TWO FUNDAMENTAL PRINCIPLES OF QUEUEING THEORY

Ward Whitt

AT&T Bell Laboratories, Room 2C-178, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA

Two fundamental principles of queueing theory are $L = \lambda W$ (Little’s Law) and PASTA (Poisson Arrivals See Time Averages). These principles are now well known and frequently applied. However, in recent years there have been some interesting new developments. This paper provides a brief overview.

1. Introduction

With the rapid change in teletraffic systems and technologies, it is comforting that the relevant theory for analyzing their performance remains much the same. Moreover, we are steadily gaining a better understanding of the fundamental principles underlying this queueing (or traffic) theory. Thus we can expect performance analysis to provide even more benefits in the future than it has in the past.

Our purpose here is to provide a brief overview of two fundamental principles of queueing theory: the celebrated $L = \lambda W$ relation (Little’s law) and the PASTA property (Poisson Arrivals See Time Averages). It is natural to discuss these principles together, because they both concern the relation between time averages and customer averages (or the state of a system at an arbitrary time in equilibrium and the state seen by an arrival in equilibrium). Indeed, these two principles can be regarded as special cases of a single principle. Expressed differently, appropriate generalizations of each principle can be used to establish the other principle. It is also natural to discuss these principles here, because the understanding we have today is due in large part to teletraffic research; e.g., fundamental early contributions were made by Palm [1]. The distribution seen by arrivals was also discussed at the Fifth ITC by Descloux [2].

The fundamental importance of these two principles is no doubt widely recognized. What may be surprising is that these principles have been the subject of so much recent research. This recent research not only enables us to better understand these principles as they were originally conceived, but it also enables us to see new aspects that evidently were not previously suspected.

There are two frameworks for expressing these fundamental principles. 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 two principles to a full investigation of the concept of statistical equilibrium or steady state. The stationary framework involves fairly sophisticated stochastic processes such as martingales and marked point processes. Looking carefully at both frameworks is very useful, because they are closely related. For example, there is an equivalence between the relatively elementary sample-path relation $H = \lambda G$ and a corresponding statement in the stationary random marked point process framework obtained via Campbell’s theorem, so that a result in one framework can be immediately translated into a result in the other framework, without giving a new proof: see §6.1-6.3 of Whitt [3]. Here we will primarily focus on the deterministic framework, but we will also discuss the stationary framework to some extent.

In §2 we discuss $L = \lambda W$ and in §3 we discuss PASTA and the bias (or covariance) formula. An expanded overview of $L = \lambda W$ and its extensions appears in [3]. For a recent overview of results in the deterministic framework, see Stidham and El-Taha [4]. For recent overviews of PASTA focusing on the stationary framework, see Bremaud [5] and Bremaud, Kannappatted and Mazumdar [6].

2. $L = \lambda W$

2.1 The Basic Result

The formula $L = \lambda W$ is a fundamental conservation law: under very general conditions (which do not require a queueing or traffic model), the time-average number in a system $L$ (e.g., the time-average queue length) is equal to the product of the arrival rate $\lambda$ and the average time spent in the system $W$ (e.g., the average waiting time per customer).

The relation $L = \lambda W$ is fairly obvious. The first fundamental insight, evidently due to Morse [7], pp. 22,75, was the recognition that it would be desirable to have a general proof. A fairly general proof was then proposed by his student Little [8]. One might think that the matter would then be closed, because a proof hardly seems necessary and, if you must have one, then Little provided it. Indeed, as Morse and Little probably recognized (but did not write), for each possible realization of the stochastic process, the integral of the number in system from 0 to $t$, say $\int_0^t Q(s) \, ds$, coincides with the sum of the waiting times of those customers to arrive by time $t$, say $\sum_{k=1}^{A(t)} W_k$, minus a remainder term, say $R(t)$, consisting of a portion of the sum of the waiting times of those customers still in the system at time $t$. This relationship is apparent from Fig. 1, where time appears on the horizontal axis, the customer index appears on the vertical axis, and a box
is placed at \([A_k, D_k] \times [k-1, k]\) if the \(k\)th customer arrives at time \(A_k\) and departs at time \(D_k\). We see that the cumulative processes associated with \(Q(t)\) and \(W_k\) correspond to areas of regions determined by boxes in Fig. 1 and the relationship follows.

