排隊論論文选集

（I）

中国科学院数学研究所运筹室

1959年7月　北京

目　次

Journal of the Royal Statistical Society
Series B (Methodological)
Vol. XIII, No. 2, 1951

Some Problems in the Theory of Queues

By David G. Kendall
Magdalen College, Oxford

{Received January, 1951}
{Read before the Research Section of the Royal Statistical Society, March 21st, 1951, Professor M. G. Kendall in the Chair}

Summary

The paper opens with a general review of some points in congestion theory, and continues with a simplified account of the Pollaczek-Khintchine "equilibrium" theory for the single-counter queue fed by an input of the Poisson type and associated with a general service-time distribution. It is pointed out that although the stochastic process describing the fluctuations in queue-size is not (in general) Markovian, it is possible to work instead with an enumerable Markov chain if attention is directed to the epochs at which individual customers depart (these epochs forming a sequence of regeneration points).

The ergodic properties of the chain are investigated with the aid of Feller's theory of recurrent events; it is found to be irreducible and aperiodic, and the further classification of the states depends on the value of the traffic intensity, ρ (measured in erlangs). When $\rho < 1$ the states are all ergodic and the system ultimately settles down to a regime of statistical equilibrium, the transition-probabilities converging to the values given by the equilibrium solution. When $\rho > 1$ the states are all transient, and when ρ has the critical value of unity the states are all recurrent-and-null. The paper closes with some further general comments, with a re-interpretation of Borel's theory of "busy periods" in terms of the Galton-Watson "branching process" of stochastic population theory and with a sketch of an argument leading to the distribution of the length in time of a busy period.

1. Introduction. - The best way of opening this discussion on the Theory of Queues would have been to present a short but balanced survey of the literature. I had initially this aim in mind, but soon found myself unequal to the task. It is not so

much the extent as the inaccessibility of the literature which makes the difficulty; many of the most important articles have appeared in technological journals, and a University Library is not the best place in which to look for them. Instead, therefore, I shall discuss one or two simple problems which are characteristic of the subject, and it is my hope that the deficiencies of the provisional list of references will be made good by those taking part in the subsequent discussion. If Fellows will let me have the details of those papers which have been omitted, it will then be possible to compile a more adequate bibliography for publication in the Journal.

The Theory of Queues has a special appeal for the mathematician interested in stochastic processes because it provides a simple example which is both (i) stationary and (ii) (in general) not Markovian. Apart from Yule's autoregressive schemes there are not many other examples which possess properties (i) and (ii) and are at the same time so easy to analyse. Thus one may expect in studying the queue to gain insight into other stochastic phenomena, and to acquire a valuable facility in the handling of the relevant techniques.

A further attractive feature of the theory is the quite astonishing range of its applications. I do not propose to mention many of them now because I am sure that their variety will become apparent from the subsequent discussion, but it would be unforgiveable not to mention that the first major study of congestion problems seems to have been that undertaken by A. K. Erlang*in 1908 at the suggestion of F. Johannsen (himself a pioneer contributor to the subject), and under the auspices of the Copenhagen Telephone Company (whose association with investigations in the mathematical theory of probability continues to-day, to our great benefit). My own interest in the subject arose from a correspondence with F. J. Anscombe and J. Howlett about queues of taxis in station yards and of customers inretail shops, and so for the most part I shall use the terminology appropriate to these homely examples.

Let me begin with a very simple problem, one which always turns up in the discussion if it is not dealt with at the outset. Suppose that taxis and (single) travellers arrive at a queueing-point in independent Poisson streams, the expected numbers of arrivals per-unit of time being respectively α and β. Starting fromzero the size of the queue at time t will be $q = m - n$ where m and n are respectively Poisson variables with mean values αt and βt. A positive q means that there is a rank of taxis, and a negative q means that there is a queue of travellers; when q is zero no one is there at all and supply and demand have reached a momentary balance. If $\alpha = \beta$ then the expected value of q is permanently equal to zero and one might expect the system to remain in a balanced state, but the variance of q is $2\alpha t$, and this increases indefinitely with the time.

