Invited paper

A review of $L = \lambda W$ and extensions

Ward Whitt

AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636, USA

Received 27 August 1990; revised 11 February 1991

A fundamental principle of queueing theory is $L = \lambda W$ (Little's law), which states that the time-average or expected time-stationary number of customers in a system is equal to the product of the arrival rate and the customer-average or expected customer-stationary time each customer spends in the system. This principle is now well known and frequently applied. However, in recent years there have been extensions, such as $H = \lambda G$ and the continuous, distributional, ordinal and central-limit-theorem versions, which show that the $L = \lambda W$ relation, when viewed properly, has much more power than was previously realized. Moreover, connections have been established between $H = \lambda G$ and other fundamental relations, such as the rate conservation law and PASTA (Poisson arrivals see time averages), which show that there is a much greater unity in the overall theory than was previously realized. This paper provides a review.

Keywords: $L = \lambda W$, Little's law, time averages and customer averages, conservation laws, limit theorems, sample-path methods, stationary marked point processes, $H = \lambda G$, rate conservation law, level crossings, the inversion formula, Campbell's formula, central limit theorems, indirect estimation.

1. Introduction

The formula $L = \lambda W$ (Little's law) expresses a fundamental principle of queueing theory: Under very general conditions, the time-average or expected time-stationary number of customers in a system, L (e.g., the average queue length) is equal to the product of the arrival rate λ and the customer-average or expected customer-stationary time each customer spends in the system, W (e.g., the average waiting time). The relation $L = \lambda W$ is very useful because the assumptions are minimal; it applies to other stochastic models besides queues; it applies to queueing networks and subnetworks as well as individual queues; it applies to subclasses as well as the entire customer population; and it is

This paper is dedicated to the memory of our colleague Professor Peter Franken (1937–1989), who contributed greatly to the subject of this paper and to queueing theory more generally.

remarkably independent of modeling details, such as service disciplines and underlying probability distributions. Moreover, there are extensions of $L = \lambda W$, such as $H = \lambda G$ and the continuous, distributional, ordinal and central-limit-theorem (CLT) versions, that enable us to analyze many seemingly unrelated problems.

The purpose of this paper is to review the work on $L = \lambda W$ and its extensions. There are two frameworks for expressing these results. The first is a *deterministic framework* involving averages over individual sample paths. The second is a *stationary framework* involving steady-state distributions. The deterministic framework is appealing because it requires only elementary arguments. It thus lays bare the essential ideas, so that we can quickly understand the principles and focus on their applied significance. The stationary framework is appealing because it leads beyond the particular issue being investigated to a full investigation of the concept of statistical equilibrium or steady state. A primary concern in the stationary framework is the connection between equilibrium at arbitrary times and equilibrium at special random times such as customer arrival epochs. Looking carefully at both frameworks is very useful, because they are closely related. Indeed, there is an equivalence between the relatively elementary sample-path relation $H = \lambda G$ in the deterministic framework and a corresponding statement in the stationary framework obtained via Campbell's formula, so that a statement in one framework can be immediately translated into a statement in the other framework; see remark 6.1.

The fundamental principle is not $L = \lambda W$, but its extension $H = \lambda G$, because $L = \lambda W$ is a special case of $H = \lambda G$ and because $H = \lambda G$ essentially embodies the full relationship between time averages and associated customer averages. Indeed, in the stationary framework, $H = \lambda G$ is essentially equivalent to the basic Palm transformation relating the steady-state distribution at arbitrary times and the steady-state distribution at random times. (We prove this equivalence in a subsequent paper.) Nevertheless, to simplify the discussion, we first consider the more familiar $L = \lambda W$. We present the deterministic framework for $L = \lambda W$ in section 2. We also state the main $L = \lambda W$ result there, which we primarily ascribe to Little [42] and Stidham [64], but also partly to Jewell [33], Newell [53] and Brumelle [8]. In section 3 we describe the steady-state version of $L = \lambda W$ involving stationary marked point processes and the Palm transformation, which we primarily ascribe to Franken [20], but also partly to Miyazawa [47] and Stidham [65–67]. In section 4 we give three illustrative applications of $L = \lambda W$ and in section 5 we discuss the literature related to $L = \lambda W$.

In section 6 we discuss the relation $H = \lambda G$, which states that a more general time-average or expected time-stationary quantity H is equal to the product of the arrival rate λ and an associated customer-average or expected customer-stationary quantity G. We discuss an average version in section 6(1), which we primarily ascribe to Stidham [60,61], Brumelle [8] and Heyman and Stidham [32], and a steady-state version in section 6(2), which we primarily ascribe to Franken

[20] and Stidham [65–67]. In section 6(3) we present an extension of $H = \lambda G$ due to Glynn and Whitt [27] to cover lump costs as well as cost rates. We also discuss the relation between $H = \lambda G$ and other basic principles such as the relation $Y = \lambda X$ of Stidham and El-Taha [68] and the rate conservation law of Miyazawa [49–51]. Recent work of Brémaud [4], Miyazawa [51] and Sigman [59] shows that these important principles as well as others are essentially equivalent. We give a few applications of $H = \lambda G$ in section 6(4), including one to give necessary and sufficient conditions for arrivals to see time averages (ASTA). This result seems to be new as stated, but it is similar to previous results by Stidham [66] and Stidham and El-Taha [68].

In section 7 we discuss indirect estimation via $L = \lambda W$ and $H = \lambda G$, which was initiated by Law [40] and continued by Carson and Law [11] and Glynn and Whitt [26]. The idea is to use a finite average related to W and the known λ to estimate L or vice versa. We also discuss the central-limit-theorem versions of $L = \lambda W$ and $H = \lambda G$ due to Glynn and Whitt [22–24,27], which provides a basis for determining which estimator is more asymptotically efficient.

We conclude in section 8 by discussing other extensions, including the distributional version due to Haji and Newell [30] and Keilson and Servi [34], the continuous (and more general) versions due to Rolski and Stidham [58] and Glynn and Whitt [27] and the ordinal version due to Halfin and Whitt [29].

For previous overviews of $L = \lambda W$ focusing on the deterministic framework, see Stidham [66], Ramalhoto, Amaral and Cochita [55], chapter 11.3 of Heyman and Sobel [31] and chapter 5.2, 5.15 of Wolff [72]. The first six sections here are quite close to Stidham [66]. For previous overviews of other kinds of sample-path analysis, see Stidham [66,67] and Stidham and El-Taha [68]. For previous overviews of the steady-state version of $L = \lambda W$ and the associated marked point process framework, see chapter 4 of Stidham [66], Franken, König, Arndt and Schmidt [21], Rolski [57], Baccelli and Brémaud [1] and chapter 7 of Walrand [69]. For overviews of the closely related ASTA topic, see Brémaud [3], Brémaud, Kannurpatti and Mazumdar [5], Melamed and Whitt [46] and section 3 of and Stidham and El-Taha [68].

2. A deterministic framework for $L = \lambda W$

It is useful and instructive to have a simple general framework for $L = \lambda W$. This is provided by a sequence of ordered pairs of real numbers $\{(A_k, D_k): k \geqslant 1\}$ satisfying $0 \leqslant A_k \leqslant A_{k+1}$ and $A_k \leqslant D_k$ for all k. In applications to probability models, this sequence corresponds to one sample path of a stochastic process; then the inequalities above and the limits discussed below are understood to hold on this particular sample path.

In applications to queues, we usually interpret A_k and D_k as the arrival and departure epochs of the kth arriving customer. We think of the kth customer being in the system during the interval $[A_k, D_k]$. However, arrival and departure should be interpreted with respect to the system under consideration. For example, if the system refers to a queue, excluding the servers, then A_k is the epoch when the kth customer to join the queue arrives (not counting customers who do not join the queue) and D_k is the epoch when the kth customer to join the queue leaves the queue (which usually occurs when the customer begins service). For many models, such as a single-server queue with the FCFS (first-come first-served) discipline, we also have $D_k \leqslant D_{k+1}$ for all k, but we do not make this assumption. For queueing applications, we think of the system as being initially empty, but other initial conditions can be introduced by setting $A_k = 0$, $1 \leqslant k \leqslant m$, for some m.

We now define associated quantities in terms of the basic sequence $\{(A_k, D_k)\colon k \geqslant 1\}$. For each $t \geqslant 0$, let $Q(t)$ be the number of k with $A_k \leqslant t \leqslant D_k$ (the number in system or queue length) and, for each $k \geqslant 1$, let $W_k = D_k - A_k$ (the time spent in the system or waiting time). Let $A(t)$ and $D(t)$ count the number of k such that A_k and D_k, respectively, are less than or equal to t.

In the context of queueing models, it is significant that we have not specified all the standard features of the model. For example, there is no mention of the service discipline. Thus, any service discipline is allowed (subject to the customer being in the system throughout the interval $[A_k, D_k]$, but this can be generalized using the extension to $H = \lambda G$; see section 6(1) below). Moreover, we do not specify customer service times (as distinct from the total time spent in the system) as part of the primitive model data. As a consequence, there are many different detailed queueing models consistent with our model. For example, *our model can always be interpreted as a standard infinite-server queueing model in which customers enter service immediately upon arrival*; simply interpret W_k as the service time of the kth customer.

In applications we are usually willing to assume that all relevant limits exist. From an applied point of view, the main issue in $L = \lambda W$ is the relation among the limits assuming that they exist. Hence, our first result is stated with an extra existence assumption. (The extra assumption below that $t^{-1}D(t) \to \lambda$ as $t \to \infty$ is not restrictive; it can be established in the proof.) A good understanding of $L = \lambda W$ and its extensions can be obtained from the simple proof here.

THEOREM 2.1

Suppose that $t^{-1}A(t) \to \lambda$ and $t^{-1}D(t) \to \lambda$ as $t \to \infty$, where $0 < \lambda < \infty$. If $k^{-1}\Sigma_{j=1}^{k} W_j \to W$ as $k \to \infty$, then $t^{-1}\int_0^t Q(s)\,\mathrm{d}s \to L$ as $t \to \infty$ and $L = \lambda W$.

Proof (sketch)

The key observation is that the integral $\int_0^t Q(s)\,\mathrm{d}s$ coincides with the sum of the waiting times of the $D(t)$ customers to depart by time t plus a portion of the sum of the waiting times of the $A(t) - D(t)$ customers who arrive by t but have

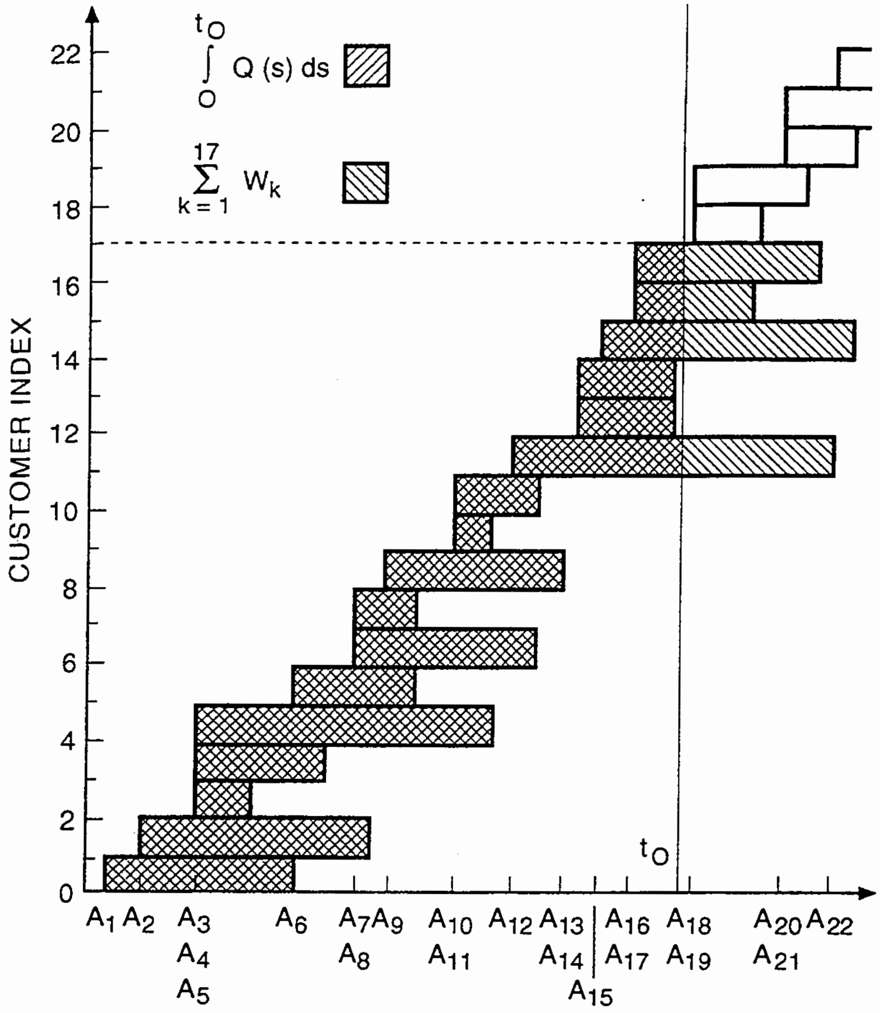

Fig. 1. The cumulative processes associated with $Q(t)$ and W_k.

not yet departed (see fig. 1), i.e., we start by establishing the relation

$$\sum_{j=1}^{D(t)} W_j \leqslant \int_0^t Q(s)\,\mathrm{d}s \leqslant \sum_{j=1}^{A(t)} W_j, \quad t \geqslant 0. \tag{2.1}$$

In fig. 1, time is on the horizontal axis and the customer index is on the vertical axis. A bar at $[k-1, k] \times [A_k, D_k]$ appears for customer k, so that W_k measure the length of the kth bar and $Q(t)$ counts the number of bars intersecting the vertical line at t. Both cumulative processes $\Sigma_{k=1}^{n} W_k$ and $\int_0^t Q(s)\,\mathrm{d}s$ thus measure the area of a set of bars and partial bars, from which (2.1) is evident.

