Equilibrium Delay Distribution for One Channel With Constant Holding Time, Poisson Input and Random Service

By PAUL J. BURKE

(Manuscript received February 19, 1959)

The equilibrium delay distribution is found for a single-server queueing system with Poisson input, random service and constant holding time. Curves are presented for various occupancy levels, and these are compared with their queued-service constant-holding-time and random-service exponential-holding-time counterparts.

In many situations involving waiting lines — for example, when customers are being served at a bargain counter in a crowded store — the ideal queue discipline (service order) of first-come first-served is not achieved. Instead, the service order tends to be at least somewhat random, and the probability of long delays is thereby increased over what it would be for the strict first-come first-served discipline. Unfortunately for the analyst, when the queue discipline is somewhere between order-of-arrival and random — as is often the case in practice — the problem of calculating the delay distribution seems to be intractable. If the service order is assumed to be actually random, however, then this problem can sometimes be solved, and the delay distribution thus found is useful as a kind of bound on the distributions to be expected in those cases where the queue discipline deviates from order-of-arrival service toward randomness.

The term "bound" as used here does not mean a bounding function in the strict sense. Actually, the delay distributions for models which differ only in queue discipline will cross each other, and hence no individual distribution can be a true bound for a family of such distributions. However, the longer delays are generally of more interest in waiting-line problems than are the shorter ones, and it is true that, other things being equal, the probability of sufficiently long delays is greater for random service than for order-of-arrival service.

Although the assumption of random service does not provide this type of bound in all situations, as would the assumption that the service order is last-come first-served, it is clearly more realistic than the latter assumption in many cases and will provide a closer bound, or approximation, to the actual delay distribution in these cases.

In addition to its usefulness as a boundary case, a queueing model which involves random service and constant holding times is of direct interest in certain telephone switching applications. For example, it is approximated, under some circumstances, at the marker connectors of the No. 5 Crossbar system, and it may have further application in electronic switching systems.

Of all possible holding-time distributions, the exponential and constant distributions have been studied most intensively in connection with queueing systems. The delay distribution was first obtained for a queueing system with constant holding times at least as early as 1909 by Erlang (Ref. 1, pp. 133–137). A solution to the delay problem for exponential holding times was published in 1917 by the same author (Ref. 1, p. 138–155). In both of these cases the service was order-of-arrival (queued) rather than random. The first attempt to obtain a delay distribution when the service order was random was published in 1942 by Mellor,[2] but this was not completely correct. A correct formulation of a random-service problem was obtained in 1946 by Vaulot,[3] the solution to which was given by Pollaczek.[4] The same problem also was solved by Palm.[5] A method for computing the delay distribution was published by Riordan in 1953.[6] The random-service problem solved by Vaulot and Pollaczek involved an exponential holding-time distribution. The present study is apparently the first attempt to combine the queue discipline of random service with holding times which are constant.

The model considered here is characterized by the following:

i. *Random input* — the probability that a call will arrive during any infinitesimal interval of length dt is proportional to dt within infinitesimals of higher order, and is independent of the state of the system, arrival times of previous calls or any other conditions whatever. It is equivalent to say that the call arrivals constitute a Poisson process.

ii. *Constant holding times* — the service time of each call is the same constant, taken here to be unity.

iii. *Random service* — if there are n calls waiting for service at the instant of a completion of service, the probability that any particular one of the calls will be served next is $1/n$. The server is never idle when there are calls waiting to be served.

iv. *No defections* — all calls wait in the system until they are served.

v. *Statistical equilibrium* — the distribution of the number of calls in the system is stationary, i.e., independent of time. Under the above assumptions, this will be the case when the arrival rate is less than one call per service interval and the system is given enough time to "settle down." Mathematically, the condition is assured by the assumption that the initial distribution of the number of calls in the system is the equilibrium distribution.

