General Stochastic Processes in Traffic Systems with One Server

By V. E. BENEŠ

(Manuscript received September 1, 1959)

Congestion in systems with one server, exemplified by simple queues, individual telephone trunks and particle counters, is considered without any restrictions on the statistical character of the incoming load. Elementary methods establish formulas and equations describing probabilities of delay and loss. These methods deemphasize special statistical models and yield a general theory. In spite of this generality, it is attempted to give intuitive proofs and extensive explanations of the physical significance of formulas, as well as rigorous derivations. The theory is applied to specific models to obtain illustrative new results.

I. INTRODUCTION AND SUMMARY

Congestion theory is the study of mathematical models of service systems such as telephone central offices, waiting lines and trunk groups. It has two practical uses: to provide engineers with specific mathematical results — curves and tables — for designing actual systems, and to provide a general framework of concepts into which new problems can be fitted and in which current problems can be solved. Corresponding to these two uses are two kinds of results: *specific results* pertaining to special models and *general theorems* valid for many models.

Most of the present literature of congestion theory consists of specific results resting on particular statistical assumptions about the traffic in the service system under study. Indeed, few results are known in congestion theory that do not depend on special statistical assumptions, such as negative exponential distributions or independent random variables. In this paper we obtain some mathematical results that are free of statistical restrictions, but which apply only to a limited class of systems. These results concern general stochastic processes in traffic systems, such as counters and simple queues, that have only one server.

The limitation to a single server is severe, and may be difficult to

overcome. We hope that appropriate analogs of the methods used here, applied to stochastic processes in several dimensions, will yield statistically general theories for systems with many servers. Some systems, for example, queues with many servers in parallel, cannot be given a theory free of special assumptions without the use of multidimensional processes; other systems have a structure that permits us, in principle, to treat them by "iterating" results for one server. In Section VIII we suggest, as a conjecture, how the general theory we give for one telephone trunk might be applied several times to give a general loss formula for a group of trunks.

The classical problems of telephone traffic engineering usually involve many servers, and so theories dealing with single servers have been of peripheral interest to traffic engineering in the past. However, modern increases in the speed of components have made possible *electronic* telephone exchanges, now under study and development. By dint of their speed, some of these systems either have a single active server (the "common control"), or may reasonably be viewed as a set of (relatively independent) single servers. Therefore, the case of one server is in fact of immediate and practical interest.

The aims of this paper are three: (a) to describe a new general approach to certain congestion problems; (b) to show that this approach, although general, can nevertheless be presented in a relatively elementary way that makes it available to many persons; (c) to illustrate how the new approach yields specific results, both new and known. What follows is written only partly as a contribution to the mathematical analysis of congestion. It is also a frankly tutorial account, aimed at increasing the public understanding of congestion by first steering attention away from special statistical models and then obtaining a general theory. Such a point of view, it is hoped, will yield new methods in problems other than congestion.

When a general theory can be given, it should be useful in several ways: in increasing our understanding of complex systems; in obtaining new specific results, curves, tables, etc; in extending theory to cover interesting cases that are known to be inadequately described by existing results. At first acquaintance, the theorems of such a general theory may not resemble "results" at all; that is, they may not seem to be facts that one could obviously and easily use to solve a real problem. A general theory is really a tool or principle expressing the essence or structure of a system; properly explained and used, this tool will yield formulas and other specifics with which problems can be treated.

Let us give examples of questions we shall be able to answer and re-

sults we shall obtain. As a first example of increased understanding, we shall be able to answer, affirmatively, the question

> Is there a small number of specific features of the incoming traffic which suffices, quite generally, to determine delay distributions and loss probabilities?

We shall exhibit these features, and give delay and loss probabilities in terms of them.

As a second example, of both increased understanding and a new case, let us consider the matter of Markov stochastic processes. A Markov process has a particularly simple probabilistic structure, in that all information about the past history of the randomly moving point is irrelevant to its future development, if its present position is known. Such processes have been very useful in congestion theory, because of the convenient functional equations (such as "statistical equilibrium" equations, continuity equations, renewal equations) that are associated with them. However, it is also true that congestion theory has occupied itself almost exclusively with Markov processes, or with the generalizations of these called regenerative, and semi-Markov, processes. Thus the following question has doubtless been asked:

> What functional equations can be derived for non-Markov problems in delay, or in loss?

In answer, we show that the probability that a queue be empty at t satisfies, quite generally, a Volterra equation of the first kind. This equation coincides with a known equation in the Markov case, as it should; similar results hold for loss.

By way of specific results, we prove that, in many cases of practical interest described by Markov processes, the solution of the Volterra equation has the intuitive form

$$\Pr\{\text{queue is empty at } t\} =$$

$$\text{average of the greater of 0 and } \left(1 - \frac{\text{total load in } [0,t)}{t}\right)$$

$$+ \frac{\text{initial load}}{t} \text{ (chance that total load in } [0,t) \text{ is at most } t).$$

In these cases, it is no longer necessary to solve the Volterra equation; one merely computes the quantities on the right-hand side of the formula just given. The "initial load" is the time the queue would take to empty if no customers arrived after time zero; the "total load in $[0,t)$"

is the initial load plus the sum of all the service-times of customers arriving in $(0,t)$.

This result is illustrated by Fig. 1, which shows the probability that a queue is empty as a function of time for two well-known cases, negative exponential service times and constant service times. In each case, arrivals are in a Poisson process at the rate λ, and the mean service time is b. The curves suggest that the approach to equilibrium is considerably faster with constant service times than with exponential service times. The "exponential" curve was computed from (48), and the "constant" curve from the formula given above. Both curves approach the same limiting value, 0.750.

The physical background of intended applications is discussed briefly in Section II, while Section III describes the cumulative *load* or *traffic*. Sections IV and V are devoted, respectively, to equations characterizing delay and loss operation in terms of the offered load; these sections are purely descriptive, and no probability is involved. Probability first enters in Section VI, where we discuss the general nature of the problems we try to solve, the methods we use and the results we obtain. Formulas and equations are stated and explained for probabilities of delay in Section VII, and for probabilities arising in loss operation in Section

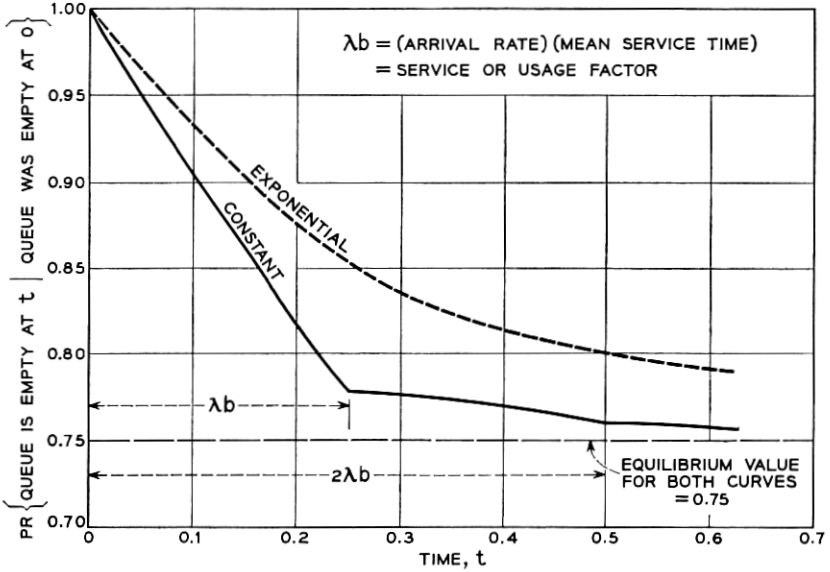

Fig. 1 — Probability that a queue, starting empty at 0, is empty at t, as a function of t for negative exponential and constant service times, with $\lambda = 1, b = 0.25$.

VIII. Proofs of, and additional comments on, the results for delay appear in Section IX, and a number of specific formulas are deduced in Section X. Precise derivations of the results for loss are in Section XI, with an additional limit theorem being presented in Section XII.

II. PHYSICAL BACKGROUND

We shall discuss results applicable to the following type of system: there is given a machine that is either idle or busy; every so often, at random instants, someone tries to use the machine; and the machine can be operated in two ways, called *loss* and *delay*. In delay operation, people who find the machine busy wait for their turns in order of arrival, and then make it busy (by using it) for the originally intended length of time; in loss operation, people who find the machine busy are sent away or "lost," while those who find the machine idle seize it and use it for a random time, the *service time*.

The system to be studied has been described in terms of people who try to use a machine, but the same structure appears in other applications. In telephony, for example, the machine to be used may be a marker, and the "people" arriving may be registers filled with dialed digits; telephone engineers usually refer to the service time as the *holding time*. In the study of radioactivity, the machine is an ionization counter, the "people" are impinging particles, and the service time (called the *dead time* of the counter) is a length of time after the registering of a particle during which the counter cannot count, so that it misses particles arriving during this interval.

The quantities of engineering interest in these models vary with the application, but the following ones seem to be of general importance: for *loss* operation, the chance of loss, i.e., the probability that an arriving customer find the machine busy, and the probability distribution of the busy period, i.e., the amount of time that must elapse until the machine next becomes idle; for *delay* operation, the distribution of waiting time.