To justify (2.1) more formally, let $I_k(t)$ be the indicator function of the interval $[A_k, D_k]$ and note that

$$W_k = \int_0^\infty I_k(t)\,\mathrm{d}t \quad \text{and} \quad Q(t) = \sum_{k=1}^\infty I_k(t). \tag{2.2}$$

Then, for example,

$$\sum_{k=1}^{D(t)} W_k = \sum_{k=1}^{D(t)} \int_0^\infty I_k(s)\,\mathrm{d}s = \sum_{k=1}^{D(t)} \int_0^t I_k(s)\,\mathrm{d}s = \int_0^t \left[\sum_{k=1}^{D(t)} I_k(s)\right] \mathrm{d}s$$
$$\leqslant \int_0^t \left[\sum_{k=1}^{\infty} I_k(s)\right] \mathrm{d}s = \int_0^t Q(s)\,\mathrm{d}s.$$

The rest is easy; e.g.,

$$t^{-1} \sum_{j=1}^{A(t)} W_j = t^{-1} A(t) (A(t))^{-1} \sum_{j=1}^{A(t)} W_j \to \lambda W \quad \text{as } t \to \infty \tag{2.3}$$

when

$$t^{-1} A(t) \to \lambda \text{ as } t \to \infty \quad \text{and} \quad k^{-1} \sum_{j=1}^{k} W_j \to W \text{ as } k \to \infty. \quad \square$$

Remark

(2.1) The content of $L = \lambda W$ is probably embodied in fig. 1 and (2.1) as much as in the statement of theorem 2.1. For example, from (2.1), we immediately see that

$$\int_0^t Q(s)\,\mathrm{d}s = \sum_{k=1}^{A(t)} W_k$$

whenever $Q(t) = A(t) - D(t) = 0$. □

The following is a mathematically more elegant (but somewhat harder to prove) version. It is a minor variant of what is in Stidham [64,66], chapter 11.3 of Heyman and Sobel [31] and chapter 5.2, 5.15 of Wolff [72].

THEOREM 2.2

Suppose that $k^{-1} A_k \to \lambda^{-1}$ as $k \to \infty$, where $0 < \lambda^{-1} < \infty$. Then

$$k^{-1} \sum_{j=1}^{k} W_j \to W \quad \text{as } k \to \infty, \tag{2.4}$$

where $W < \infty$, if and only if

$$k^{-1} W_k \to 0 \text{ as } k \to \infty \quad \text{and} \quad t^{-1} \int_0^t Q(s)\,\mathrm{d}s \to L \text{ as } t \to \infty, \tag{2.5}$$

where $L < \infty$. If (2.4) and (2.5) hold, then $L = \lambda W$.

Proof (sketch)

As in theorem 2.1, the key is (2.1) and the related relation

$$\int_0^{A_k} Q(s)\,\mathrm{d}s \leqslant \sum_{j=1}^{k} W_j \leqslant \int_0^{D_k'} Q(s)\,\mathrm{d}s, \quad k \geqslant 1, \tag{2.6}$$

where $D_k' = \max\{D_j\colon 1 \leqslant j \leqslant k\}$. Next we note that $t^{-1}A(t) \to \lambda$ as $t \to \infty$ since $k^{-1}A_k \to \lambda^{-1}$ as $k \to \infty$. (The two limits are easily seen to be equivalent.) We then show, given that $k^{-1}A_k \to \lambda^{-1}$ as $k \to \infty$, that the following are equivalent: $k^{-1}D_k \to \lambda^{-1}$ as $k \to \infty$, $k^{-1}W_k \to 0$ as $k \to \infty$ and $k^{-1}D_k' \to \lambda^{-1}$ as $k \to \infty$. We then show that if $k^{-1}D_k \to \lambda^{-1}$ as $k \to \infty$, then $t^{-1}D(t) \to \lambda$ as $t \to \infty$. (The two limits are not equivalent; see theorem 2 of [22].) The rest is as in (2.3). □

Remarks

(2.2) As far as the minor issue of existence of limits is concerned, interest centers on the asymmetry ((2.5) does not imply (2.4) if we remove the limit on $k^{-1}W_k$ from (2.5)). Stated differently, (2.4) implies that $k^{-1}W_k \to 0$ as $k \to \infty$, but $t^{-1}\int_0^t Q(s)\,\mathrm{d}s \to L$ as $t \to \infty$ does not imply that $t^{-1}Q(t) \to 0$ as $t \to \infty$. The asymmetry was evidently discovered by Brumelle [8] and was clearly expressed by Stidham [64]. The implication (2.5) → (2.4) follows from lemma 2.2 and theorem 2.4 of Stidham [66]. A few additional results appear in theorem 2 of Glynn and Whitt [22] and theorem 5 of Glynn and Whitt [27].

(2.3) If we assume that $k^{-1}W_k \to 0$ as $k \to \infty$, then theorem 2.2 remains valid with $W = \infty$ and $L = \infty$. □

Theorems 2.1 and 2.2 immediately imply corollaries for stochastic models when we introduce probability structure and associate "with probability one (w.p.1)" with all statements. To state this corollary to theorem 2.2, now suppose that $\{(A_k, D_k)\colon k \geqslant 1\}$ is a random sequence having the specified properties w.p.1. Then $Q(t)$, λ^{-1}, L and W become random variables. Of course, typically λ^{-1}, L and W are deterministic, but that is not necessary.

COROLLARY

Suppose that $k^{-1}A_k \to \lambda^{-1}$ as $k \to \infty$ w.p.1, where $P(0 < \lambda^{-1} < \infty) = 1$. Then

$$k^{-1}\sum_{j=1}^{k} W_j \to W \quad \text{as} \quad k \to \infty \text{ w.p.1},$$

where $P(W < \infty) = 1$, if and only if

$$k^{-1}W_k \to 0 \text{ as } k \to \infty \quad \text{and} \quad t^{-1}\int_0^t Q(s)\,\mathrm{d}s \to L \text{ as } t \to \infty \text{ w.p.1},$$

where $P(L < \infty) = 1$. If the w.p.1 limits hold, then $P(L = \lambda W) = 1$.

3. A stationary framework for steady-state quantities

In applications to probability models, we are often interested in steady-state means. Theorems 2.1 and 2.2 typically apply because in great generality L coincides with the expected steady-state (continuous-time) number of customers in the system, while W coincides with the expected steady-state (discrete-time) length of time each customer spends in the system.

To be more precise, let $\Rightarrow$ denote convergence in distribution and let $\{(A_k, W_k)\colon k \geqslant 1\}$ and $\{Q(t)\colon t \geqslant 0\}$ be stochastic processes. If we are thinking about steady-state, then we typically are willing to assume that

$$Q(t) \Rightarrow Q(\infty) \text{ as } t \to \infty \quad \text{and} \quad W_k \Rightarrow W(\infty) \text{ as } k \to \infty \tag{3.1}$$

as well as

$$t^{-1}\int_0^t Q(s)\,\mathrm{d}s \to EQ(\infty) \text{ as } t \to \infty \quad \text{and}$$
$$k^{-1}\sum_{j=1}^{k} W_j \to EW(\infty) \text{ as } k \to \infty \text{ w.p.1.} \tag{3.2}$$

Under (3.1) and (3.2), if $k^{-1}A_k \to \lambda^{-1}$ w.p.1, then the corollary to theorem 2.2 implies that

$$L = EQ(\infty) = \lambda EW(\infty) = \lambda W. \tag{3.3}$$

For practical purposes, we are usually willing to assume (3.1) and (3.2), so that theorem 2.2 captures the steady-state version of $L = \lambda W$ in (3.3) as well as the sample-path-average version in section 2. However, it is also instructive to directly construct appropriate stationary versions of the stochastic processes and then obtain the steady-state version of $L = \lambda W$ in (3.3). When this approach is followed, (3.3) is only one of many consequences.

We now present a steady-state version of $L = \lambda W$, drawing on the theory of stationary marked point processes. (The rest of this section and section 6(2) are somewhat more technical than the rest of this paper; they are not essential for reading the rest of this paper.) For more details, see Franken et al. [21], Rolski [57] and Baccelli and Brémaud [1]. Our account is similar to the introduction in chapter 7 of Walrand [69]. However, there is an important difference. The stationary-process literature presents the steady-state version of $L = \lambda W$ in the context of one special model, in particular, the standard $G/G/s/\infty$ model with s servers working in parallel and the FCFS service discipline, whereas we (following Stidham [65–67]) present a steady-state version of $L = \lambda W$ in the natural stationary-process analog of the much more general framework in section 2.

The reason for the more restrictive framework in the stationary-process literature is that the primary concern there is the construction of all stochastic processes of interest in terms of the primitive data of the model, which consist of the arrival times and service times as well as the service discipline and related detailed full specifications of system operation. In contrast, as in section 2, we assume that we are given only a partial specification of the system, in particular, only the arrival and departure times (or, equivalently, the interarrival times and waiting times) with appropriate stationary structure. The stationary-process literature indicates how to construct the process we start with in the standard $G/G/s$ case. Moreover, as we observed in section 2, the $G/G/s$ results in the stationary-process literature apply to our model, because our model can always be interpreted as a $G/G/\infty$ model. Thus what we present, both here and in section 6(2), is an easy consequence of Franken [20] and Franken et al. [21].

We assume that we have a strictly stationary stochastic sequence $\{(A_k - A_{k-1}, W_k)\colon -\infty < k < \infty\}$ with $A_0 = 0$, $A_k \leqslant A_{k+1}$ and $W_k \geqslant 0$ w.p.1 for all k; i.e., the distribution of the sequence $\{(A_{j+k} - A_{j+k-1}, W_{j+k})\colon -\infty < k < \infty\}$ is independent of $j \geqslant 1$. (We interpret A_k as the arrival epoch of customer k and W_k as the time he spends in the system.) From this, we obtain the customer-stationary means, i.e., $\lambda^{-1} = E(A_k - A_{k-1})$ and $W = EW_k$. We assume that $0 < \lambda^{-1} < \infty$. Then $\{(A_k, W_k)\colon -\infty < k < \infty\}$ is called a synchronous (or customary-stationary) stationary marked point process with A_k being the kth point and W_k the kth mark.

We now indicate how to construct the associated time-stationary marked point process, say $\{(A'_k, W'_k)\colon -\infty < k < \infty\}$, associated with $\{(A_k, W_k)\colon -\infty < k < \infty\}$. It is important to note that this is not a sample-path construction. We construct the probability law of $\{(A'_k, W'_k)\}$ in terms of the probability law of $\{(A_k, W_k)\}$. For this purpose, we assume that the point process is *simple*, i.e., $P(A_k - A_{k-1} > 0) = 1$. Of course, this condition makes this section less general than section 2. Miyazawa [51] shows how to treat multiple points.

