A Proof for the Queuing Formula: L= $\lambda $W

John D. C. Little

Operations Research, Volume 9, Issue 3 (May - Jun., 1961), 383-387.

Stable URL:
http://links.jstor.org/sici?sici=0030-364X%28196105%2F196106%299%3A3%3C383%3AAPFTQF%3E2.0.CO

A PROOF FOR THE QUEUING FORMULA: $L=\lambda W$

John D. C. Little

*Case Institute of Technology, Cleveland, Ohio**

(Received November 9, 1960)

In a queuing process, let $1/\lambda$ be the mean time between the arrivals of two consecutive units, L be the mean number of units in the system, and W be the mean time spent by a unit in the system. It is shown that, if the three means are finite and the corresponding stochastic processes strictly stationary, and, if the arrival process is metrically transitive with nonzero mean, then $L=\lambda W$.

HEURISTIC arguments are sometimes given to show that, in a steady-state queuing process, the following formula holds:

$$L=\lambda W, \tag{1}$$

where L = expected number of units in the system
W = expected time spent by a unit in the system
$1/\lambda$ = expected time between two consecutive arrivals to the system.

Expression (1) is of interest because it is sometimes easier to find L than W (or vice versa) in solving a queuing model.

A brief plausibility argument for rather general validity of (1) is given by MORSE (reference 1, p. 22). He goes on to prove it in a number of specific models. GALLIHER[2] establishes it for the case of Poisson arrivals which have a rate independent of queue length and which come to a multiple channel facility having a first-come, first-served discipline. We shall prove it under assumptions considerably more general.

By a *queuing process* will be meant a mathematically specified operation in which units arrive, wait, and then leave. It is presumed that the operation thereby generates three well-defined stochastic processes:

$\{n_t, -\infty<t<\infty\}$ = the number of units in the system at time t
$\{w_r, -\infty<r<\infty\}$ = the time spent in the system by the rth arriving unit
$\{\tau_r, -\infty<r<\infty\}$ = the time between the arrivals of the rth and $(r+1)$st units to the system.

These processes are defined on some space Ω and any point $\omega\in\Omega$ selects a

* This work was supported in part by the Air Development Center, Wright-Patterson Air Force Base, United States Air Force, under Contract No. AF33(616)-6446.

function and two sequences,

$$n_t(\omega), \qquad \{w_r(\omega)\}, \qquad \{\tau_r(\omega)\},$$

which represent a specific realization of the queuing operation over all time. The random variables n_t, w_r, and τ_r are nonnegative.

The time of arrival of the rth unit will be denoted t_r and is defined by

$$t_{r+1}(\omega)=t_r(\omega)+\tau_r(\omega).$$

For convenience we choose

$$t_1(\omega)\geqq 0, \qquad t_0(\omega)<0.$$

The following relation is taken to be part of the definition of a queuing process: Let

$$u(x)=\begin{cases}1 & \text{for} \quad x\geqq 0,\\ 0 & \text{for} \quad x<0;\end{cases}$$

then, for any ω, $$n_t=\textstyle\sum_{-\infty}^{+\infty} u(t-t_j)\, u(t_j+w_j-t). \tag{2}$$

This relation says that the number in the system at t is the number of units whose time of arrival is before (or equal to) t and time of departure is after (or equal to) t.

THEOREM 1: *If, in a queuing process, (i) each of the stochastic processes n_t, w_r, and τ_r is strictly stationary with finite mean, and (ii) the τ_r process is metrically transitive with mean $T\equiv 1/\lambda>0$, and, if we let*

$$L(\omega)=\lim_{t\to\infty}\frac{1}{t}\int_0^t n_s(\omega)\,ds, \qquad W(\omega)=\lim_{m\to\infty}\frac{1}{m}\sum_1^m w_j(\omega),$$
$$T(\omega)=\lim_{m\to\infty}\frac{1}{m}\sum_1^m \tau_j(\omega), \tag{3}$$

then, with probability 1, *the limits in* (3) *exist, are finite, and satisfy*

$$W(\omega)=T(\omega)\,L(\omega). \tag{4}$$

The existence and finiteness of the limits is an immediate consequence of the ergodic theorems for strictly stationary stochastic processes (*see* DOOB, reference 3, pp. 465 and 515).