III. THE CUMULATIVE LOAD, $K(t)$

Before we can study the two modes of operation, loss and delay, we must describe the offered load of work, or the arriving traffic. This is done most easily by using a step function, $K(t)$, which is nondecreasing and left-continuous. The locations of the jumps are the epochs of arrival of customers, and the magnitudes of the jumps are the service times or the lengths of time the machine is made busy. Equivalently, the offered load is completely determined by the arrival epoch t_k and the service-

time S_k of the kth arriving unit, $k = 1, 2, \cdots$. This situation is depicted in Fig. 2. We shall use $K(t)$ to describe the load for both loss and delay operation.

IV. INTEGRAL EQUATION DESCRIBING DELAY OPERATION

As a mathematical description of the delays to be encountered under delay operation, we use the virtual waiting time, $W(t)$, which can be defined as the time a unit would have to wait for "service" if it arrived at time t. At the epoch t_n of arrival of the nth unit, $W(t)$ jumps upward discontinuously an amount equal to S_n, the work or service time of the unit. Otherwise, $W(t)$ has slope -1 if it is positive; if it reaches zero, it stays equal to zero until the next jump of the load function $K(t)$. Corresponding graphs of $K(t)$ and $W(t)$ appear in Fig. 3.

If $K(t)$ is interpreted as the work offered in the interval $(0,t)$, then $W(t)$ can be thought of as the amount of work remaining to be done at time t. In terms of this interpretation, it can be seen that

work remaining at t = total work load offered up to t − elapsed time + total time during which machine was idle in $(0,t)$.

The machine is idle when, and only when, $W(u) = 0$. Then, formally, $W(t)$ is defined in terms of $K(t)$ by the integral equation

$$W(t) = K(t) - t + \int_0^t U[-W(u)]\,du, \qquad t \geqq 0, \tag{1}$$

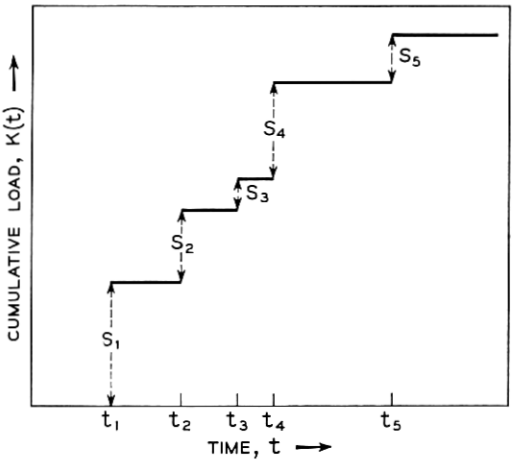

Fig. 2 — Cumulative load.

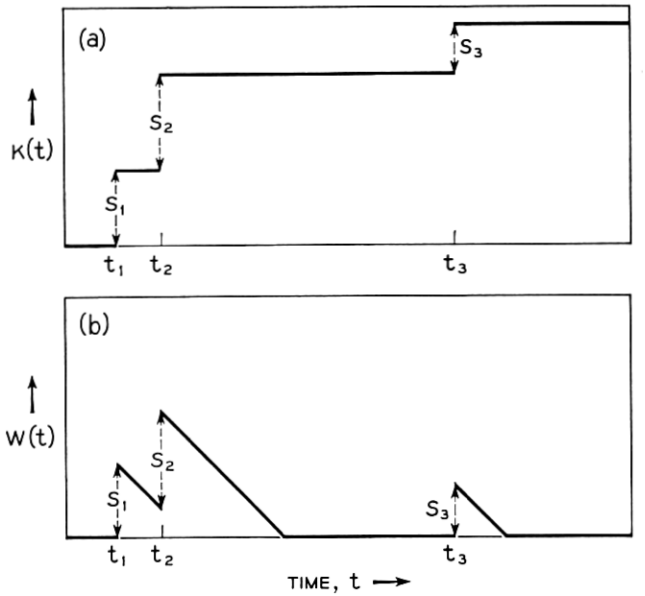

Fig. 3 — (a) Graph of $K(t)$ vs. t; (b) graph of $W(t)$ vs. t.

where $U(t)$ is the unit step function; i.e., $U(x) = 1$ for $x \geqq 0$, and $U(x) = 0$ otherwise. For simplicity, we have set $W(0) = K(0)$.

It has been shown by Reich[1] that the solution of (1) is, if $K(x) - x$ has a zero in $(0,t)$,

$$W(t) = \max_{0 \leqq x \leqq t} [K(t) - K(x) - t + x].$$

This fact may be interpreted physically as follows: the quantity in brackets, $[K(t) - K(x) - t + x]$, is, if positive, the excess of arriving load in the interval (x,t) over the elapsed time $t - x$; it is therefore the *overload* in (x,t). Reich's formula then says essentially that

$$\text{delay at } t = \max_{0 \leqq x \leqq t} [\text{overload in } (x,t)].$$

The relationship between the waiting time $W(t)$ and the offered traffic $K(t)$ can be further elucidated graphically by reference to Fig. 4. The light solid line shows $K(t) - t$, the traffic offered up to time t minus the traffic that could have been served if the server had been kept busy throughout the interval $(0,t)$. It is supposed in Fig. 4 that the server starts busy at $t = 0$. It is busy until $t = a$. At this point, the server becomes idle and $K(t) - t$ turns negative, and its negative value is the

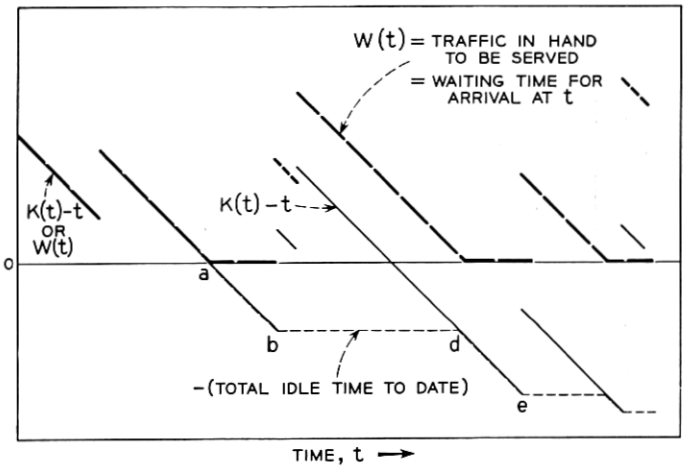

Fig. 4 — Relationship between waiting time and offered traffic.

negative of the idle time. At $t = b$, more traffic is offered and $K(t) - t$ jumps up.

The heavy dashed line represents the waiting time at t, $W(t)$; W(t) can also be thought of as the traffic in hand and yet to be served. It can never go negative. It is equal to $K(t) - t$ before a, and is zero from a to b. At b, it jumps up, remaining above and parallel to $K(t) - t$ until $t = d$, when the server goes idle again. At b, $W(t)$ is above $K(t) - t$ by exactly the amount by which $K(t) - t$ was most negative. At d, when $W(t)$ reaches zero, $K(t) - t$ is just reaching its previous minimum, and $K(d) - d = K(b) - b$.

During the interval (d,e), $W(t)$ remains at zero, while $K(t) - t$ becomes more negative, establishing new minima as it goes and building up more idle time. At $t = e$, $K(t) - t$ and $W(t)$ both jump up; $W(t)$ is again parallel to $K(t) - t$, but it is now above it by an amount equal to the negative of the last minimum, $K(e) - e$.

In Fig. 4,

$$\min_{a \leqq x \leqq t} [K(x) - x], \quad t \geqq a,$$

is shown as a light dashed line. It is a monotone, nonincreasing function of t, and is the negative of the total idle time up to time t. To account for the period $t < a$, when $K(t) - t$ has not yet become negative and the server has not yet been idle, we write

$$W(t) = K(t) - t - \min \{0, \min_{0 \leqq x \leqq t} [K(x) - x]\},$$

and thus obtain another representation for the delay, i.e., a solution of (1).

In a manner similar to that of Fig. 4, Fig. 5 depicts, simultaneously, the offered load $K(t)$ in a light solid line; the waiting-time $W(t)$ in a heavy solid line; the negative of the accumulated idle time in a heavy dashed line; and the "load-time excess" $K(t) - t$, when it does not coincide with the negative of the idle time, in a light dashed line. The terminology in Fig. 5 has been purposely chosen to suggest an interpretation in terms of inventory or storage theory: $W(t)$ is the *real backlog* (of orders, say,), $K(t)$ is the *cumulative amount ordered* and $K(t) - t$ might be termed the load time excess or the *virtual backlog*. Then

$$\text{real backlog} = \text{virtual backlog} + \text{accumulated idle time.}$$

V. INTEGRAL EQUATION DESCRIBING LOSS OPERATION

For loss operation, we use $A(t)$, the service time remaining at t, as an indicator of the condition of the machine. We may define $A(t)$ as