The difference between two independent Poisson variables has been discussed by Irwin (1937) and by Skellam, (1946) and their results in the present notation show that the chance of there being a positive or negative or zero "queue" not greater in absolute value than Q is

$$e^{-2\alpha t} \sum_{q=-Q}^{Q} I_{|q|}(2\alpha t) \simeq \frac{2Q+1}{\sqrt{(4\pi\alpha t)}}, \qquad (1)$$

which tends to zero when t tends to infinity for every fixed Q, however large. This simple example displays a behaviour characteristic of many stochastic phenomena; when supply and demand exactly balance, the expectation-values of the variables are liable to tell a misleading story.

I turn now to an abstract description of a simple queueing system of much wider importance. Here one is concerned, not with the matching of supply and demand, but with a series of customers demandind service at a single counter and waiting in turn to be served. ("Customer" is used here in a technical sense; in practive it might, for example, have to be e uated with "aircraft".) To complete the specification one must give a careful account of: -

(a) the input-process,
(b) the queue-discipline,
(c) the service-mechanism.

The simplest hypothesis about the input is one which states that the customers arrive "at random" (i.e., in a Poisson process), the number of arrivals in time t being a Poisson variable of expectation t/a; say. The time-interval u between two consecutive arrivals will then have the negative-exponential distribution

$$e^{-u/a}\,du/a \qquad (0<u<\infty), \qquad (2)$$

P.2 *Erlang's papers (translated into English) are included in a memoir by Brockmeyer, Halstrom and Jensen (1948). This valuable work also contains much interesting historical material, and a detailed study (by Jensen) of Erlang's work from the standpoint of the modern theory of stochastic processes. An account of Erlang's work is also to be found in the excellent survey of congestion problems given by Fry (1928).

---

and successive u-variables will be statistically independent. This hypothesis will be adopted throughout the present paper save for some remarks about more general types of input to be made in §5.

The queue-discipline is the rule or moral code determining the manner in which the customers form up into a queue and the manner in which they behave while waiting. In the simplest case they line up before the (single) counter and await their respective turns (this is the hypothesis which will be adopted here). It should be noted, however, that one may wish to introduce the possibility of "extraordinary departures" (Palm, 1937; Jensen, 19[illegible], pp. 74-5) of customers who have become tired of waiting (or of aircraft which have been directed to try to land on another airfield). I shall have very little to say about the much more difficult problem which arises when there are several (say N) counters, and it is only in these circumstances that much variety in the specification of the queue-discipline becomes possible. The customers may then form a single queue, the man at the head moving as soon as possible to a vacant counter. Alternatively (though there are yet other possibilities) each customer on arrival may be handed a ticket carrying his serial number m and directed to join the queue formed up in front of the jth counter, where

$$m \equiv j \quad (\mathrm{mod}\ N) \quad . \quad . \quad . \quad . \quad (3)$$

This means that the total input is divided among the several subqueues in a relatively simple manner, and it is then possible (if the total input is Poissonian) to discuss the behaviour of one of the sub-queue in isolation from the rest of the system, its own input-process being of a more general type (see §5). The rule (3) is rather artificial, but its use introduces a considerable simplification into the mathematical theory. In this paper, however, except where the contrary is explicitly stated, I shall be concerned only with queues formed in front of a single counter ($N = 1$).

Lastly, the service-mechanism: let v denote the time which elapses while a particular customer is being served; I shall call this the service-time, although in telephone-traffic theory it is more usual to call it the holding-time. It is natural to suppose that the successive service-times are statistically independent of one another and of the input, and that each enjoys the same probability distribution

$$dB(v) \quad (0 < v < \infty), \quad . \quad . \quad ; \quad (4)$$

say. There are two special cases of importance:

(i) (negative-exponential service-time)

$$dB(v) \equiv e^{-v/b}\, dv/b \quad . \quad . \quad . \quad . \quad (5)$$

(ii) (constant service-time)

$$B(v) \equiv 0 \ (v < b); \quad B(v) \equiv 1 \ (v > b) \quad (6)$$

In either case b is the mean value of the service-time; I shall reserve the symbol b for this use alone, and I shall always assume that b is finite and positive. Both of (i) and (ii) were considered by Erlang, and he alsoemployed an intermediate hypothesis in which the service-time distribution has the χ^2-form

$$\left(\frac{k}{b}\right)^{k} \frac{1}{\Gamma(k)} e^{-kv/b} v^{k-1} dv \qquad (0 < v < \infty) \cdot\cdot (7)$$

