estimate the average waiting time by $W = L/\lambda$, using our estimates of L and λ. We can also estimate confidence intervals. See [5, 6] for further discussion.

References

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

[2] Glynn, P. W., W. Whitt. 1986. A central-limit-theorem version of $L = \lambda W$. *Queueing Systems* **1** 191–215.

[3] Hall, R.W. (1991), *Queueing Methods for Services and Manufacturing*, Englewood Cliffs, NJ: Prentice Hall.

[4] Jewell, W. S. 1967. A simple proof of $L = \lambda W$. *Oper Res.* **15** 1109–1116.

[5] Kim, S.-H., W. Whitt. (2013) Statistical Analysis with Little's Law. *Operations Research* **61** 1030–1045.

[6] Kim, S.-H., W. Whitt. (2013a) Estimating Waiting Times with the Time-Varying Little's Law. *Probability in the Engineering and Informational Sciences*, **27** 471–506

[7] Little, J. D. C. 1961. A proof of the queueing formula: $L = \lambda W$. *Oper. Res.* **9** 383–387.

[8] Little, J. D. C. 2011. Little's law as viewed on its 50^{th} anniversary. *Oper. Res.* **59** 536–539.

[9] Mandelbaum, A. 2011. Little's law over a finite horizon. Pages 17.1-17.6 in Teaching notes on Little's law in a course on Service Engineering, October 2011. Available at: http://iew3.technion.ac.il/serveng/Lectures/lectures.html (Accessed August 3, 2012)

[10] Nozari, A. and W. Whitt. 1988. Estimating Average Production Intervals Using Inventory Measurements: Little's Law for Partially Observable Processes. *Oper. Res.* **36** (2) 208-223.

[11] Ross, S. M. (2010), *Introduction to Probability Models*, 10^{th} edition, Academic Press.

[12] Rudin, W. (1976), *Principles of Mathematical Analysis*, 3^{rd} edition, McGraw-Hill.

[13] Stidham, S., Jr. 1974. A last word on $L = \lambda W$. *Oper. Res.* **22** 417–421.

[14] Whitt, W. (1991), A review of $L = \lambda W$. *Queueing Systems* **9** 235–268.

[15] Whitt, W. (1992), Correction note on $L = \lambda W$. *Queueing Systems* **12** 431–432.

[16] Whitt, W. (2002), Preservation of Pointwise Convergence. *Internet Supplement to Stochastic-Process Limits*
Available at: http://www.columbia.edu/∼ww2040/supplement.html

[17] Whitt, W. (2012), Extending the FCLT Version of $L = \lambda W$. *Operations Research Letters* **40** 230–234.