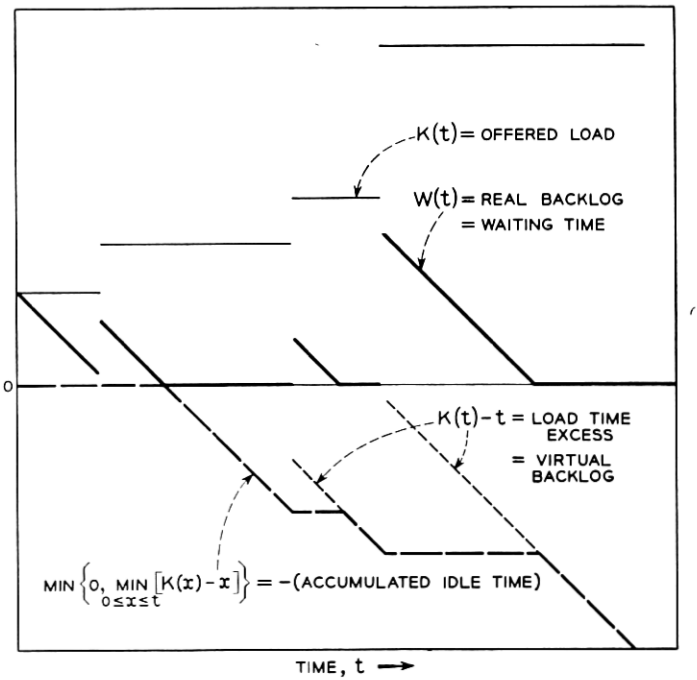

Fig. 5 — Relationships among offered load, waiting time, negative of accumulated idle time and "load time excess".

the amount of time remaining until the machine next becomes idle, with the convention that $A(t) = 0$ means that the machine is idle at t. At epochs of arrivals, $A(t)$ jumps upward discontinuously if the arrival in question finds the machine idle, the amount of jump being the service time of the arriving customer. Arrivals that find the machine busy do not affect $A(t)$. While $A(t)$ is positive, it has slope -1; when it reaches zero, it stays equal to zero until the next jump. A graph of $A(t)$ consists of isosceles "right triangles" laid on their sides; corresponding graphs of $K(t)$ and $A(t)$ appear in Fig. 6. For the study of $A(t)$, it is more convenient to define $K(t)$ to be right-continuous.

In terms of the interpretation of $K(t)$ as cumulative work load it is easy to see that

work remaining to be done at t = total work load offered up to t

− load missed in $(0,t)$ because machine was busy

− elapsed time + total time during which machine was idle in $(0,t)$.

This means that $A(t)$ is defined in terms of $K(t)$ by the equation

$$A(t) = K(t) - \int_{0+}^{t} \{1 - U[-A(u)]\}\, dK(u) - t + \int_{0}^{t} U[-A(u)]\, du, \tag{2}$$

where, as before, $U(t)$ is the unit step function, and we have set $A(0) = K(0)$.

VI. CHARACTER OF THE GENERAL RESULTS

Models of waiting lines and telephone traffic usually contain explicit assumptions about the statistical nature of the offered load $K(t)$. For instance, Erlang's original models[2] amount to assuming that the interarrival times $(t_n - t_{n-1})$ are all independent with the same negative exponential distribution; and similarly for the service times. These assumptions give a class of models whose parameters are the means of the negative exponential distributions.

A broader class of models is specified by retaining the assumptions that the interarrival times be independent and identically distributed (and similarly for service times), but allowing any distribution, not just the negative exponential. The interarrival and service time distributions may still be said to "parametrize" this broader class of models, since their choice determines a model in the class.

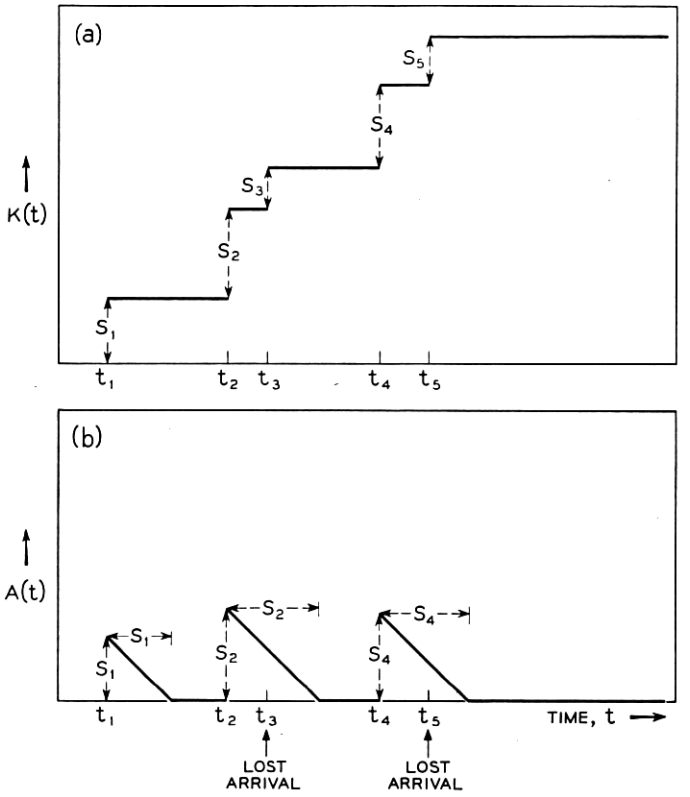

Fig. 6 — Graphs of (a) $K(t)$ vs. t and (b) $W(t)$ vs. t for loss operation.

As more general assumptions even than those of the last paragraph are considered, it becomes extremely laborious first to specify the model and then to compute interesting quantities such as distributions of delay, probabilities of loss, etc. So, instead of looking for ways of exactly characterizing the model, we can try to search directly for simple ways of expressing the quantities of interest in terms of the model. Since the probabilities

$$\Pr\{W(t) \leqq w\}, \tag{3}$$

$$\Pr\{A(t) \leqq w\} \tag{4}$$

are what we actually wish to compute from the model, the question arises whether this calculation can be made without first specifying the entire probabilistic structure of $K(t)$. The following intuitive argument

can be adduced for answering "yes". Both $W(t)$ and $A(t)$ are defined in terms of the load by very special relationships that are expressed in the integral equations we have given. Hence, no matter what are the statistical features of $K(t)$, it is likely that the distributions of $W(t)$ and $A(t)$ depend only on some very particular, physically interpretable statistical functions associated with $K(t)$. It is not obvious that such an economy can be made in the generality we desire.

Two principal results, described in the next sections, state that the probabilities (3) and (4) can in fact be given fairly simple expressions that are generally valid for any load process. These expressions depend only on certain special functions obtainable from the statistical structure of the load $K(t)$. Each function has a definite intuitive or physical significance, to be given later. These statistical functions achieve the desired economy of description, because we can state that the desired probabilities depend only on the features of $K(t)$ expressed in the functions. For the purposes of calculating (3) and (4), we do not need the entire probabilistic structure of $K(t)$, but only the small relevant part.

Another way of putting the problem that we have attempted to solve is: for general load processes $K(t)$, what small amount of information about the statistical nature of $K(t)$ will determine $\Pr\{W(t) \leqq w\}$ for all t and w, and how is this computation to be made? [A similar question arises for $A(t)$.] The statistical functions are then the *information about* $K(t)$; the formulas for the probabilities (3) and (4) indicate the *method of computation*. Specifically, (1) defines a transformation of a stochastic process, $K(t)$, of service times and arrival epochs into another stochastic process, $W(t)$, of waiting times. For each t, there is an operator or formula that gives the distribution of $W(t)$ in terms of suitable fundamental statistical functions associated with $K(u)$ for $u \leqq t$. The principal problem is to find the form of the operator and the character of the fundamental functions. The answer to this problem should depend only on the integral equation (1) and on the fact that $K(t)$ is a nondecreasing step function. It should depend on no special features of the probability measure for $K(t)$ except those implied by this last property. Accordingly, we shall assume at first that $K(t)$ is a random, nondecreasing step function; its *only* statistical peculiarity is that it is a nondecreasing step function.

There already exists an extensive literature dealing with probabilistic models of telephone trunks, electronic counters and similar machines; we shall attempt to relate our approach to this literature. The models used to date have usually included strong hypotheses of independence or else used special distribution functions, such as the negative exponen-

tial. The results obtained depend methodologically, almost without exception, on the possibility of writing Kolmogorov equations for a Markov process, or integral equations of renewal type for probabilities or expectations associated with a regenerative process (Smith[3]). The extensive work of Takács[4,5,6] falls entirely in this category (see also the references therein).

Our approach is, in a sense, an inversion of the usual method described above. The latter consists in first doing probability theory to set up Kolmogorov or renewal equations, and then doing analysis to solve the equations. We can, however, achieve greater generality by taking maximum advantage of the fact that the processes of interest, $W(t)$ and $A(t)$, already satisfy (1) and (2), respectively. This we do by careful analysis of $W(t)$ and $A(t)$ themselves, by in effect performing some of our analysis in the domain of random variables, and taking averages only at convenient points.

VII. PROBABILITIES OF DELAY

It has been shown[7] that $\Pr\{W(t) \leqq w\}$ can be expressed in terms of the two statistical functions

$$\Pr\{K(t) \leqq w\}, \tag{5}$$

$$R(t,u,w) = \Pr\{K(t) - K(u) - t + u \leqq w \mid W(u) = 0\}. \tag{6}$$

We first clarify the physical significance of the functions (5) and (6) entering into the representation of $\Pr\{W(t) \leqq w\}$. The first is simply the probability distribution of the cumulative load, $K(t)$. The second, written as

$$R(t,u,w) = \Pr\{[K(t) - t] - [K(u) - u] \leqq w \mid W(u) = 0\}$$

is the (conditional) probability distribution of the change in the "overload" that occurs in the interval (u,t), given that the server was idle at time u. Here we have used the term "overload" to mean the amount of work that would have been left still undone even if the traffic had been so spaced that there was no idle time. This overload is represented by $K(t) - t$. A negative value of the overload represents the negative of the idle time that would have occurred if the traffic spacing had permitted the server to finish all of the work by time t.