It is evident that the remainder term \(R(t)\) should disappear when we consider limiting averages, after which we easily obtain the desired result; i.e.,

\[
L = \lim_{t \to \infty} \int_0^t Q(s) \, ds = \lim_{t \to \infty} \int_0^t A(t) \sum_{k=1}^{A(t)} W_k - R(t) = \lambda W
\]

for each sample path, provided that

\[
\lim_{t \to \infty} t^{-1} A(t) = \lambda, \quad \lim_{t \to \infty} t^{-1} \sum_{k=1}^{A(t)} W_k = W
\]

and

\[
\lim_{t \to \infty} t^{-1} R(t) = 0.
\]

This simple argument is the main idea. For the basic result, the rest is establishing appropriate sufficient conditions for (2.2).

However, Little's version of \(L = \lambda W\) in [8] is not entirely satisfactory. First, although Little probably understood the simple argument above, he was evidently looking for a clean formal proof that did not require assuming that \(t^{-1} R(t) \to 0\). In any case, Little did not present the simple argument above. Secondly, in the version he did present, the conditions are complicated and restrictive (being only for steady-state). Finally, his result has a serious flaw, which was pointed out by Brumelle [9]. In particular, Little assumes that the continuous-time process \(\{Q(t) : t \geq 0\}\) and the associated discrete-time process \(\{W_k : k \geq 1\}\) are simultaneously stationary. In fact, this is not possible unless the arrival process is a Poisson process. In general, the continuous-time and discrete-time processes should be related as the Palm transformation in an appropriate framework involving marked point processes, as in Franken et al. [10].

The first valid proof of \(L = \lambda W\) was evidently provided by Jewell [11], but it was restricted to the situation in which the queue empties infinitely often and regenerates each time it does (as in the GI/G/1 queue with \(p < 1\)). Brumelle [9] then provided a proof in a framework similar to Little's [8], but without requiring the continuous-time process be stationary.

However, it seemed to many researchers that the various stochastic assumptions were superfluous, and that it should be possible to give a relatively simple sample-path proof of \(L = \lambda W\) closely related to the intuitive argument above. A nice sample-path proof of \(L = \lambda W\) was finally provided by Stidham [12]. Suppose that the arrival counting process \(A(t)\) satisfies \(t^{-1} A(t) \to \lambda\) w.p.1 (with probability one), 0 < \(\lambda < \infty\). Stidham showed that (2.2) holds, and thus (2.1) holds, if \(\sum_{k=1}^{A(t)} W_k \to W\) where \(W < \infty\), all w.p.1.

The sample-path version of \(L = \lambda W\) established by Stidham [12] relates the long-run averages. This relation for long-run averages also applies to steady-state means when they are properly defined, but this requires some care. A satisfactory direct proof of a steady-state version of \(L = \lambda W\) was first established by Franken [13]; see §4.2 of Franken et al. [10] although it was not stated as generally as it could be. This seems to have first been done by Stidham [14]; see §3 of [3]. Franken [13] and Stidham [14] seem to have provided a proper treatment of what Little intended in [8]. Moreover, establishing a proper stationary-process framework has many significant applications beyond \(L = \lambda W\). However, for \(L = \lambda W\) itself, we feel that the sample-path version of Stidham [12] captures the essence.

The primary interest in subsequent work is in extensions, i.e., important new statements not embodied in (2.1). We turn to these now.

