

Application of the Theory of Probability to Telephone Trunking Problems

By EDWARD C. MOLINA

IF telephone plants were provided in such quantities that when a subscriber makes a call there would be immediately available such switching arrangements and such trunks or paths to the desired point as may be necessary to establish the connection instantly, it would require that paths and switching facilities be provided to meet the maximum demand occurring at any time, with the result that there would be a large amount of plant not in use most of the time.

Obviously, this would result in high costs, particularly in cases of long circuits or where the switching arrangements are complex.

Sound telephone engineering requires, therefore, that we approach this condition only in so far as it is practical and economical to do so, consistent with good telephone service to the subscriber. To take an extreme case—if enough toll lines were provided between New York and San Francisco so that no call would ever be delayed because of busy lines or busy switching arrangements, the rates it would be necessary to charge would be prohibitive, although the speed of service would be very good. Obviously, it is necessary to adopt a compromise between the number of circuits and amount of equipment and the time required to complete a call.

Handling traffic on any other than an instantaneous basis is generally spoken of as handling it on a "delayed basis," even though this delay may be, and generally is, inappreciable to the subscriber. While most of the traffic is handled practically on a no-delay basis, there are certain kinds that are handled on a "delayed basis," such as

1. Calls handled by toll lines, when all toll lines happen to be busy.
2. Calls served by senders or line finders in machine switching systems.
3. Calls handled by operators; the delays implied here being due to the time required by an operator to perform her functions apart from delays due to limitations of equipment.

For traffic handled in this manner it is desirable to have formulas or curves for determining the percentage of calls delayed and the average delay on calls delayed. The product of these two figures will give, of course, the average delay on all calls.

The object of this paper is to present for consideration the results of some theoretical studies made with reference to calls handled on a delay basis.

It is felt that these results may be applied with but slight modifications to many of those traffic problems in which calls are subjected to delay rather than loss when idle mechanisms for advancing them are not immediately available. Such items as service to the subscribers, wear on selecting mechanisms, reliability of circuit operation, etc., are often dependent on the magnitudes of the delays encountered in such cases. Their application to problems in manual traffic is probably less immediate and precise due to the human element which enters into the reckoning. Such factors as an operator's ability to speed up at times of heavy traffic and the facility with which she may reach distant signals appearing before other positions make the problems rather more involved than those dealing with mechanisms whose reactions under various circumstances are more possible to predict. Nevertheless in such cases as these, as well as in problems relating to the delays encountered in clearing trouble conditions, installing telephones, awaiting elevator service, and many other problems of interest to engineers in general, the results and methods discussed here, though probably not directly applicable, may prove highly suggestive in a qualitative way.

Obviously the average number of calls to be handled per unit of time, the average length of holding time per call, and the number of trunks assigned to handle the traffic are factors entering into the mathematical results. A knowledge of these three quantities is not sufficient, however, for the solution of the problem. Quite different results will be secured according to the assumption made as to how the holding times of individual calls vary about their given average, i.e., one set of results follows from the assumption that holding times are all of the same length—other sets of results if holding times vary, the precise set depending on the particular law of variation assumed.