This has an interesting interpretation (of which Erlang was well aware) in terms of k successive "phases" or "states" of service ; the same idea can be applied to a problem of population growth (Kendall, 1948). It will be noticed that (5) and (6) are the limiting forms of (7) when $k = 1$ and when k tends to infinity respectively

A parameter of special importance is the traffic-intensity, $\rho = b/a$; in a long period of time, T , there will be approximately T/a calls for service, each of mean length b , and so ρ is the expected service demanded per unit of time, in "erlangs". (For the definition of this dimensionless unit see Jensen(1950). Properly it is not a unit at all, but an indication of how the preceding figure has been calculated; in this respect it resembles the octave, the decibel and the stellar magnitude.)

When the service-time distribution has the form (5), (6) or (7) the "equilibrium" theory is due to Erlang, and his results are to be found in the memoir by Brockmeyer, Halstrom and Jensen, and in the book by Fry(1928). The solution to the problem of the simple queue in statistical equilibrium and with a general service-time distribution was given by Pollaczek (1930a) and Khintchine (1932) Pollaczek's paper involves very heavy analysis and makes difficult reading, and although Khintchine achieved a considerable simplification of the argument his paper is in the Russian language and an English translation appears not to be availbale; an account of their important results may therefore be found useful, and one will be given in §2. Before proceeding to this, however, I want to make one or two comments on an interesting qualitative feature of the problem.

Suppose that the state of a stochastic system at time t is described by a random function $x(t)$; the stochastic process is then said to be of the Markov type if a knowledge of the present value of $x(t)$ makes all information about its past history irrelevant to a prediction of its future behaviour. Such a process may become non-Markovian if part of the information contained in $x(t)$ is suppressed, for a knowledge of the previous behaviour may enable some of the suppressed information about the present state to be recovered. (The importance of this fact has been pointed out by Bartlett (1950); see also Kendall (1950).) The congestion process (with a Poissonian input and a single counter) is Markovian if the present state of the process is described by the pair of random variables q and v , where q is the instantaneous queue-size and v is the expended service-time of the customer at the head of the queue, but in general it ceases to be Markovian if the state of the process is measured by the queue-size alone.* The only exception to this statement occurs when the service-time has a negative-exponential distribution; a special and quite well-known characteristic property of this distribution then ensures that a knowledge of the expended service-time

* Cf. Feller (1949a) (§ 8 and 9).

t_0 is of no predictive value.

Even with a general distribution of service-time, however, the stochastic process describing the flucturations in queue-size is of the type which Bartlett and I (1951) have called regenerative. "Regenerative process" if our (free) translation of Palm's expression, "Prozesse mit begrenzter Nachwirkung", and a stochastic process will be said to be of this type if and if only if it possesses "regeneration points" (Palm : "Gleichgewichtpunkte"). An epoch is a point of regeneration if a knowledge of the state of the process at that particular epoch has the characteristic Markovian consequence that a statement of the past history of the process loses all its predictive value. A Markov process, therefore, is precisely a process for which every epoch is a point of regeneration.

In analysing a regenerative process the identification of its regeneration points is of fundamental importance, and a classification of existing techniques fromthis pointof view is of considerable interest (see, for example, Bartlett and Kendall, 1951; see also Ramakrishnan, 1951). These ideas have, of course, been current in an implicit form for many years, but the first clear formulation appears to have been that given in an important monograph by Palm (1943). More recently Feller (1949b, 1950) has developed a closely related theory of processes admitting "recurrent events".

For the congestion process, when the input is Poissonian and when the state of the system is measured by the instantaneous queue-size, the regeneration points are the epochs at which customers leave, together with the epochs at which the counter happens to be free. An appreciation of this fact will throw a good deal of light on the subsequent arguments.