Let $\overset{\mathrm{d}}{=}$ denote equality in distribution. Let a and b be positive numbers. We stipulate that

$$(\{(A'_k, W'_k)\} \mid A'_0 = -a, A'_1 = b) \overset{\mathrm{d}}{=} (\{(A_k - a, W_k)\} \mid A_1 = a + b), \tag{3.4}$$

$$P(A'_1 - A'_0 \leqslant t) = \frac{1}{EA_1}\left[tP(A_1 \leqslant t) - \int_0^t P(A_1 \leqslant u)\,\mathrm{d}u\right], \quad t \geqslant 0, \tag{3.5}$$

and

$$P(-A'_0 \leqslant s \mid A'_1 - A'_0 = t) = \frac{s}{t}, \quad 0 \leqslant s \leqslant t. \tag{3.6}$$

It is useful to think of constructing $\{(A'_k, W'_k)\}$ from $\{(A_k, W_k)\}$ by first shifting all the points A_k to the left by a common amount, so that $A'_0 < 0 < A'_1$. (We use the assumption that the point process is simple here.) Then the conditional distribution of $\{(A'_k, W'_k)\}$ given A'_0 and A'_1 is expressed in terms of the conditional distribution of the shifted sequence $\{(A_k - a, W_k)\}$ given the lengths of the first interval, as in (3.4). The length $A'_1 - A'_0$ of the interval covering the origin is then given the equilibrium spread distribution associated with A_1 as in (3.5), which has the familiar length-bias density $tf(t)/EA_1 = \lambda tf(t)$ if $P(A_1 \leqslant t)$ has a density $f(t)$; e.g., see p. 66 of Wolff [72]. The conditional distribution of $-A'_0$ given $A'_1 - A'_0$ is then uniform in the available interval.

To express this relationship more concisely, let (A, W) represent the sequence $\{(A_k, W_k)\}$ in $(\mathbb{R}^2)^\infty$ and similarly for (A', W'). Moreover, let $(A, W) + t$ represent $\{(A_k - t, W_k)\}$, corresponding to moving the time origin to t. Then (3.4)–(3.6) leads to the *inversion formula*

$$\begin{aligned} Ef(A', W') &= \int_0^\infty \int_0^t E[f(A', W') \mid A'_1 - A'_0 = t,\ A'_0 = -s] \\ &\qquad \times P(A'_1 - A'_0 \in \mathrm{d}t,\ -A'_0 \in \mathrm{d}s) \\ &= \int_0^\infty \int_0^t E[f((A, W) + s) \mid A_1 = t] P(A'_1 - A'_0 \in \mathrm{d}t,\ -A'_0 \in \mathrm{d}s) \\ &= \int_0^\infty \int_0^t E[f((A, W) + s) \mid A_1 = t] \lambda \ \mathrm{d}s P(A_1 \in \mathrm{d}t) \\ &= \lambda E \int_0^{A_1} f((A, W) + s) \ \mathrm{d}s \end{aligned} \tag{3.7}$$

for all nonnegative measurable real-valued functions f on $(\mathbb{R}^2)^\infty$.

Upon reflection, it should be intuitively clear that $(A', W') \equiv \{(A'_k, W'_k)\}$ so constructed is a stationary stochastic process in the sense that its distribution is independent of shifts by time t, and this is easily verified.

LEMMA 3.1

If f is bounded, then for each $t > 0$,

$$Ef((A', W') + t) = Ef(A', W').$$

Proof

By the assumed stationarity of $\{(A_k - A_{k-1}, W_k)\}$, the interval of integration $[0, A_1]$ in (3.7) can be replaced by $[A_k, A_{k+1}]$ for any k. Hence, by (3.7),

$$\begin{aligned} Ef((A', W') + t) &= \frac{\lambda}{k} E \int_0^{A_k} f((A, W) + s + t) \ \mathrm{d}s \\ &= \frac{\lambda}{k} E \int_t^{A_k + t} f((A, W) + s) \ \mathrm{d}s, \end{aligned}$$

so that, if f is bounded by M, then

$$| Ef((A', W') + t) - Ef(A', W') | \leqslant \frac{\lambda}{k} 2tM. \tag{3.8}$$

Letting $k \to \infty$ in (3.8) completes the proof. □

Remark

(3.1) The standard time-stationary marked point process is (A', W') above, while the associated synchronous or Palm version is (A, W). In the literature it is customary to characterize these processes, not in terms of the random variables of random sequences, but in terms of the underlying probability measures on a common measurable space, say $(\Omega, \mathscr{F})$. The probability measure on $(\Omega, \mathscr{F})$ supporting (A', W') is denoted by P, while the probability measure supporting (A, W) is denoted by P^0 and called the Palm probability measure. For our problem, we may take $\Omega = (\mathbb{R}^2)^\infty$ and let $\mathscr{F}$ be its associated Borel σ-field (using the product topology). If we use the identity map from $(\mathbb{R}^2)^\infty$ to $(\mathbb{R}^2)^\infty$ to map into (A, W), then the Palm probability measure P^0 coincides with the probability law of (A, W). Then probability measure P can be defined on the same underlying measurable space in terms of P^0 according to (3.4)–(3.6). The image of the identity map on $(\mathbb{R}^2)^\infty$ with P then has the same probability law as (A', W'). With this construction, (A, W) and (A', W') can be regarded as a single function on an underlying measurable space with respect to two different probability measures, P^0 and P. Expectations of functionals can then be computed with respect to either probability measure, and it is customary to use the notation E^0 and E, respectively. This discussion shows that what we have done is no less general. □

The associated time-stationary queue-length process $\{Q'(t): -\infty < t < \infty\}$ is defined in terms of $\{(A'_k, W'_k)\}$ by

$$Q'(t) = \sum_{k=-\infty}^{k=\infty} \mathbf{1}_{\{A'_k \leqslant t \leqslant A'_k + W'_k\}}, \quad -\infty < t < \infty. \tag{3.9}$$

Since $\{(A'_k, W'_k)\}$ is time-stationary, $\{Q'(t): -\infty < t < \infty\}$ is a stationary process and it suffices to evaluate $EQ'(0)$.

The key to relating $EQ'(0)$ to λEW_1 is Campbell's theorem, which we now state in a simple form suitable for our application; see p. 229 of Walrand [69], p. 11 of Baccelli and Brémaud [1] and p. 20 of Franken et al. [21]. Campbell's theorem is a special case of Mecke's theorem or the generalized Campbell theorem, which gives an expression for the law of (A, W) in terms of the law of (A', W').

THEOREM 3.1

For any nonnegative measurable function g on $\mathbb{R}^2$,

$$E\left[\sum_{k=-\infty}^{\infty} g(A'_k, W'_k)\right] = \lambda \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(a, w) P(W_0 \in \mathrm{d}w)\, \mathrm{d}a. \tag{3.10}$$

Proof (sketch)

First consider functions g that are products of indicator functions, i.e., $g(a, w) = \mathbf{1}_{B_1}(a)\mathbf{1}_{B_2}(w)$. Afterwards extend to general g by the monotone convergence theorem. For g of this indicator form, the left hand side of (3.10) corresponds to the expected number of pairs (A'_k, W'_k) such that $A'_k \in B_1$ and $W'_k \in B_2$, which can easily be shown to be λ times the Lebesgue measure of B_1 times $P(W_1 \in B_2)$; see pp. 23, 107 of [21]. This in turn coincides with the right hand side of (3.10). □

We now apply theorem 3.1 to establish the steady-state version of $L = \lambda W$.

THEOREM 3.2

If $\{(A_k - A_{k-1}, W_k)\colon -\infty < k < \infty\}$ is a stationary sequence with $A_0 = 0$ and $EA_1 = \lambda^{-1}$, $0 < \lambda^{-1} < \infty$, then $\{Q'(t)\colon -\infty < t < \infty\}$ defined by (3.4)–((3.6) and (3.9)) is a stationary process with $EQ'(0) = \lambda EW_0$.

Proof

In (3.10) let $g(a, w)$ be the indicator function of $\{a \leqslant 0 \leqslant a + w\}$. By (3.9) and (3.10),

$$\begin{aligned} EQ'(0) &= \lambda \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(a, w) P(W_0 \in \mathrm{d}w)\, \mathrm{d}a \\ &= \lambda \int_{-\infty}^{0} \int_{-a}^{\infty} P(W_0 \in \mathrm{d}w)\, \mathrm{d}a \\ &= \lambda \int_{0}^{\infty} P(W_0 \geqslant a)\, \mathrm{d}a = \lambda EW_0. \quad \square \end{aligned}$$

Remark

(3.2) We can also obtain theorem 3.2 via theorem 2.2 and ergodic theorems, if we assume ergodicity. First, the stationary marked point process is ergodic if and only if the synchronous marked point process is: p. 28 of [1] and p. 35 of [21]. Moreover, the limits are then the same; e.g.,

$$k^{-1}A_k \to \lambda^{-1} \quad \text{and} \quad k^{-1} \sum_{j=1}^{k} W_j \to W \quad \text{w.p.1 as } k \to \infty$$

if and only if

$$k^{-1}A'_k \to \lambda^{-1} \quad \text{and} \quad k^{-1} \sum_{j=1}^{k} W'_j \to W \quad \text{w.p.1 as } k \to \infty,$$

and

$$t^{-1}\int_0^t Q(s)\, \mathrm{d}s \to L \text{ as } t \to \infty \quad \text{if and only if} \quad t^{-1}\int_0^t Q'(s)\, \mathrm{d}s \to L \text{ as } t \to \infty;$$

i.e., we have the cross-ergodic theorems, p. 29 of [1] and p. 36 of [21]. Hence, in the presence of stationarity and ergodicity, theorem 3.2 is equivalent to theorem

2.2. Of course, the real advantage of theorem 2.2 is that we can obtain $L = \lambda W$ for averages without discussing stationarity. On the other hand, if we want to discuss steady-state quantities, then we need to know that they are well defined, so that the stationary framework is an appropriate general framework (which does not require ergodicity).

4. Applications of $L = \lambda W$

There are two rather distinct ways to apply $L = \lambda W$ in the deterministic framework: first, to calculate one of the three limiting averages L, λ or W given that all the limits exist and two of them are known and, second, to prove the existence of one limit given that the existence of the other two limits has been established. $L = \lambda W$ seems to be much more useful (as well as more interesting) as an aid in calculation than it is an aid in proving existence. Typically, if we can establish w.p.1 convergence for either the time average to L or the customer average to W, then by a minor modification of the same argument we can directly establish the w.p.1 convergence of the other average. However, even for proving existence, theorem 2.2 can simplify matters a little.

Example 4.1. The M / G / 1 queue with a non-FIFO service discipline

Consider a standard $M/G/1$ queue with a non-FIFO (first in, first out) service discipline (rule for selecting which waiting customer starts service next), such as LIFO (last in, first out) or ROS (random order of service), for which the distribution of the queue length process is the same as it is for FIFO. The steady-state queue length distribution and its mean are thus well known from the FIFO theory, but the steady-state waiting time distribution and its mean are in general hard to calculate directly. Thus, we can apply $L = \lambda W$ to calculate the long-run average waiting time W.

In particular, we can apply the corollary to theorem 2.2. However, to verify existence of the limit, we must show that $k^{-1}W_k \to 0$ w.p.1. For this step, let B_k be the length of the kth busy period and let N_k be the index of the busy period in which the kth customer is served. Note that B_k and N_k are the same as for the FIFO discipline. Since the successive busy periods are i.i.d. with finite mean, obviously $k^{-1}B_k \to 0$ as $k \to \infty$ w.p.1. Hence,

$$\frac{W_k}{k} \leqslant \frac{B_{N_k}}{k} \leqslant \frac{B_{N_k}}{N_k} \to 0 \quad \text{as } k \to \infty \text{ w.p.1.} \tag{4.1}$$

Example 4.2. Sojourn times in a Jackson network

Consider an open Markovian Jackson queueing network with a product-form steady-state distribution for the number of customers at each queue. In general, the expected equilibrium sojourn time for a customer in the network is difficult to calculate directly, so that $L = \lambda W$ is very handy. As with example 4.1, we can use (4.1) and regenerative structure to establish the extra condition in the corollary to theorem 2.2. Here B_k is the kth busy period for the whole network, which is the interval during which there is at least one customer in the network.