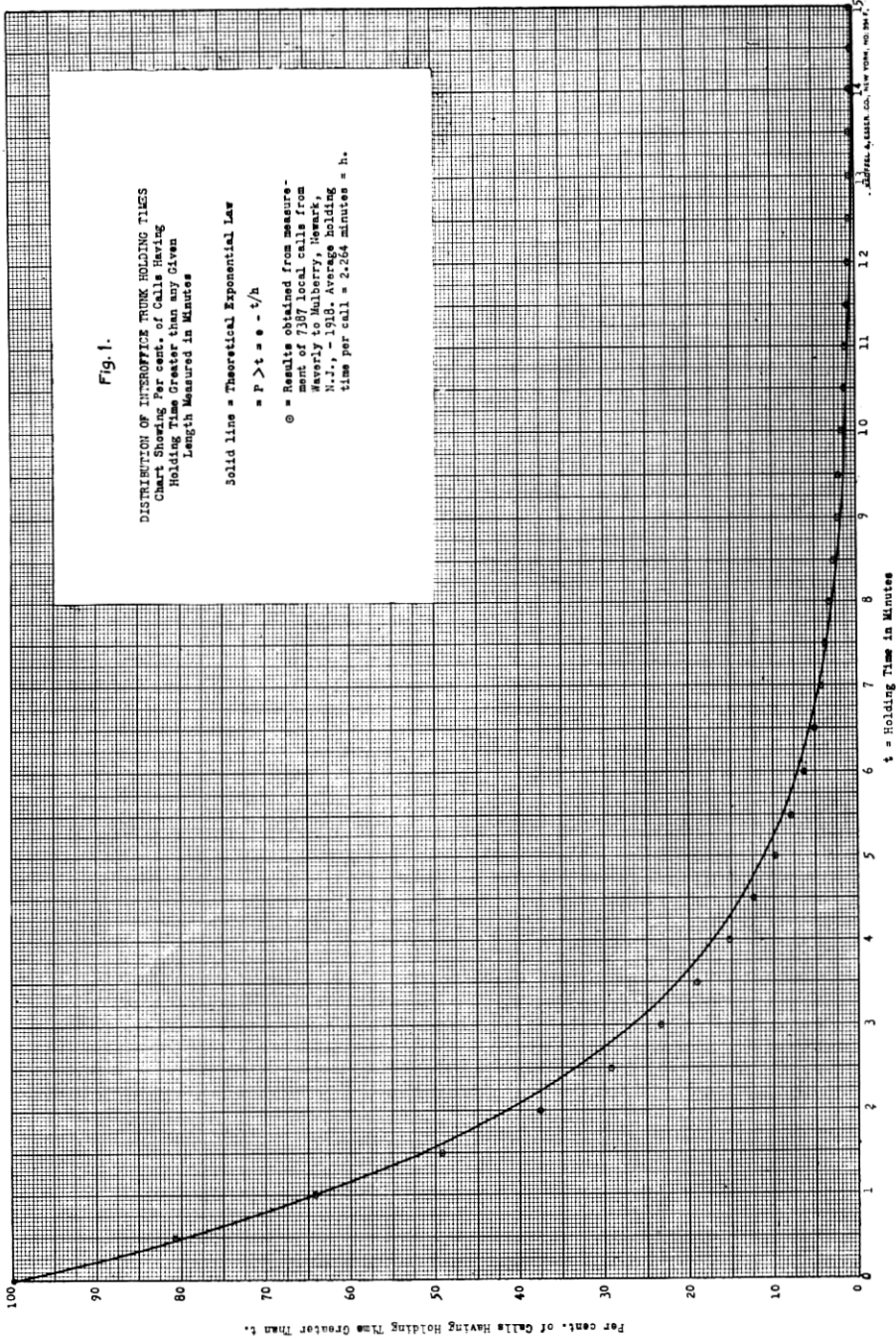
The choice made of laws representing holding time variations must be governed by two considerations:

1. An assumed law must agree at least approximately with the points obtained if we plot the way holding times vary as found from observations.

2. The form of the law must lend itself to the mathematical solution of the delay problem.

CASE NO. 1

An assumption which permits of an easy and exact mathematical solution of the problem may be stated as follows: If a call is picked



Bell Telephone Laboratories, New York, N.Y.

at random, the probability that its holding time is greater than an interval of time of length t is $e^{-t/h}$, where e is the base of natural logarithms and h is the average holding time of all calls. Fortunately there are cases in practice where the variations in length of holding times are closely represented by this exponential law. This is clearly shown by the following Fig. 1 entitled "Distribution of Interoffice-Trunk Holding Times." The points shown on the figure are the plots of actual holding times obtained from pen register records made on a group of trunks running from Waverly to Mulberry in Newark, New Jersey.

CASE NO. 2

Another assumption covered in this memorandum, because it checks closely with cases arising in practice, is that all calls have exactly the same holding time. The precise solution of the delay service problem becomes extremely difficult on this assumption of a constant holding time. An approximate solution is presented in this paper. Cases in practice where holding times are essentially constant are those of sender holding times of key indicator trunk groups and cordless B boards.

