Working Curves for Delayed Exponential Calls Served in Random Order

By ROGER I. WILKINSON

(Manuscript received December 19, 1952)

Working curves of delays for waiting calls served at random are given for a considerable range of loads and group sizes. Exponential holding time calls are assumed originating at random, and served by a simple group of paths. Results of a number of throwdown tests are given to illustrate the effect on call delays of several modes of service, and particularly of service on a random basis. For random service, these results verify the theory recently developed by J. Riordan; perhaps more interestingly they show the effects on delays of certain blends of queued and random service which approximate methods of handling delayed calls in practical use (such as gating and limited storage circuits). The use of random and queued delay theory is illustrated by a number of examples. To remind the reader that these results are not limited to telephony, department store and vehicular traffic problems are included.

A theory for predicting the delays which telephone calls (or other corresponding types of traffic such as vehicular, aircraft, people waiting in line, etc.) having exponentially distributed holding times would encounter when the delayed calls are served in a random order was published in a recent issue of this JOURNAL* by John Riordan. Mr Riordan's mathematical analysis involved a determination of the first several moments of the delay distributions. He then devised a method of combining elementary exponential curves in such a way as to satisfy the moments previously calculated.

Since a limited number of moments were used in the above determinations the curves derived are approximate only, but at the same time they are believed to be good approximations. The critical cases are those of paths carrying very heavy loads, in the occupancy ranges of $\alpha = 0.80$ or higher.

* Bell System Technical Journal, January, 1953, pages 100–119.

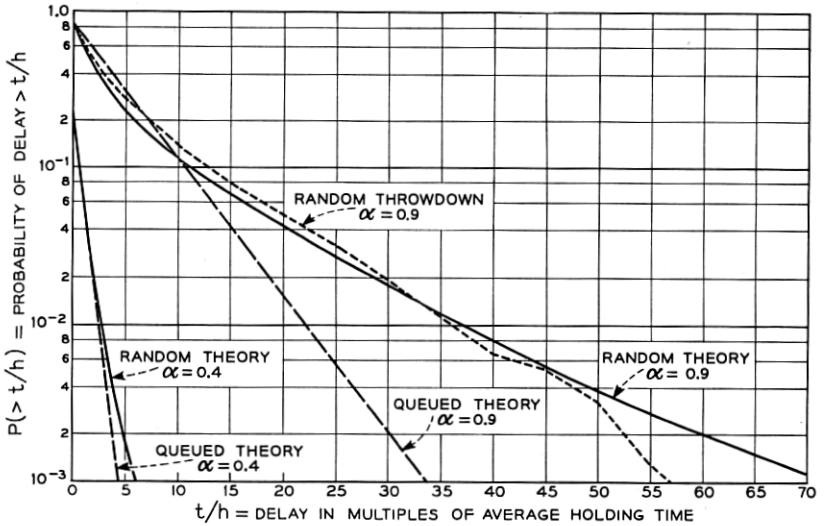

Fig. 1 — Distribution of delays. Theory versus throwdown, delayed calls handled at random, $c = 2$ paths, $\alpha = 0.90$, 3000 throwdown calls.

THROWDOWN CHECKS

Before calculating a field of curves for working purposes it was thought desirable to make at least a modest throwdown test, or traffic simulation, at these high occupancies to observe the agreement of theoretical delays with those determined by a trial in which the theoretical assumptions would be closely followed. This has now been performed at two trunk group sizes, $c = 2$ paths, loaded by approximately $a = 1.8$ erlangs or an occupancy of $\alpha = 0.90$, and $c = 10$ paths at an occupancy of approximately $\alpha = 0.80$.

For these throwdowns, random origination times were obtained through use of Tippett's Random Numbers. An hour was visualized as being composed of 100,000 (or, as in one case, 1 million) consecutive discrete intervals, numbered serially. Choosing 5 (or 6) digit random numbers then provided the start times of the subscribers' bids for service.

Likewise holding times were chosen by random numbers from an exponential universe by dividing it into 100 equal probability segments and assigning each a number from 00 to 99. A central value of holding time was chosen to represent the range of cases within each segment. The last segment, number 99, on the long tail was further subdivided into 100 parts in order to give more definition in the long call lengths which are believed to be critical.

A comparison of the proportion of traffic expected to suffer delays beyond various multiples of the average holding time as given by Riordan's theory for delayed calls served in random order, and by the throwdown results, is given in Figs. 1 and 2. As discussed below, the cases studied are considered to give satisfactory assurance as to the adequacy of the approximations involved in the theory.

The two trunk case based on 3000 calls submitted shows fairly good agreement with the theoretical distribution out to delays as large as 50 multiples of an average holding time which includes more than 99.5

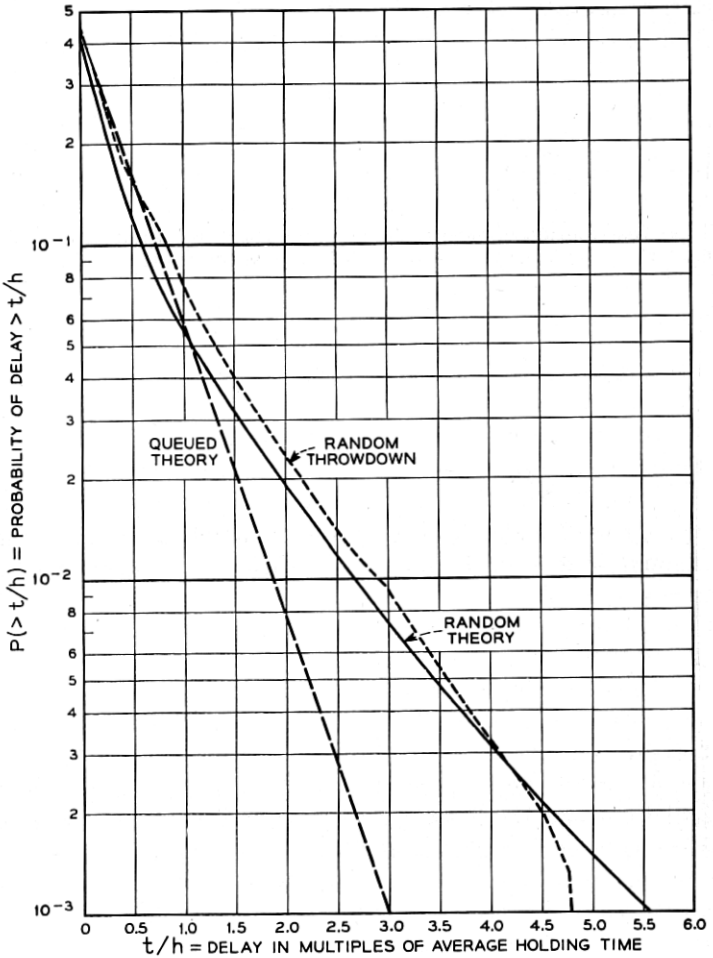

Fig. 2 — Distribution of delays. Theory versus throwdown, delayed calls handled at random, c = 10 paths, α = 0.80, 1500 throwdown calls.