Example 4.3. Conservation of load in a loss model

A nice application of $L = \lambda W$ is to characterize the blocking probability in a model with finite waiting space. Assume that customers arrive at a service facility at rate α and that a fraction β of them are blocked (do not enter the system). Assume that each customer who is not blocked first waits in a waiting space (perhaps for zero time) and then enters a service mechanism to receive service, which continues uninterrupted until service is complete. Assume that there is a bound on the number of customers that can be waiting. (Any arrivals when the waiting space is full are necessarily blocked). Let the average time spent in service per customer be τ and let ν be the long-run time-average number of customers in service. (For example, if we are considering a conventional $G/G/s/r$ model with s identical servers in parallel, r extra waiting spaces and the first-come first-served discipline, then τ is the long-run average service time per customer and ν is the long-run time-average number of busy servers.)

From $L = \lambda W$, we easily obtain

$$\beta = 1 - \frac{\nu}{\alpha\tau}. \tag{4.2}$$

Just let the system be the service mechanism, so that $L = \nu$, $\lambda = \alpha(1 - \beta)$ and $W = \tau$. (Note that we have assumed all limits exist.) Since the waiting space is bounded, the rate of arrivals to the service mechanism is the same as the flow rate into the system, i.e., $\alpha(1 - \beta)$. Hence, any three components of the vector $(\alpha, \beta, \tau, \nu)$ determine the fourth.

It is interesting that (4.2) was established directly by Rice [56] using a sample-path argument much like the proof of theorem 2.1. For other applications of $L = \lambda W$, see Heyman and Sobel [31] and Wolff [72].

5. Historical commentary

The relation $L = \lambda W$ is one of many fundamental conservation laws, like the conservation of mass, which help us understand the physical world; e.g., see Krakowski [39]. The relation $L = \lambda W$ may seem so obvious that a proof hardly seems worthwhile. Indeed, $L = \lambda W$ was evidently applied frequently without proof. (An early example cited by Maxwell [43] is Cobham [13].) However, we should recall how the early Greeks developed new levels of understanding (and the mathematical method) by proving "obvious" things, e.g., that a circle is bisected by any diameter; see chapter 1 of Eves and Newsom [18].

The first fundamental insight, evidently due to Morse [52], pp. 22, 75, was the recognition that it would be desirable to have a general proof of $L=\lambda W$. A fairly general proof was then proposed by Morse's student Little [42], in what has become one of the most frequently referenced papers in applied probability and operations research.

Given this celebrity, it is interesting that Little's [42] version of $L=\lambda W$ actually has a serious flaw, which was first noted by Brumelle [8], p. 511; see also p. 140 of [21]. In particular, Little's [42] assumptions are much more restrictive than they first appear, because they require that all the processes be defined on the same sample space and be simultaneously stationary. In general, stationary versions of the discrete-time process $\{(A_k - A_{k-1}, W_k)\colon k \geqslant 1\}$ and the continuous-time process $\{Q(t)\colon t \geqslant 0\}$ need to be related by the Palm transformation as in (3.4)–(3.7) and (3.9). (Little [42] worked with (3.9), but omitting the primes on A_k and W_k.) However, Brumelle [8] showed how this difficulty can be circumvented by not assuming that the continuous-time process $\{Q(t)\colon t \geqslant 0\}$ is stationary.

Franken [20] then established the stationary-process version for the standard $G/G/s$ model by exploiting the Palm transformation and applying Campbell's theorem; see (4.2.2) on p. 137 of [21]. Closely related results were obtained by Miyazawa [47,48]. As noted in section 3, if we apply the $G/G/\infty$ representation of our model in section 2, then Franken's [20,21] results for $G/G/s$ apply to it, thus yielding theorem 3.2. With this interpretation, we evidently have a proper treatment of what Little [42] intended.

The stationary-process framework for $L=\lambda W$ proposed by Little [42] was not entirely satisfactory, however, not only because it was not quite correct, but also because it requires steady-state conditions and is technically complicated. This motivated Jewell [33] to establish a version of $L=\lambda W$ based on regenerative structure. The key idea here is that $\int_0^t Q(s)\,\mathrm{d}s = \sum_{k=1}^{A(t)} W_k$ whenever $Q(t)=0$, as noted in remark 2.1. The result then follows from (2.3). Jewell's [33] version is appealing because it does not require steady-state conditions, it applies to most queueing models of interest, and it is easier to understand. It also was the first correct proof. Examples 4.1 and 4.2 above illustrate that repeated empty epochs often help establish the conditions for existence.

However, to many researchers, it appeared that stochastic conditions are not really essential. Indeed, one immediately gets this impression from reading Little [42], but toward the end (see p. 386) his proof actually relies on the special stochastic assumptions. Quick sample-path arguments were proposed by Eilon [16] and Maxwell [43], but these did not address the key issues. A sample-path analog of Jewell's [33] regenerative proof was given by Newell [53]. A truly satisfactory sample-path proof of $L=\lambda W$ was finally obtained by Stidham [64]. More complicated sample-path proofs also follow from Brumelle [8] and Stidham [62].

Closely related to the sample-path approach to $L=\lambda W$ is the operational analysis of Buzen [10] and Denning and Buzen [15], which has played an important role in computer system performance analysis; e.g., see chapter 3 of Lazowska, Zahorjan, Graham and Sevcik [41]. Operational analysis provides an elementary finite-time analog of $L = \lambda W$ as well as other results. The finite-time version of $L = \lambda W$ is primarily a rediscovery of the fact that, for each sample path, $\int_0^t Q(s)\,\mathrm{d}s = \Sigma_{k=1}^{A(t)} W_k$ whenever $Q(t) = 0$, which was exploited earlier by Jewell [33] and others. Given a t for which $Q(t) = 0$, we obtain the finite-time version of $L = \lambda W$ by simply *defining* L, λ and W to be $t^{-1}\int_0^t Q(s)\,\mathrm{d}s$, $t^{-1}A(t)$ and $A(t)^{-1}\Sigma_{k=1}^{A(t)} W_k$. If $Q(t) \neq 0$, then we define W to be L/λ with L and λ defined as above. This finite-time result is remarkably easy to understand (there need be no limits or stochastic processes), as is illustrated by chapter 3 of Lazowska et al. [41]. (The brief remark on p. 43 there is perhaps sufficient warning about the approach.) Thus operational analysis has helped make $L = \lambda W$ well known and more frequently applied. However, the finite-time "measurement" version leaves open the question of how the finite-time averages are related to the limits (e.g., prediction).

6. The relation $H = \lambda G$

The well known relation $L = \lambda W$ which we have been discussing is actually a special case of a much more fundamental and far reaching relation, $H = \lambda G$. We develop $H = \lambda G$ in the deterministic and stationary frameworks in section 6(1) and 6(2), respectively. We use the same notation to emphasize the fact that a statement in one framework can immediately be translated into a statement in the other framework.

In section 6(3) we present a more general version of $H = \lambda G$ due to Glynn and Whitt [27] to cover lump costs as well as cost rates. We then use this version of $H = \lambda G$ to derive related conservation laws. In section 6(4) we discuss a few applications of $H = \lambda G$.

(1) THE DETERMINISTIC FRAMEWORK

Motivated by design problems for queues with nonlinear waiting costs, Stidham [60,61] evidently first noted that the relation $L = \lambda W$ can be extended to relate a general time average H to an associated customer average G. The most significant developments toward a general theory were provided by Brumelle [8,9], based on his thesis with Wolff, and Heyman and Stidham [32], with other contributions by Stidham [60,61,63,66] and Maxwell [43].

To state an $H = \lambda G$ analog of theorem 2.2, we augment the model of section 2 by a sequence of nonnegative integrable real-valued functions $\{f_k(t)\colon t \geqslant 0\}\colon k \geqslant 1\}$, with $f_k(t)$ being interpreted as a cost rate associated with customer k at time t. As illustrated by the applications of $H = \lambda G$, the functions $f_k(t)$ allow us to incorporate additional model details, such as the service times. The following is a slight extension of the version of $H = \lambda G$ in Heyman and Stidham [32]. It follows from Glynn and Whitt [27] (and theorem 6.3 below). For related results (e.g., when $f_k(t)$ is not assumed to be nonnegative or 0 outside a finite interval), see chapter 3 of Stidham [66].

THEOREM 6.1

Suppose that

$$k^{-1}A_k \to \lambda^{-1} \quad \text{and} \quad k^{-1}D_k \to \lambda^{-1} \quad \text{as } k \to \infty, \tag{6.1}$$

where $0 < \lambda^{-1} < \infty$ and, for each $k \geqslant 1$,

$$f_k(t) = 0 \quad \text{if } t \notin [A_k, D_k]. \tag{6.2}$$

Then

$$G_k \equiv k^{-1} \sum_{j=1}^{k} \int_0^\infty f_j(t)\, \mathrm{d}t \to G \quad \text{as } k \to \infty \tag{6.3}$$

if and only if

$$H_t \equiv t^{-1} \int_0^t \sum_{j=1}^{\infty} f_j(s)\, \mathrm{d}s \to H \quad \text{as } t \to \infty. \tag{6.4}$$

If (6.3) and (6.4) hold, then $H = \lambda G$.

A variant of theorem 2.2 is obtained from theorem 6.1 by simply letting $f_k(t)$ be the indicator function of the interval $[A_k, D_k]$, as in (2.2). An obvious application of theorem 6.1 is to treat the case in which a customer may enter and leave the system more than once before finally departing for good, as in a subset of the queues in a queueing network. To apply theorem 6.1, we let $f_k(t)$ be the indicator function of the set of times that the kth customer is in the system.

(2) THE STATIONARY FRAMEWORK

Paralleling section 3, it is easy to obtain a steady-state version of $H = \lambda G$ in section 6(1). The idea is to generalize the marks in section 3, replacing the real-valued random variable W_k by a stochastic process $\{X_k(t)\colon t \geqslant 0\}$ having sample paths coinciding with $f_k(t + A_k)$ for $f_k(t)$ in section 6(1). As a regularity condition, we assume that, in addition to being nonnegative, the sample paths of $\{X_k(t)\colon t \geqslant 0\}$ are right-continuous with limits from the left, so that we can regard the space of sample paths as a complete separable metric space, i.e., the function space $D[0, \infty)$ with the Skorohod topology; see Ethier and Kurtz [17]. We assume that $\{(A_k - A_{k-1}, \{X_k(t)\colon t \geqslant 0\})\colon k \geqslant 1\}$ is a stationary sequence with $A_0 = 0$. We then extend it to a stationary sequence on $-\infty < k < \infty$.

Paralleling (6.3), we are interested in the "cumulative cost" associated with an arbitrary customer, i.e.,

$$Z_k = \int_{A_k}^{\infty} X_k(t - A_k)\,\mathrm{d}t \tag{6.5}$$

and its expected value $EZ_0 \equiv G$.

We construct the associated stationary marked point process $\{(A'_k, \{X'_k(t): t \geqslant 0\}): -\infty < k < \infty\}$ just as in (3.4)–(3.6). The time-stationary quantity of interest to us is the total "instantaneous cost rate" of all customers at an arbitrary time t, i.e.,

$$Y'(t) = \sum_{k=-\infty}^{k=+\infty} X'_k(t - A'_k), \quad t \geqslant 0, \tag{6.6}$$

with $X'_k(t) = 0$ for $t \leqslant 0$, and its expected value $EY'(0) \equiv H$.

Like theorem 3.2, the following result is an immediate consequence of Campbell's theorem (theorem 3.1 with general marks). Thus, it is primarily a consequence of Franken [20]; see chapter 4.2 of Franken et al. [21]. The explicit statement was made by Stidham [65,66].

THEOREM 6.2

If $\{(A_k - A_{k-1}, \{X_k(t): t \geqslant 0\}: -\infty < k < \infty\}$ is a stationary sequence with $A_0 = 0$ and $EA_1 = \lambda^{-1}$, $0 < \lambda^{-1} < \infty$, then $\{Y'(t): t \geqslant 0\}$ defined by (6.6) is a stationary process with $H \equiv EY'(0) = \lambda EZ_0 \equiv \lambda G$.