Consider a specific point $\omega\in\Omega$. Let t_m denote the length of the interval $[0, t_m(\omega))$. Define

$$L_m(\omega)=\frac{1}{t_m}\int_0^{t_m} n_s(\omega)\,ds, \qquad W_m(\omega)=\frac{1}{m}\sum_1^m w_j(\omega),$$
$$T_m(\omega)=\frac{1}{m}\sum_0^m \tau_j(\omega). \tag{5}$$

In order to take the limits of (5) simultaneously, we first show that as $m \rightarrow \infty$, $t_m \rightarrow \infty$ w. p. 1 (with probability one). By the ergodic theorem, the metric transitivity of the τ_r process, and its nonzero mean, we have $1/T_m(\omega) \rightarrow 1/T(\omega) = 1/T < \infty$ w. p. 1. Let $\alpha = \tau_0(\omega) - t_1(\omega)$. We see that $0 < \alpha < \infty$ w. p. 1. Then $1/T_m(\omega) = m/(t_m + \alpha) \rightarrow \lim m/t_m$ w. p. 1; $\lim m/t_m < \infty$ w. p. 1; so that $m \rightarrow \infty$ implies $t_m \rightarrow \infty$ w. p. 1.

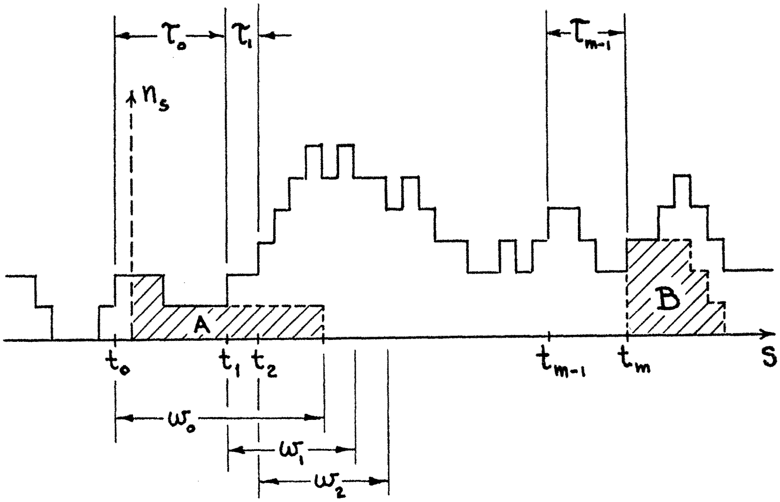

Fig. 1. Part of a specific queuing realization ω, showing the number in the system at time s, n_s; the wait in the system for the rth arrival, w_r; and the interarrival time started by the rth arrival, τ_r. (The figure is drawn for the case of departure in order of arrival, but this is not required for the proofs in the text.)

Integrating (2) for fixed ω gives

$$\int_0^{t_m} n_s \, ds = \sum_1^m w_j + \sum_{j \leqq 0} v(w_j + t_j) - \sum_{j \leqq m} v(w_j + t_j - t_m), \qquad (6)$$

where $v(x) = x$ for $x > 0$ and $v(x) = 0$ for $x \leqq 0$. The situation is illustrated in Fig. 1. The area under the curve n_s from 0 to t_m is, except for certain carry-over effects at the ends of the interval, the sum of the waiting times of the units that arrived during the interval. These carry-over effects are indicated by the areas A and B, which correspond to the last two sums on the right in (6).

Dividing by m and using (5) gives

$$W_m - T_m L_m = (1/m) \textstyle\sum_{j \leqq m} v(w_j + t_j - t_m) \\ - (\alpha/m) L_m - (1/m) \textstyle\sum_{j \leqq 0} v(w_j + t_j).$$

The last two terms on the right can be shown to go to zero w. p. 1 as

$m \rightarrow \infty$: In the last term the sum consists of a finite number (n_0) of finite terms except on the union of (n_0+1) ω-sets of probability zero. Thus the sum is finite w. p. 1, and, since it is independent of m, the desired limit is zero w. p. 1. In the next to last term, $L_m \rightarrow L(\omega) < \infty$ and $\alpha/m \rightarrow 0$ w. p. 1. Thus

$$W(\omega)-T(\omega)L(\omega)=\lim(1/m)\ \sum\nolimits_{j \leqq m} v(w_j+t_j-t_m) \geqq 0 \text{ w. p. 1.}$$