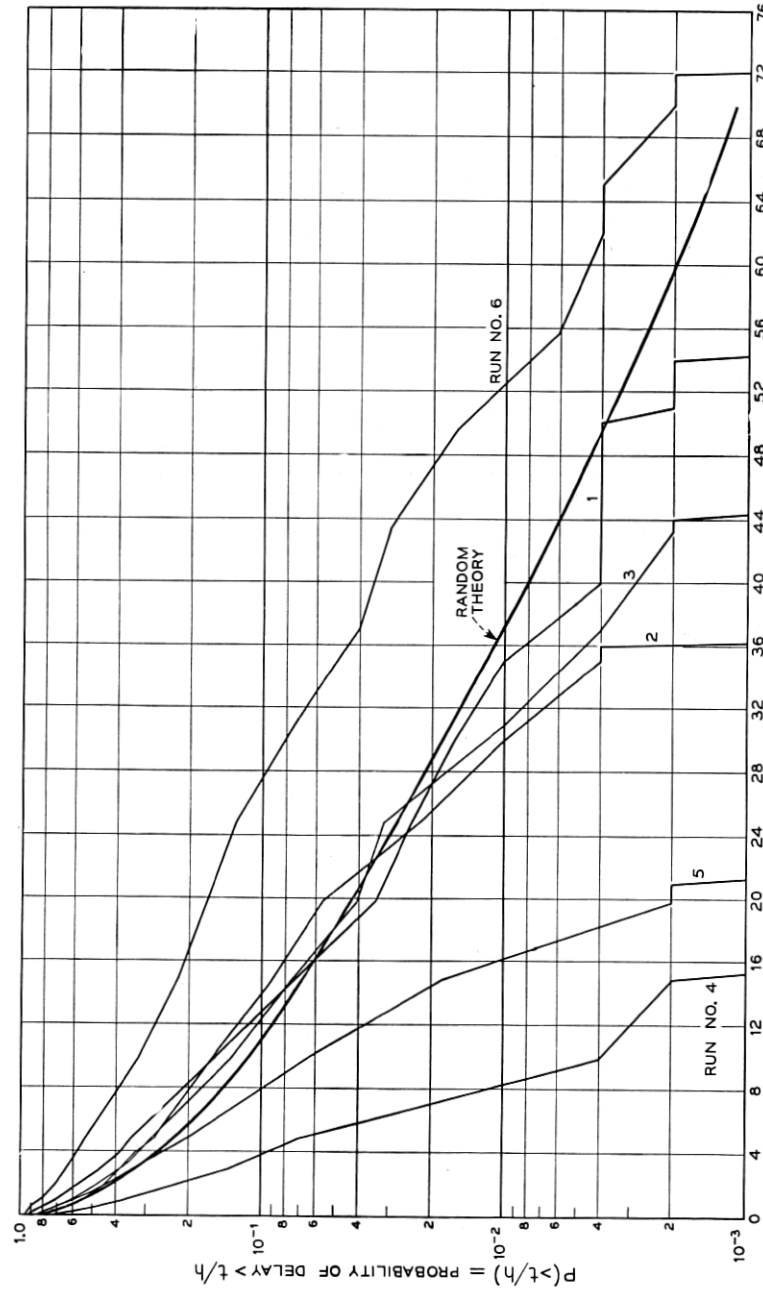

Fig. 3 — Distributions of delays on 2 paths by groups of 500 calls in random delay throwdowns, designed load $\alpha = 0.90$.

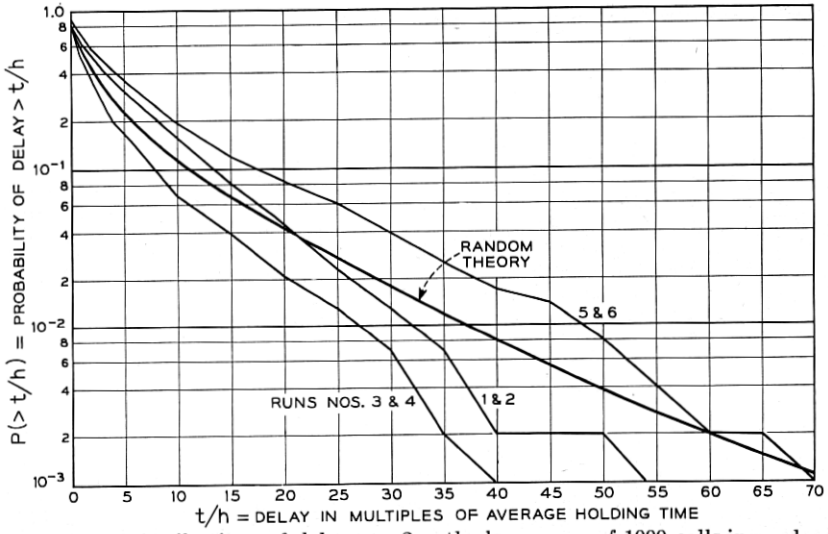

Fig. 4 — Distributions of delays on 2 paths by groups of 1000 calls in random delay throwdowns, designed load $\alpha = 0.90$.

per cent of all the calls delayed. The 10 trunk case based on 1500 calls at 0.8 occupancy also shows good agreement to 99.5 per cent of the calls delayed.

In making throwdown tests of this sort, the criterion for deciding when one has proceeded long enough is rather vague. The usual practice is to summarize the delays at regular intervals and observe at what point it seems likely that making additional tests would not change the results by a sensible amount. For the $c = 2$ trunk case, six runs of 500 calls each produced the several very different broken line curves of Fig. 3 shown superposed on the theoretical delay distribution for $\alpha = 0.90$. Clearly no one of these by itself could be given much weight.

Consecutive runs were paired to form three runs of 1000 calls each, as shown in Fig. 4. As one would expect, their spreads have narrowed appreciably. Combining these three runs yielded the dotted curve of Fig. 1, which, of course, has a correspondingly smaller likelihood of sampling error in it. On the basis of such a succession of narrowing spreads, one can, with some feeling of assurance, estimate within what narrow band about the observed curve the true unknown curve (approachable by many more tests) must lie.

On Figs. 1 and 2 the shapes and positioning of the total throwdown and theoretical curves seldom differ more than 20 per cent on the probability scale down to the $P = 0.005$ probability level. The dis-

parities measured along the delay axis in the higher ranges of the variable, are, of course, considerably less. A comparison of the theoretical and observed proportions of calls delayed, and the average delays on all calls is shown in the following table:

Trunk Group Size, c	Occupancy α	No. Calls in Throwdown	Proportion of Calls Delayed		Average Delay on All Calls in Multiples of Average Hold Time	
			Theory	Throwdown	Theory	Throwdown
2	0.9	3000	0.853	0.855	4.30	4.71
10	0.8	1500	0.409	0.444	0.205	0.254

These differences between theory and observation are well within the variations which would be expected with the lengths of throwdown runs made.

Further reassurance that the traffic submitted in the two throwdown tests originated in a manner reasonably similar to that assumed in the theory was obtained by making "switch counts" at regular intervals during the throwdowns from which frequency distributions, $f(x)$, of the number x of calls simultaneously present were constructed. These are shown in Figs. 5 and 6 for the two throwdown cases. The solid

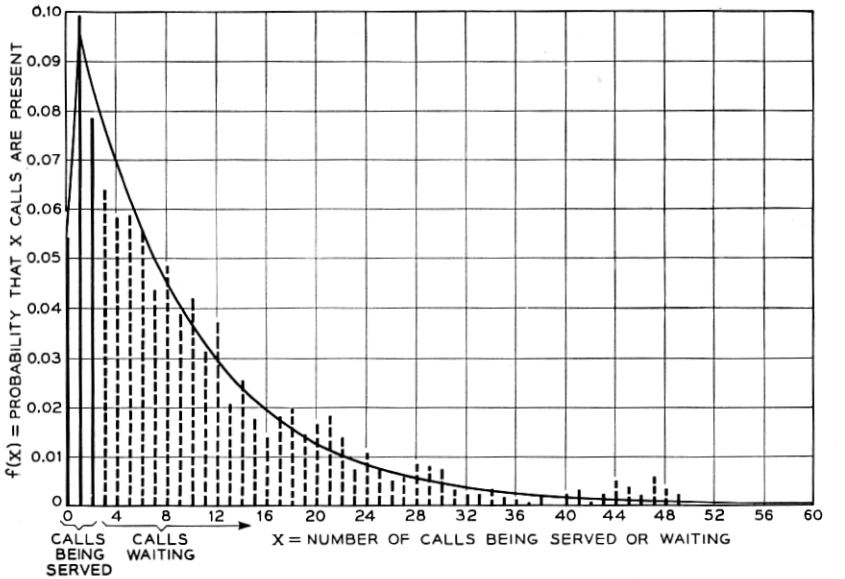

Fig. 5 — Distribution $f(x)$ of simultaneous calls. Theory versus throwdown, $c = 2$ paths, $\alpha = 0.90$, 3000 throwdown calls.

line spikes correspond to observations when all calls in the system were being served, that is $x \leqq c$. The dotted spikes show those proportions of observations when one or more calls were waiting, that is $x > c$. The theoretical values of $f(x)$ are indicated by the smooth curves where they pass over discrete values of x. The theory and observations are seen to be in quite good agreement.

Referring again to the theoretical delays (and the throwdown checks) on Figs. 1 and 2, very much larger delays can obviously be obtained when delayed calls are handled at random than when they are handled in a strict first-come-first-served, or queued, order, the latter distri-

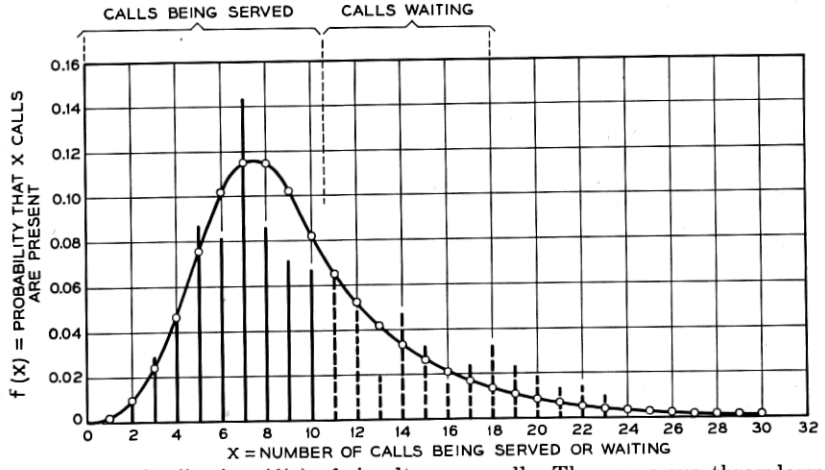

Fig. 6 — Distribution ($f(x)$) of simultaneous calls. Theory versus throwdown, $c = 10$ paths, $\alpha = 0.80$, 1500 throwdown calls.