Proof (sketch)

As for theorem 3.2, we apply theorem 3.1, but extended to allow a general mark space. From (6.5), we see that

$$H \equiv EY'(0) = E \sum_{k=-\infty}^{\infty} g(A'_k, \{X'_k(t): t \geqslant 0\})$$

for $g(a, \{x(t): t \geqslant 0\}) = x(-a)$. Hence, by the extended (3.10),

$$H = \lambda \int_{-\infty}^{\infty} \int_{D[0,\infty)} x(-a) P(X_0 \in \mathrm{d}x)\,\mathrm{d}a$$

$$= \lambda E \int_{-\infty}^{\infty} X_0(-a)\,\mathrm{d}a = \lambda EZ_0 = \lambda G,$$

where $X_0 \equiv \{X_0(t): t \geqslant 0\}$ and $x \equiv \{x(t): t \geqslant 0\}$ is a typical element of $D[0, \infty)$. □

Remark

(6.1) From the above, it should be clear that there is essentially an equivalence between steady-state applications of $H = \lambda G$ and sample-path average applications. Hence, once we have a result within one framework, we can easily obtain the corresponding result in the other framework, provided that we check a few additional regularity conditions. The arrival epochs come from $\{A_k\}$ in both cases and the sample paths of $\{X_k(t)\colon t \geqslant 0\}$ coincide with $\{f_k(t+A_k)\colon t \geqslant 0\}$; e.g., given $X_k(t)$ we let $f_k(t) = X_k(t-A_k)$, $t \geqslant 0$. Of course, the conditions in theorems 6.1 and 6.2 are not identical, so there are additional conditions to check in such a conversion, but these can be expected to hold, i.e., they are typically only technical regularity conditions. For example, as noted at the end of section 3, if we assume ergodicity in the stationary-process framework, then we can apply theorem 6.1 and the cross-ergodic theorems to obtain the corresponding steady-state statement in theorem 6.2. □

(3) AN EXTENSION OF $H = \lambda G$ AND OTHER CONSERVATION LAWS

We now present an extension of $H = \lambda G$ in section 6(1) due to Glynn and Whitt [27] to represent lump costs as well as cost rates. Instead of the functions $f_k(t)$ in section 6(1), consider nondecreasing nonnegative functions $F_k(t)$, with $F_k(t)$ being interpreted as the cumulative cost associated with customer k at time t. Section 6(1) is the special case in which

$$F_k(t) = \int_{-\infty}^{t} f_k(s)\,\mathrm{d}s. \tag{6.7}$$

Let A_k and D_k be as in section 2.

THEOREM 6.3

Suppose that (6.1) holds and

$$F_k(t) = 0 \text{ for } t < A_k \quad \text{and} \quad F_k(t) = F_k(D_k) \text{ for } t \geqslant D_k. \tag{6.8}$$

Then

$$G_k \equiv k^{-1} \sum_{j=1}^{k} F_j(\infty) \to G \quad \text{as } k \to \infty \tag{6.9}$$

if and only if

$$H_t \equiv t^{-1} \sum_{j=1}^{\infty} F_j(t) \to H \quad \text{as } t \to \infty. \tag{6.10}$$

If (6.9) and (6.10) hold, then $H = \lambda G$.

Proof

We indicate how to apply theorem 4 of [27]. We first specify the quantities T_n and S_n considered there. For any $\epsilon > 0$, let $T_n = A_n - \epsilon$, $W_n = D_n - A_n$ and $S_n = W_n + \epsilon$. By (6.1), $n^{-1}T_n \to \lambda^{-1}$ as $n \to \infty$, $0 < \lambda^{-1} < \infty$, and $t^{-1}N(t) \to \lambda$ as $t \to \infty$, $0 < \lambda < \infty$. By (6.8), $F_k(T_n) = 0$ for all $k \geqslant n$ as in (21) of [27], so that (17) of [27] holds. Moreover, $F_k(\infty) - F_k(T_k + S_k) = 0$ for all k as in (22) of [27] so that (18) of [27], holds if and only if $S_n/T_n \to 0$ as $n \to \infty$. However, by (6.1), $W_n/n \to 0$ as $n \to \infty$, which implies $S_n/T_n \to 0$ as $n \to \infty$. □

Remark

(6.2) It is almost possible to apply theorem 6.1 to prove theorem 6.3. Let $f_k(t) = F_k'(t)$, where $F_k'(t)$ is the (nonnegative) derivative of the absolutely continuous part of $F_k(t)$. Then theorem 6.1 applies to $f_k(t)$. Let $J_k = F_k(\infty) - \int_0^\infty F_k'(t)\,\mathrm{d}t$, so that G_k and H_t in (6.12) and (6.13) can be expressed as

$$G_k = k^{-1} \sum_{j=1}^{k} \int_0^\infty f_k(t)\,\mathrm{d}t + k^{-1} \sum_{j=1}^{k} J_j$$

and

$$H_t = t^{-1} \int_0^t \sum_{j=1}^{\infty} f_j(s)\,\mathrm{d}s + t^{-1} \sum_{j=1}^{A(t)} J_j.$$

To apply theorem 6.1, we must treat the two components of G_k and H_t separately. However, it is possible for $H_t \to H$ without having the two components converge, and similarly for G_k. To establish the equivalence of $t^{-1}\Sigma_{j=1}^{A(t)} J_j \to \lambda J$ and $k^{-1}\Sigma_{j=1}^{k} J_j \to J$, we can apply theorem 6.4 below. □

We now present a "sample path version of the renewal reward theorem" called $Y = \lambda X$, which was used by Stidham and El-Taha [68] to derive many different sample path results; see p. 134 of [68]. This result has an easy direct proof, but we show that it also can be regarded as an elementary consequence of $H = \lambda G$ in the form of theorem 6.3.

THEOREM 6.4

Suppose that $Y(t)$ is a nondecreasing real-valued function of a real variable, $0 \leqslant A_k \leqslant A_{k+1}$ for all k and $k^{-1}A_k \to \lambda^{-1}$ as $k \to \infty$ with $0 < \lambda^{-1} < \infty$. Then $t^{-1}Y(t) \to Y$ as $t \to \infty$ if and only if $k^{-1}Y(A_k) \to X$ as $k \to \infty$, in which case $Y = \lambda X$.

Proof

We apply theorem 6.3, letting $D_k = A_{k+1}$ and

$$F_k(t) = \begin{cases} 0, & t < A_k, \\ Y(t) - Y(A_k -), & A_k \leqslant t < A_{k+1}, \\ Y(A_{k+1} -) - Y(A_k -), & t \geqslant A_{k+1}. \end{cases} \tag{6.11}$$

Since $k^{-1}A_k \to \lambda^{-1}$ as $k \to \infty$, (6.1) holds. By (6.11), (6.8) holds. Note that

$$G_k \equiv k^{-1} \sum_{j=1}^{k} F_j(\infty) = k^{-1}[Y(A_{k+1} -) - Y(A_1 -)] \tag{6.12}$$

and

$$H_t \equiv t^{-1} \sum_{j=1}^{\infty} F_j(t) = t^{-1}Y(t). \tag{6.13}$$

The rest is an easy consequence of theorem 6.3, using $k^{-1}Y(A_k -) \leqslant k^{-1}Y(A_k) \leqslant k^{-1}Y(A_{k+1})$. □

Remark

(6.3) The nondecreasing property in theorem 6.4 plays an essential role. To see this, let $Y(t) = \sum_{j=1}^{[t]} J_j$ and $A_k = 2k$. If $J_{2j} = -J_{2j-1} = -j$, $j \geqslant 1$, then $Y(2j) = 0$ while $Y(2j-1) = j$, $j \geqslant 1$. Hence, $t^{-1}Y(t)$ does not converge, while $k^{-1}Y(A_k) = 0$ for all k. □

We next state an elementary deterministic-framework version of the stationary-framework *rate conservation law* (RCL) of Miyazawa [49–51]. This deterministic RCL comes from Sigman [59]. Closely related analysis of a queueing workload process was done by Zazanis [73]. A minor modification can be considered a consequence of theorem 6.4 (and thus $H = \lambda G$ in theorem 6.3).

THEOREM 6.5

Suppose that $x(t)$ is a real-valued function of a real variable such that

$$x(t) = x(0) + \int_0^t x'(s)\,\mathrm{d}s + \sum_{k=1}^{A(t)} J_k, \quad t \geqslant 0, \tag{6.14}$$

where $0 \leqslant A_k < A_{k+1}$ for all k with $A(t) = \max\{k \geqslant 0\colon A_k \leqslant t\}$. Also suppose that $t^{-1}x(t) \to 0$ and $t^{-1}A(t) \to \lambda$ as $t \to \infty$ with $0 < \lambda < \infty$. Then

$$t^{-1}\int_0^t x'(s)\,\mathrm{d}s \to \alpha \quad \text{as } t \to \infty \tag{6.15}$$

if and only if

$$k^{-1}\sum_{j=1}^{k} J_j \to \beta \quad \text{as } k \to \infty, \tag{6.16}$$

in which case $\alpha = -\lambda\beta$.

Proof

Since $A_k < A_{k+1}$ for all k and $t^{-1}A(t) \to \lambda$ as $t \to \infty$, it is easy to see that (6.16) holds if and only if

$$t^{-1}\sum_{k=1}^{A(t)} J_k \to \lambda\beta \quad \text{as } t \to \infty. \tag{6.17}$$

From (6.14) and the various conditions, (6.15) is easily seen to be equivalent to (6.17) with $\alpha = -\lambda\beta$. □

Remarks

(6.4) The equivalence between (6.16) and (6.17) follows from theorem 6.4

when $J_k \geqslant 0$ for all k. A modified version of theorem 6.5 with (6.16) replaced by

$$k^{-1} \sum_{j=1}^{k} J_j^{+} \to \beta^{+} \quad \text{and} \quad k^{-1} \sum_{j=1}^{k} J_j^{-} \to \beta^{-} \quad \text{as } t \to \infty,$$

where $x^{+} = \max\{x, 0\}$, $x^{-} = -\min\{x, 0\}$ and $\beta = \beta^{+} - \beta^{-}$, thus follows from theorem 6.4. Such convergence usually would hold in practice.

(6.5) A sufficient condition for (6.14) is for x to be of bounded variation on $[0, t]$ for each t with the continuous component being absolutely continuous. Then $x = y - z$, where y and z are nondecreasing and bounded on $[0, t]$ for each t. Thus x can be written uniquely as a convex combination of discrete, singular and absolutely continuous components; see pp. 9–10 of Chung [12]. The discrete component is $x(0) + \sum_{j=1}^{A(t)} J_j$ and the absolutely continuous component is $\int_0^t x'(s)\,\mathrm{d}s$.

(6.6) Sigman [59] showed that the version of $H = \lambda G$ in theorem 6.1 can be obtained as a corollary to the RCL in theorem 6.5. It suffices to let

$$x(t) = \sum_{n=1}^{\infty} \mathbf{1}_{[A_k, D_k]}(t) \int_t^{\infty} f_n(s)\,\mathrm{d}s, \tag{6.18}$$

where $\mathbf{1}_A(t)$ is the indicator function of A. Then

$$x(t) = x(0) + \int_0^t \left[\sum_{n=1}^{\infty} f_n(s) \right] \mathrm{d}s + \sum_{j=1}^{A(t)} \int_{-\infty}^{\infty} f_j(s)\,\mathrm{d}s. \quad \square$$

In the stationary framework of Miyazawa [49,50], $x(t)$ in (6.14) corresponds to the sample path of a stochastic process, α in (6.15) corresponds to the expected time-stationary value of the derivative, while β corresponds to the expected customer-stationary value of a jump. Using the terminology of remark 3.1 (which is used by Miyazawa [49,50]), we would write the conclusion of the RCL in the stationary framework as

$$EX'(0) = -\lambda E^0[X(0+) - X(0-)]. \tag{6.19}$$

Since the RCL in theorem 6.5 can be considered a consequence of theorem 6.4 and theorem 6.4 can be considered a consequence of theorem 6.3, we can regard $H = \lambda G$ as the fundamental principle. On the other hand, arguments such as theorem 6.4 are used to prove theorem 4 of [27], which is used to prove theorem 6.3. In any case, it appears that several different principles can be regarded as fundamental.

Recently several different principles have been shown to be equivalent in the stationary framework. Miyazawa [49] applied the inversion formula (3.7) to establish the RCL in the stationary framework, but Brémaud [4] showed that the inversion formula can also be deduced from the RCL. Moreover, Miyazawa [51] showed that Mecke's formula, which implies Campbell's formula (3.10) and the inversion formula, also can be deduced from the RCL. Hence, all these various principles can be regarded as equivalent. In applications we can use any one that is convenient.