The significance of $R(t,u,w)$ can be made more apparent by describing how to measure it. We look at many identical copies of the system at a time u when the server is idle, and we measure the totaled service times

of all customers arriving in the next interval (u,t), for $t > u$; the fraction of these numbers that is in the range $(0, t - u + w)$ is $R(t,u,w)$.

The delay $W(t)$ is never negative, and so

$$\Pr\{W(t) \leqq w\} = 0, \qquad \text{if } w < 0. \tag{7}$$

For $w > 0$, $\Pr\{W(t) \leqq w\}$ is given by

$$\Pr\{W(t) \leqq w\} = \Pr\{K(t) - t \leqq w\} - \frac{\partial}{\partial w} \int_0^t R(t,u,w) \Pr\{W(u) = 0\}\, du. \tag{8}$$

The chance, $\Pr\{W(u) = 0\}$, that the server will be idle at time u satisfies the Volterra equation of the first kind, for $-t \leqq w \leqq 0$,

$$\text{average of } \max[0,w - K(t) + t] = \int_0^{t+w} R(t,u,w,) \Pr\{W(u) = 0\}\, du, \tag{9}$$

and the left side of (9) is expressible in terms of (5) as

$$E\{\max[0,w - K(t) + t]\} = \int_0^{t+w} \Pr\{K(t) \leqq u\}\, du,$$

where $E\{u\}$ is the expectation of u.

Once the "basic" functions (5) and (6) are known, the computation of $\Pr\{W(t) \leqq w\}$ proceeds by first solving the integral equation (9) for the chance $\Pr\{W(u) = 0\}$ that the server will be idle at time u. This probability can then be used in (8) to give $\Pr\{W(t) \leqq w\}$.

Proofs for, and extensive explanations of, these results for delay have been deferred until Section IX, while we continue by discussing results for loss operation. The proofs to be given are new, and much simpler than those of Ref. 7. They also make it easier to exhibit the physical significance of the formulas and functions arising.

VIII. FORMULAS FOR LOSS OPERATION

We shall show that $\Pr\{A(t) \leqq w\}$ can be expressed in terms of two kernels, $R(t,u)$ and $Q(t,u)$, as follows:

$$\Pr\{A(t) \leqq w\} = \Pr\{A(0) \leqq t + w\} - \int_0^t [1 - R(t + w,u)]\, dE\{S(u)\}, \tag{10}$$

where $E\{S(u)\}$ is the average number of units that arrive in $(0,u]$ and find the machine idle. It satisfies the integral equation

$$E\{S(t)\} = \Pr\{y_1 \leqq t\} + \int_0^t Q(t,u)\, dE\{S(u)\}, \tag{11}$$

where y_1 is the epoch of the first successful arrival.

The kernel $R(t,u)$ may be interpreted as a rigorous version of

> Pr{service time of a successful unit arriving at u is $\leqq t - u$ | a successful unit has arrived at u},

and $Q(t,u)$ may be thought of as a rigorous version of

> Pr{next successful arrival after u occurs before t | a successful unit has arrived at u}.

The first term on the right in (10) is self-explanatory. Precise definitions of $R(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ are given in Section XI.

To explain (10) itself, we observe that $A(0) > t + w$ implies $A(t) = A(0) - t > w$, and hence

$$\Pr\{A(t) \leqq w\} \leqq \Pr\{A(0) \leqq t + w\}.$$

The integral term in (10) is therefore a correction to the *overestimate*, $\Pr\{A(0) \leqq t + w\}$. From the interpretation of $R(t,u)$, and that of $dE\{S(u)\}$ as the "density" of successful arrivals at u, we see that (10) can be rendered in words as

$$\Pr\{\text{work time left at } t \leqq w\} = \Pr\{\text{work time left at } 0 \leqq t + w\}$$
$$-\Pr\{\text{some successful arrival during } (0,t] \text{ stays beyond time } t + w\}.$$

It is reasonable to suspect that, if the load process $K(t)$ has some weak stationarity properties, and if certain "averages" exist, then

$$\Pr\{A(t) \leqq w\}$$

has, for each $w \geqq 0$, a nonzero limit as $t \to \infty$. For $w = 0$, one expects intuitively that this limit will have the form

$$\Pr\{A(\infty) = 0\} = \frac{a}{a + b},$$

where a,b are constants which can be interpreted as follows: if we watch the process $A(t)$, we notice that periods of time during which $A(t) = 0$ alternate with those during which $A(t) > 0$; then a is the average length of a period during which $A(t) = 0$, and b is the average length of a period during which $A(t)$ exceeds 0.

These conjectures are justified in Section XII, where it is shown that, if the kernels $R(t,u)$ and $Q(t,u)$ used in (10) and (11) are functions of

$(t - u)$ only, and if

$$1 - R(x) \tag{12}$$

is integrable over $(0, \infty)$, then

$$\lim_{t \to \infty} \Pr\{A(t) \leqq w\}$$

exists and has the form

$$\frac{a + \int_0^w [1 - R(u)]\, du}{a + b},$$

with

$$b = \int_0^\infty [1 - R(u)]\, du,$$

$$a + b = \int_0^\infty [1 - Q(u)]\, du.$$

We have already pointed out that $R(t,u)$ can be interpreted as

$$\Pr\{\text{service time of a successful unit which arrives at } u \text{ is} \leqq t - u \mid \text{a successful unit arrived at } u\}. \tag{13}$$

Then $R(t,u) = R(t - u)$ states that (13) does not depend on the first and third occurrences of u therein; i.e., that service times have the same distribution no matter when they begin. Thus, the dependence of $R(t,u)$ on $(t - u)$ only is a weak sort of stationarity property, and $R(x)$ can be interpreted as the probability distribution of service times of successful units. Then

$$b = \int_0^\infty [1 - R(u)]\, du = \{\text{average service time of successful units}\},$$

since when the mean of a positive variate exists it equals the integral of the "tail" of its distribution.

In a similar way, the fact that $Q(t,u)$ is a difference kernel may be interpreted as a stationarity condition, and $Q(x)$ can be thought of as the distribution function of the intervals between successful arrivals, so that

$$a + b = \int_0^\infty [1 - Q(x)]\, dx$$

$$= \{\text{average interval between successful arrivals}\}$$

and

$$a = \text{average length of an idle period.}$$

The preceding discussion suggests that we use

$$\text{chance of loss} = \frac{b}{a + b} \tag{14}$$

as a natural measure of the probability of loss for the single telephone trunk or particle counter that we are considering. Of course, any engineer would have used (14) to describe loss, justifying it by intuitive arguments. This fact does not detract from our result, which gives some idea of the weak assumptions that are sufficient for *proving* (14).

Suppose that, instead of having only one machine, we had $N \geqq 1$ machines, and used them in a fixed serial order of preference. That is, arrivals finding the first n machines busy try the $(n + 1)$th. Such a situation arises in telephony, for instance: the "machines" are telephone trunk lines, the "arriving units" are attempts to place a call and the work or service times are the holding times of calls. Each trunk is then receiving the traffic overflowing the previous trunks in the ordering. Our theory then applies to each trunk considered by itself, and if the conditions for the validity of (14) obtain for each trunk, the chance of loss for the whole group must have the form

$$\prod_{n=1}^{N} \frac{b_n}{a_n + b_n},$$

where

$$b_n = \left\{\begin{matrix}\text{average service time of units that find the first (n} - 1)\\ \text{trunks busy, the } n\text{th idle}\end{matrix}\right\},$$

$$a_n + b_n = \left\{\begin{matrix}\text{average time interval between calls accommodated on}\\ \text{the } n\text{th trunk}\end{matrix}\right\}.$$

Formulas (10) and (11) have been essentially proved[8] by involved arguments using the integral equation (2) defining $A(t)$. We shall give a simple heuristic derivation in this section and a rigorous one in Section XI.

In order to explain the results, we first recall that the process $A(t)$ consists of alternating intervals during which $A(t)$ is first zero, then positive with slope -1, then zero again, and so on. The next arrival to find the machine idle always makes it busy again.

We are interested in expressing $\Pr\{A(t) \leqq w\}$, and so we search for

other ways of specifying the event $\{A(t) \leqq w\}$. We first consider those cases in which $A(0) = 0$; that is, the system starts empty, as in Fig. 7. Then it is not hard to see that $A(t)$ will be less than or equal to w only if the number $S(t)$ of successful arrivals during $(0,t]$ equals the number of successful arrivals in $(0,t]$ that have left the system by time $t + w$. In other words, if the machine is idle at $t = 0$, then the work $A(t)$ remaining at time t is less than or equal to w if and only if all people who arrived to find it idle in $(0,t]$ are finished with it by time $t + w$. We shall set

$$\theta(t,t + w) = \text{number of successful arrivals in } (0,t] \text{ who have left the system by time } t + w.$$

Then, if $A(0) = 0$, the events

$$\{A(t) \leqq w\} \quad \text{and} \quad \{S(t) = \theta(t,t + w)\} \tag{15}$$

are the same.