With reference to either Case No. 1 or Case No. 2 consider a group of a certain number of operators handling a certain number of calls having a certain average holding time. If we double the average holding time and halve the number of calls (so that the operators are busy for the same per cent of time on the average), the per cent of calls delayed will not change, but the average delay on calls delayed as measured in seconds will exactly double. Suppose, for example, we wish to obtain the same average speed of answer to line signals with two different groups or teams of operators, the first handling traffic which requires only a short operator holding time or work interval, and the second handling traffic which requires a longer work interval. If the teams are equal in number and ability, we must allow the second team a larger proportion of idle time than the first. If, on the other hand, the proportion of idle time is to be the same for both teams, the second team must be larger or more capable than the first. This general effect is well known, but it is hoped that the results herewith presented will supply more exact knowledge of the subject.

Before proceeding further it will be helpful to give here the notation used on the delay curves following this paper.

h = average holding time per call.

c = number of trunks in a straight multiple.

a = average number of calls originating per interval of time h .

$P(>t)$ = probability that a call is delayed for an interval of time greater than t . In other words, 100 times $P(>t)$ gives the per cent of calls which will, on the average, be delayed for an interval greater than t .

CHARTS

The two series of charts following this paper embody, for Case No. 1 and Case No. 2, respectively, curves giving for different values of c and the ratio a/c the probability that a call will be delayed to an extent which will exceed a given multiple of the average holding time. These curves may consequently be read to determine what proportion of the calls will be delayed on the average by an interval as measured in holding times. For example, consider the particular varying holding time chart which corresponds to $c = 10$; we see from the curve marked $a/c = 0.50$ that there is a probability of 0.001 that a call will be delayed for an interval of time which will exceed 0.72 of the average holding time; or, put another way, that 0.1 per cent of the calls will be, on the average, delayed this amount. Again there is a probability of only 0.000019 for a call being delayed at least 1.5 times the average holding time; or 0.0019 per cent of the calls will be delayed by this amount. If, on the same chart, we consider the curve marked $a/c = 0.70$, we find that 22 per cent of the calls will be delayed, 1 per cent will be delayed at least 1.04 times the average holding time, 0.01 per cent will be delayed 2.58 times the average holding time, or more.

The dotted line on each chart gives, at its points of intersection with the curves, the average delay on calls delayed as a multiple of the average holding time interval. For example, on the $c = 10$ varying holding chart we note that for $a/c = 0.70$ the *average delay on calls delayed* is $0.33h$. To obtain the average delay on all calls we multiply by the proportion of calls delayed, $P(>0) = 0.22$, and obtain 0.073 times the average holding time.

A glance at the formulas, given in the Appendix, for the average delay on calls delayed shows that this delay does not reduce to zero when a/c becomes zero. It approaches a lower limit which has the value h/c in Case No. 1 and the value $h/(c + 1)$ in Case No. 2. The latter limit may readily be anticipated from physical considerations as follows. Assume that the group consists of a single trunk; we have to show that the average delay when a call is delayed approaches the limit $h/(1 + 1) = h/2$ as the load approaches zero. Now when the load is very low, those cases where two or more calls have to wait for the trunk to become idle are quite negligible; we only have to

consider the delay incurred by a single call originating while the trunk is busy. But as calls originate at random, the delayed call is just as likely to have fallen near the beginning as toward the end of the constant interval h during which the trunk is busy. In other words, on the average, the delayed call will have originated in the middle of the constant interval h and thus the average delay incurred will be $h/2$. This lower limit for the average delay on calls delayed is indicated in the lower left-hand corner of each sheet of curves by the point where the axis of abscissæ is intersected by a short vertical line.

It will be noted that the constant holding time delay curves change their direction at those points for which the abscissæ are exact multiples of the holding time interval h . No such discontinuities in slope