Work related to the RCL includes the level crossing analysis of Brill and Posner [6,7] and Cohen [14], the intensity conservation principle of König and Schmidt [36,37] and recent work by Stidham and El-Taha [68], Brémaud [4], Ferrandiz and Lazar [19], Mazumdar, Kannurpatti and Rosenberg [44], Sigman [59] and Zazanis [73].

(4) APPLICATIONS OF $H = \lambda G$

Perhaps the best known application of $H = \lambda G$ is Brumelle's [8] formula relating the expected time-stationary workload, say V, in the general $G/G/s$ queue to the expected customer-stationary service time, say S, and waiting time before beginning service, say W, i.e.,

$$EV = \lambda\left[E(SW) + E(S^2)/2\right]. \tag{6.20}$$

The sample-path version of (6.20) is obtained from theorem 6.1 by setting

$$f_k(t) = \begin{cases} S_k, & A_k < t \leqslant A_k + W_k, \\ S_k - (t - W_k - A_k), & A_k + W_k \leqslant t \leqslant A_k + W_k + S_k, \\ 0, & \text{otherwise,} \end{cases} \tag{6.21}$$

where A_k is the arrival epoch, S_k is the service time and W_k is the waiting time (before beginning service) of customer k. The steady-state version in (6.20) is obtained from theorem 6.2 by setting $X_k(t) = f_k(t + A_k)$ for $f_k(t)$ in (6.21); see (4.2.4) on p. 107 of Franken et al. [21].

From (6.21), it is apparent that the model can be quite general. Customer k is in the queue waiting from A_k until $A_k + W_k$ and then is in service from $A_k + W_k$ until $A_k + W_k + S_k$. The remaining workload associated with this customer is S_k in the interval $[A_k, A_k + W_k]$ and then decreases at rate 1 while he is in service. To apply theorem 6.1, we need $k^{-1}A_k \to \lambda^{-1}$ and $k^{-1}(W_k + S_k) \to 0$ as $k \to \infty$.

Often a customer's service time is independent of his waiting time, in which case $E(SW) = E(S)E(W)$ in (6.20). If, in addition, the model is a single server queue with the FIFO discipline, then V is the virtual waiting time, i.e., W is the customer quantity associated with V. Moreover, if ASTA holds [46,70], then V is distributed the same as W, so that (6.20) yields the Pollaczek–Khintchine formula; see pp. 408–412 of Heyman and Sobel [31].

The relation of $H = \lambda G$ can also be applied to derive the equilibrium excess, age and spread distributions, although they are also easy to obtain directly. These are well known within the stationary-process framework; see p. 27 of [21]. The sample-path equivalent via $H = \lambda G$ has been given by Wolff [71,72].

We also mention Brumelle's [9] application of $H = \lambda G$ to relate the higher moments of the time-stationary number in queue in a $G/G/s$ model to the higher moments of the customer-stationary waiting time; see also Miyazawa [48]. McKenna [45] has recently obtained interesting generalizations of these moment relations for closed, product-form queueing networks. These network results have yet to be "explained" by being treated in either a general stationary-process framework or a sample-path framework.

An interesting class of applications of $H = \lambda G$ are those that lead to relations between the steady-state distribution of a continuous-time stochastic process and the steady-state distribution of an embedded sequence obtained by evaluating this stochastic process at times of an associated point process. The first such application of $H = \lambda G$ appears in section 2 of Heyman and Stidham [32]. It is extended in chapter 5.3 of Stidham [66] and will be further extended here using similar reasoning. Stidham and El-Taha [68] obtain related results via an application of theorem 6.4.

To state our application of $H = \lambda G$, we introduce a function $Z \equiv \{Z(t): t \geqslant 0\}$ (sample path of a stochastic process) mapping the interval $[0, \infty)$ into a separable metric space S with limits from the left and right. We suppose that our sequence $\{A_k: k \geqslant 0\}$ is a subsequence of another arrival sequence $\{B_k: k \geqslant 0\}$ and let $D_k = A_{k+1}$, $k \geqslant 0$. In particular, we let $A_k = B_m$ if $Z(B_j -) \in C_1$ for the kth time when $j = m$. We then let $f_k(t) = 1$ if $Z(t) \in C_2$ and $A_k \leqslant t \leqslant D_k$, and 0 otherwise. (Of course, C_1 and C_2 must be measurable subsets of S.)

We assume that $k^{-1}B_k \to \lambda_B^{-1}$ as $k \to \infty$ and that the proportion of j such that $Z(B_j -) \in C_1$ converges to $\pi(C_1)$. Hence, $k^{-1}A_k \to (\lambda_B \pi(C_1))^{-1}$ as $k \to \infty$ and $t^{-1}A(t) \to \lambda_B \pi(C_1)$ as $t \to \infty$. Then, assuming that the limits exist, $H \equiv H(C_2)$ is the long-run proportion of time that $Z(t) \in C_2$, which we denote by $p(C_2)$, and $G \equiv G(C_1, C_2)$ is the long-run average time spent in C_2 per A_k – arrival (between successive B_k arrivals for which $Z(B_k -) \in C_1$). Hence, by $H = \lambda G$, we have

THEOREM 6.6

The customer average $\pi(C_1)$ is related to the time average $p(C_2)$ by

$$p(C_2) \equiv H(C_2) = \lambda_B \pi(C_1) G(C_1, C_2). \tag{6.22}$$

For example, let $\{Z(t): t \geqslant 0\}$ be in a system in which arrivals and departures occur one at a time. Let $\{B_k: k \geqslant 0\}$ be the sequence of arrival times. Then, if we let $C_2 = \{m\}$ and $C_1 = \{m - 1\}$, we obtain the relation

$$r(k)p(k) = \lambda \pi(k-1), \quad k \geqslant 1, \tag{6.23}$$

where λ is the arrival rate and $r(k)$ is the departure rate while in state k (the limit as $t \to \infty$ of the ratio of the number of departures while in state k during $[0, t]$ and the time spent in k during $[0, t]$), as given in theorem 5.4 of Stidham [66]. To justify (6.23), note that there is precisely one departure in state k between A_j and A_{j+1} for each j. Hence, $G(k-1, k) = 1/r(k)$ as in (5.6) of [66]. Of course, in general it is hard to evaluate $r(k)$, but for the special case of the $GI/M/c/k$ queue, $r(k) = \min\{k, c\}\mu$ and we have the special case considered by Heyman and Stidham [32].

Now, for a new application of $H = \lambda G$ to determine when arrivals see time averages (ASTA), consider (6.22) and let $C_1 = C_2 = C$ and $G(C) = G(C, C)$. Then (6.23) becomes

$$p(C) = \lambda_B \pi(C) G(C), \tag{6.24}$$

so that

$$p(\cdot) = \pi(\cdot) \text{ if and only if } G(C) = \lambda_B^{-1} \text{ for all } C \text{ such that } p(C) > 0. \tag{6.25}$$

Note that $G(C)$ in (6.24) is the long-run average time spent in C between successive arrivals finding the system in C. Hence $\lambda(C) \equiv 1/G(C)$ is the long-run rate of arrivals in C; i.e.,

$$\lambda(C) = \lim_{t \to \infty} \frac{\sum_{j=1}^{B(t)} \mathbf{1}_{\{Z(B_j-) \in C\}}}{\int_0^t \mathbf{1}_{\{Z(s) \in C\}} \, \mathrm{d}s} = \lim_{t \to \infty} \frac{A(t)}{\int_0^t \mathbf{1}_{\{Z(s) \in C\}} \, \mathrm{d}s}, \tag{6.26}$$

where $B(t)$ counts the number of k such that $B_k \leqslant t$. From above, (6.25) is equivalent to

THEOREM 6.7

The customer-average distribution $\pi(\cdot)$ coincides with the time-average distribution $P(\cdot)$ if and only $\lambda(C) = \lambda_B$ for all C such that $p(C) > 0$.

We regard (6.25) and theorem 6.7 as sample-path versions of ASTA generalizing the discrete-state sample-path versions of ASTA given in theorem 3.5 and corollary 3.6 of Stidham and El-Taha [68]. Moreover, the proof is similar.

Unfortunately, the stochastic implications of theorems 6.6 and 6.7 are not so clear. It is intuitively obvious that if $\{A(t): t \geqslant 0\}$ is a Poisson process satisfying a lack of anticipation assumption (LAA) as in Wolff [70], then $\lambda(C) = \lambda(S) = \lambda_B$ for all C such that $p(C) > 0$, so that $p(\cdot) = \pi(\cdot)$, but a rigorous proof that $\lambda(C) = \lambda(S)$ for all measurable C in S requires further argument. For example, theorem 2 of Melamed and Whitt [46] implies that $p(\cdot) = \pi(\cdot)$, so that indeed $\lambda(C) = \lambda_B$ for all C such that $p(C) > 0$ by theorem 6.7. However, theorems 6.6 and 6.7 do show how the same conclusions can hold without the customary assumptions on the local behavior of the stochastic processes.

7. Estimation of CLTs

Maxwell [43] seems to have been the first to point out the significance of $L = \lambda W$ for estimation, e.g., by simulation. Operational analysis [10,15] also uses

$L = \lambda W$ for estimation. Assuming that we know λ, we can apply $L = \lambda W$ to estimate both L and W if we can estimate either one (and similarly for $H = \lambda G$).

It is natural then to ask whether it is more efficient to estimate L and W directly or indirectly via the other. Such an investigation was first carried out by Law [40] for the $M/G/1$ queue, based on his thesis with Wolff, and then extended to the $GI/G/s$ queue by Carson and Law [11] and to general models in the framework of chapter 2 by Glynn and Whitt [22–24,26,27].

An important insight of Glynn and Whitt was that the relations $L = \lambda W$ and $H = \lambda G$ can be generalized to provide central-limit-theorem (CLT) versions, which provide a basis for comparing the asymptotic efficiency (the size of confidence intervals with large samples) of the estimators. The probability structure underlying $L = \lambda W$ becomes crucial when we want to compare finite averages to their limits. The CLTs can be interpreted as providing rates of convergence associated with the limiting averages.

To illustrate the CLT versions, we state one result in the setting of section 2. For this purpose, let $\Rightarrow$ denote convergence in distribution. The following is a consequence of theorem 1 of Glynn and Whitt [24]. Let $A_0 = 0$ and let $[x]$ be the integer part of x.

THEOREM 7.1

If $\{(A_k - A_{k-1}, W_k): k \geqslant 1\}$ is a stationary sequence of nonnegative random vectors satisfying

$$k^{-1/2}\left(A_k - \lambda^{-1}k,\ \sum_{j=1}^{k} W_j - Wk\right) \Rightarrow (X, Y) \quad \text{as } k \to \infty, \tag{7.1}$$

where λ^{-1} and W are constants satisfying $0 < \lambda^{-1} < \infty$ and $W < \infty$, and (X, Y) is an arbitrary random vector, then $E(A_k - A_{k-1}) = \lambda^{-1}$, $EW_k = W$ and

$$t^{-1/2}\left[A_{[\lambda t]} - t,\ \sum_{j=1}^{[\lambda t]} W_j - \lambda Wt,\ \sum_{j=1}^{A(t)} W_j - \lambda Wt,\ \int_0^t Q(s)\,\mathrm{d}s - \lambda Wt\right]$$
$$\Rightarrow \lambda^{1/2}(X, Y, Y - WX, Y - WX) \quad \text{as } t \to \infty. \tag{7.2}$$

Moreover, if (X, Y) has a bivariate normal distribution with zero means and covariances (σ_{ij}^2), then $\lambda^{1/2}(Y - WX)$ is normally distributed with zero mean and variance $\lambda(\lambda^2 W^2 \sigma_{11}^2 - 2\lambda W \sigma_{12}^2 + \sigma_{22}^2)$.