2. An elementary derivation of the formula for the mean waiting-time in statistical equilibrium. - Suppose that a queue formed in front of a single counter is fed by a random (Poissonian) input, and suppose that the traffic-intensity $\rho\,(=b/a)$ is less than unity, so that the input does not saturate the system. The epochs at which the customers leave are points of regeneration. Consider such an epoch, and let q be the size of queue which the departing customer leaves behind him (this number does not include himself, but it does include the person next to be served in his place; of course q could be zero). Let the next person to be served have a service-time v, and during this time suppose that r new customers arrive. Then conditionally r is a Poisson variable of mean value v/a, and v has the service-time distribution. When the next person leaves let q' be the size of the queue which he leaves behind him. Then in statistical equilibrium (if such an equilibrium solution can exist) the random variables q and q' must have the same marginal distribution, and in particular their mean and mean square values must be equal.

The variables q' and q are related by t[illegible]

$$q' = \max(q-1, 0) + r,$$

which it is more convenient to write as

$$q' = q - 1 + \delta + r, \quad . \quad . \quad . \quad . \quad . \qquad (9)$$

where $\delta = \delta(q)$ is zero for all non-zero q, and $\delta(0) = 1$. It is important to note the following consequences of the definition of the function $\delta(q)$:

$$\delta^2 = \delta, \quad \text{and} \quad q(1-\delta) = q \quad . \quad . \qquad (10)$$

It will now be assumed (i) that an equilibrium solution exists, and (ii) that the equilibrium values of $E(q)$ and $E\{q^2\}$ are finite. If these facts are accepted as intuitively obvious then the following derivation of the mean waiting-time is valid; otherwise a more detailed investigation is necessary, and this will be given later.

On forming the expectations of both sides of (9) it will be found that

$$E[\delta] = 1 - E[r] = 1 - b/a = 1 - \rho; \quad . \quad . \quad . \qquad (11)$$

this is the chance that δ is non-zero; i.e., it is the chance that a departing customer leaves an empty counter behind him. On squaring both sides of (9) and making use of (10) one obtains

$$q'^2 = q^2 - 2q(1-r) + (r-1)^2 + \delta(2r-1),$$

and now on forming expectations (and noting that r is independent of q and δ) it follows that

$$E[q] = E[r] - \frac{E[r(r-1)]}{2\{1 - E[r]\}},$$

$$= \frac{b}{a} + \frac{Var(v) + b^2}{2a(a-b)} \quad . \quad . \quad . \quad . \quad . \qquad (12)$$

It is tempting to think that the first expression for $E[q]$ may be valid for more general input-processes, but this seems to be a fallacy; the argument breaks down because r and q will not then be statistically independent (some further remarks on more general input-processes will be found in ≢5).

From (12) it is quite easy to pass to the expression for the mean waiting-time. Suppose that a departing customer leaves q customers behind him and let his own waiting- and service-times be respectively w and v. Then q is the number of arrivals in a total time $w + v$, and so

$$E(q) = \{E[w] + E[v]\}/a \quad . \quad . \quad . \qquad (13)$$

From this it will be found that

$$\frac{E\{w\}}{E\{v\}} = \frac{\rho}{2(1-\rho)}\{1 - \mathrm{Var}(u/b)\} \quad . \quad . \quad . \qquad (14)$$

and this is equivalent to tne Pollaczek-Khintchine formula.

It is convenient to express the result in this way, because the ratio on the left-hand side is a useful "figure of demerit" for the system (it is the ratio of the mean time spent waiting to the mean service-time waited for). If the mean service-time is kept constant and if the frequency of calls for service remains the same then the values of b and a will be fixed, and in these circumstances it follows from (14) that maximum efficiency will be obtained if and only if there is no variation in the service-time. With a negative-exponential distribution of service-times the ratio (14) is equal to twice the minimum possible value, and so in this sense a system characterized by a negative-exponential service-time distribution is working at only 50 per cent. efficiency.

For a fixed form of service-time distribution, on the other hand, the "congestion-ratio" (14) is a constant multiple of $\rho/(1-\rho)$, and the situation can then only be improved by a reduction of the traffic intensity ρ (i.e., by a reduction in the mean service-time or by a reduction in the frequency of calls for service). This implies a corresponding increase in the fraction of time ($1-\rho$) during which the counter is unused*, so that while a reduction in the value of ρ increases the efficiency in one respect, it automatically reduces it in another.