2.2 The Extension \(H = \lambda G\)

Motivated by design problems for queues with nonlinear waiting costs, Stidham [15] evidently first noted that the relation \(L = \lambda W\) can be extended to relate a general time average \(H\) to an associated customer average \(G\). A general theory was first developed by Brumelle [9,16]. Then a sample-path version of \(H = \lambda G\), generalizing Stidham [12], was established by Heyman and Stidham [17]. A corresponding stationary-process version of \(H = \lambda G\) follows from results for the standard G/G/s queue with the first-come first-served discipline by Franken [13], but seems to have been first stated for general systems by Stidham [14]; see §6.2 of [3].

The relation \(H = \lambda G\) can be regarded as a simple extension of \(L = \lambda W\), obtained by replacing \(Q(t)\) by \(Q^*(t) = \sum_{k=1}^{A(t)} f_k(t)\) and \(W_k\) by \(W_k^* = \int_{t_k}^t f_k(t) \, dt\), where \(f_k(t)\) is a nonnegative function for each \(k\) which is 0 outside of the interval \([A_k, D_k]\).

As before, the key idea is that \(\int_0^t Q^*(s) \, ds\) is approximately equal to \(\sum_{k=1}^{A(t)} W_k^*\). Then \(L = \lambda W\) in (2.1) is just the special case
in which $f_\lambda(t)$ is the indicator function of the interval $[A_k, D_k]$.

A familiar application of $H = \lambda G$ (using the stationary framework) is Brumelle's [9] formula relating the time-stationary workload $V(0)$ in a general stationary $G/G/s$ queue to the customer-stationary waiting time $W_0$ and service time $S_0$, i.e.,

$$E V(0) = \lambda E^0(W_0 S_0 + \frac{S_0^2}{2}). \quad (2.3)$$

(The expectation operator on the right in (2.3) is written $E^0$ to emphasize that it is with respect to the Palm measure.)

Another application of $H = \lambda G$ is to obtain results related to PASTA, i.e., relations between time-stationary distributions and associated customer-stationary distributions. This was first done in §2 of Heyman and Stidham [17] and was continued in §§3.4.3 of [10], for example. It can be shown that

$$\min\{k, s\} \mu p(k) = \lambda \pi(k - 1), \quad 1 \leq k \leq s + r, \quad (2.4)$$

in a $G/M/n/r$ model with $s$ parallel servers and $r$ extra waiting spaces, where $p(k)$ and $\pi(k)$ are the time-stationary and customer-stationary probability mass functions, respectively, $\lambda$ is the arrival rate and $\mu$ is the service rate. See §4.3 of [17] for related results obtained in the stationary framework. See Theorem 6.4 of [3] for necessary and sufficient conditions for ASTA based on $H = \lambda G$.

Brumelle [16] also applied $H = \lambda G$ to relax the higher moments of the time-stationary number in a $G/G/s$ model to higher moments of the customer-stationary waiting time. McKenna [18] recently obtained interesting generalizations of these moment relations for closed, product-form queueing networks.

2.3 Other Extensions

A different extension is to continuous and more general versions of $L = \lambda W$ and $H = \lambda G$; e.g., we allow the input to be continuous as in fluid storage models or we allow lump costs in addition to cost rates. These extensions are treated by Rolski and Stidham [19] and Glynn and Whitt [20].

Another extension is to obtain central-limit-theorem (CLT), weak-law-of-large-numbers (WLLN) and law-of-the-iterated-logarithm (LIL) versions of $L = \lambda W$ and $H = \lambda W$. These have been obtained by Glynn and Whitt; see [21] and references cited there. These results stem from the fact that $\int_0^t \lambda^* (s) ds$ is approximately equal to $\sum_{k=1}^t W_k$. The CLT versions of $L = \lambda W$ have important applications to estimation questions. In particular, if the relation $L = \lambda W$ indicate that we can estimate any one of the averages $L$, $\lambda$, and $W$ in terms of estimates for the other two. The CLT version of $L = \lambda W$ provides a basis for computing the asymptotic efficiency of these different estimators; see Glynn and Whitt [22].