If $A(0) > 0$, the system starts busy, and the graph of $A(t)$ appears as in Fig. 8. Assume first that $A(0) > t$; then, of course, $A(t) = -t + A(0)$, because the machine has been busy since $t = 0$, and is not yet finished. If $t \geqq A(0)$, though, the machine became idle at $t = A(0)$. In the first instance, $S(t) = \theta(t,t + w) = 0$, because there have not yet been any successful arrivals, and $A(t) \leqq w$ if and only if

$$t < A(0) \leqq t + w.$$

In the second instance, the argument we used for the case $A(0) = 0$ applies, and (15) holds. We now average these cases according to their

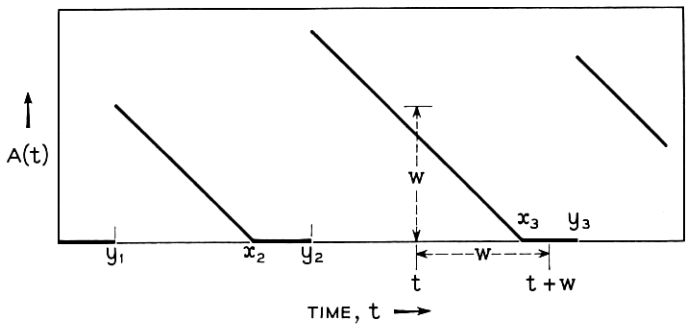

Fig. 7 — Graph of $A(t)$ vs. t when system starts empty.

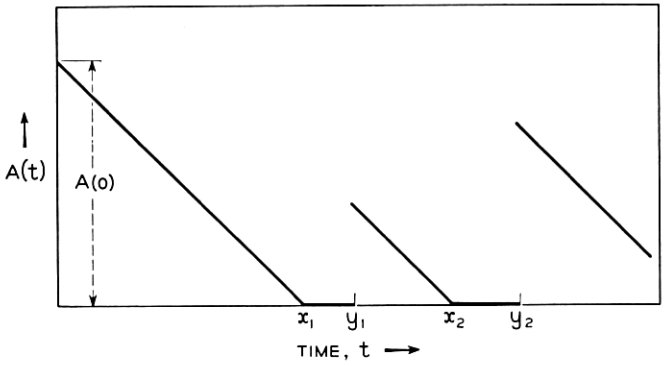

Fig. 8 — Graph of $A(t)$ vs. t when system starts busy.

probabilities; since either $S(t) - \theta(t,t+w) = 0$ or $S(t) - \theta(t,t+w) = 1$, we find

$$\Pr\{A(t) \leqq w\} = \Pr\{t < A(0) \leqq t+w \quad \text{and} \quad S(t) = \theta(t,t+w)\} + \Pr\{A(0) \leqq t \quad \text{and} \quad S(t) = \theta(t,t+w)\}. \tag{16}$$

Where $\{\cdots\}$ is an event, let $\chi\{\cdots\}$ be its *characteristic function*, defined as 1 if the event happens, and 0 otherwise. Then the first term on the right in (16) is

$$\text{average of } \big(\chi\{t < A(0) \leqq t+w\} \cdot [1 - S(t) + \theta(t,t+w)]\big),$$

and the second is

$$\text{average of } \big(\chi\{A(0) \leqq t\} \cdot [1 - S(t) + \theta(t,t+w)]\big),$$

because

$$1 - S(t) + \theta(t,t+w) = 0 \qquad \text{if} \quad S(t) \neq \theta(t,t+w).$$

Therefore

$$\Pr\{A(t) \leqq w\} = \Pr\{A(0) \leqq t+w\} - E\{S(t)\} + E\{\theta(t,t+w)\}. \tag{17}$$

Thus, we have expressed $\Pr\{A(t) \leqq w\}$ in terms of the initial distribution and the average values of $S(t)$ and $\theta(t,t+w)$.

To prove (10), it remains to express the average of $\theta(t,t+w)$ as the integral of a kernel with respect to the average of $S(t)$. To do this, let u be a point in $(0,t]$, and suppose that a successful arrival occurs at u. Such an arrival can only contribute to the average of $\theta(t,t+w)$ if it leaves before $t+w$. The proportion of such arrivals is just

$$\begin{aligned}&\Pr\{\text{service time of a successful unit which arrives at } u \text{ is}\\ &\leqq t + w - u \mid \text{a successful arrival occurred at } u\}\\ &= R(t + w,u).\end{aligned} \tag{18}$$

The "density" of successful arrivals at u is $dE\{S(u)\}$; therefore,

$$E\{\theta(t,t + w)\} = \int_0^t R(t + w,u)\, dE\{S(u)\},$$

which proves (10).

We continue with a heuristic derivation of the integral equation (11) for $E\{S(t)\}$. First we notice that the event $\{A(t) = 0\}$ can occur in two ways: either some successful arrival has occurred in $(0,t]$, or none has. If none has, then either $A(\cdot)$ started idle at 0 and is still idle at t, or else it started busy at 0, became idle at the point $A(0) < t$ and is still idle at t. If t_1 is the first arrival in $(0,\infty)$ and y_1 the first successful arrival, the chance that no successful arrivals occurred in $(0,t]$ and $A(t) = 0$ is

$$\Pr\{A(0) = 0 \quad \text{and} \quad t_1 > t\} + \Pr\{0 < A(0) \leqq t \quad \text{and} \quad y_1 > t\}.$$

Assuming some successful arrival did occur in $(0,t]$, suppose it occurred at u. Such an arrival is only relevant to the event $\{A(t) = 0\}$ if it is the last such arrival in $(0,t]$, and if the service time of the customer then arriving is at most $(t - u)$. The proportion of such "relevant" arrivals is

$$\begin{aligned}&\Pr\{\text{service time of successful arrival occurring at } u \text{ is} \leqq t - u\\ &\text{and no more customers arrive in the time interval between}\\ &\text{his departure and } t \mid \text{a successful arrival occurred at } u\}\\ &= G(t,u).\end{aligned} \tag{19}$$

As before, we now argue that the density of successful arrivals at u is $dE\{S(u)\}$ and so, using (19) and noting that $y_1 = t_1$ if $A(0) = 0$, we find

$$\Pr\{A(t) = 0\} = \Pr\{A(0) \leqq t, y_1 > t\} + \int_0^t G(t,u)\, dE\{S(u)\}. \tag{20}$$

By combining this result with (10) for $w = 0$, we obtain an integral equation for $E\{S(u)\}$:

$$E\{S(t)\} = \Pr\{y_1 \leqq t\} + \int_0^t Q(t,u)\, dE\{S(u)\}, \tag{21}$$

where the kernel $Q(t,u) = R(t,u) - G(t,u)$. By examining the interpretations (18) and (20) of $R(t,u)$ and $G(t,u)$, it can be seen that

$$Q(t,u) = \Pr\{\text{next successful arrival after } u \text{ occurs before } t \mid \text{a successful arrival occurred at } u\}.$$

Hence, $Q(t,u)$ should be a distribution function in t. Proof of this, together with discussions of (18) and (20), appears in Section XI.

IX. PROOF AND DISCUSSION OF THE RESULTS FOR DELAY

The proof and explanation of (8) and (9) depend on two simple preliminary results. The first of these is as follows: let x be any nonnegative random variable; then, for $y \geqq 0$,

$$\int_0^y \Pr\{x \leqq u\}\, du = E\{\max(0, y - x)\}. \tag{22}$$

This formula states that the area under the (cumulative) distribution of x to the left of y is just the average value of the greater of zero and $y - x$. This is easily seen from an integration by parts:

$$\int_0^y \Pr\{x \leqq u\}\, du = u \Pr\{x \leqq u\} \Big|_0^y - \int_0^y u\, d \Pr\{x \leqq u\}.$$

To begin the proof we note that the total idle time $T(t)$ represented by the term

$$T(t) = \int_0^t U[-W(u)]\, du$$

in the integral equation (1), is always nonnegative, so that

$$W(t) \geqq K(t) - t, \tag{23}$$

and also

$$\Pr\{W(t) \leqq w\} \leqq \Pr\{K(t) - t \leqq w\}.$$

Now in Fig. 9 the larger area represents the event $\{K(t) - t \leqq w\}$ and the smaller one inside it represents the event $\{W(t) \leqq w\}$. This latter event is included in the former because, if $W(t) \leqq w$, then

$$K(t) - t \leqq w,$$

by (23). The difference between the two areas represents the event

$$\{K(t) - t \leqq w < K(t) - t + T(t)\},$$

and so

$$\begin{aligned} \Pr\{W(t) \leqq w\} = {} & \Pr\{K(t) - t \leqq w\} \\ & - \Pr\{K(t) - t \leqq w < K(t) - t + T(t)\}. \end{aligned} \tag{24}$$

We next observe that (8) is the derivative with respect to w of

$$\int_0^w \Pr\{W(t) \leqq u\}\, du = \int_0^w \Pr\{K(t) - t \leqq u\}\, du - \int_0^t R(t,u,w)\, \Pr\{W(u) = 0\}\, du\,, \tag{25}$$

which may be written, using (24) and taking the condition inside, as

$$E\{\max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t]\} = -\int_0^t \Pr\{K(t) - K(u) - t + u \leqq w \quad \text{and} \quad W(u) = 0\}\, du. \tag{26}$$

Thus (8) and (9) are established if we can prove (26).

To establish (26) we need the second preliminary result, a general property of monotone continuous functions.