butions being shown by the straight lines which start at nearly the same ordinates at delay 0 as the random handling curves, and cut down across the lower part of the charts.* Although fewer very short delays occur

* Delay curves for exponentially distrubuted holding time calls in systems where delayed calls are handled in order of arrival, are given by E. C. Molina in "Application of the Theory of Probability to Telephone Trunking Problems," Bell System Technical Journal, Vol. 6, p. 461, July, 1927. They are calculated from the Erlang equation

$$P(>t) = P(>0)e^{-(c-a)t} = \frac{\dfrac{a^c e^{-a}}{c!}\dfrac{c}{c-a}}{\displaystyle\sum_{x=0}^{c-1}\frac{a^x e^{-a}}{x!} + \frac{a^c e^{-a}}{c!}\frac{c}{c-a}}\, e^{-(c-a)t} \qquad (1)$$

where the delay t is expressed in multiples of the average holding time. Values of $P(>0) = C(c,a)$ can be read approximately from Figure 21.

with this method of handling than when a random selection of the waiting calls is followed, the very long delays are markedly reduced, and on this account the queueing procedure is generally preferred. These effects are particularly evident at the higher occupancies. As illustrated in Fig. 1, the "queued" and "random" delay curves at an occupancy of $\alpha = 0.4$ show little difference down to the $P = 0.001$ delay level.

IMPERFECT QUEUEING

Interest has often centered in questions as to what form the delay curves might take in a system in which queueing of the calls is maintained to a limited extent, and beyond which the record of order of arrival would be lost. Such an instance might occur with a team of toll recording operators who were able to keep well in mind the order of arrival of signals up to a certain number waiting, whereupon they would lose track and not regain this ability until the number of waiting calls had again dropped below some small number. Other situations with actual or equivalent limited delay storage arrangements can readily be imagined.

To study a case of limited queueing, a short subsidiary throwdown was next run on the $c = 2$ case, using the 1000 calls of Runs 1 and 2 of Figs. 3 and 4 (which comprised the 1000-call sequence most closely approaching the theoretical distribution). Three rules for delayed call handling were tested:

(1) Delayed calls are served in random order.

(2) Delayed calls are queued (served in order of arrival).

(3) Delayed calls are queued until more than w are waiting at which time their arrival order is lost and they are served at random. When the number waiting again drops below w, newly arriving calls are queued behind those randomized calls still waiting. Note that case 1 corresponds to $w = 0$, and case 2 to $w = \infty$.

The comparative results are shown on Fig. 7, with w given successively values of 0, 8, 20, 25, 30 and ∞. The $w = 0$ curve, of course, is taken directly from Fig. 4 for Runs 1 and 2 combined. Although this curve does not agree particularly well with theory (Curve A), its movement with changes in w is nevertheless instructive. As seen, queueing as far as $w = 8$ waiting calls produced practically no improvement in the delay distributions. (Perhaps with the occurrence of such large numbers of waiting calls, reaching a maximum of 35, one could not expect queueing of so few as 8 to have much effect.) The next selection of $w = 20$, however, still showed only a relatively slight improvement, particularly in

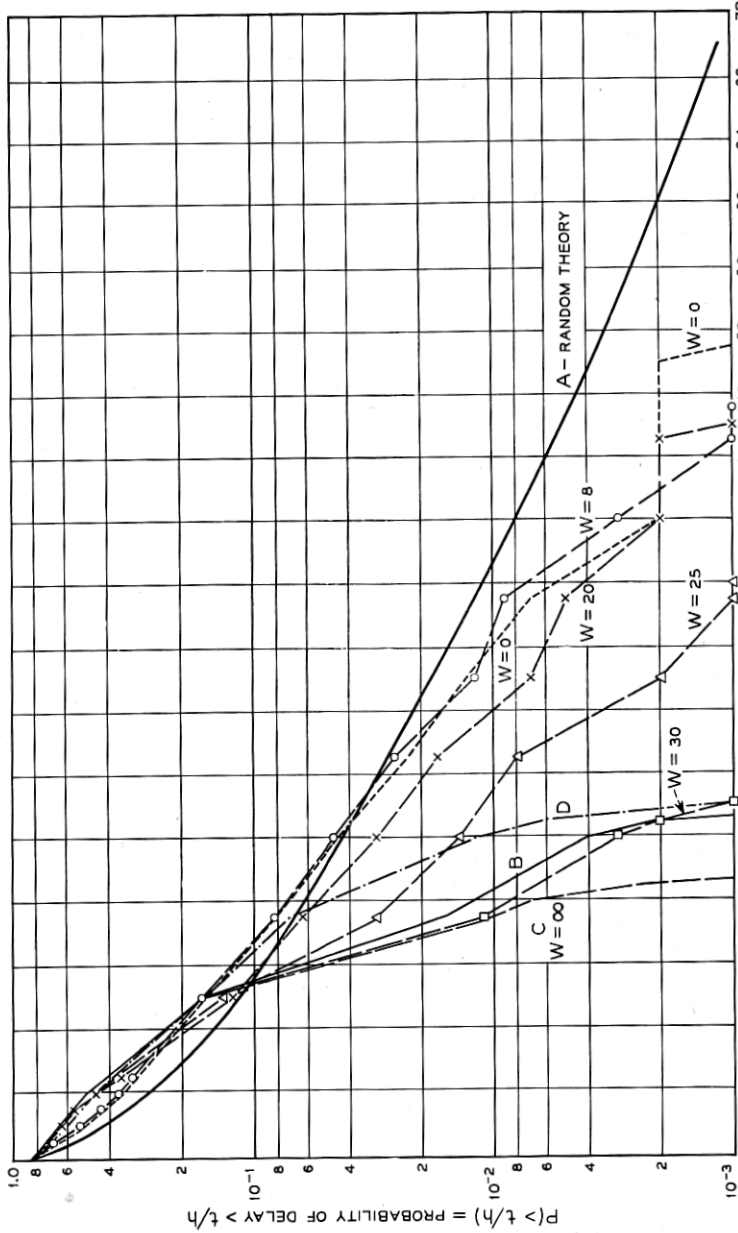

Fig. 7 — Throwdown delay distributions with imperfect queueing, $c = 2$ paths, $\alpha = 0.90$, 1000 throwdown calls.

the long delays occurring on a few of the more unfortunate calls. Even choosing $w = 25$ moved the delay curve hardly more than half way from the completely random to the fully queued curve of $w = \infty$. The $w = 30$ selection shows the accomplishment of nearly fully queued results, the latter being given by curve C ($w = \infty$). Thus one would apparently find little value in instructing a team of two operators working at an occupancy of 90 per cent to try to remember the order of arrival of waiting calls unless they could keep track of an unexpectedly large number.

Electrical storing circuits have long been used to assist the ordering of waiting calls. They have the especial advantage of not becoming confused and losing the order of the calls which have engaged them up to the limit of their storage capacity. In the Bell System two methods of approaching true queueing are in common use. In one method, such as found, for instance, in the No. 3 Information Desk, a number of storage circuits are provided so that as a waiting call is served from the number one storage position, all the others waiting on storage circuits drop down one position. If s such circuits are provided, and more than s calls have been waiting, one of the excess will then be chosen at random to occupy the newly vacated sth storage circuit.

The second method used widely in both local and toll systems is known as gating. In its simplest form a gate opens into a "corral" where the operators or other service media are located. So long as calls simultaneously demanding service do not exceed the number of operators (trunks, markers, etc.) the gate is ineffective. As soon as one call has to wait, the gate closes until that call obtains service, and then admits to the corral all calls which have accumulated on the outside. The gate again closes until all calls within the corral are served; and so on. Thus the calls are admitted in bunches to the corral. Between bunches there is strict queueing but within bunches when they get inside the gate the calls are substantially served at random. As long as the bunches are small the effect of true queueing is approached. In any event a strong safeguard against excessively long delays on a few unlucky calls is introduced. In the Bell System, a variety of gating plans are found such as double gates, gates with additional preferences for certain types of calls, and schemes for placing calls outside the gate again if they cannot be served immediately. Each of these must be studied with its own peculiar characteristics in mind.