The over-all delay distribution is obtained below by decomposing it into a weighted sum of conditional delay distributions, depending on the state of the system at the epoch (instant) of the first departure (completion of service) following the arrival of the call. It suffices to define the state of the system at the departure epochs as the number of calls remaining in the system. (The call just completing service is not counted.) Each delay consists of two parts. The first part of the delay is the time from the arrival of the call in question until the first departure epoch following the arrival. The second part is the time from this departure epoch until the call in question gains service. The first part has a continuous distribution over the interval $[0 - 1]$; the second part is distributed over the nonnegative integers.

Thus, in Fig. 1 the call that arrives at time a_2 suffers a delay $d_1 - a_2$, which may vary from zero to a full holding time, until the first departure after its arrival. At time d_1 the call that arrived at a_2 will surely gain service, since it is the only call in the system at that time, and hence the integral part of the delay for this call will be zero units of time with probability one. In contrast, the call which arrives at a_3 will have to compete for service at d_2 with another call, and hence the integral part of its delay will have a probability of one-half to be zero and an equal probability to be greater than zero. In general, the integral part of the delay will have a (discrete) probability distribution that depends

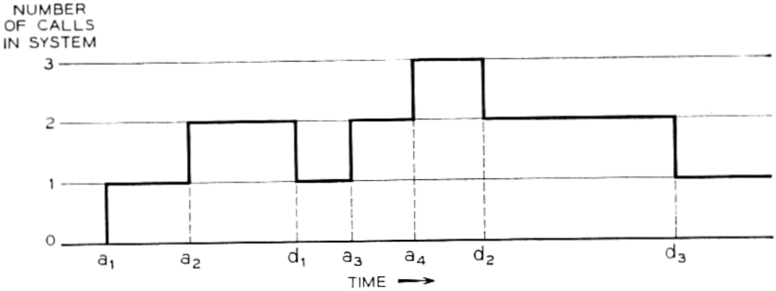


Fig. 1 — Number of calls in the system as a function of time.

on the number of calls in the system at the first opportunity that the call in question has to gain service. The determination of this probability distribution is the major portion of the task of evaluating the over-all delay distribution.

It might be pointed out that the equilibrium state probabilities are the same whether they are considered over the whole process or only over the set of discrete instants of time consisting of the departure epochs. This is shown by the fact that the generating function for the state probabilities given by Kendall,[7] which refers to the departure epochs, is the same, when specialized to constant holding times, as that of Crommelin[8] specialized to one server; and that the latter refers to the entire process.

What is needed now is the probability that a call, conditional on its being delayed, will be one of n calls remaining in the system at the first departure epoch after its arrival. It turns out that the latter probability is the same as the unconditional probability of $n - 1$ calls in the system, as shown by the following argument.

An arriving delayed call will be one of n calls in the system just after the next departure following its arrival in one of $n + 1$ mutually exclusive ways: there were k calls in the system at the last previous departure epoch before the arrival of the call in question and $n - k$ other calls arrived during the service interval, $k = 1, \cdots, n$; or there were zero calls in the system at the last previous departure epoch and $n - 1$ other calls (besides the call being served) arrived during the service interval. If $\Pr\{n \mid \lambda\}$ represents the desired probability and $P_k(\lambda)$ represents the unconditional probability of k calls in the system when λ is the arrival rate,

$$\Pr\{n \mid \lambda\} = [P_0(\lambda) + P_1(\lambda)]\, p(n-1,\lambda) + \cdots + P_n(\lambda)\, p(0,\lambda), \qquad (1)$$

where $p(k,\lambda)$, is an individual Poisson term, i.e., the probability of k arrivals during a service-time interval. However, (1) is exactly Crommelin's equation for P_{n-1} in the one-server case. Therefore

$$\Pr\{n \mid \lambda\} = P_{n-1}(\lambda). \qquad (2)$$

With the dependence on λ suppressed hereafter, let the conditional probability that the delay is not greater than t, given that the delayed call is one of n calls waiting for service at the first post-arrival departure epoch, be denoted $G(t \mid n)$. Let the resultant delay distribution for delayed calls be denoted $F(t)$. Then

$$F(t) = \sum_{n=1}^{\infty} P_{n-1}\, G(t \mid n). \qquad (3)$$

It remains to evaluate $G(t \mid n)$.