Finally, we mention two other extensions of $L = \lambda W$: an ordinal version in Hafifin and Whitt [23] and a distributional version in Haji and Newell [24] and Keilson and Servi [25]; see (3.19) and (3.21) of Miyazawa [26] and §8.4 of [3]. The distributional version relates the time-stationary distribution of the number in system $G(0)$ to $A(W_0)$, the time-stationary number of arrivals during the customer-stationary waiting time. If PASTA holds and the service distribution is first-come first-served, then these distributions coincide with $(A(t) : t \geq 0)$ being Poisson and independent of $W_0$.

3. PASTA and the Bias Formula

3.1 The Basic PASTA Result

The PASTA (Poisson Arrivals See Time Averages) property is a fundamental principle that rivals $L = \lambda W$ in its applicability. It concerns an arbitrary continuous-time real-valued stochastic process $(X(t) : t \geq 0)$ with sample paths having limits from the left and right, such as a workload or queue-length process, and an associated discrete-time stochastic process $(A_k : k \geq 1)$ with $0 \leq A_k \leq A_{k+1}$ for all $k$. (We typically think of $(A_k : k \geq 1)$ being an arrival process, but that is not necessary.) To state the result in the deterministic framework, for each $t \geq 0$, let $A(t)$ count the number of $k$ such that $A_k \leq t$, and form the time average

$$V(t) = r^{-1} \int_0^t X(s) ds \quad (3.1)$$

and the customer average

$$W(t) = A(t)^{-1} \sum_{k=1}^{A(t)} X(A_k -) \quad (3.2)$$

The basic PASTA result due to Wolff [27] states that, if the counting process $(A(t) : t \geq 0)$ is a Poisson process satisfying a lack of anticipation assumption (LAA), then the time average $V(t)$ in (3.1) converges w.p.1 if and only if the customer average $W(t)$ in (3.2) does, and the limits agree. The LAA stipulates that the future of the counting process $(A(t) : t \geq 0)$ beyond $s$ should be independent of the process $(X(t) : t \geq s)$ before $s$, for each $s$.

Wolff's [27] statement of PASTA closely parallels Stidham's [22] elegant treatment of $L = \lambda W$, but this version of PASTA seems to require a harder proof. In particular, Wolff exploits martingale theory. Recent work by Georgiadis [28], Bremaud [5], Makowski, Melamed and Whitt [29], Rosenkrantz and Simha [30] and Bremaud, Kannappuri and Mazumdar [6] confirms that the essential ingredient for Wolff's [27] line of reasoning is the strong law of large numbers for martingales (or local martingales).

Paralleling Franken's [13] treatment of $L = \lambda W$, König and Schmidt [31] had previously established a PASTA result in the stationary framework; see Theorem 1.6.6 of Franken et al. [10]. Moreover, they showed that the Poisson property is actually not needed. They showed that something weaker than LAA suffices; in particular, it suffices for the forward residual time until the next point in $(A(t) : t \geq 0)$ after 0 to be independent of the state $X(0)$.

Unfortunately, the proofs of PASTA based on martingales and stationary process theory are both relatively complicated. Thus it is of interest that Melamed and Whitt [32] discovered that it is possible to give a much more elementary proof if we assume that all limits exist and focus on conditions for the limits to be equal. This leaves open some mathematically interesting questions, but it addresses the issue of primary concern.

In particular, Melamed and Whitt's approach in [32] is based on assuming that
Suppose for $A(t)$ the Poisson arrivals that see time averages (e.g., flows in certain
open queueing networks), which are not covered by the König-Schmidt [31] independence condition or the weak LAA in §2 of [32]. Hence, it remained to find more general conditions for arrivals to see time averages (ASTA). Second, it was of interest to know the relation between $\tilde{V}(t)$ and $V(t)$ over the limiting behavior of $\tilde{V}(t) = E[V(t)]$ to the limiting behavior of
\[ E[V(t)] = \lim_{t \to \infty} v \quad \text{w.p.1} \tag{3.3} \]
for $V(t)$ in (3.1) and $0 < v < \infty$.