To illustrate the effectiveness of the storage circuit type of automatic queueing arrangement, the 1000 calls of Runs 1 and 2, for the two path case, were processed by a throwdown through a two operator, 20 storage

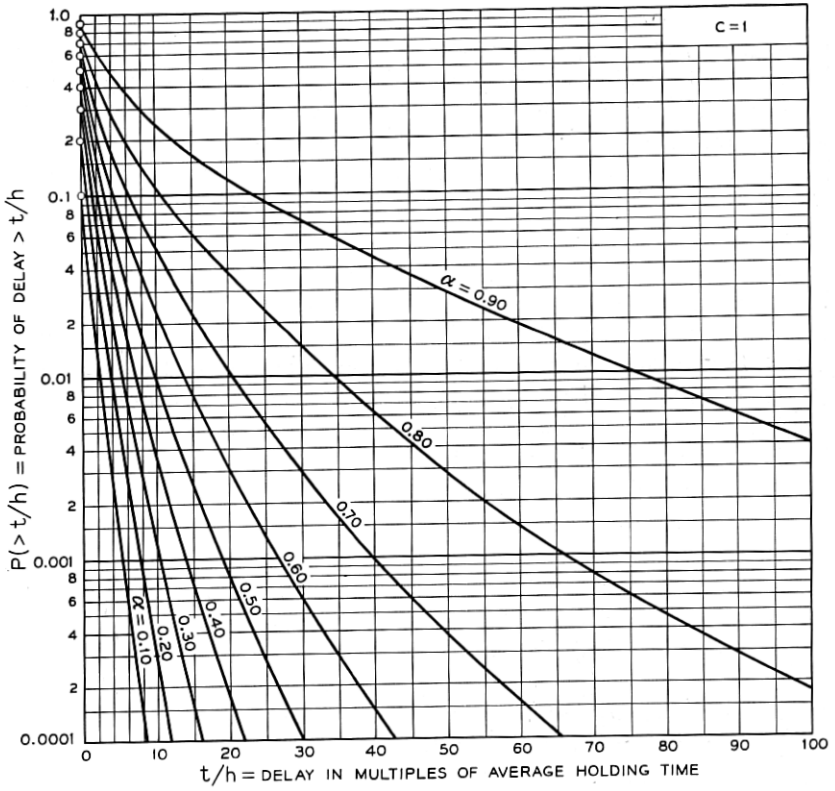

Fig. 8 — Delayed traffic served in random order, exponential holding times, $c = 1$.

circuit system. The resultant delay distribution is shown as Curve B on Fig. 7. (It is appreciated that this hardly represents a tolerable normal operating situation, but rather illustrates what the performance might be under extremely heavy traffic conditions.) The results are very close to those obtained with perfect queueing (Curve C) and show in striking fashion the gains in service to be made in certain delay situations by providing a limited storage apparatus with a memory not subject to confusion during moments of heavy overload.

When the 1000 calls of Runs 1 and 2 are submitted to the 2 paths through a simple gate in order to produce approximate queueing, the resultant delays are shown by Curve D on Fig. 7. Large improvements again occur in reducing the very long delays found with random handling. In fact by use of this simple (and usually relatively inexpensive) gating

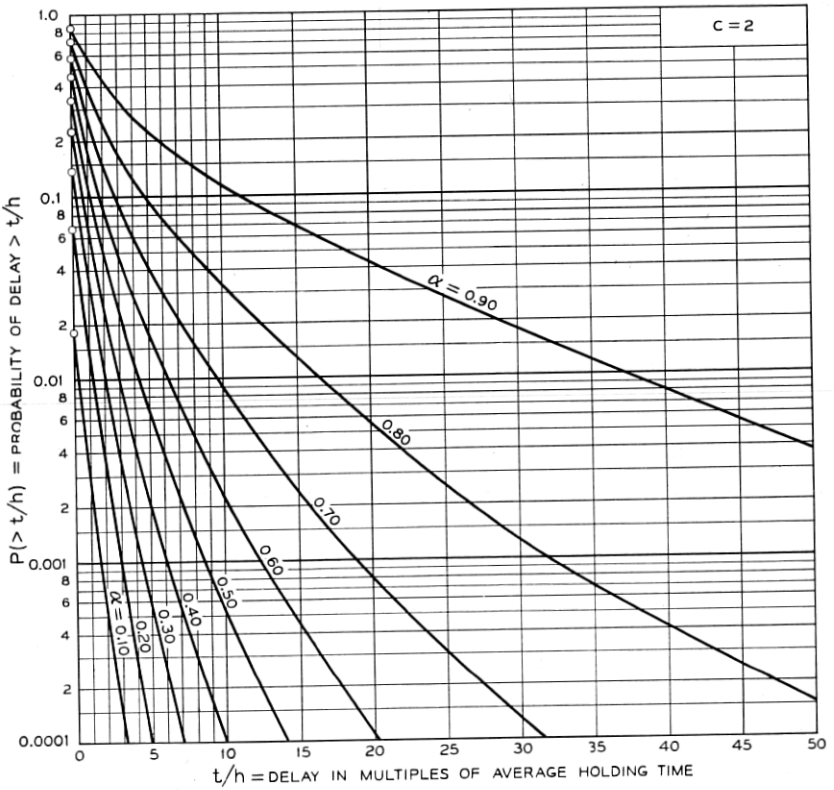

Fig. 9 — Delayed traffic served in random order, exponential holding times, $c = 2$.

scheme, delay results are obtained nearly as good as those realized by the provision of 20 storage circuits (Curve B).

WORKING CURVES

The adequacy of the Riordan theory when delayed exponential calls are served at random is believed to have been established and that it may be used with confidence to solve those practical problems where the underlying assumptions are well satisfied.

For working purposes, curves showing distributions of delays expected for occupancies up to $\alpha = 0.90$ and for group sizes of $c = 1, 2, 3, 4, 5, 6, 8, 10, 20, 50$ and 100, are shown in Figs. 8 to 18. These are plotted in the customary fashion with delay in multiples of average holding time

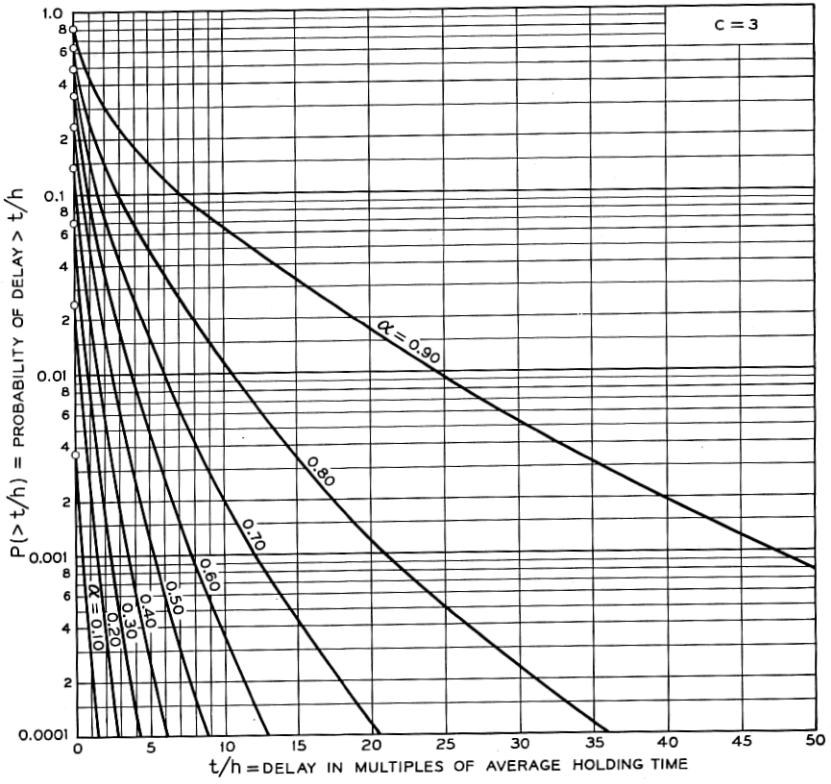

Fig. 10 — Delayed traffic served in random order, exponential holding times, $c = 3$.

as abscissa, and $P(>t/h)$, the probability of a random call meeting a delay greater than t/h, as ordinate.