Owing to the queue discipline of random service, the delay suffered by a call between the instant of its arrival and the first post-arrival departure epoch is independent of the delay subsequent to this epoch. Also, it is known that the initial delay of a fractional part of a holding time is uniformly distributed over a service interval, owing to the property of random arrivals. Furthermore, conditioning on the number in the system at the first post-arrival departure epoch does not affect the independence property or the uniform distribution of the fractional delay (as would, for example, conditioning on the number in the system at the instant of arrival). Let the delay be represented by T, the integral part of T by T' and the fractional part by T''. Similarly, let the quantity t be decomposed into t' and t''. Then

$$\begin{aligned} G(t \mid n) &= \Pr\{T \leqq t \mid n\} \\ &= \Pr\{T' < t' \mid n\} + \Pr\{T' = t' \mid n\} \Pr\{T'' \leqq t''\} \end{aligned} \tag{4}$$

because of the independence of the integral and fractional parts of the delay. Or

$$G(t \mid n) = \sum_{i=0}^{t'-1} \Pr\{T' = i \mid n\} + t'' \Pr\{T' = t' \mid n\} \tag{5}$$

because of the uniform distribution of T''.

It is not difficult to write a formula for $\Pr\{T' = i \mid n\}$. First,

$$\Pr\{T' = 0 \mid n\} = \frac{1}{n},$$

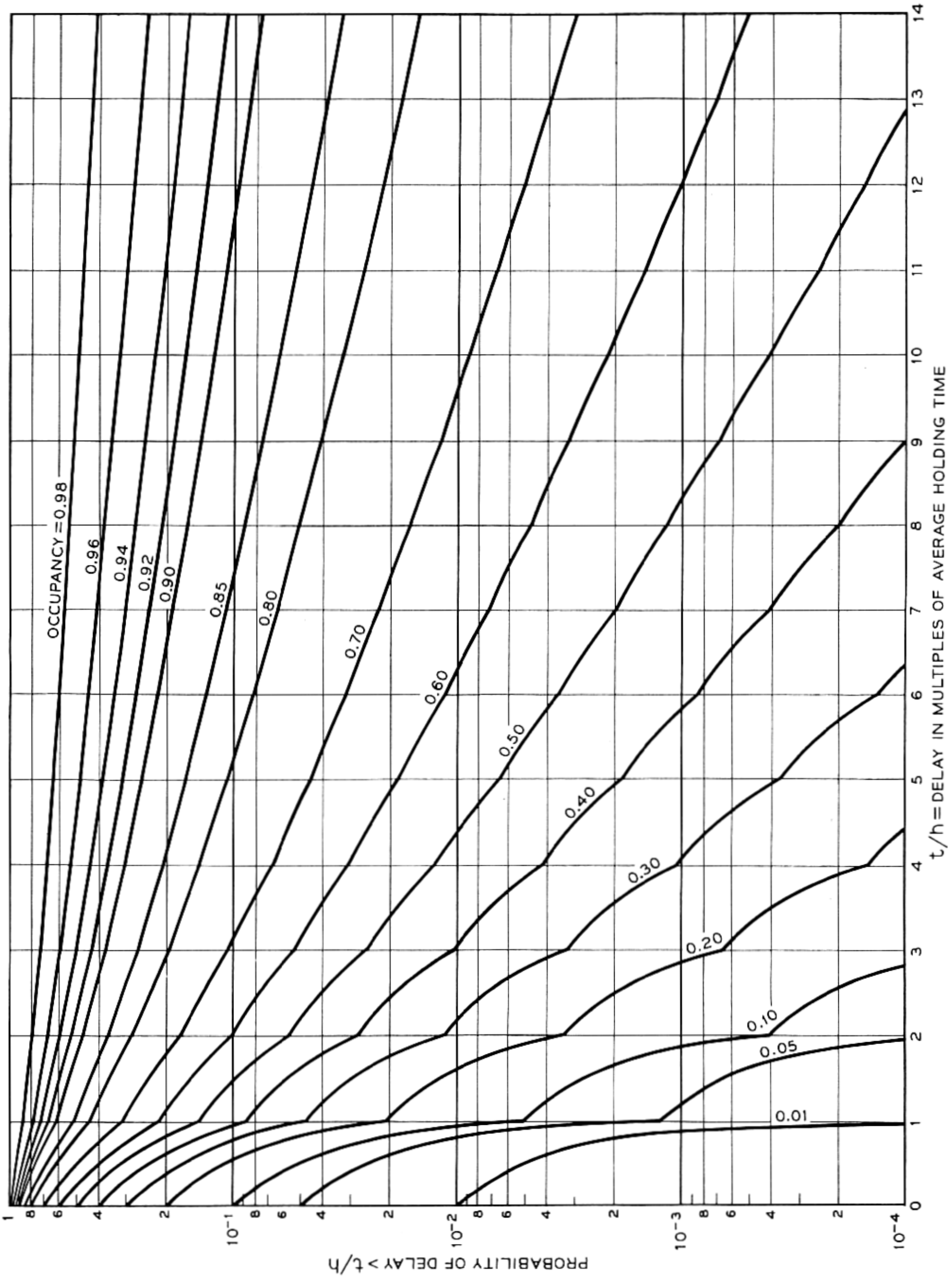
since, at the first post-arrival departure epoch, the delayed call is one of n calls equally likely to be served. Also