\[
\lim_{t \to \infty} t^{-1} E \left[ \sum_{k=1}^{A(t)} X(A_k -) \right] = \lim_{t \to \infty} t^{-1} \sum_{k=1}^{A(t)} X(A_k -) = u \quad \text{w.p.1} \tag{3.4}
\]

and
\[
\lim_{t \to \infty} t^{-1} E[A(t)] = \lim_{t \to \infty} t^{-1} A(t) = \lambda \quad \text{w.p.1} \tag{3.5}
\]
for $0 < u < \infty$ and $0 < \lambda < \infty$, so that
\[
\lim_{t \to \infty} E[A(t)]^{-1} E \left[ \sum_{k=1}^{A(t)} X(A_k -) \right] = \lim_{t \to \infty} W(t) = \lambda u \tag{3.6}
\]
for $W(t)$ in (3.2).

Given (3.3)-(3.6), it suffices to relate $v$ to $\lambda u$ or, equivalently, it suffices to relate the limiting behavior of $\tilde{V}(t) = E[V(t)]$ to the limiting behavior of
\[
\tilde{W}(t) = (E[A(t)])^{-1} E \left[ \sum_{k=1}^{A(t)} X(A_k -) \right]. \tag{3.7}
\]

In Section 2 of [32], Melamed and Whitt show that $\tilde{V}(t) = \tilde{W}(t)$ for all $t$ under weaker conditions than Wolff’s PASTA assumptions, by using only elementary properties of Riemann-Stieltjes integrals. (In particular, the Poisson property is not assumed directly; the sufficient condition is for $A(t + u) - A(t)$ to be uncorrelated with $X(t)$ for all $t$ and all sufficiently small positive $u$.)

3.2 ASTA and the Bias Formula

The PASTA results of König and Schmidt [31], Wolff [27] and §2 of Melamed and Whitt [32] left open two fundamental questions. First, it was known that there are cases of non-Poisson arrivals that see time averages (e.g., flows in certain open queueing networks), which are not covered by the König-Schmidt [31] independence condition or the weak LAA in §2 of [32]. Hence, it remained to find more general conditions for arrivals to see time averages (ASTA). Second, it was of interest to know the relation between the limiting averages (or the corresponding steady-state distributions) when they are not equal.

Both these questions were answered (at least for the most part) independently by Brémaud [5], Melamed and Whitt [32,33] and Stidham and El-Taha [4]. The answer lies in what Melamed and Whitt [32] call the covariance formula and we now call the bias formula, which is perhaps most naturally stated in a stationary framework. For this purpose, let $\mu(t)$ be a conditional intensity of a point from the counting process $\{A(t): t \geq 0\}$ at time $t$ conditional upon $X(t)$, e.g.,
\[
\mu(t) = \lim_{u \to 0} u^{-1} E[A(t+u) - A(t) \mid X(t)]. \tag{3.8}
\]
Suppose that $\{\mu(t), X(t): t \geq 0\}$ is a stationary process and let $\tilde{X}$ have the (discrete-time) stationary distribution of $\{X(A_k -): k \geq 1\}$. Then the bias formula is
\[
E[f(\tilde{X})] = \frac{E[f(\mu(0), f(X(0)))]}{E[\mu(0)]} = \frac{E[f(X(0))] + \text{cov}(\mu(0), f(X(0)))}{E[\mu(0)]} \tag{3.9}
\]
where $f$ is any measurable real-valued function and $\text{cov}$ is the covariance. Obviously, (3.9) shows how the distributions of $\tilde{X}$ and $X(0)$ are related. Moreover, (3.8) and (3.9) imply that $\tilde{X}$ is distributed the same as $X(0)$ if and only if the conditional intensity $\mu(0)$ is independent of $X(0)$ or, equivalently (from (3.8)), if and only if $\mu(0) = E[\mu(0)]$; see Theorem 4 of [32].