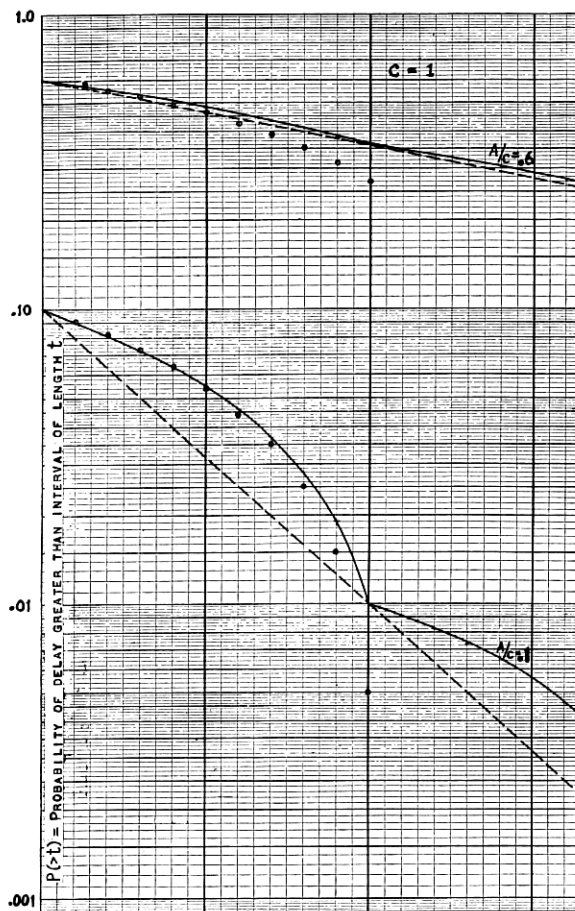


FIG. 4.

appear on the varying holding time curves. This difference in the two classes of curves should occasion no surprise. In the varying holding time case the quantity h has no physical significance; it is merely a numerical value obtained by an algebraic process called averaging. In the other case the quantity h represents a physical characteristic of each and every call.

As stated on page 464 the solution presented in this paper for the case where holding times are all of constant length is not exact. It is desirable therefore to have some idea of the degree of approximation attained.

Figure 4 shows a comparison between our tentative solution and true values which Erlang of Copenhagen, Denmark, succeeded in obtaining by a method which unfortunately becomes impracticable for values of c greater than 3. Our results are shown by the solid curve and Erlang's results by the small circles. We have also indicated on the $c = 1, 2$ and 3 constant holding time charts, Erlang's points for $a/c = 0.50$. For Erlang's work, reference may be had to the *Elektrotechnische Zeitschrift*, December 19, 1918, page 504.

It may be noted that for the higher values of the ratio a/c the curves are practically straight lines. They depart materially from straight lines for the lower values of the ratio a/c , particularly if c itself is not very large.

ASSUMPTIONS MADE IN MATHEMATICAL THEORY

The mathematical theory back of the curves accompanying this paper is based on the following assumptions:

1. Calls originating independently of each other, and at random with reference to time, have complete access to a single group of trunks.
2. The probability of a call originating during a particular infinitesimal interval, dt , is practically independent of the number of trunks busy or number of waiting calls at the beginning of said interval. This assumption implies that the total interval of time during which the calls fall at random is very large compared with the average holding time per call and that the total number of calls under consideration is very large compared with the number of calls originating per average holding time interval.
3. Calls are served in the order in which they originate. This restriction does not apply to the *average* delays obtained.
- 4A. The average holding time being h , the holding times of individual calls vary around this average in such a way that $e^{-t/h}$ is the probability that for a call taken at random the holding time is greater than t .

- 4B. The holding times of all calls are equal to a constant h .
 5B. If, at any instant, s of the c trunks are busy, the distribution in time of the instants at which said s busy trunks were seized is identical with the distribution of s points picked individually at random on a straight line of length h .

Assumptions 1 and 2 together imply that the number of *sources* of calls is so large that any blocking of calls due to limitation of sources need not be considered. Assumptions 1, 2, 3 and 4A were made in deriving the formulas for Case No. 1. Assumptions 1, 2, 3, 4B and 5B were made in deriving the formulas for Case No. 2. Assumption 5B is not strictly compatible with the physics of the constant holding time case. The distribution in time of the s calls mentioned in 5B will, to a certain extent, depend on the history of previous calls. It is because this dependence is ignored that the solution for Case No. 2, presented in the Appendix, is only approximate.

APPENDIX

MATHEMATICAL THEORY OF DELAY FORMULAS

The following mathematical analysis is based on the assumptions given above on pages 467 and 468.

Consider the state of affairs at the instant a particular call "X" originates. Suppose call "X" encounters x other calls; if x is less than c , call "X" will be served immediately but, if not, "X" will have to wait. Our problem is to determine the probability that the delay which "X" may suffer shall have a specified value.