$$\Pr\{T' = 1 \mid n\} = \left(1 - \frac{1}{n}\right) \sum_{j_1=0}^{\infty} p(j_1) \frac{1}{n - 1 + j_1},$$

where $p(j_1)$ represents the Poisson probability of j_1 arrivals in a service interval, since, if the delayed call is not served at the first opportunity, any number of calls from zero upward may arrive during the next complete service interval. Extending this reasoning, one has for $i > 0$,

$$\begin{aligned} \Pr\{T' = i \mid n\} = \left(1 - \frac{1}{n}\right) & \sum_{j_1 + \cdots + j_k \geqq k-n+1} \prod_{k=1}^{i-1} p(j_k) \\ & \cdot \left[1 - \frac{1}{n - k + \sum_1^k j_r}\right] \frac{p(j_i)}{n - i + \sum_1^i j_r}. \end{aligned} \tag{6}$$

Although (6) can be written down directly, it is more convenient


OCCUPANCY = 0.98
0.96
0.94
0.92
0.90
0.85
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.05
0.01
t/h = DELAY IN MULTIPLES OF AVERAGE HOLDING TIME
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
PROBABILITY OF DELAY > t/h
1
8
6
4
2
10^{-1}
8
6
4
2
10^{-2}
8
6
4
2
10^{-3}
8
6
4
2
10^{-4}

computationally to use a recursive formula for the probabilities. (This was pointed out to the author by W. S. Hayward, Jr.) Let