Lemma: If $F(t)$ is continuous and monotone increasing and $F(0) = 0$, then, for any $x \geqq 0$ and $t \geqq 0$,

$$\max[0,x - F(t)] = x - \int_0^t U[x - F(y)]\, dF(y), \tag{27}$$

where $U(y) = 1$ for $y \geqq 0$, and $U(y) = 0$ for $y < 0$.

Proof: We note that, as y increases, the integrand in (27) is unity until either $x = F(y)$ or $y = t$, whichever occurs first, and it is zero thereafter. If $x = F(y)$ occurs first, then the integral equals x; if $y = t$ occurs first,

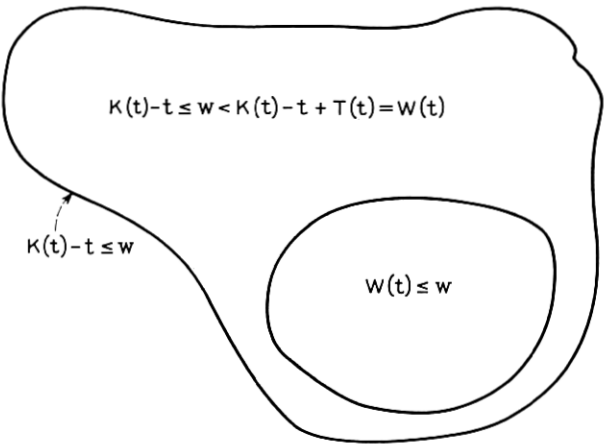

Fig. 9 — Graphical representation of events $\{K(t) - t \leqq w\}$ and $\{W(t) - t \leqq w\}$.

the integral equals $F(t)$. Hence,

$$\int_0^t U[x - F(y)]\, dF(y) = \min[x,F(t)]. \tag{28}$$

For $x \geqq 0$ it can be seen that

$$\min[x,F(t)] = x - \max[0,x - F(t)],$$

and this proves the lemma.

To use this result in proving (26) we interpret $F(t)$ in the lemma as $T(t)$, the total idle time during $(0,t)$, which is a continuous increasing function, with $T(0) = 0$. We next consider the expressions, for $w \geqq -t$,

$$A = w - K(t) + t - \int_0^t U[w - K(t) + t - T(u)]\, dT(u),$$

$$B = w - K(t) + t + T(t) - \int_0^t U[w - K(t) + t + T(t) - T(u)]\, dT(u).$$

If $w - K(t) + t \geqq 0$, then, by the lemma,

$$A - B = \max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t]. \tag{29}$$

However, since $w - K(t) + t \geqq 0$, we see that the integrand in B is always unity, so that B equals $w - K(t) + t$, and

$$A - B = -\int_0^t U[w - K(t) + t - T(u)]\, dT(u).$$

Hence $w - K(t) + t \geqq 0$ implies

$$\max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t] = -\int_0^t U[w - K(t) + t - T(u)]\, dT(u). \tag{30}$$

However, if $w - K(t) + t < 0$, then both sides of (30) vanish and so (30) holds generally, for any $w - K(t) + t$.

From (1) it is evident that, while $W(u)$ is zero, then

$$T(u) = u - K(u);$$

so the integral in (30) is

$$\int_0^t U[w - K(t) + t + K(u) - u] \frac{dT(u)}{du}\, du, \tag{31}$$

where $dT(u)/du$ is one if $W(u) = 0$, and is zero if not; i.e.,

$$\frac{dT(u)}{du} = U[-W(u)]. \tag{32}$$

We recall that, if x and y are random variables, then

$$\Pr\{x \leqq \lambda_1 \quad \text{and} \quad y \leqq \lambda_2\} = E\{U[\lambda_1 - x]U[\lambda_2 - y]\}; \tag{33}$$

i.e., the joint distribution is the average of the product

$$U[\lambda_1 - x] \cdot U[\lambda_2 - y].$$

If we now average (30) and bear in mind (31), (32) and (33), we obtain (26) for all $w \geqq -t$. For negative w,

$$E\{\max[0,w - K(t) + t - T(t)]\} = 0,$$

so that formula (26) takes the form, for $-t \leqq w \leqq 0$,

$$E\{\max[0,w - K(t) + t]\} = \int_0^{t+w} \Pr\{K(t) - K(u) - t + u \leqq w \quad \text{and} \quad W(u) = 0\}\, du. \tag{34}$$

This is (9), and we have completed the proof of the results stated at the beginning of this section.

It can be seen from (8) and (24), and from Fig. 9, that

$$\frac{\partial}{\partial w} \int_0^t \Pr\{K(t) - K(u) - t + u \leqq w \quad \text{and} \quad W(u) = 0\}\, du \tag{35}$$

is the correction term to the overestimate $\Pr\{K(t) - t \leqq w\}$ for $\Pr\{W(t) \leqq w\}$. [See (24).] It is the probability of the event

$$\{K(t) - t \leqq w < K(t) - t + T(t)\},$$

represented by the difference between the areas in Fig. 9. Now, the presence of the derivative $\partial/\partial w$ in (24) is explained by the fact that (24) is the derivative of (25) for $w \geqq 0$, and thus is due to our use of (22). However, it is not obvious intuitively why, in (35), the rest of the term after the $\partial/\partial w$, should be a *time integral.*

An explanation of this can be obtained from (26), which expresses the average of the random variable

$$\alpha = \max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t].$$

It can be seen that

$$\alpha = \begin{cases} -T(t) & \text{if } w > K(t) - t + T(t) = W(t) \\ -w + K(t) - t & \text{if } K(t) - t < w < W(t) \\ 0 & \text{if } w < K(t) - t. \end{cases}$$

A graph of α as a function of w is shown in Fig. 10. The point $K(t) - t$ at which α starts downward may, of course, be negative, although it is positive in the figure.

Thus, α is a negative quantity whose magnitude is no greater than $T(t)$, the total idle time prior to t. Now, if α were in fact equal to $-T(t)$, we could write its average as

$$-E\{T(t)\} = -\int_0^t \Pr\{W(u) = 0\}\, du.$$

But α may be smaller in magnitude than $T(t)$; this fact explains the presence of the conditional probability in the integrand of

$$E\{\alpha\} = -\int_0^t \Pr\{K(t) - K(u) - t + u \leqq w \mid W(u) = 0\} \Pr\{W(u) = 0\}\, du.$$

The kernel in this expression is a probability, so it reduces the magnitude of the integrand whenever it is less than one.

X. DELAY EXAMPLE: POISSON ARRIVALS, GENERAL SERVICE TIMES

For a first example, we assume that customers arrive in a Poisson process of intensity λ, and that service times are mutually independent, with a general distribution function $B(x)$. Such a system has been treated before,[9,10,11] but few explicit formulas are known except for the

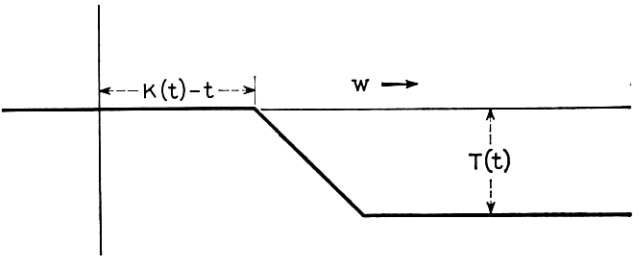

Fig. 10 — Graph of $\alpha(w)$.

exponential case

$$B(x) = \begin{cases} 1 - e^{-\mu x} & x \geqq 0, \\ 0 & x < 0, \end{cases}$$

considered by Ledermann and Reuter[11] and Bailey.[9] The author[10] has sketched a method for calculating $\Pr\{W(t) = 0 \mid W(0)\}$ for general distributions $B(x)$ of service time, but gave no explicit results. He proved that the Laplace transform of $\Pr\{W(t) = 0 \mid W(0)\}$ is given by

$$\int_0^\infty e^{-\tau t} \Pr\{W(t) = 0 \mid W(0)\}\, dt = \frac{e^{-\eta(\tau) W(0)}}{\eta(\tau)}, \qquad \mathrm{Re}(\tau) > 0, \quad (36)$$

where $\eta(\tau)$ is the unique root in the right half-plane of the equation

$$\tau - \eta + \lambda = \lambda B^*(\eta), \qquad \mathrm{Re}(\tau) > 0,$$

with

$$B^*(s) = \int_0^\infty e^{-st}\, dB(t).$$

It was also shown that any function of $\eta(\tau)$, analytic in the right half-plane, could be expanded in a Lagrange series. We shall now derive these results (by a quite different way) directly from the general integral equation (9), and then obtain some specific new results.