Note that the standard version of $L = \lambda W$ is reflected by the centering constants in (7.1) and (7.2). The joint convergence in (7.2) implies that $t^{-1/2}(\sum_{j=1}^{A(t)} W_j - \lambda Wt)$ and $t^{-1/2}(\int_0^t Q(s)\,\mathrm{d}s - \lambda Wt)$ not only have the same distribution asymptotically as $t \to \infty$, but assume the same values in the limit; i.e.,

$$t^{-1}\sum_{j=1}^{A(t)} W_j = t^{-1}\int_0^t Q(s)\,\mathrm{d}s + o_P(t^{-1/2}), \tag{7.3}$$

which means that modulo a term that converges in probability to zero after dividing by $t^{-1/2}$, the averages $t^{-1}\Sigma_{j=1}^{A(t)}W_j$ and $t^{-1}\int_0^t Q(s)\,\mathrm{d}s$ have the same value; equivalently,

$$t^{1/2}\left[t^{-1}\sum_{j=1}^{A(t)} W_j - t^{-1}\int_0^t Q(s)\,\mathrm{d}s\right] \Rightarrow 0 \quad \text{as } t \to \infty. \tag{7.4}$$

The key to obtaining relations among the classical limit theorems (including laws of iterated logarithm [22,25] and weak laws of large numbers as well as CLTs and strong laws of large numbers) is the recognition that the fundamental relation behind $L = \lambda W$ is the relation among cumulative processes in (2.1) and (2.6).

Theorem 7.1 has the drawback of requiring stationarity, but the stationarity can be omitted if we use functional central limit theorems (FCLTs) instead of ordinary CLTs; see [22,23]. Indeed, the FCLT versions of $L = \lambda W$ and $H = \lambda G$ are much easier to prove than theorem 7.1 (using sample path arguments as in theorem 2.1 and the continuous mapping theorem); see [22] and [27]. Unlike all the previous theorems in this paper, we know of no easy proof of theorem 7.1.

To apply the CLT version of $L = \lambda W$ in theorem 7.1, we need to establish the joint CLT in (7.1). We now give a sufficient condition in terms of regenerative structure from [23]. We suppose that there is a sequence of i.i.d. nonnegative integer-valued random variables $\{C_k\colon k \geqslant 1\}$ with $EC_1 = m$, $0 < m < \infty$, such that the random vectors $(C_k, X_{\beta_{k-1}+1}, \ldots, X_{\beta_k})$ are i.i.d., where $\beta_k = C_1 + \ldots + C_k$, $\beta_0 = 0$, and $X_k = (A_k - A_{k-1}, W_k)$. The variables β_k constitute the regeneration points. Let $U_k = A_k - A_{k-1}$,

$$S_k = U_{\beta_{k-1}+1} + \ldots + U_{\beta_k} \quad \text{and} \quad T_k = W_{\beta_{k-1}+1} + \ldots + W_{\beta_k}. \tag{7.5}$$

As a consequence of the i.i.d. assumptions above, (C_k, S_k, T_k) are i.i.d. Let

$$\lambda^{-1} = m^{-1}ES_1 \quad \text{and} \quad W = m^{-1}ET_1, \tag{7.6}$$

and assume that $0 < \lambda < \infty$.

THEOREM 7.2

If, in addition to the regenerative structure assumptions above, EC_1^2, ES_1^2 and ET_1^2 are finite, then (7.1) holds with the limit (X, Y) there having a bivariate normal distribution with zero means and

$$\begin{aligned}
\sigma_1^2 &\equiv \operatorname{Var} X = m^{-1} \operatorname{Var}\left(S_1 - \lambda^{-1}C_1\right),\\
\sigma_2^2 &\equiv \operatorname{Var} Y = m^{-1} \operatorname{Var}\left(T_1 - WC_1\right),\\
\sigma_{12}^2 &\equiv \operatorname{Cov}(X, Y) = m^{-1} \operatorname{Cov}\left(S_1 - \lambda^{-1}C_1, T_1 - WC_1\right).
\end{aligned} \tag{7.7}$$

As a consequence of the CLT-versions of $L = \lambda W$, Glynn and Whitt [26] show that an indirect estimator for L, using the natural estimator for W and the known arrival rate λ is more asymptotically efficient than a natural estimator for L using the same data, provided that the interarrival and waiting times are negatively correlated. (Note that this is the typical situation: when the interarrival times become shorter, the waiting times typical become larger.) For example, suppose that we use data available from the interval $[0, t]$. Then $\lambda\hat{W}_t \equiv \lambda D(t)^{-1}\Sigma_{j=1}^{D(t)} W_j$ is more asymptotically efficient than $\hat{L}_t \equiv t^{-1}\int_0^t Q(s)\,\mathrm{d}s$ under this condition. (Note that we do not work with $\lambda A(t)^{-1}\Sigma_{j=1}^{A(t)} W_j$ because the time spent in the system by customers still present at time t is typically not known.) An estimator that is even more asymptotically efficient (in some sense, the most asymptotically efficient) is the linear control estimator $\lambda\hat{W}_t + \hat{a}_t(\hat{\lambda}_t^{-1} - \lambda^{-1})$, where $\hat{\lambda}_t = t^{-1}D(t)$ and $\hat{a}_t$ is a consistent estimator for a constant times the ratio of covariance matrix elements. These gains in asymptotic efficiency can be realized because we know the arrival rate λ. In [26] it is also shown that $L = \lambda W$ does not change the asymptotic efficiency when the arrival rate λ needs to be estimated as well.

Remark

(7.1) Glynn, Melamed and Whitt (forthcoming paper) have recently obtained a joint central limit theorem for customer and time averages related to PASTA that is similar in spirit to the CLT versions of $L = \lambda W$.

8. Other extensions

(1) $L = \lambda W$ FOR PARTIALLY OBSERVABLE PROCESSES

Our discussion of estimation in section 7 focused on asymptotic efficiency, where asymptotic efficiency referred to the width of confidence intervals with large samples. However, in estimation it is important to consider the effort required to obtain an estimator as well as its statistical precision; e.g., see Glynn and Whitt [28]. Thus, even when the confidence intervals should be smaller with one estimator, it may be desirable to use another estimator.

For example, in many manufacturing settings it is much easier to count the work in process (WIP, L) at any given time than it is to measure production intervals (the time spent in the system by each product, W). Thus we may want to apply $L = \lambda W$ to estimate W using L, even though the statistical precision would be better using W. However, there is a fundamental difficulty with this approach. The observation of WIP does not take into account such features as partial yields, changing lot sizes and reconstituted lots. We may want to estimate the average time from start to finish, conditional on the item turning out to be good product. Thus, what we want to observe to apply $L = \lambda W$ is only the WIP that will eventually be good, but this eventually good WIP is not directly observable. Nozari and Whitt [54] investigate this problem and suggest estimating the expected amount of good product associated with current WIP. However, such estimation seems to require considerable care.

(2) CONTINUOUS AND MORE GENERAL VERSIONS

Rolski and Stidham [58] extended $L = \lambda W$ and $H = \lambda G$ to situations in which the input can be continuous as well as discrete, as in fluid storage models. Then one relates the time-average system content to the average time spent in the system per particle. Glynn and Whitt [27] further extended $L = \lambda W$ and $H = \lambda G$ to the setting of general two-dimensional cumulative processes that need not be expressible as integrals or sums. Theorem 6.3 is an application of this generalization to cover lump costs as well as cost rates. The generalization also can be used to treat stochastic integrals; see section 1.7 of [27]. The generalization also leads to higher dimensions; see section 1.8 of [27].

(3) AN ORDINAL VERSION OF $L = \lambda W$

Halfin and Whitt [29] established a relation which can be interpreted as an ordinal version of $L = \lambda W$. The idea is to measure time solely in terms of arrival indices, so that the waiting time becomes the number of arrivals during the time a customer is in the system. The conclusion is that the long-run average number of customers in a queueing system at an arrival epoch is equal to the long-run average number of arrivals during a customer's sojourn time in the system. (This result is fairly obvious for a single-server queue with the FCFS discipline, but that is not assumed.)

To be precise, consider the setting of section 2 and let X_k be the number of customers with indices greater than k that arrive while customer k is in the system, and let N_k be the number of customers with indices less than k that are in the system at the arrival epoch of customer k. (The qualifications on the indices are included to cope with multiple events occurring at the same time.)

The following is theorem 2 of [29].

THEOREM 8.1

The following are equivalent:

(a) $k^{-1}\Sigma_{j=1}^{k} X_j \to x$ as $k \to \infty$;

(b) $k^{-1}\Sigma_{j=1}^{k} N_j \to x$ and $k^{-1}X_k \to 0$ as $k \to \infty$.

Theorem 8.1 is proved in [29] by applying theorem 2.1; let $A_k = k$ (so that $\lambda = 1$) and $W_k = X_k$. Then theorem 8.1 corresponds to a discrete-time version of $L = \lambda W$, which is an easy consequence of theorem 2.1 or, directly,

$$\int_0^k Q(s)\,\mathrm{d}s = \sum_{j=1}^{k} N_j, \quad k \geqslant 1.$$

Theorem 8.1 is applied in [29] to obtain a conservation law for single-server queues. The conservation law is applied to establish extremal properties for the

FIFO service discipline: In a $G/GI/1$ system, the FIFO discipline minimizes (maximizes) the long-run average sojourn time per customer among all work-conserving disciplines that are non-anticipating with respect to the service times (may depend on completed service times, but not on residual service times) when the service-time distribution is NBUE (NWUE), i.e., new better (worse) than used in expectation. Among the disciplines in this class are round robin, processor sharing and shortest expected remaining processing time.

(4) A DISTRIBUTIONAL VERSION OF $L = \lambda W$

From the perspective of the average version of $L = \lambda W$ in section 2, the CLT-version of $L = \lambda W$ discussed in section 7 might reasonably be regarded as the natural distributional generalization, but from the perspective of the steady-state version of $L = \lambda W$ in section 3 a natural distributional version is the one proposed by Haji and Newell [30], an especially interesting special case of which was recently discovered by Keilson and Servi[34]. See also (3.19) and (3.21) of Miyazawa [48], theorem 4.2.7 and p. 140 in Franken et al. [21] and Bertsimas and Nakazato [2]. Keilson and Servi [34,35] also present many interesting applications.

The distributional version of $L = \lambda W$ seems most useful when the arrival process is a Poisson process, which is the case considered by Keilson and Servi [34]. The Poisson version of the distributional version of $L = \lambda W$ is said to hold if the time-stationary number in $Q'(0)$ is related to the customer-stationary time in system W_0 by

$$Q'(0) \stackrel{\mathrm{d}}{=} \Pi(\lambda W_0), \tag{8.1}$$

where $\{\Pi(t): t \geqslant 0\}$ is a Poisson process with rate 1 that is independent of W_0. It is reasonable to regard (8.1) as a distributional version of $L = \lambda W$, because it connects the distributions of $Q'(0)$ and W_0. We obtain $L = \lambda W$ when we take expected values in (8.1), but we should not make too much of this, because from the general theory we know that we obtain $L = \lambda W$ whether or not (8.1) holds.

As indicated above, a similar relation also may hold when the arrival process is not Poisson; then $\Pi(\lambda\,\cdot)$ in (8.1) is replaced by the time-stationary arrival counting process $A(\cdot)$. The starting point for the more general statement is to observe that, in the stationary process framework of section 3 in which $A'_0 < 0 < A'_1$, $Q'(0) \geqslant n$ holds if and only if $W'_{-(n-1)} > A'_{-(n-1)}$ when the discipline is FCFS; see p. 618 of Haji and Newell [30] and (3.19) of Miyazawa [48]. Note that when the discipline is FCFS a general model can always be represented as a stationary $G/G/1$ model by redefining the service times. Thus, the specific model considered by Miyazawa [48] and Franken et al. [21] can be generalized.

The following is a minor variant of the Poisson result from Keilson and Servi [34]. (A lack of anticipation assumption has been added to ensure that ASTA holds.)

THEOREM 8.2

In the steady-state setting of section 3, (8.1) holds if: (i) $\{A(t)\colon t \geqslant 0\}$ is a Poisson process with $\{A(t+u)-A(t)\colon u \geqslant 0\}$ independent of $Q(t)$ for each t, (ii) $\{D(t)\colon t \geqslant 0\}$ has jumps of size one only w.p.1, (iii) $D_k \leqslant D_{k+1}$ for all k w.p.1 and (iv) W_k is independent of the arrival process after A_k.