$$\Pr\{T' = i \mid n\} = Q_i(n).$$

Then

$$Q_0(n) = \frac{1}{n}$$

and

$$Q_i(n) = [1 - Q_0(n)] \sum_{j=0}^{\infty} p(j) Q_{i-1}(n + j - 1). \tag{7}$$

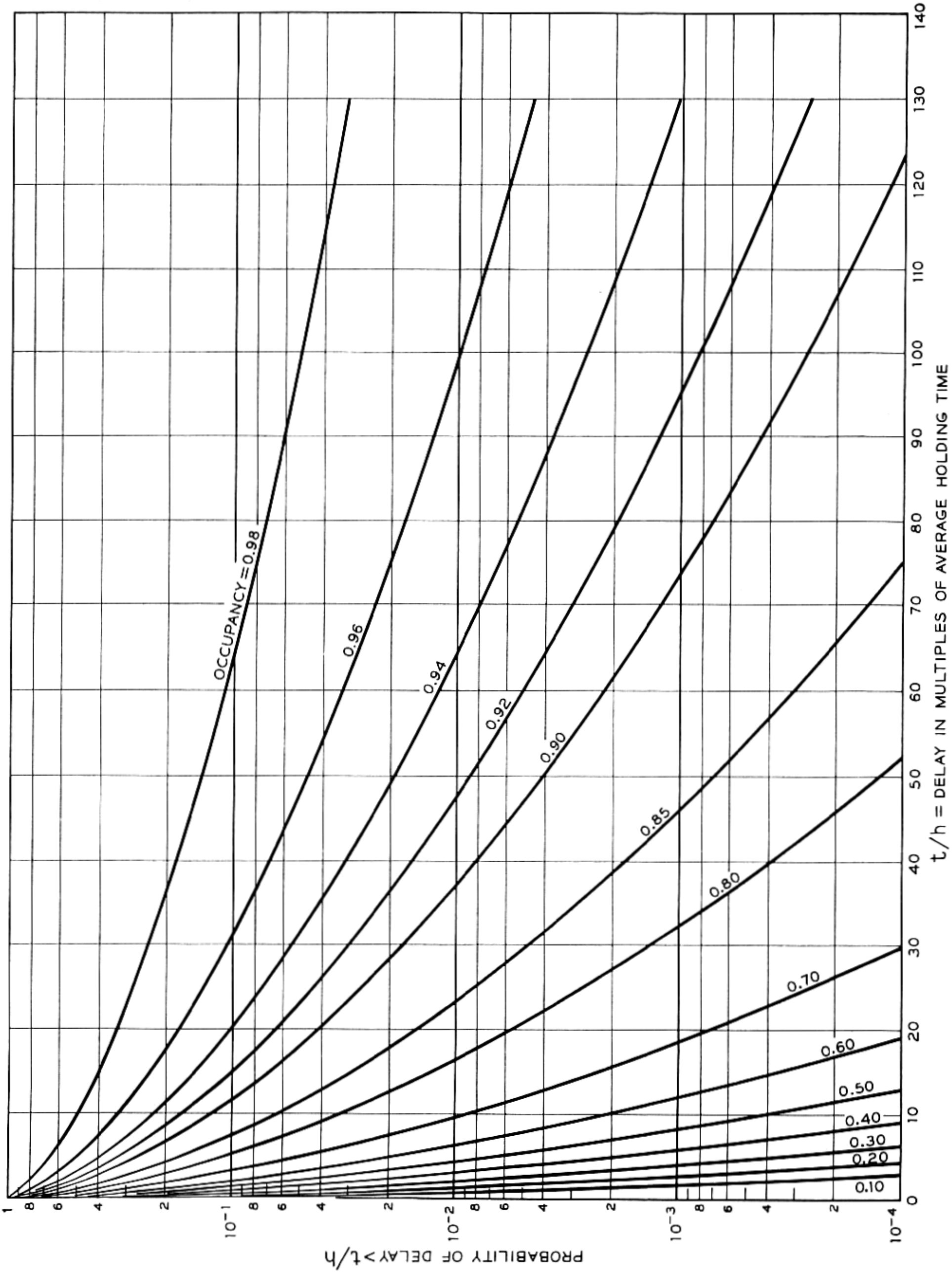
It is clear that (6) is the solution of (7). The delay distribution, $F(t)$, is obtained by substituting (6) into (5) and the latter into (3). The values of P_{n-1} necessary for evaluating (3) are obtained recursively from (1) and (2), together with the relation $P_0 = 1 - \lambda$.

The results of the calculations are shown as falling distributions of delays for all calls. That is, $\lambda[1 - F(t)]$ is plotted rather than $F(t)$, in keeping with custom in the field of delay theory. The distributions are shown in Fig. 2 for delays up to 14 holding times and, in Fig. 3 on a compressed scale, for delays up to 130 holding times.