We begin by determining the relative frequency with which the number of calls encountered by "X" has the value x . Let $f(x)$ be the relative frequency with which x calls are encountered by "X." At an instant of time t , x calls will be encountered if at the preceding instant $(t - dt)$ either x , $(x + 1)$ or $(x - 1)$ calls would have been encountered. We ignore as of too rare occurrence the cases where more than $(x + 1)$ or less than $(x - 1)$ calls would have been encountered at time $(t - dt)$ with x at time t . Now in passing from time $(t - dt)$ to time t the probability of an *increase* of one call is proportional to the difference in time, dt , and to the average number of calls, a , falling per holding time interval. Likewise the probability of a *decrease* of one call is proportional to the time difference, dt , and to the number of calls occupying trunks (a decrease must be due to a busy trunk becoming idle). Therefore

$$f(x) = f(x - 1) \frac{adt}{h} + f(x + 1) \frac{(x + 1)dt}{h} + f(x) \left[1 - \frac{adt}{h} - \frac{xdt}{h} \right]$$

when $x < c$, and

$$f(x) = f(x-1) \frac{adt}{h} + f(x+1) \frac{cdt}{h} + f(x) \left[1 - \frac{adt}{h} - \frac{cdt}{h} \right]$$

when $x \geq c$.

For these two equations we may substitute the simpler equations

$$xf(x) \frac{(dt)}{(h)} = af(x-1) \frac{(dt)}{(h)}$$

or

$$cf(x) \frac{(dt)}{(h)} = af(x-1) \frac{(dt)}{(h)}.$$

The solution of these equations gives

$$f(x) = f(0) \left[\frac{a^x}{c^{x-c}} \right], \quad x < c$$

and

$$f(x) = f(0) \frac{a^x}{x}, \quad x \geq c,$$

where $f(0)$ is the arbitrary constant entering in the integration of the finite difference equations. But we must have, evidently,

$$\sum_{x=0}^{x=\infty} f(x) = 1.$$

Substituting in this equation the values for $f(x)$ given above, we obtain

$$1/f(0) = e^a \left[1 - P(c, a) + \frac{a^c e^{-a}}{c} \left(\frac{c}{c-a} \right) \right].$$

Since call "X" will be delayed whenever the number x of calls he encounters is equal or greater than c , we have

$$P(> 0) = \sum_{x=c}^{x=\infty} f(x) = \frac{\left(\frac{a^c e^{-a}}{c} \right) \left(\frac{c}{c-a} \right)}{1 - P(c, a) + \left(\frac{a^c e^{-a}}{c} \right) \left(\frac{c}{c-a} \right)}.$$

The next question is to determine the probability, $P(> t)$, of a delay which is greater than an interval of length t .

We will get one answer to this question if we make use of assumption 4A, and a different answer on the basis of assumptions 4B and 5B. Therefore, from here on, it will be necessary to treat separately the varying and constant holding time cases.

Case No. 1—Holding Times Vary Exponentially

$P(> t)$ for this case was obtained by Erlang of Copenhagen. In 1917 he published the formula without its proof. The following deduction of his formula is therefore submitted.

The particular call "X" considered above will be delayed if the number of calls he encountered, x , is equal to or greater than the number, c , of trunks in the group. Suppose that the $x = c + (x - c)$ calls encountered by "X" are handled by the trunks in the manner indicated in the following Fig. 2, where $m_1 + m_2 + \dots + m_c = x - c$.

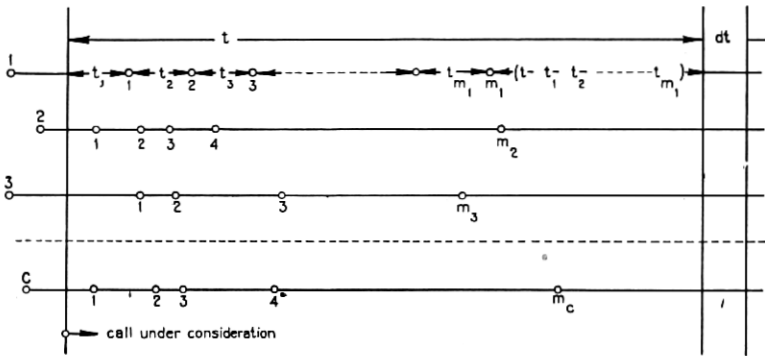


FIG. 2.