Since arrivals are Poisson and service times are independent with distribution $B(x)$, then the load process $K(t)$ is the compound Poisson process, and

$$E\{e^{-s[K(t)-K(u)]}\} = e^{\lambda(t-u)[B^*(s)-1]}.$$

The kernel $R(t,u,0)$ of (9) is

$$\sum_{n=0}^{\infty} \frac{e^{-\lambda(t-u)}\lambda^n (t-u)^n}{n!} B_n(t-u) = \Pr\{K(t) - K(u) \leqq t - u\}, \quad (37)$$

where $B_n(x)$ is the convolution of $B(x)$ with itself n times, i.e., the distribution of the sum of n service times, and $B_0(x)$ is 1 for $x \geqq 0$ and is 0 otherwise. In fact, the Poisson term in (37) is just the chance that n customers arrive in (u,t), and $B_n(t-u)$ is the chance that their combined service time is not more than $(t-u)$. Similarly, we find that

$$\int_0^t \Pr\{K(t) \leqq u\}\, du = E\{\max[0,t - K(t)]\}$$

$$= \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \int_0^t \max[0,t - x - W(0)]\, dB_n(x). \quad (38)$$

Equation (9) for this example is therefore

$$\sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \int_0^t \max(0,t - x - W(0))\, dB_n(x) = \int_0^t \sum_{n=0}^{\infty} \frac{e^{-\lambda(t-u)}\lambda^n (t-u)^n}{n!} B_n(t-u)\Pr\{W(u) = 0 \mid W(0)\}\, du. \tag{39}$$

Since the right-hand side is a convolution, we take Laplace transforms. That of the kernel, (37), can be written as

$$\int_0^{\infty} e^{-\tau t} R(t)\, dt = \frac{1}{2\pi i} \int_0^{\infty} \int_{c-i\infty}^{c+i\infty} e^{-t[\tau - s + \lambda - \lambda B^*]} \frac{ds}{s}\, dt, \qquad c > 0$$
$$= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{\frac{ds}{s}}{\tau - s + \lambda - \lambda B^*}, \tag{40}$$

since the order of integration can be interchanged.

Let S_R be the semicircle that is the right-hand half of the circle $|s - c| = R$. It can be seen that on this semicircle

$$\frac{s^{-1}}{\tau - s + \lambda - \lambda B^*} = O(R^{-2}),$$

so that

$$\lim_{R\to\infty} \frac{1}{2\pi i} \int_{S_R} \frac{s^{-1}\, ds}{\tau - s + \lambda - \lambda B^*} = 0$$

It has been shown[10] that, for $\mathrm{Re}(\tau) > 0$, the function $\tau - s + \lambda - \lambda B^*(s)$ has a unique zero, $\eta(\tau)$, in the right half-plane. Hence,

$$\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{s^{-1}\, ds}{\tau - s + \lambda - \lambda B^*} = \frac{1}{\eta(\tau)} [\text{residue of } (\tau - s + \lambda - \lambda B^*)^{-1} \text{ at } s = \eta(\tau)], \tag{41}$$

when $c < \mathrm{Re}\ \eta(\tau)$.

The Laplace transform of $E\{\max[0,t - K(t)]\}$ can be written in the form

$$\frac{1}{2\pi i} \int_0^{\infty} \int_{c-i\infty}^{c+i\infty} e^{-t(\tau - s + \lambda)} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} (B^*)^n \frac{e^{-sW(0)}}{s^2}\, ds\, dt. \tag{42}$$

Formula (42) simplifies to

$$\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{s^{-2} e^{-sW(0)}\, ds}{\tau - s + \lambda - \lambda B^*},$$

which, by arguments like those already used for (40), can be shown to equal

$$[\eta(\tau)]^{-2}e^{-\eta(\tau)W(0)}[\text{residue of } (\tau - s + \lambda - \lambda B^*)^{-1} \text{ at } s = \eta(\tau)]. \quad (43)$$

It follows from (39), (40) and (43) that the Laplace transform of $\Pr\{W(t) = 0 \mid W(0)\}$ is given by

$$\int_0^\infty e^{-\tau t} \Pr\{W(t) = 0 \mid W(0)\}\, dt = \frac{e^{-\eta(\tau)W(0)}}{\eta(\tau)}, \quad (44)$$

where $\eta(\tau)$ satisfies $\tau - \eta + \lambda = \lambda B^*(\eta)$.

As shown,[10] any function F of $\eta(\tau)$, analytic in $\mathrm{Re}(\tau) > 0$, may be expanded in the Lagrange series

$$F[\eta(\tau)] = F(\tau + \lambda) + \sum_{n=1}^{\infty} \frac{(-\lambda)^n}{n!} \frac{d^{n-1}}{ds^{n-1}} \left[\frac{dF}{ds} (B^*)^n\right]_{s=\tau+\lambda}.$$

We can use this expansion to invert the transform given by (44). In some cases, this inversion gives an explicit expression for

$$\Pr\{W(t) = 0 \mid W(0)\}$$

in terms of the kernel and forcing function of the integral equation (39). Setting $F(x) = x^{-1}e^{-xW(0)}$ in the expansion and inverting the resulting transform, we find

$$\Pr\{W(t) = 0 \mid W(0)\} = e^{-\lambda t}U[t - W(0)] + \sum_{n=1}^{\infty} \frac{e^{-\lambda t}\lambda^n t^{n-1}}{n!} \left\{\int_0^t B_n[u - W(0)]\, du + W(0)B_n[t - W(0)]\right\}. \quad (45)$$

By rearranging terms in (45), comparing with (37) and (38), and recalling that $W(0) = K(0)$ by convention, we can put (45) into the form

$$\Pr\{W(t) = 0 \mid W(0)\} = t^{-1}E\{\max[0,t - K(t)]\} + \frac{W(0)}{t} \Pr\{K(t) \leqq t\}. \quad (46)$$

Note that $E\{\max[0,t - K(t)]\}$ is the left-hand side (the forcing function) of the integral equation (39) and that, when $W(0) = 0$, (46) gives an explicit representation of the solution of (39) *in terms of the forcing function alone.* The intuitive meaning of (46) can be expressed as follows: the chance that the system is empty, conditional on the initial load $W(0)$ is equal to

average of greater of 0 and $1 - t^{-1}K(t)$

$$+ \frac{W(0)}{t} \text{ (chance that } W(0) \text{ plus load arriving in } (0,t) \text{ is at most } t).$$

It is easy to obtain new specific formulas from (46). For example, suppose that service times have the fixed length b. In this case of "constant" service times, for $t > W(0)$

$$E\{\max[0,t - K(t)]\} = \sum_{nb \leqq t-W(0)} \frac{e^{-\lambda t}(\lambda t)^n}{n!} [t - nb - W(0)],$$

$$\Pr\{K(t) \leqq t\} = \sum_{nb \leqq t-W(0)} \frac{e^{-\lambda t}(\lambda t)^n}{n!},$$

and hence

$$\Pr\{W(t) = 0 \mid W(0)\} = 1 - \lambda b - P(T,\lambda t) + \lambda b P(T - 1,\lambda t),$$

where $bT = t - W(0)$ and

$$P(c,a) = \sum_{n \geqq c} \frac{e^{-a}a^n}{n!}$$

is the cumulative term (the "tail") of the Poisson distribution with mean a. This formula for $\Pr\{W(t) = 0 \mid W(0)\}$ for constant service times was used to compute the curve of Fig. 1.

By rewriting (46) in terms of inversion integrals, we obtain another representation of $\Pr\{W(t) = 0 \mid W(0)\}$. This one is more useful because, from it, we can find explicit formulas for new cases (by evaluating the inversion integrals). From (42) we see that

$$E\{\max[0,t - K(t)]\} = \frac{e^{-\lambda t}}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{st - sW(0) + \lambda t B^*(s)} \frac{ds}{s^2},$$

and (45) yields

$$\Pr\{K(t) \leqq t\} = \frac{e^{-\lambda t}}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{st - sW(0) + \lambda t B^*(s)} \frac{ds}{s}.$$

Therefore, from (46),

$$\Pr\{W(t) = 0 \mid W(0)\} = \frac{e^{-\lambda t}}{2\pi i t} \int_{c-i\infty}^{c+i\infty} e^{st - sW(0) + \lambda t B^*(s)} \frac{sW(0) + 1}{s^2} ds. \quad (47)$$

As we shall see, this result is useful because it is often easier to evaluate the complex integral than to sum the series (37) or (38).

For example, if service times have the negative exponential distribution with mean u, then it has essentially been shown[9,11] that

$$\frac{d}{dt}\Pr\{W(t) = 0 \mid W(0) = 0\} = -\left(\frac{\lambda}{\mu}\right)^{\frac{1}{2}} e^{-(\mu+\lambda)t}\frac{I_1[2(\mu\lambda)^{\frac{1}{2}}t]}{t}.\dagger$$

Hence

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = 1 - \rho\int_0^{\mu t} \rho^{-\frac{1}{2}} e^{-(1+\rho)x} I_1(2\rho^{\frac{1}{2}}x)\frac{dx}{x}, \quad (48)$$

where $\rho = \lambda/\mu$ = traffic intensity, and μt is the time measured in mean holding times. This result can be obtained directly from (44) by solving a quadratic equation, and it can be put into another form by using (47). Thus,

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = \frac{e^{-(\lambda+\mu)t}}{2\pi i t}\int_{c+\mu-i\infty}^{c+\mu-i\infty} e^{ut+\lambda\mu t/u}(u-\mu)^2\, du.$$

Writing the integrand as $e^{ut}(u-\mu)^2(e^{\lambda\mu t/u} - 1 + 1)$ and using Ref. 12, p. 244, no. 31, we find

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = e^{-\lambda t}\left[1 + \int_0^{t/(4\lambda\mu)} e^{-y^2/(4\lambda\mu t)}\left(1 - \frac{y^2}{4\lambda\mu t^2}\right) I_1(y)\, dy\right].$$

For another example, suppose that service times have the "gamma" probability density

$$\frac{t^{-\frac{1}{2}}\mu^{\frac{1}{2}}e^{-\mu t}}{\Gamma(\frac{1}{2})}$$

whose Laplace transform is $[\mu/(\mu+s)]^{\frac{1}{2}}$. Then (47), with a change of variable, gives

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = \frac{e^{-(\lambda+\mu)t}}{2\pi i t}\int_{c-i\infty}^{c+i\infty} e^{ut+\lambda t(\mu/u)^{\frac{1}{2}}}(u-\mu)^2\, du,$$

and

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = t^{-1}e^{-\lambda t}\int_0^t e^{-\mu u}G(u)(1+\mu u)\, du\},$$

---

† This formula is a simplification of that of Ref. 9 obtained by using standard Bessel function relations; $I_1(x)$ is the Bessel function of imaginary argument, of order one.

where

$$G(u) = u^{-5/2}\alpha^{-3/4}\int_0^\infty x^{3/2}e^{-x^2/\alpha u^2}I_1(x)\,dx$$

and $\alpha = 4\lambda\mu^{\frac{1}{2}}$.