As an example of the use of the curves, suppose a single marker whose holding time is 0.1 second serves calls at random and that it is desired to limit the probability of a delay greater than 2 seconds at this marker to 0.0001. It is required to find the permissible occupancy. Since a delay of 20 holding times is involved, Fig. 3 should be consulted. On this chart, the occupancy for a probability equal to 0.0001 of delay $t/h = 20$ is found to be just above 0.60, roughly 0.61. Thus, the marker can handle a random input averaging 6.1 calls per second.

In some applications in which service is not precisely order-of-arrival, it may be presumed that the delay distribution will lie between those for random and queued service. In such cases, the delay distributions will fall in a band bounded by random service and queued-service (Crommelin) curves. Examples of such bands are shown on Fig. 4. It should be noted that the bounding curves for any occupancy must cross, since the average delay is independent of the queue discipline.

It is of some interest also to compare the random-service delay distributions for constant and exponential holding times. It is conjectured that a pair of such curves for a given occupancy defines a band containing all random-service delay distributions, for that occupany, where the holding-time distribution is of gamma type in which the coefficient of


OCCUPANCY = 0.98
0.96
0.94
0.92
0.90
0.85
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
t/h = DELAY IN MULTIPLES OF AVERAGE HOLDING TIME
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
PROBABILITY OF DELAY > t/h
1
8
6
4
2
10^{-1}
8
6
4
2
10^{-2}
8
6
4
2
10^{-3}
8
6
4
2
10^{-4}