Estimates of average delays, $\bar{t}$ (which are the same for queued and random service), are also commonly desired, and these are shown in Fig. 19. They are calculated from the equation

$$\bar{t}/h = P(>0)/(c - a) \tag{2}$$

If one wishes instead the average delay, $\bar{\bar{t}}$, *on calls delayed*, it may be obtained from

$$\bar{\bar{t}}/h = \frac{\bar{t}/h}{P(>0)} = \frac{1}{c - a} \tag{3}$$

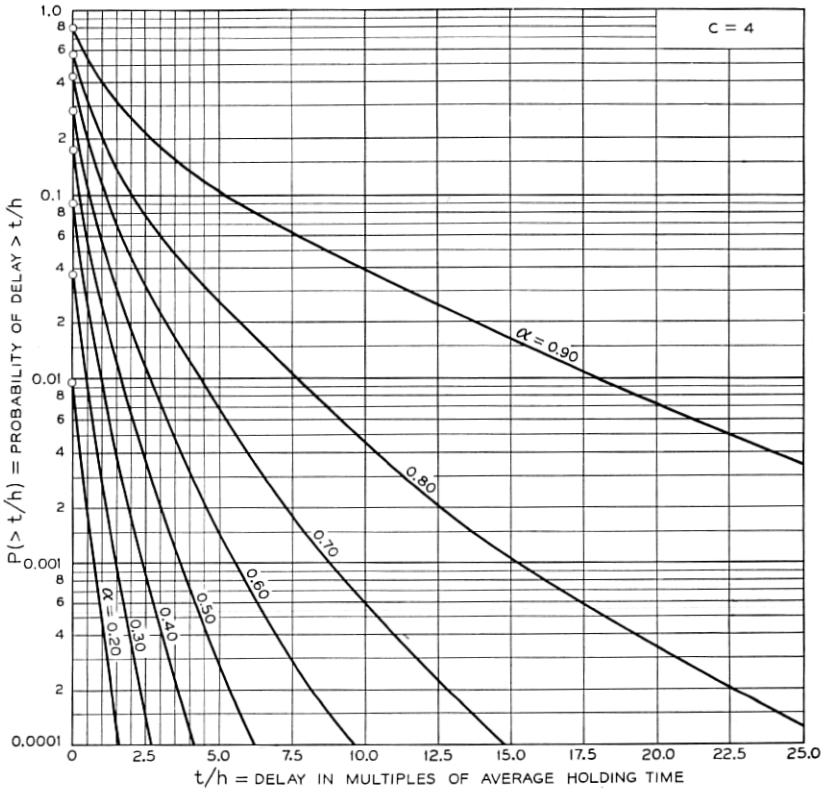

Fig. 11 — Delayed traffic served in random order, exponential holding times, $c = 4$.

ILLUSTRATIVE EXAMPLES

Example No. 1

A 20-trunk toll route carrying exponentially distributed holding time calls with average length of 5 minutes, is loaded with 16.0 erlangs of traffic, and any calls delayed will be served in random order. What per cent of all calls will be delayed? What per cent will be delayed more than 5 minutes? More than 10 minutes?

Solution. Enter Fig. 16 (the $c = 20$ chart) and read on the $\alpha = \frac{16}{20} = 0.80$ occupancy curve. At $t/h = 0$, the per cent of all calls delayed is found to be 26 per cent. At $t/h = 1$, the calls having delays exceeding 1 holding time, or 5 minutes, are 1.2 per cent, and at $t/h = 2$, the calls with delays exceeding 10 minutes are 0.2 per cent.

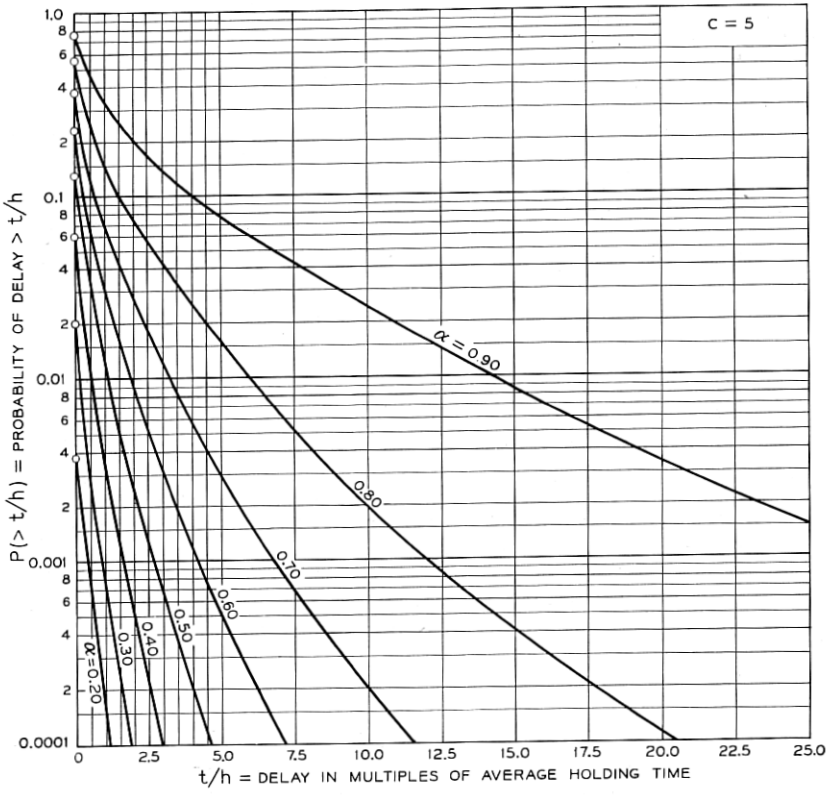

Fig. 12 — Delayed traffic served in random order, exponential holding times, $c = 5$.

Example No. 2

How many operators will be required at a department store telephone order desk to handle 225 calls per hour with an average delay not longer than half a minute, and with no more than 20 per cent of the calls delayed over 1 minute? Assume the average operator work time per call is 100 seconds, and waiting calls are handled in indiscriminate order.

Solution. The load to be carried is $a = (225)(100)/3600 = 6.25$ erlangs. The average delay, $\bar{t}$, is not to exceed $30/100 = 0.3$ holding time. Reading on Fig. 19, opposite an ordinate of 0.3 we select several trial values of trunks (operators) c, versus occupancy α, and form Table I, calculating the last column from the first two:

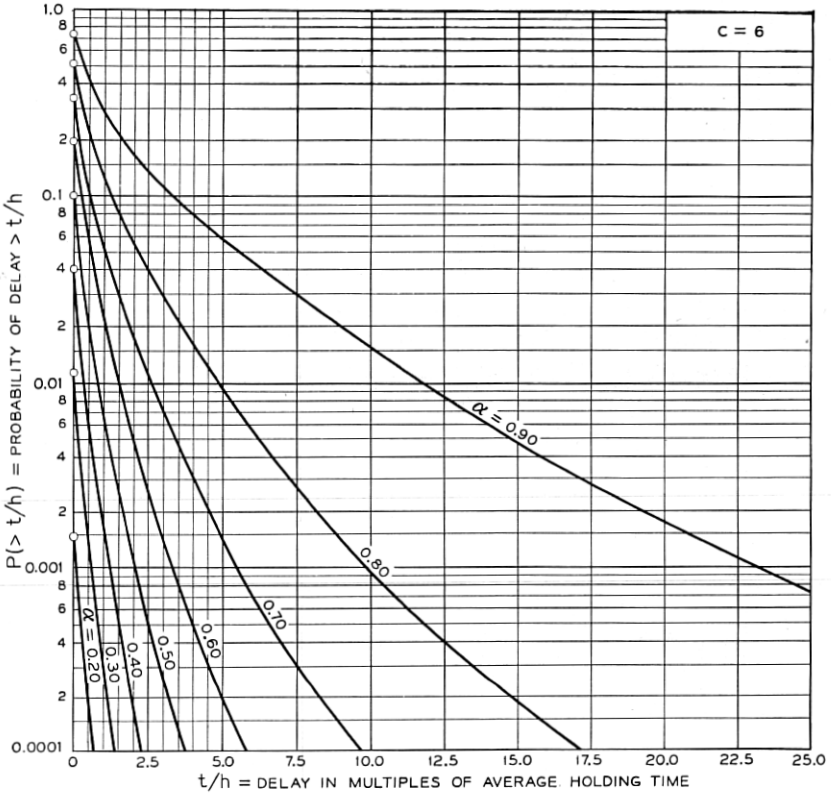

Fig. 13 — Delayed traffic served in random order, exponential holding times, $c = 6$.

To carry the 6.25 erlangs of traffic and meet the average delay requirement we see that 8 operators will be needed. Will 8 operators also fulfill the no more than 20 per cent delay over 1 minute requirement? Enter Fig. 14 (the $c = 8$ chart) with an occupancy of $\alpha = 6.25/8 = 0.78$. The per cent of calls exceeding a delay of $60/100 = 0.6$ holding time is about 12 per cent. A provision of 8 operators satisfies both requirements.

