M. A. Leibowitz

A Note on Some Fundamental Parameters of Multiqueue Systems

In a previous paper, an approximate method was developed for treating a class of multiqueue problems.[1] These problems are of considerable interest in the design of communication systems with many terminals. The purpose of this note is to obtain expressions for some parameters often used to characterize such systems and to draw an important distinction between two different types of averages—a distinction which can lead to serious consequences if overlooked.

The situation considered in Ref. 1 is this: N queues of unrestricted length are served in cyclic order by a single server. At any queue, the server serves all units which were on that queue at the moment that it arrived. The input to each queue is Poisson with average interarrival time $1/\lambda$, and $s(t)$, $w(t)$ are the probability densities of the *service time* for a unit and the *walking time* (i.e., the time required by the server to go from one queue to the next).

In Ref. 1, the quantity of interest was p_n, the stationary probability that there are n units on a given queue *at the moment of the server's arrival.* This probability must be distinguished from another and generally more useful quantity: namely, r_n, the probability that there are n units on the queue at some *randomly chosen instant.* The following alternative definitions may make this distinction clearer. Consider k successive arrivals of the server at the same queue taking place over a period of time S. Then for k and S sufficiently large, p_n will be very nearly the fraction of the k arrivals at which the server finds the queue to be of length n, while r_n will be the fraction of time S that the queue is of this length.

To determine r_n it is necessary to introduce another quantity which is in itself important. Let $P(T)$ be the probability density of the *scan time*, i.e., the time required to serve each queue once. $P(T)$ is analogous to p_n in that it is related to the frequency at which scans of a given length occur and not to the proportion of time they occupy.

Since the scan time, T, is also the time between successive arrivals of the server at the same queue,

$$p_n = \int_0^\infty e^{-\lambda T} \frac{(\lambda T)^n}{n!} P(T)dT \tag{1}$$

and

$$G(x) = \sum_{n=0}^{\infty} p_n x^n = \int_0^\infty e^{-\lambda T(1-x)} P(T)dT\,. \tag{2}$$

Relations between the moments of $\{p_n\}$ and $P(T)$ are easily obtained by differentiating Eq. (2) with respect to x, and explicit expressions can then be found by using the results of Ref. 1. We have, e.g., for the expected value of T,

$$\overline{T} = \frac{\bar{n}}{\lambda} = \frac{N\overline{w}}{1 - N\lambda/\mu}\,, \tag{3}$$

where $1/\mu$ and $\overline{w}$ are the mean service and walking times, respectively.

The *efficiency*, E, can be defined as the fraction of time spent by the server in actually servicing the queues. Thus,

$$E = \frac{\overline{T} - N\overline{w}}{\overline{T}} = N\lambda/\mu\,. \tag{4}$$

Note that E is *independent* of the walking time.

To find the mean *waiting time*, $\overline{W}$, of a unit, it is tempting to reason as follows: When a given unit arrives, the server could be at any queue in the system, and on the average this queue will be half-way removed from the unit. Thus half a scan of the system will be required before the unit can be served, so that $\overline{W} = T/2$. This leads to the remarkable conclusion that if the walking time were zero, units would never have to wait (cf. Eq. 3 above). However, this conclusion and the argument behind it is *wrong.*

To see this, consider how $\overline{T}$ would be determined experimentally. One would measure the time T_j ($j = 1, 2, \cdots, k$) of k *successive* scans of the system

(beginning with some specified queue) and form $(T_1 + T_2 + \cdots + T_k)/k$ which for sufficiently large k would be nearly $\overline{T}$. Another average, $\overline{T}'$, is obtained if we observe the system at k *random* instants, record the times T_j' $(j = 1, 2, \cdots k)$ of the scans then occurring, form $(T_1' + T_2' + \cdots + T_k')/k$ and let T' be the limit of this quotient as $k \to \infty$. Unless the scan time is always the same (i.e., deterministic) $\overline{T}' > \overline{T}$. The reason for this is that if the instants of observation are chosen at random, there is a greater tendency to observe the system during the longer scan times than the shorter, simply because the former last longer. In fact, if $R(T)$ is the probability density of scan times chosen by random observation as above, then

$$R(T) = TP(T)/\overline{T}. \tag{5}$$

Indeed, consider a very long period of time S. Then $R(T)dT$ is that fraction of the time S which is occupied by scans whose lengths lie in the small interval $[T, T + dT]$; $R(T)$, therefore, is analogous to $\{r_n\}$. Now in the time S approximately $S/\overline{T}$ scans occur. Of these, $SP(T)dT/\overline{T}$ will have lengths between T and $T + dT$, and hence they will require a time $STP(T)dT/\overline{T}$. But this is just $SR(T)$.

It follows from (5) that

$$\overline{T}' = \overline{T^2}/\overline{T}. \tag{6}$$

Now units enter the queues at *random* times. Since these times can equally well be regarded as random instants of observation, the correct relation between the waiting time and the scan time is

$$\overline{W} = \overline{T}'/2 = \overline{T^2}/2\overline{T} = \frac{1}{\lambda}(\overline{n^2} - \bar{n})/\bar{n} \tag{7}$$

by Eqs. (2) and (6).

Because of the complex interactions of the queues, an exact calculation of $\overline{W}$ seems prohibitively complicated for $N > 2$. However, approximate expressions can be found from the results of Ref. 1. Thus, correct to terms $O(\lambda^2)$ as $\lambda \to 0$, we have (if the variance $\delta^2(w) \equiv \overline{w^2} - \bar{w}^2 = 0$ and the time scale is chosen so that $\mu = 1$)

$$\overline{W} = \frac{1}{2}\left[N\lambda\overline{s^2} + \frac{N\bar{w}}{1 - N\lambda}\right]. \qquad \text{(cf. Ref. 1, Eq. 25)}$$

Note that the waiting time does *not* tend to zero as the walking time becomes indefinitely small.

Finally we obtain an expression for the probability r_n in terms of $P(T)$. Let $P(n, t)$ be the probability density that if one observes a given queue at some random instant (a) a scan beginning with this queue has been in progress for exactly a time $t/2$, and (b) the queue contains n units. Since the probability density of (a) is

$$\int_t^\infty P(T)dT/\overline{T},$$

$P(n, t)$ is given by[3]

$$P(n, t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}\int_t^\infty P(T)dT/\overline{T}.$$

Hence, since

$$r_n = \int_0^\infty P(n, t)dt,$$

it follows that

$$H(z) \equiv \sum_{n=0}^{\infty} r_n z^n = \int_0^\infty \sum_{n=0}^{\infty} z^n P(n, t)dT$$

$$= \frac{1}{\overline{T}}\int_0^\infty e^{-\lambda t(1-z)}dt \int_t^\infty P(T)dT. \tag{8}$$

From this equation, the moments of the distribution $\{r_n\}$ can be related to those of $P(T)$ and $\{p_n\}$.

References and footnotes

1. M. A. Leibowitz, *IBM Journal* **5**, 204 (1961).
2. P. M. Morse, *Queues, Inventory, and Maintenance*, John Wiley and Sons, New York, 1958.
3. The following equations assume that at the beginning of the scan the queue is empty, i.e., we do not count the units which are on the queue when the server arrives. We can suppose, e.g., the server takes them off the queue and serves them elsewhere. This is actually the case in many communication systems.

Received May 29, 1962