By assumption 4A and taking h as the unit of time, the probability that trunk No. 1 will be busy during an interval of time t is

$$(e^{-t_1} dt_1)(e^{-t_2} dt_2) \dots (e^{-t_{m_1}} dt_{m_1}) e^{-(t-t_1-t_2 \dots t_{m_1})}$$

Giving $(t_1, t_2 \dots t_{m_1})$ all positive values consistent with their sum $\geq t$, we obtain (see Todhunter's "Integral Calculus," sixth edition, art. 276)

$$e^{-t} \frac{(t^{m_1})}{(m_1)!}$$

Therefore the compound probability that all c trunks are busy during the interval t with the x calls existing at the instant under consideration and that then one of the trunks becomes idle in the interval dt is

$$(e^{-t})^c \left[\sum \left(\frac{t^{m_1} t^{m_2} \dots t^{m_c}}{m_1! m_2! \dots m_c!} \right) \right] (cdt),$$

where \sum means that we are to give $m_1, m_2 \dots m_c$ all values such that

$$m_1 + m_2 + \dots + m_c = x - c.$$

By the multinomial theorem

$$\sum \left(\frac{m_1 + m_2 + \dots + m_c}{\underbrace{m_1} \underbrace{m_2} \dots \underbrace{m_c}} \right) (a^{m_1} b^{m_2} \dots K^{m_c}) = (ct)^{x-c}$$

for $a = b = \dots K = t$ and $(m_1 + m_2 + \dots + m_c) = x - c$.

The expression above reduces to

$$(e^{-ct}) \left[\frac{(ct)^{x-c}}{x-c} \right] c dt.$$

Now all this is on the assumption that "X" encountered x other calls. Therefore we must multiply by $f(x)$ and sum for all values of x from c to ∞ . We obtain

$$\begin{aligned} \sum_{x=c}^{\infty} f(x) e^{-ct} \frac{(ct)^{x-c}}{x-c} c dt &= \\ \sum_{x=c}^{\infty} f(0) \frac{a^x}{c^{x-c} c} e^{-ct} \frac{(ct)^{x-c}}{x-c} c dt &= \\ f(0) \left(\frac{a^c}{c} \right) c e^{-ct} dt \sum_{x=c}^{\infty} \frac{(at)^{x-c}}{x-c} &= \\ f(0) \left(\frac{a^c}{c} \right) c dt e^{-ct} e^{at} = f(0) \left(\frac{a^c}{c} \right) c e^{-(c-a)t} dt. \end{aligned}$$

This is the probability that "X" will be delayed for an interval of length t . To obtain the probability that the delay will be greater than t we must integrate with reference to t from t to ∞ . But

$$\int_t^{\infty} e^{-(c-a)t} dt = \frac{e^{-(c-a)t}}{c-a}.$$

Thus, finally,

$$\begin{aligned} P(> t) &= f(0) \frac{a^c}{c} \left(\frac{c}{c-a} \right) e^{-(c-a)t} \\ &= P(> 0) e^{-(c-a)t}. \end{aligned}$$

This formula for $P(> t)$ has been deduced by taking h as the unit of time. Evidently we would have obtained

$$P(> t) = P(> 0) e^{-(c-a)t/h}$$

if h had not been taken as unit of time.

For the average delay on all calls we have

$$\bar{t} = \int_0^{\infty} t \frac{dP(> t)}{dt} dt = P(> 0) \left(\frac{h}{c-a} \right),$$

and the average delay on calls delayed is

$$\left(\frac{h}{c - a} \right).$$

Case II—Holding Times Constant

Write x , the number of calls encountered by "X," in the form

$$x = nc + m - 1,$$

where n and m are positive integers such that $m \succ c$. In the Fig. 3 below these $nc + m - 1$ calls are shown in groups arranged according

