THE DEPENDENCE OF DELAYS IN TANDEM QUEUES

By P. J. Burke

Bell Telephone Laboratories, Holmdel, N. J.

Edgar Reich [1] proved that in single-server tandem queues, the durations of time spent by a customer in successive systems are independent. In this connection Reich stated "... if the waiting times are defined so as *not* to include the service times ... the question of mutual independence of these quantities ... is apparently an open problem."

It is proved below that these waiting, or delay, times are not mutually independent.

We assume two single-server queueing systems Q_1 and Q_2 in tandem that have exponential service-time distributions with respective service rates μ_1 and μ_2. The customers arrive in a Poisson process at Q_1, and as soon as their service is completed in Q_1 they enter Q_2. The arrival rate, or parameter of the Poisson arrival process, is λ; and $\lambda < \mu_1$, $\lambda < \mu_2$. We assume further that statistical equilibrium obtains with respect to the distributions of the numbers of customers in Q_1 and in Q_2. Under these conditions, it has been proved that the input to Q_2 is a Poisson process with parameter λ ([2] or [3] p. 45). Although the further assumption of order-of-arrival service is required for Reich's proof of the independence of the times spent in the successive systems, the present result holds for all queue disciplines that do not allow defections or pre-emption.

The method, in essence, is a comparison of the conditional delay distribution in Q_2, given that there is no previous delay in Q_1, with the unconditional (marginal) delay distribution in Q_2.

Let W_j be the delay of a customer in Q_j. For W_1 to be independent of W_2, it is necessary that their joint distribution function be factorizable into the marginal distribution functions. In particular, if $\Pr\{\cdot\}$ designates the probability of the event in the braces, it is necessary that $\Pr\{W_1 = 0, W_2 = 0\} = \Pr\{W_1 = 0\} \cdot \Pr\{W_2 = 0\}$, or $\Pr\{W_2 = 0\} = \Pr\{W_2 = 0 \mid W_1 = 0\}$.

It will be proved that in fact $\Pr\{W_2 = 0\} < \Pr\{W_2 = 0 \mid W_1 = 0\}$, and hence that the necessary condition for the independence of W_1 and W_2 is contradicted.

We know that $\Pr\{W_2 = 0\} = 1 - \lambda/u_2$. We shall show that $\Pr\{W_2 = 0 \mid W_1 = 0\} > 1 - \lambda/u_2$.

Consider, in Q_1, an arrival epoch, T, of an undelayed call. By Jackson's theorem [4] the number of customers in Q_2 at T has the unconditional equilibrium state distribution. If N_2 designates the number of customers in Q_2 at T,

$$\Pr\{N_2 = k\} = (1 - \lambda/\mu_2)(\lambda/\mu_2)^k.$$

Therefore the time for the customers in Q_2 at T to complete service in Q_2 has the distribution given by a mass at the origin, $1 - \lambda/\mu_2$, and a density function,

Received 6 May 1963; revised 13 November 1963.

$$D(t) = \sum_{k=1}^{\infty} (1 - \lambda/\mu_2)(\lambda/\mu_2)^k [\mu_2 e^{-\mu_2 t}]_k ;$$

where $[\cdot]_k$ represents the k-fold convolution of the density in the square brackets (with itself). It is well known that

$$[\mu_2 e^{-\mu_2 t}]_k = \mu_2 e^{-\mu_2 t} (\mu_2 t)^{k-1}/(k-1)!$$

Hence

$$D(t) = (\lambda/\mu_2)(\mu_2 - \lambda) e^{-(\mu_2 - \lambda)t}. \tag{1}$$

(This result can be written down almost at once if it is realized that Jackson's theorem implies the independence of the virtual delay in Q_2 and the state of Q_1 at any instant.)

The customer entering Q_1 at T will be undelayed in Q_2 if and only if his service time in Q_1 is greater than the total time required to serve all customers who were in Q_2 at T. The service time in Q_1 has the density $\mu_1 e^{-\mu_1 t}$; and hence the probability that this service time exceeds the virtual delay in Q_2 at T is given by

$$P = 1 - \frac{\lambda}{\mu_2} + \int_0^{\infty} e^{-\mu_1 t} D(t)\, dt = 1 - \frac{\lambda}{\mu_2} + \frac{\lambda}{\mu_2} \frac{(\mu_2 - \lambda)}{\mu_1 + \mu_2 - \lambda};$$

and since $\lambda < \mu_2$, $P > 1 - \lambda/\mu_2$.

But since $P = \Pr\{W_2 = 0 \mid W_1 = 0\}$, this contradicts the necessary condition for the independence of the delays, namely that P be equal to $1 - \lambda/\mu_2$.

It should be noted that the assumption of order-of-arrival service (or that of any other queue discipline) does not enter the above argument. The only random variables which are relevant are the service time in Q_1 and the virtual delay in Q_2, neither of which depends on queue discipline.

If we now add the assumption of order-of-arrival service in Q_2, we can easily calculate the conditional delay distribution in Q_2 for calls undelayed in Q_1.

The delay in Q_2 will exceed some number t, given that the delay in Q_1 was zero, if and only if the service time in Q_1 is less than the virtual delay in Q_2 (at the arrival of the call in question at Q_1) minus t. Therefore

$$\Pr\{W_2 > t \mid W_1 = 0\} = \int_t^{\infty} [1 - e^{-\mu_1(x-t)}] \frac{\lambda}{\mu_2} (\mu_2 - \lambda) e^{-(\mu_2 - \lambda)x}\, dx$$

$$= \frac{\lambda}{\mu_2} \left(\frac{\mu_1}{\mu_1 + \mu_2 - \lambda} \right) e^{-(\mu_2 - \lambda)t},$$

and further routine calculation gives

$$\Pr\{W_2 > t \mid W_1 > 0\} = \mu_1(\mu_1 + \mu_2 - \lambda)^{-1} e^{-(\mu_2 - \lambda)t}.$$

REFERENCES

[1] Reich, Edgar (1963). Note on queues in tandem. *Ann. Math. Statist.* **34** 338–341.

[2] Burke, Paul J. (1956). The output of a queueing system. *Operations Research.* **4** 699–704.

[3] Takács, Lajos (1962). *Introduction to the Theory of Queues.* Oxford Univ. Press.

[4] Jackson, R. R. P. (1954). Queueing systems with phase type service. *Operational Research Quart.* **5** 109–120.