XI. PRECISE DERIVATION OF LOSS RESULTS

The proof of (10) given in Section VIII is rigorous up to and including (17), so it suffices to give a precise construction of the kernels $R(t,u)$, $G(t,u)$ and $Q(t,u)$. Let x_n be the nth epoch at which $A(t)$ becomes equal to 0, and y_n be the arrival time of the nth successful unit. It is readily seen that

$$\begin{aligned} E\{\theta(t,t+w)\} &= \sum_{n=1}^{\infty} \Pr\{y_n \leqq t \quad \text{and} \quad x_{n+1} \leqq t+w\} \\ &= \sum_{n=1}^{\infty} \int_0^t \Pr\{x_{n+1} \leqq t+w \,|\, y_n = u\}\, d\Pr\{y_n \leqq u\}. \end{aligned}$$

Similarly,

$$E\{S(t)\} = \sum_{n=1}^{\infty} \Pr\{y_n \leqq t\}. \tag{49}$$

Thus, each $\Pr\{y_n \leqq \cdot\}$ measure, $n = 1, 2, \cdots$, is absolutely continuous with respect to (49). We can therefore express $E\{\theta(t,t+w)\}$ as the integral of a kernel $R(t+w,u)$ against $E\{S(\cdot)\}$ measure, as in

$$E\{\theta(t,t+w)\} = \int_0^t R(t+w,u)\, dE\{S(u)\},$$

where the kernel is defined in terms of the indicated Radon-Nikodym derivatives by

$$R(y,u) = \sum_{n=1}^{\infty} \Pr\{x_{n+1} \leqq y \,|\, y_n = u\} \frac{d\Pr\{y_n \leqq u\}}{dE\{S(u)\}}.$$

In a like manner, we can write

$$\Pr\{A(t) = 0 \text{ and some successful arrival occurred in } (0,t]\} = \sum_{n=1}^{\infty} \Pr\{x_{n+1} \leqq t < y_{n+1}\}. \tag{50}$$

Introducing the kernel

$$G(y,u) = \sum_{n=1}^{\infty} \Pr\{x_{n+1} \leqq y < y_{n+1} \,|\, y_n = u\} \frac{d\Pr\{y_n \leqq u\}}{dE\{S(u)\}},$$

we can render (50) as

$$\int_0^t G(t,u)\, dE\{S(u)\}\,.$$

The kernel $Q(t,u)$, finally, can be defined as $R(t,u) - G(t,u)$ or as

$$\sum_{n=1}^{\infty} \Pr\{y_{n+1} \leqq t \mid y_n = u\} \frac{d \Pr\{y_n \leqq u\}}{dE\{S(u)\}}\,.$$

The sense in which $Q(\cdot,u)$ is a "distribution function" is given by the following result: for almost all u with respect to $E\{S(\cdot)\}$ measure, $Q(\cdot,u)$ is a distribution function. We show first that the derivatives

$$\varphi_n(u) = \frac{d \Pr\{y_n \leqq y\}}{dE\{S(u)\}}$$

have the properties

$$\sum_{n=1}^{\infty} \varphi_n(u) = 1, \tag{51}$$

$$0 \leqq \varphi_n(u) \leqq 1, \tag{52}$$

almost everywhere in $E\{S(\cdot)\}$. Now (51) is true by definition. Suppose (52) failed on a set B of positive measure; then either $E\{S(\cdot)\}$ is not a positive measure, or else

$$\Pr\{y_n \epsilon B\} > \int_B dE\{S(u)\}\,,$$

both of which are impossible. It is readily seen, by an elementary decomposition of the events, that, for each $n \geqq 1$,

$$\Pr\{x_{n+1} \leqq t \mid y_n = u\} - \Pr\{x_{n+1} \leqq t < y_{n+1} \mid y_n = u\} = \Pr\{y_{n+1} \leqq t \mid y_n = u\}.$$

Except on a set C_n of $\Pr\{y_n \leqq \cdot\}$ measure zero, this is a distribution function in t, and the result follows from

$$\int_C dE\{S(u)\} = 0\,, \qquad \text{if} \quad C = \bigcup_{n=1}^{\infty} Cn.$$

XII. A LIMIT THEOREM

We shall prove that dependence of the kernels $R(t,u)$ and $Q(t,u)$ on $(t - u)$ only and existence of the "mean service time of successful arrivals" are sufficient to guarantee that

$$\Pr\{A(t) \leqq w\}$$

approaches a limit as $t \to \infty$. It is natural to study cases in which only difference kernels occur, because of the Volterra equation derived in Section IX. The result to be proved requires no restriction on the "arrival rate" of units — an upper bound is unnecessary because of the loss operation; i.e., if the arrival rate increases, the rate at which successful units leave the system can only increase to a limit.

Theorem: If the kernels $R(t,u)$ and $Q(t,u)$ only depend on $(t - u)$, if the average

$$b = \int_0^\infty [1 - R(x)]\, dx = \text{(average service time of a successful unit)}$$

exists, and if $Q(\cdot)$ is not a lattice distribution, then

$$\lim_{t\to\infty} \Pr\{A(t) \leqq w\} = \frac{a + \int_0^w [1 - R(u)]\, du}{a + b}, \tag{53}$$

where

$$(a + b) = \int_0^\infty [1 - Q(x)]\, dx,$$

the limit being one if $(a + b) = \infty$.

Proof: By the remarks at the end of Section XI, both $R(x)$ and $Q(x)$ may be taken to be the distribution functions of positive variates. Equation (11) becomes a renewal equation, and $E\{S(u)\}$ is essentially the renewal function $H(\cdot)$ of Smith.[13] The integrand of (10) is

$$1 - R(t + w - u),$$

and is nonincreasing and integrable. Also, $Q(\cdot)$ is not of lattice type. Hence, by Smith's Theorem 1:

$$\lim_{t\to\infty} \int_0^t [1 - R(t + w - u)]\, dE\{S(u)\} = \frac{\int_w^\infty [1 - R(u)]\, du}{\int_0^\infty x\, dQ(x)}, \tag{54}$$

which gives (53) upon rearrangement if (54) is taken to be zero if the mean of $Q(x)$ does not exist.

XIII. ACKNOWLEDGMENTS

The author is indebted to E. B. Ferrell for suggesting many improvements in exposition, and to J. Riordan and W. O. Turner for reading the draft.

REFERENCES

1. Reich, E., On the Integrodifferential Equations of Takács, I., Ann. Math. Stat., **29,** 1958, p. 563.
2. Jensen, A., An Elucidation of Erlang's Statistical Works Through the Theory of Stochastic Processes, in Brockmeyer, E., et al., *The Life and Works of A. K. Erlang*, Copenhagen Telephone Co., Copenhagen, 1948, p. 23.
3. Smith, W. L., Regenerative Stochastic Processes, Proc. Royal Soc. (London) A, **232,** 1955, p. 6.
4. Takács, L., On the Generalization of Erlang's Formula, Acta Math. Acad. Sci. Hung., **7,** 1956, p. 419.
5. Takács, L., On a Coincidence Problem Concerning Telephone Traffic, Acta Math. Acad. Sci. Hung., **9,** 1958, p. 45.
6. Takács, L., A Telefon-forgalom Elméletének Néhány Valószinüségszámitási Kérdéséröl, A. Magyar Tud. Akad. (Math. and Phys.), **8,** 1958, p. 151.
7. Beneš, V. E., Combinatory Methods and Stochastic Kolmogorov Equations in the Theory of Queues with One Server, to be published.
8. Beneš, V. E., General Stochastic Processes in the Theories of Counters and Telephone Traffic, to be published.
9. Bailey, N. T. J., A Continuous Time Treatment of a Simple Queue Using Generating Functions, J. Royal Stat. Soc. B, **15,** 1954, p. 288.
10. Beneš, V. E., On Queues with Poisson Arrivals, Ann. Math. Stat., **28,** 1957, p. 670.
11. Ledermann, W. and Reuter, G. E. H., Spectral Theory for the Differential Equations of Simple Birth and Death Processes, Phil. Trans. Royal Soc. (London) A, **246,** 1954, p. 321.
12. Erdelyi, A., et al., *Tables of Integral Transforms*, McGraw-Hill Book Co., New York, 1954.
13. Smith, W. L., Asymptotic Renewal Theorems, Proc. Royal Soc. (Edinburgh) A, **64,** 1954, p. 9.