TABLE I

c	α	$a = c\alpha$
7	0.78	5.46
8	0.81	6.48
9	0.82	7.38

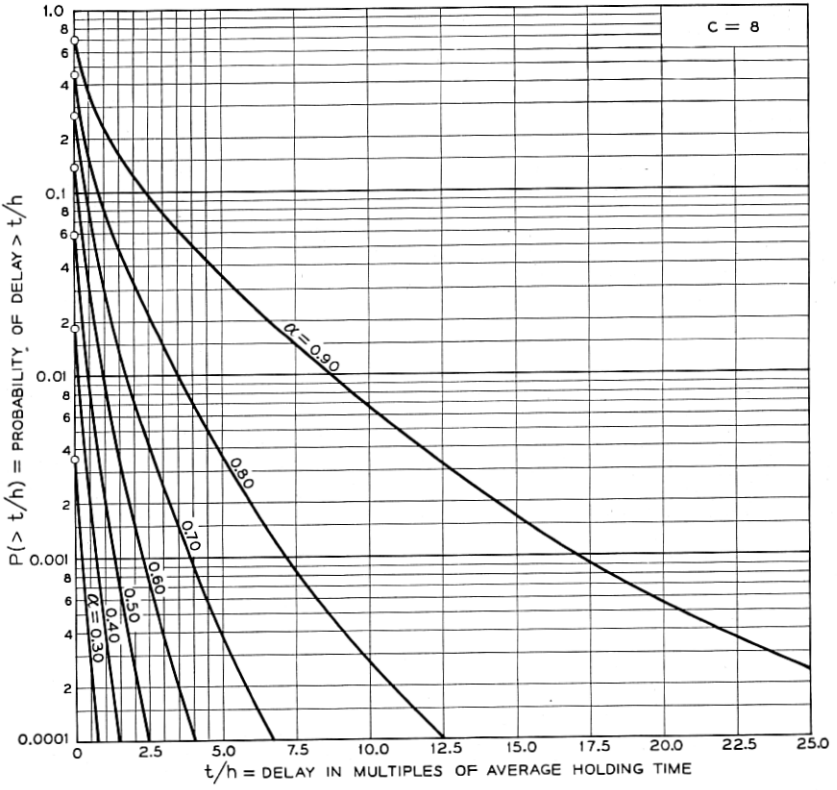

Fig. 14 — Delayed traffic served in random order, exponential holding times, $c = 8$.

Example No. 3

Suppose in Example 2, the second requirement had been that no more than one of 1000 customers should be required to wait over 3 minutes. Would 8 operators then suffice?

Solution. Reading on Fig. 14, with $\alpha = 0.78$ and $t/h = 180/100 = 1.8$, $P(>t/h) = 0.027$. Thus 27 in 1000 calls would be expected to experience delays over 3 minutes, and therefore more than 8 operators will be required. Consulting the $c = 10$ curves of Fig. 15, we find that with $\alpha = 0.625$, and $t/h = 1.8$, $P(>3$ minutes delay$) = 0.0012$ which closely meets the one in a thousand requirement. Ten operators would then be needed; and this would, of course, (from Fig. 19) reduce the average

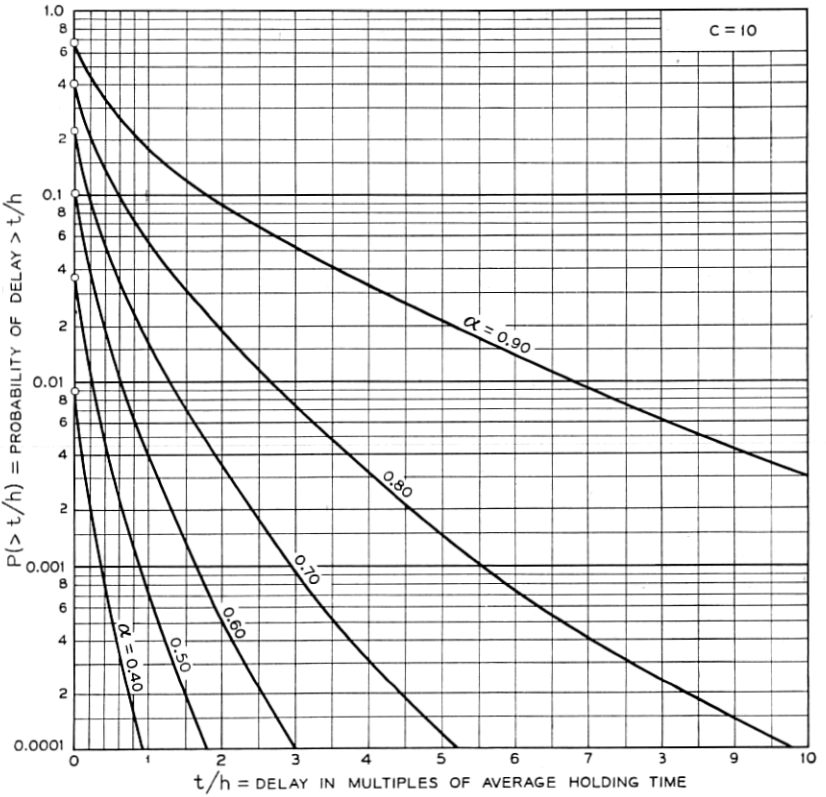

Fig. 15 — Delayed traffic served in random order, exponential holding times, $c = 10$.

delay on all calls to 0.035 (100) = 3.5 seconds, an improvement in this characteristic of 7 to 1 over the 8 operator service.*

Example No. 4

How much improvement in the delay service would be obtained in Examples 2 and 3 by purchasing storing or gating equipment which would substantially insure calls being handled in order of arrival?

Solution. With 8 operators working at an occupancy of 0.78, the pro-

* Had some number of operators been required other than those for which working charts, Figs. 8 to 18, are supplied, intermediate values could be obtained by graphical interpolation, or better still by employing the basic Riordan chart, Fig. 20, combined with $P(>0)$ found on Fig. 21, to obtain delay versus load for any desired number of paths or facilities. This latter process is described in the Appendix.

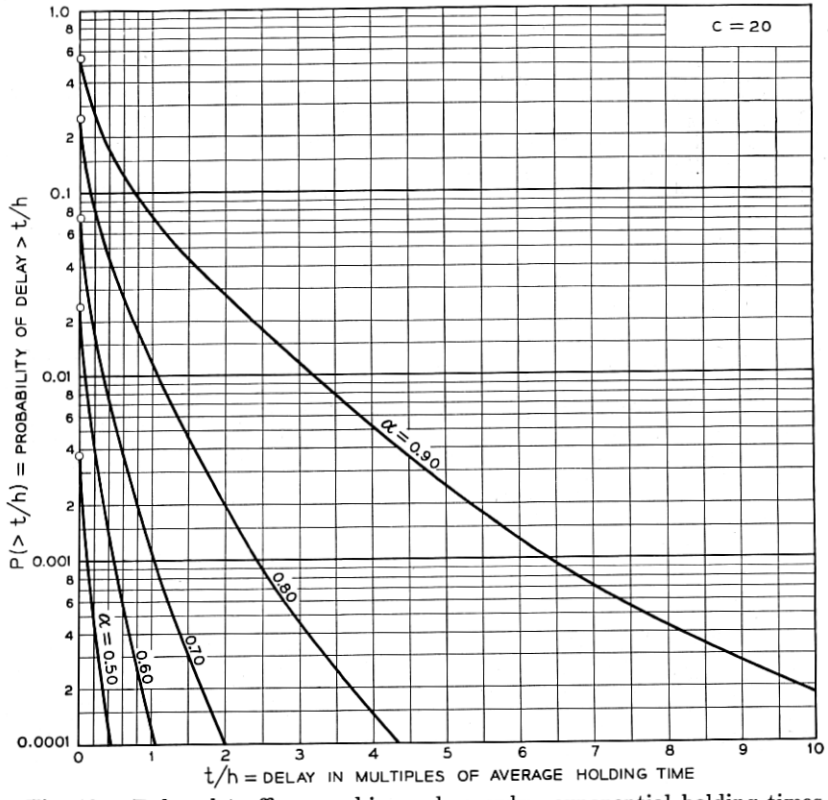

Fig. 16 — Delayed traffic served in random order, exponential holding times, $c = 20$.

portion of calls delayed is found to be $P(>0) = 0.41$ (Fig. 14). The probabilities of exceeding delays of $t/h = 0.6$ and 1.8 holding times are calculated for calls served in order of arrival by equation (1), in the following table:

t (Min.)	t/h	Queued $P(>0) = P(>0)e^{-(c-a)\,t/h}$	Random Handling
1	0.6	0.143	0.12
3	1.8	0.019	0.027

Comparing the queued and random handling of delayed calls one finds the perhaps unexpected result that with random handling some 2 per cent *fewer* calls are delayed longer than 1 minute than if perfect queueing had been present. This is due to the characteristic shapes of the two types

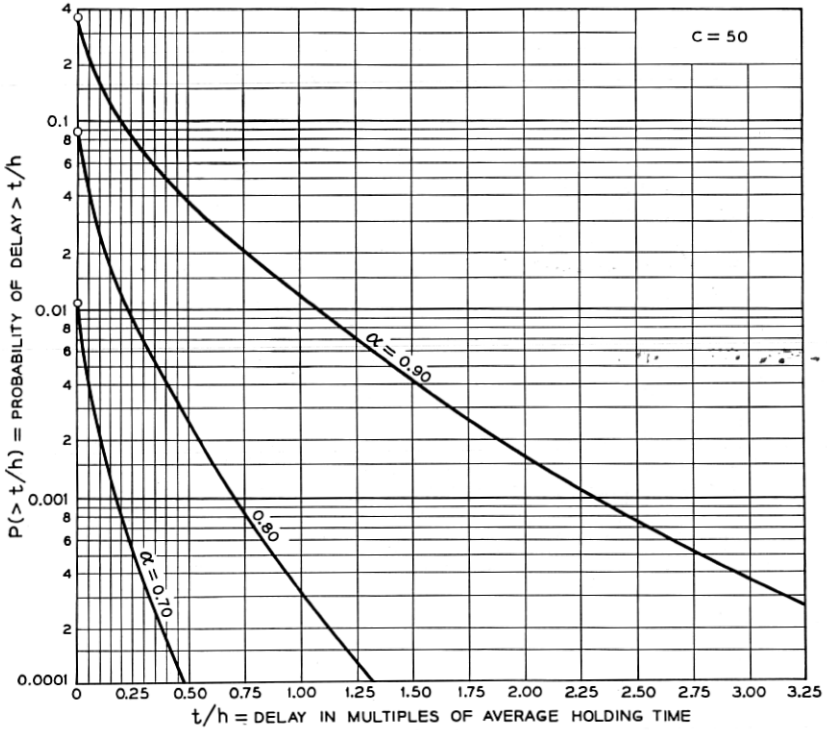

Fig. 17 — Delayed traffic served in random order, exponential holding times, $c = 50$.