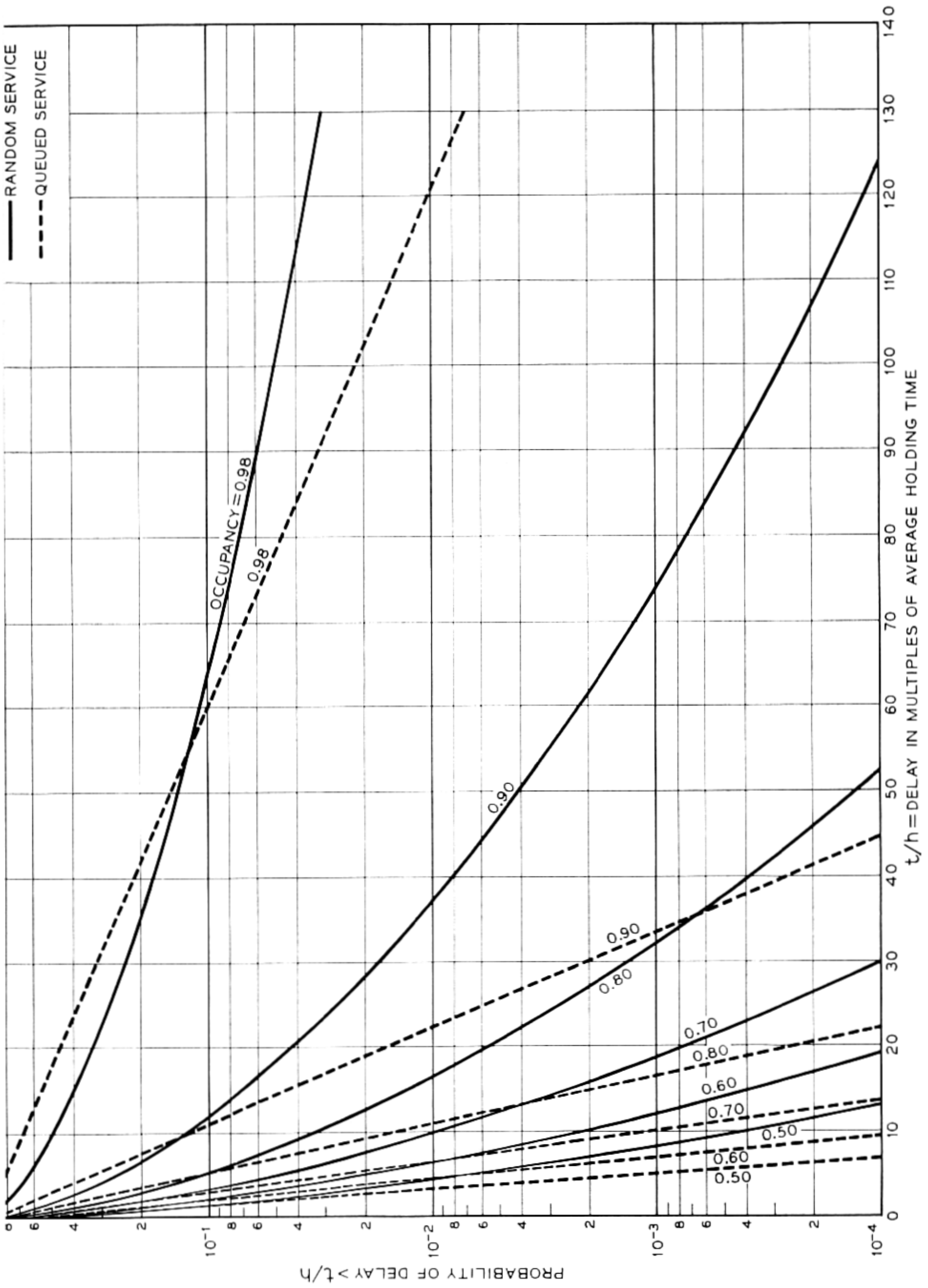
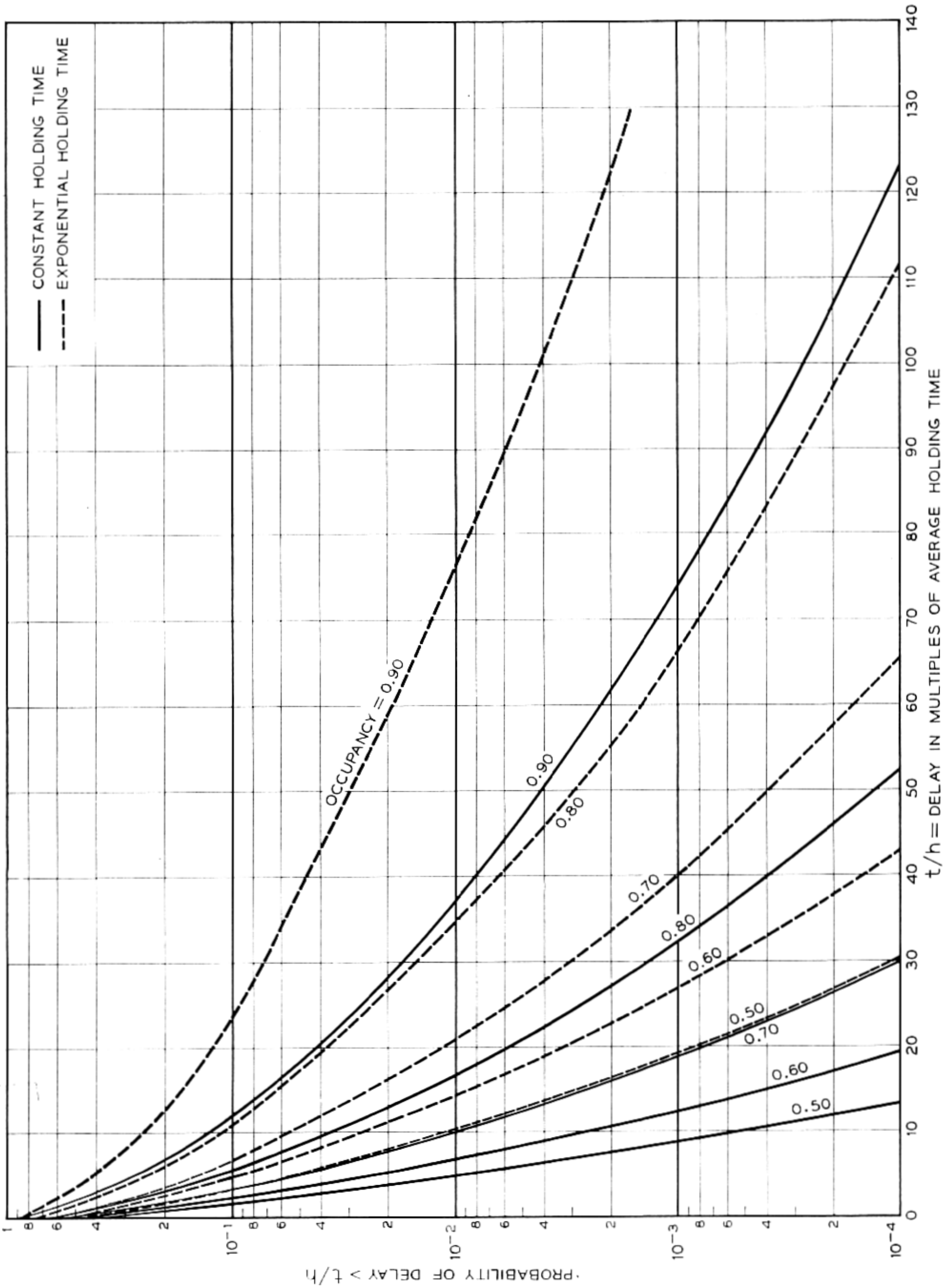