If now we consider the interval $(t_{-m}(\omega), 0]$ and define L_{-m}, W_{-m}, and T_{-m} analogously to their counterparts above, e.g.,

$$L_{-m}=[1/(-t_{-m})] \int_{t-m}^{0} n_s(\omega)\, ds,$$

then the symmetry of the ergodic theorems with respect to time and arguments the same as used previously yield

$$W(\omega)-T(\omega)L(\omega)=-\lim(1/m)\sum\nolimits_{j \leqq -m} v(w_j+t_j-t_{-m}) \leqq 0 \text{ w. p. 1.}$$

Therefore, $$W(\omega)=T(\omega)L(\omega) \text{ w. p. 1}$$
as was to be shown.

THEOREM 2: *Let*

$$L=\mathrm{E}\{n_0\}, \qquad W=\mathrm{E}\{w_0\}, \qquad T=\mathrm{E}\{\tau_0\},$$

then, under the hypotheses of Theorem 1,

$$W=TL.$$

The ergodic theorems state that for almost all ω the limits (3) are the conditional expectations:

$$L(\omega)=\mathrm{E}\{n_0|\mathfrak{I}_a\}, \qquad W(\omega)=\mathrm{E}\{w_0|\mathfrak{I}_b\}, \qquad T(\omega)=\mathrm{E}\{\tau_0|\mathfrak{I}_c\}$$

where $\mathfrak{I}_a$, $\mathfrak{I}_b$, and $\mathfrak{I}_c$ are the Borel fields of invariant subsets for the corresponding processes. Since the τ_r process is metrically transitive,

$$T(\omega)=T,$$

and (4) becomes $$W(\omega)=TL(\omega) \text{ w. p. 1.}$$

Integration over Ω gives, by definition of conditional expectation,

$$W=TL$$

as was to be shown.

DISCUSSION

THEOREM 2 is the principal result for applications and shows that (1) is a valid relation among phase averages. Theorem 1, on the other hand, is perhaps more basic for it shows that an equivalent of (1) using time averages holds with probability one for any specific realization of the queuing process.

The results are remarkably free of specific assumptions about arrival and service distributions, independence of interarrival times, number of channels, queue discipline, etc. A requirement is made for strict stationarity (although this is probably not the weakest requirement possible), but the steady state in most current queuing models would appear to be strictly stationary. Similarly, in cases of practical interest, the arrival process is likely to be metrically transitive.

Notice that the definition of what constitutes the 'system' is left flexible. In conventional usage, the number of units in the system refers to the number in queue plus those in service. The theorem here, however, only requires consistency of meaning in the phrases, 'number of units in the system,' 'time spent in the system,' and 'arrival to the system.' Thus, if we choose to label the queue as the system and let L_q and W_q refer to the mean number and mean wait in queue, we obtain

$$L_q = \lambda W_q.$$

Similarly, if we have a model with priority classes $i=1, 2, \cdots, p$, and let L_i be the mean number of priority i units present, W_i the mean wait of a priority i unit, and $1/\lambda_i$ the mean interarrival time for priority i units, then

$$L_i = \lambda_i W_i.$$

Morse (reference 1, p. 75) asks when (1) does not hold. As an example, we cite a type of model, used in his book and elsewhere, in which arrivals come with rate λ but not all arrivals join the system. Then (1) does not hold. However, inspection of the theorem shows that (1) will hold if λ is redefined to include only those arrivals that join the system. Alternatively, we can say that the units that do not join have a zero waiting time in the system and include them in the calculation of W. This too will make (1) hold.

THE AUTHOR thanks DR. H. NEWTON GARBER for several constructive suggestions about the proof. In particular, equation (2) and its explicit definitional implications are his.

REFERENCES

1. P. M. MORSE, *Queues, Inventories and Maintenance*, Wiley, New York, 1958.
2. H. P. GALLIHER, *Notes on Operations Research 1959*, Chap. 4, Technology Press, Cambridge, 1959.
3. J. L. DOOB, *Stochastic Processes*, Wiley, New York, 1953.