of delay distributions, random handling producing more quite short and very long delays than does queueing. When a criterion of service is set at a relatively short delay, one may often expect it to be met more easily by not providing storing or gating circuits. On the other hand a criterion of service based on relatively long delays can nearly always be more readily met by the use of devices insuring partial or total queueing. In the example above the per cent of calls delayed longer than 3 minutes would be cut by a third through the use of queueing devices.

Example No. 5

Automobiles are parked in a large area adjacent to a State Fair grounds. There is one main exit through which two cars can pass at the same time. Upon leaving, drivers pay according to their parking time; and it requires, on the average, 20 seconds to complete the payment. If cars wish to leave during the afternoon busy period at a rate of 5.4 per

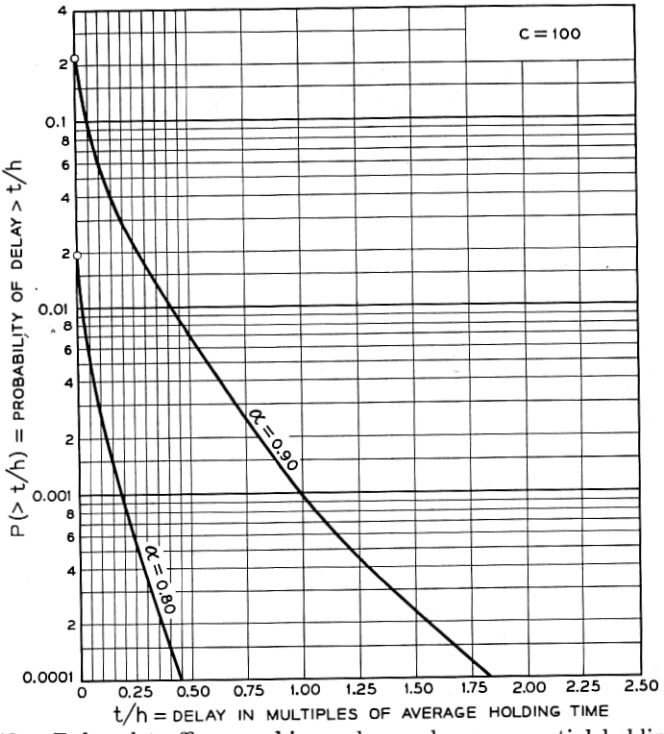

Fig. 18 — Delayed traffic served in random order, exponential holding times, $c = 100$.

minute, what per cent of the cars will be delayed more than 5 minutes? What will be the average delay for all cars?

Solution. Assume there is no traffic supervision and cars converge on the gate from many directions. Service in random order (or worse) among those delayed might then be approximated. Also the distribution of times for calculating and collecting the charge might be roughly exponential. We have then,

$$c = 2 \text{ paths}$$

$$\alpha = (5.4)(20)/(60)(2) = 0.90$$

$$t/h = \frac{5(60)}{20} = 15$$

Enter Fig. 9 at $t/h = 15$, read to the $\alpha = 0.90$ curve, opposite which find $P = 0.069$. Hence 7 per cent of the cars would be expected to have to wait 5 minutes or more. To obtain the average delay for all cars, enter

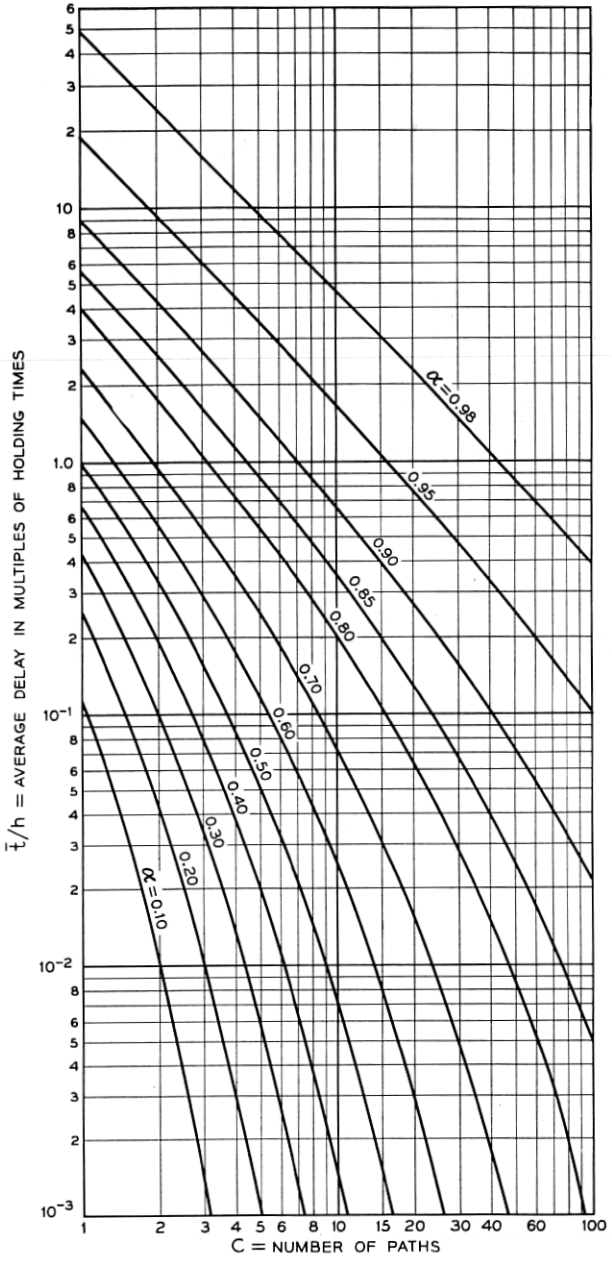

Fig. 19 — Average delay on all calls, exponential holding times.

Fig. 19 with the abscissa of 2 paths, read to the $\alpha = 0.90$ curve and find the average delay = 4.25 average holding times = 85 seconds. Or, one may obtain the same answer by substituting in equation (2),

$$\bar{t} = P(>0)h/(c - a) = (0.85)(20)/(2 - 1.80) = 85 \text{ seconds.}$$

Example No. 6

Suppose in Example 5, an efficient corps of police had been directing traffic toward the exit so that good queueing was maintained. What per cent of the cars would then be delayed more than 5 minutes?

Solution. We may now refer to other published delay curves for queued operation*, or, more generally, calculate the well known equation (1). In the present case we can read the answer from the "queued" curve of Fig. 1 as 4.2 per cent. Thus serving customers in the order of arrival nearly halves the occurrence of very long delays. (Note that the average delay for all cars remains unchanged at 85 seconds.) If a partial queueing were maintained the improvement would be intermediate, perhaps comparable with one of the "limited queueing" distributions shown on Fig. 7.

The author is indebted to Miss C. A. Lennon for constructing the working delay curves, and to Misses C. J. Durnan and J. C. McNulta for performing the throwdown checks.

Appendix

Calculation of delay values not found on the working curves of figs. 8–18, for delayed exponential calls served in random order

A master chart, Fig. 20, reproduced from Riordan†, gives in condensed form the *proportion $F(u)$ of delayed calls delayed* longer than u, where the delay is now expressed in multiples of the h/c ($u = ct/h$), and

c = number of paths (trunks, operators, etc.) provided

h = average holding time

t = delay time

To obtain the probability $P(>t/h)$ of any call being delayed longer than

* E. C. Molina, Ibid.
† Loc. cit.