Fig. 4 — Comparison of random and queued service, with constant holding times.


CONSTANT HOLDING TIME
EXPONENTIAL HOLDING TIME
OCCUPANCY = 0.90
0.90
0.80
0.70
0.80
0.60
0.50
0.70
0.60
0.50
t/h = DELAY IN MULTIPLES OF AVERAGE HOLDING TIME
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
PROBABILITY OF DELAY > t/h
1
8
6
4
2
10^{-1}
8
6
4
2
10^{-2}
8
6
4
2
10^{-3}
8
6
4
2
10^{-4}

variation is not greater than unity. (In particular, the χ^2 distributions with two or more degrees of freedom are of this type.) Several such pairs of curves are shown on Fig. 5. (The exponential-holding-time curves are based on Wilkinson.[9]) Here, of course, the curves do not cross — the exponential-holding-time curves always (except at $t = 0$) lie above their constant-holding-time counterparts.

ACKNOWLEDGMENT

The extensive computations required for Figs. 2 through 5 were programmed by Miss Catherine Lennon, to whom the author expresses his gratitude.

REFERENCES

1. Erlang, A. K., *The Life and Works of A. K. Erlang*, The Copenhagen Telephone Co., Copenhagen, 1948.
2. Mellor, S. D., Delayed Call Formulae When Calls Are Served in a Random Order, P.O.E.E.J., **35,** 1942, p. 53.
3. Vaulot, E., Delais d'attente des appels téléphoniques traités au hazard, Comptes Rend., **222,** 1946, p. 268.
4. Pollaczek, T., La loi d'attente des appels téléphoniques, Comptes Rend., **222,** 1946, p. 353.
5. Palm, C., Väntetider Vid Slumpvis Avverkad Kö, Tekniska Meddelanden Från Kungl, Telegrafstyrelsen, Specialnummer för Teletrafikteknik, Stockholm, 1946, p. 70. (Translated in Tele, English Edition, No. 1, 1957, p. 68.)
6. Riordan, J., Delay Curves for Calls Served at Random, B.S.T.J., **32,** 1953, p. 100.
7. Kendall, D. G., Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains, Ann. Math. Stat., **24,** 1953, p. 338.
8. Crommelin, C. D., Delay Probability Formulae When the Holding Times Are Constant, P.O.E.E.J., **25,** 1932, p. 41.
9. Wilkinson, R. I., Delayed Exponential Calls Served in Random Order, B.S.T.J. **32,** 1953, p. 360.