Proof

Condition (i) yields PASTA [70,72], i.e., $Q'(0) \overset{d}{=} Q(0-)$ where $Q(0-)$ is the customer-stationary number in system just prior to an arrival. Let $Q^*(0+)$ be the customer-stationary number in system after the departure D_0, which by (iii) coincides with $Q(W_0)$. By (ii), $Q^*(0+) \overset{d}{=} Q(0-)$; e.g., see p. 112 of [21]. However, since service is in a FIFO manner by (iii), $Q^*(0+)$ is just the number of arrivals during the time W_0 that customer 0 spends in the system, i.e., $Q(W_0) = A(W_0)$. Finally, by (iv), W_0 is independent of $\{A(t)\colon t \geqslant 0\}$. Hence, (8.1) holds. □

A useful consequence of (8.1) is the relation

$$\operatorname{Var} Q'(0) = \lambda^2 \operatorname{Var} W_0 + \lambda E W_0. \tag{8.2}$$

Note that theorems 8.1 and 8.2 have some similarities. In the steady-state setting of section 3, theorem 8.1 implies that

$$EQ(0-) = EA(W_0), \tag{8.3}$$

whereas in the proof of theorem 8.2 it is shown that

$$Q(0-) \overset{d}{=} A(W_0) \overset{d}{=} \Pi(\lambda W_0) \tag{8.4}$$

under the stronger conditions there. From (8.3), we obtain necessary and sufficient conditions for an expected-value version of ASTA for the process Q.

THEOREM 8.3

In the steady-state setting of section 3, $EQ'(0) = EQ(0-)$ if and only if $EA(W_0) = \lambda E W_0$.

Proof

By theorem 8.1, (8.3) holds. By theorem 3.2, $EQ'(0) = \lambda E W_0$. □

Acknowledgement

I am grateful to Peter Glynn, Shlomo Halfin, Benjamin Melamed and Ardavan Nozari for joint work related to $L = \lambda W$, and to Muhammad El-Taha, Daniel Heyman, Masakiyo Miyazawa, Richard Serfozo, Karl Sigman and Shaler Stidham, Jr., for helpful discussions.

References

[1] F. Baccelli and P. Brémaud, *Palm Probabilities and Stationary Queueing Systems*, Lecture Notes in Statistics 41 (Springer, New York, 1987).

[2] D.J. Be:tsimas and D. Nakazato, The general distributional Little's law, Operations Research Center, MIT (1990).

[3] P. Brémaud, Characteristics of queueing systems observed at events and the connection between stochastic intensity and Pálm probability, Queueing Systems 5 (1989) 99–112.

[4] P. Brémaud, An elementary proof of Sengupta's invariance relation and a remark on Miyazawa's conservation principle, J. Appl. Prob., to appear.

[5] P. Brémaud, R. Kannurpatti and R. Mazumdar, Event and time averages: a review and some generalizations, Adv. Appl. Prob. 23 (1991), to appear.

[6] P.H. Brill and M.J.M. Posner, Level crossing in point processes applied to queues: single-server case, Oper. Res. 25 (1977) 662–674.

[7] P.H. Brill and M.J.M. Posner, The system point method in exponential queues: a level crossing approach, Math. Oper. Res. 6 (1981) 31–49.

[8] S.L. Brumelle, On the relation between customer and time averages in queues, J. Appl. Prob. 8 (1971) 508–520.

[9] S.L. Brumelle, A generalization of $L = \lambda W$ to moments of queue length and waiting times, Oper. Res. 20 (1972) 1127–1136.

[10] J.P. Buzen, Fundamental operational laws of computer system performance, Acta Informatica 7 (1976) 167–182.

[11] J.S. Carson and A.M. Law, Conservation equations and variance reduction in queueing simulations, Oper. Res. 28 (1980) 535–546.

[12] K.L. Chung, *A Course in Probability Theory*, 2nd ed. (Academic Press, New York, 1974).

[13] A. Cobham, Priority assignment in waiting line problems, Oper. Res. 2 (1954) 70–76.

[14] J.W. Cohen, On up- and downcrossings, J. Appl. Prob. 14 (1977) 405–410.

[15] P.J. Denning.and J.P. Buzen, The operational analysis of queueing network models, Comp. Surveys 10 (1978) 225–261.

[16] S. Eilon, A simpler proof of $L = \lambda W$, Oper. Res. 17 (1969) 915–917.

[17] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence* (Wiley, New York, 1986).

[18] H. Eves and C.V. Newson, *An Introduction to the Foundations and Fundamental Concepts of Mathematics*, 2nd ed. (Holt, Rinehart and Winston, New York, 1965).

[19] J.M. Ferrandiz and A.A. Lazar, Rate conservation for stationary processes, J. Appl. Prob., to appear.

[20] P. Franken, Some applications of the theory of stochastic point processes in queueing theory, Math. Nachr. 70 (1976) 303–319 (in German).

[21] P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*, (Akademie-Verlag, Berlin, 1981, and Wiley, New York, 1982).

[22] P.W. Glynn and W. Whitt, A central limit theorem version of $L = \lambda W$, Queueing Systems 1 (1986) 191–215.

[23] P.W. Glynn and W. Whitt, Sufficient conditions for functional-limit-theorem versions of $L = \lambda W$, Queueing Systems 1 (1987) 279–287.

[24] P.W. Glynn and W. Whitt, Ordinary CLT and WLLN versions of $L = \lambda W$, Math. Oper. Res. 13 (1988) 674–692.

[25] P.W. Glynn and W. Whitt, An LIL version of $L = \lambda W$, Math. Oper. Res. 13 (1988) 693–710.

[26] P.W. Glynn and W. Whitt, Indirect estimation via $L = \lambda W$, Oper. Res. 37 (1989) 82–103.

[27] P.W. Glynn and W. Whitt, Extensions of the queueing relations $L = \lambda W$ and $H = \lambda G$, Oper. Res. 37 (1989) 634–644.

[28] P.W. Glynn and W. Whitt, The asymptotic efficiency of simulation estimators, Oper. Res. 40 (1992), to appear.

[29] S. Halfin and W. Whitt, An extremal property of the FIFO discipline via an ordinal version of $L = \lambda W$, Commun. Statist.-Stochastic Models 5 (1989) 515–529.

[30] R. Haji and G.F. Newell, A relation between stationary queue and waiting-time distributions, J. Appl. Prob. 8 (1971) 617–620.

[31] D.P. Heyman and M.J. Sobel, *Stochastic Models in Operations Research*, vol. 1 (McGraw-Hill, New York, 1982).

[32] D.P. Heyman and S. Stidham, Jr., The relation between customer and time averages in queues, Oper. Res. 28 (1980) 983–994.

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

[34] J. Keilson and L.D. Servi, A distributional form of Little's law, Oper. Res. Lett. 7 (1988) 223–227.

[35] J. Keilson and L.D. Servi, The distributional form of Little's law and the Fuhrmann–Cooper decomposition, Oper. Res. Lett. 9 (1990) 239–247.

[36] D. König and V. Schmidt, Imbedded and non-imbedded stationary characteristics of queueing systems with varying service rate and point processes, J. Appl. Prob. 17 (1980) 753–767.

[37] D. König and V. Schmidt, Relationships between time and customer stationary characteristics of queueing systems, in: *Point Processes and Queueing Systems*, eds. P. Bartfai and J. Tomko (North-Holland, Amsterdam, 1981) pp. 181–225.

[38] D. König and V. Schmidt, EPSTA: the coincidence of time-stationary and customer-stationary distributions, Queueing Systems 5 (1989) 247–264.

[39] M. Krakowski, Conservation methods in queueing theory, Rev. Franc. Automat. Inform. Rech. Oper. 7 (1973) 63–83.

[40] A.M. Law, Efficient estimators for simulated queueing systems, Manag. Sci. 22 (1975) 30–41.

[41] E.D. Lazowska, J. Zahorjan, G.S. Graham and K.C. Sevcik, *Quantitative System Performance, Computer System Analysis Using Queueing Network Models* (Prentice–Hall, Englewood Cliffs, NJ, 1984).

[42] J.D.C. Little, A proof for the queueing formula: $L = \lambda W$, Oper. Res. 9 (1961) 383–387.

[43] W.L. Maxwell, On the generality of the equation $L = \lambda W$, Oper. Res. 18 (1970) 172–174.

[44] R. Mazumdar, R. Kannurpatti and C. Rosenberg, On rate conservation law for non-stationary processes, J. Appl. Prob., to appear.

[45] J. McKenna, A generalization of Little's law to moments of queue lengths and waiting times in closed, product-form queueing networks, J. Appl. Prob. 26 (1989) 121–133.

[46] B. Melamed and W. Whitt, On arrivals that see time averages, Oper. Res. 38 (1990) 156–172.

[47] M. Miyazawa, Time and customer processes in queues with stationary inputs, J. Appl. Prob. 14 (1977) 349–357.

[48] M. Miyazawa, A formal approach to queueing processes in the steady state and their applications, J. Appl. Prob. 16 (1979) 332–346.

[49] M. Miyazawa, The derivation of invariance relations in complex queueing systems with stationary inputs, Adv. Appl. Prob. 15 (1983) 874–885.

[50] M. Miyazawa, The intensity conservation law for queues with randomly changed service rate, J. Appl. Prob. 22 (1985) 408–418.

[51] M. Miyazawa, Derivation of Little's and related formulas by rate conservation law with multiplicity, Department of Information Sciences, Science University of Tokyo (1990).

[52] P.M. Morse, *Queues, Inventories and Maintenance* (Wiley, New York, 1958).

[53] G.F. Newell, *Applications of Queueing Theory* (Chapman and Hall, London, 1971).

[54] A. Nozari and W. Whitt, Estimating average production intervals using inventory measurements: Little's law for partially observable processes, Oper. Res. 36 (1988) 308–323.

[55] M.G. Ramalhoto, J.A. Amaral, and M.T. Cochita, A survey of J. Little's formula, Int. Stat. Rev. 51 (1983) 255–278.

[56] S.O. Rice, Single-server systems – I. relations between some averages, Bell Sys. Tech. J. 11 (1962) 269–278.

[57] T. Rolski, *Stationary Random Processes Associated with Point Processes*, Lecture Notes in Statistics 5 (Springer, New York, 1981).

[58] T. Rolski and S. Stidham, Jr., Continuous versions of the queueing formulas $L = \lambda W$ and $H = \lambda G$, Oper. Res. Lett. 2 (1983) 211–215.

[59] K. Sigman, A note on a sample-path rate conservation law and its relationship with $H = \lambda G$, Department of Industrial Engineering and Operations Research, Columbia University (1990).

[60] S. Stidham, Jr., *Static Decision Models for Queueing Systems with Non-Linear Waiting Costs*, Ph.D. dissertation and Tech. Rept. 9, Department of Operations Research, Stanford University (1968).

[61] S. Stidham, Jr., On the optimality of the single-server queueing system, Oper. Res. 18 (1970) 708–732.

[62] S. Stidham, Jr., $L = \lambda W$: a discounted analogue and a new proof, Oper. Res. 20 (1972) 1115–1126.

[63] S. Stidham, Jr., Regenerative processes in the theory of queues, with applications to the alternating-priority queue, Adv. Appl. Prob. 4 (1972) 542–577.

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

[65] S. Stidham, Jr., On the relation between time averages and customer averages in stationary random marked point processes, Technical Report 79-1, Department of Industrial Engineering, North Carolina State University (1979).

[66] S. Stidham, Jr., Sample-path analysis of queues, in: *Applied Probability – Computer Science: The Interface*, vol. 2, eds. R.L. Disney and T.J. Ott (Birkhäuser, Boston, 1982) pp. 41–70.

[67] S. Stidham, Jr., On the relation between time averages and customer averages in queues, in: *Variational Methods and Stochastic Analysis*, eds. H.-J. Kimn and D.M. Chung, *Proc. Workshop in Pure Mathematics 9* (1990) pp. 243–278.

[68] S. Stidham, Jr. and M. El-Taha, Sample-path analysis of processes with imbedded point processes, Queueing Systems 5 (1989) 131–166.

[69] J. Walrand, *An Introduction to Queueing Networks* (Prentice Hall, Englewood Cliffs, NJ, 1988).

[70] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. 30 (1982) 232–233.

[71] R.W. Wolff, Sample-path derivations of the excess, age and spread distributions, J. Appl. Prob. 25 (1988) 432–436.

[72] R.W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).

[73] M.A. Zazanis, Sample path analysis of level crossings for the workload process, Department of Industrial Engineering and Management Sciences, Northwestern University.