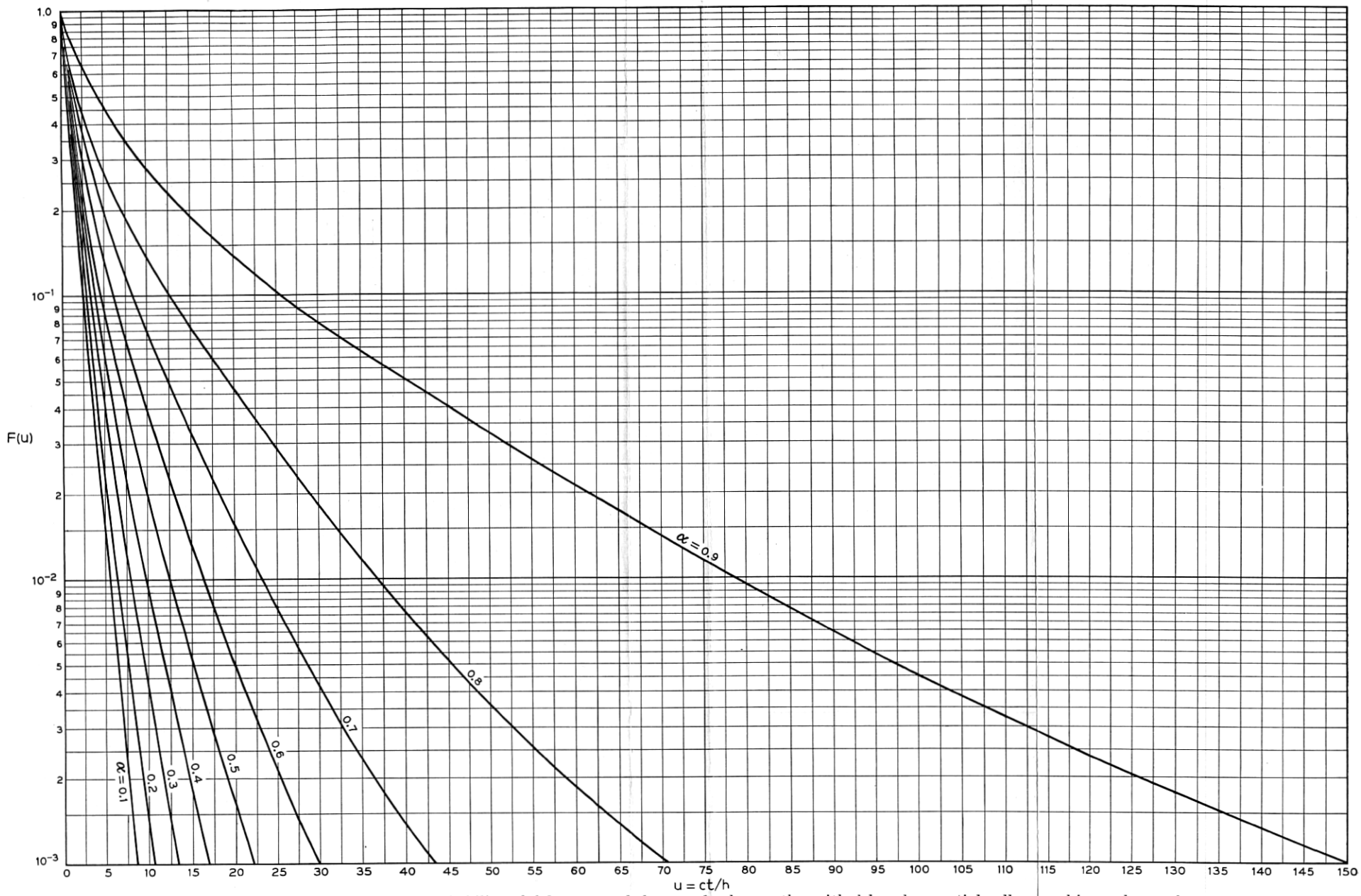

Fig. 20 — General relationship, probability of delays exceeded versus load on paths, with delayed expontial calls served in random order.

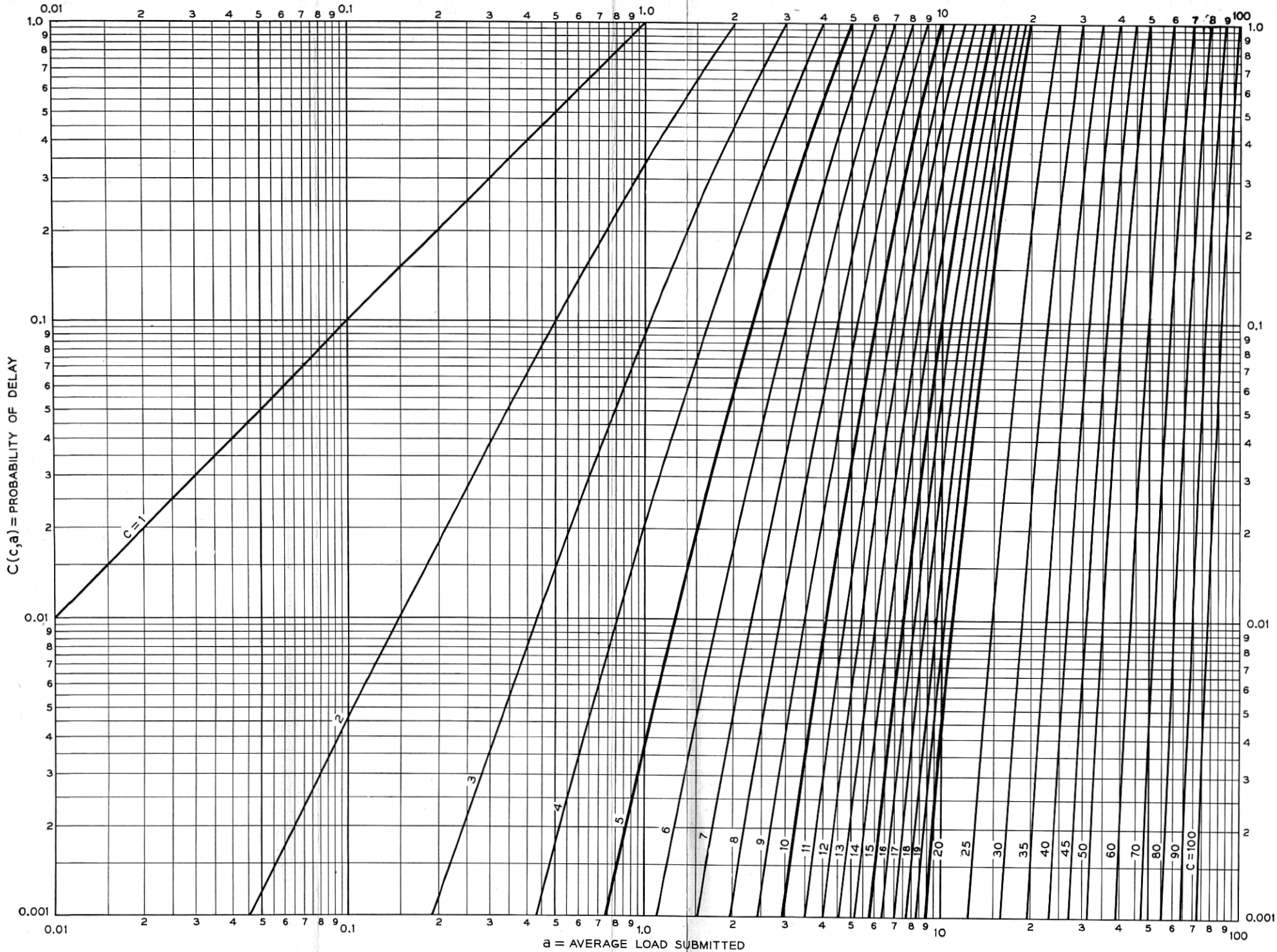

Fig. 21 — Probability of delay for exponential holding time calls handled in a "delayed" basis.

t/h, we have

$$P(>t/h) = P(>0)\, F(u) = C(c, a)\, F(u) \tag{4}$$

Values of $P(>0) = C(c, a)$ are given for a wide range of a and c in Fig. 21. The application of equation (4) is quite simple.

Illustration 1. Suppose it is desired to obtain the probability of a call being delayed more than 3 holding times on a 10 trunk group without storage or gating circuits, and which carries $a = 9$ erlangs. Here $t/h = 3.0$, $c = 10$, $\alpha = 0.9$. Then $u = ct/h = 30$, and reading on Fig. 20 with this value of u, and $\alpha = 0.9$, we find $F(u) = 0.080$. Fig. 21 provides $C(c, a) = 0.67$ for $a = 9$ and $c = 10$. Substituting in equation (4),

$$P(>3 \text{ hold times}) = 0.67\ (0.080) = 0.053,$$

which checks the value read directly from the $c = 10$ curves of Fig. 15.

Illustration 2. With an occupancy of $\alpha = 0.65$ on 15 paths what is the probability of meeting a delay greater than one holding time when delayed calls are served in random order? Calculate $u = ct/h = 15$. Enter with this abscissa on Fig. 20, and interpolating between the $\alpha = 0.6$ and 0.7 curves, read $F(u) = 0.022$. Fig. 21 shows for $a = 0.65(15) = 9.75$ and $c = 15$, $C(c, a) = 0.085$. Hence

$$P(>1 \text{ hold time}) = 0.085(0.022) = 0.0019.$$