A technical difficulty with (3.9) is the conditional intensity $\mu(t)$ in (3.8) and its properties. The conditional intensity is easy to define and often fairly reasonable to evaluate when $X(t) = g(Z(t))$ for some function $g$ and Markov process $\{Z(t): t \geq 0\}$ and the process $\{A(t): t \geq 0\}$ counts designated jump transitions in $\{Z(t): t \geq 0\}$, as can be seen from the examples in §5 of [32]. Alternatively, it is possible to use the stochastic intensity of martingale theory, as in [33]. Then we can obtain an average version of (3.9), which does not require a stationary framework.

Brémaud [5] independently arrived at the bias formula (3.9) without assuming (3.3)-(3.6). In particular, Brémaud (1989) uses a stationary framework together with martingale structure. He shows that (3.9) holds in this framework, where $\mu(t)$ is the stochastic intensity. Indeed, Brémaud notes that the bias formula lies at the heart of the connection between the martingale and stationary-process theories of point processes, because (3.9) is equivalent to an expression for the stochastic intensity derived by Papangelou [34] as a consequence of his theorem stating that an underlying probability measure $P$ supporting a stationary point process $N$ is absolutely continuous with respect to the associated Palm measure $P^0$ on the sigma field $\mathcal{F}_N$ if and only if $N$ admits a stochastic intensity.

Stidham and El-Taha [4], pp.141-2, also obtained a sample-path-average analog of (3.9) in a discrete-state setting via a sample-path argument. (An extension to general state spaces appears in §6.3 of [3], as was mentioned in §2.2 above.) Their argument is especially appealing because it provides the natural analog for PASTA of the simple statement and proof of $L = \lambda W$ in Stidham [12]. Moreover, their proof is very elementary. However, additional stochastic arguments are needed to translate their result into (3.9).

Further work on PASTA, ASTA and the bias formula appears in Makowski et al. [29], König and Schmidt [35,36], Miyazawa and Wolff [37] and Brémaud et al. [6]. Makowski et al. [29] and Brémaud et al. [6] show how a w.p.1 limiting average version of (3.9) can be established without assuming (3.3)-(3.6), using the SLLN for martingales. König and Schmidt [35,36], Miyazawa and Wolff [37] and Brémaud et al. [6] establish general conditions for an intensity to exist and for ASTA when an intensity need not exist.

A significant application of the bias formula (3.9) beyond establishing ASTA is to give a new proof of the arrival theorems for open and closed product-form queueing networks; see §5 of Melamed and Whitt [32] and p. 108 of Brémaud [5].
3.3 Other Extensions

A useful extension of PASTA is conditional-PASTA, first established by van Doorn and Regterschot [38] and Georgiadis [28]. An extension to conditional-ASTA is given in §6 of Melamed and Whitt [32]. Further discussion appears in König and Schmidt [36] and Brémaud et al. [6].

Another question of interest is when ASTA implies that the counting process must be Poisson, which has been called anti-PASTA by Green and Melamed [39]. Conditions for anti-PASTA are given in §3 of Melamed and Whitt [33], Wolff [40], §4 of Makowski, et al. [29], Miyazawa and Wolff [37], §6 of Brémaud et al. [6] and references cited there. This is related to more general questions about characterizing flows; see Serfozo [41].

Discrete-time analogs of PASTA have also been considered; see Halin [42], Georgiadis [28], Makowski et al. [29], §5 of Brémaud et al. [6] and references cited there.

When ASTA does not hold, it is sometimes possible to make stochastic comparisons; see König and Schmidt [43], §4.3 of Franken et al. [10], Whitt [44], Niu [45], and Corollary 2, p. 164, of Melamed and Whitt [32].

Finally, Yao [46] and Glynn, Melamed and Whitt [42] establish CLTs related to ASTA and discuss applications to estimation, parallelising Glynn and Whitt’s [21,22] work related to $L = \lambda W$.

References