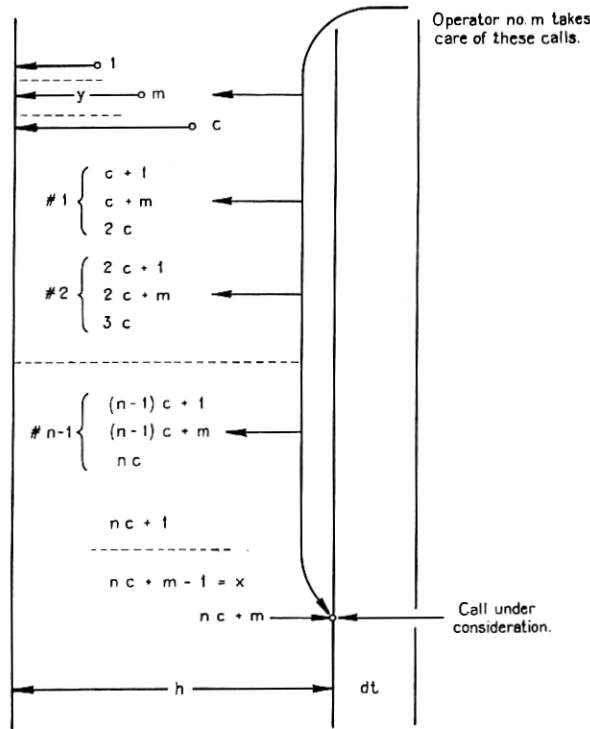


FIG. 3

to the order in which they originated. The first c calls are occupying the trunks. Then $n - 1$ groups, each consisting of c waiting calls, are shown and finally a remainder of $m - 1$ waiting calls; our particular call "X" is the $(nc + m)$ th in order of time.

Now evidently, as indicated in the figure, trunk number m will

serve call "X" after said trunk has served the call occupying it and also the m th call in each of the $n - 1$ waiting groups. Therefore our call will suffer a delay of length

$$(n - 1)h + y,$$

where y is the time which elapsed between the beginning of the interval h and the instant at which the m th trunk was seized by the call occupying it. The probability of this delay is a compound one made up of two factors.

1st—The probability that x calls are encountered. This probability is, as derived above,

$$\begin{aligned} f(x) &= f(0) \frac{a^x}{c^{x-c} \underline{c}} \\ &= \frac{f(0)c^c(a/c)^{nc}}{\underline{c}} \left(\frac{a}{c}\right)^{m-1}, \end{aligned}$$

since $x = nc + m - 1$.

2d—The probability that the distance from the beginning of the interval h to the instant at which the m th trunk was seized is y or, in more precise terms, lies between y and $y + dy$. This probability is, on the basis of assumption 5B,

$$c \left(\frac{dy}{h}\right) \left[\binom{c-1}{m-1} \left(\frac{y}{h}\right)^{m-1} \left(1 - \frac{y}{h}\right)^{(c-1)-(m-1)} \right].$$

The product of these two probabilities gives, writing $y = uh$, $(a/c) = R$,

$$\frac{f(0)c^{c+1}(R)^{nc}}{\underline{c}} \left[\binom{c-1}{m-1} \left(\frac{uR}{1-u}\right)^{m-1} (1-u)^{c-1} \right] du.$$

But the subscribers' interest in a delay of magnitude $(n - 1)h + y$ is totally independent of what value m might have. Therefore the last probability expression must be summed for all permissible values of m , that is from $m = 1$ to $m = c$. We then obtain for the total probability of a delay of extent between $(n - 1)h + y$ and $(n - 1)h + y + dy$:

$$\frac{f(0)c^{c+1}R^{nc}(1-u)^{c-1}}{\underline{c}} \left[\sum_{m=1}^c \binom{c-1}{m-1} \left(\frac{uR}{1-u}\right)^{m-1} \right] du =$$

$$\frac{f(0)c^{c+1}R^{nc}(1-u)^{c-1}}{c} \left[1 + \frac{uR}{1-u} \right]^{c-1} du =$$

$$\frac{f(0)c^{c+1}R^{nc}}{c} [1 - (1-R)u]^{c-1} du =$$

$$P(>0)(c-a)R^{(n-1)c} [1 - (1-R)u]^{c-1} du.$$

We now obtain by integration with reference to u , summation with reference to n , or both, the following results.

Character of Delay	Probability of Delay
From $(n-1+u)h$ to nh	$P(>0)R^{(n-1)c}([1 - (1-R)u]^c - R^c)$
From $(n-1)h$ to nh	$P(>0)R^{(n-1)c}(1 - R^c)$
Greater than $(n-1)h$	$P(>0)R^{(n-1)c}$
Greater than $(n-1+u)h$	$P(>0)R^{(n-1)c}[1 - (1-R)u]^c$
Average delay on all calls	$P(>0) \left(\frac{h}{c-a} \right) \left(\frac{c}{c+1} \right) \left(\frac{1-R^{c+1}}{1-R^c} \right)$