In practive these simple principles will not always apply, or at least their working may be obscured by a number of second-order effects. For example, in some circumstances a reduction in the mean service-time may result in a proportionate increase in the number of calls for service, so that ρ would tend to stay constant. When this situation obtains the only way of improving the system is to modify the form of the service-time distribution (in the direction of complete concentration at the mean value), but the simple theory would, of course, cease to apply if the disgruntled customer marched immediately back to the end of the queue.

Other second-order effects which provide an automatic control reducing congestion below the theoretical value are:

(a) the premature departure of customers who become tired of waiting;

(b) the disconraging effect of a long queue on an incoming customer;

(c) the tendency to serve more rapidly a customer who has a long queue waiting behind him.

* In a long period of time, T , the number of arrivals will be approximately T/a . A fraction ($1-\rho$) of these will on departing leave an empty counter behind them and the ensuring "slack periods" will be of mean length a . This makes the total "Slack" time approximately equal to $T(1-\rho)$.

Similar remarks apply when the other applications are being considered, although there are often instructive differences of detail.

If one continues to assume that the system can maintain itself in statistical equilibrium, then it is possible to develop the preceding argument in a fairly obvious way so as to obtain the stationary distributions for the waiting-time w of an arbitrary customer and the queue q which he leaves behind him. The generating function for the q-distribution is

$$H(z) = E[z^q] = \frac{(1-\rho)(1-z)}{1-z/K(z)}, \quad . \quad . \quad . \quad . \quad (15)$$

and the Laplace transform of the w-distribution is accordingly given by

$$\gamma(s) = E[e^{-sw}] = \int_0^\infty e^{-sw}\, dC(w)$$

$$= (1-\rho)\left\{1-\rho\left[\frac{1-\beta(s)}{bs}\right]\right\}^{-1}, \quad . \quad . \quad . \quad (16)$$

where

$$\beta(s) = E[e^{-sw}] = \int_0^\infty e^{-sw}\, dB(v) \quad . \quad . \quad . \quad (17)$$

is the Laplace transform of the service-time distribution $dB(v)$, and

$$K(z) = \beta\left(\frac{1-z}{a}\right) \quad . \quad . \quad . \quad . \quad . \quad (18)$$

From these formulae one can obtain (as Khintchine did) expressions for the higher moments (in particular, the variances) of the respective random variables. In any given case the function $\beta(s)$ will be specified, and then the determination of the distributions themselves raises only inversion problems of a standard type.

It is worth noting that when s tends to infinity through real positive values the expression on the right-hand side of (16) tends to the non-zero limit $1-\rho$. This indicates the existence of a probability-concentration of intensity $1-\rho$ at the point $w=0$; i.e., $1-\rho$ is the probability that an incoming customer will find the counter free and will accordingly have a zero waiting-time. It is interesting that the probability of a zero waiting-time should be independent of the form of the service-time distribution; the form of the latter distribution will, however, affect the way in which the remaining amount of probability (ρ) is distributed over the open interval $0 < w < \infty$. When the service-time distribution is negative-exponential in form the continuous component of the waiting-time distribution is

$$e^{-\rho w/c}\rho\, dw/c \qquad (0 < w < \infty),$$

where

$$c = b^2/(a-b).$$

This is one of the formulae of Erlang, and this and his correspondig solution when the service-times are constant will be found in the memoir by Brockmeyer et al. (1948) and in the book by Fry (1928).

I now leave the equilibrium theory and turn to an investigation of the ergodic properties of the congestion process. The difficulties associated with the non-Markovian character of the process will be evaded by concentrating attention on a Markov chain which describes the behaviour of the process at an enumerable set of regeneration points.

3. The approach to equilibrium. - Intuition suggests that when the traffic intensity ρ is less than unity the system should possess the characteristic "ergodic" property of settling down after the lapse of a sufficiently long period of time (or after the passage of a sufficiently large number of customers) into an equilibrium mode of behaviour independent of its initial state, and that when $\rho > 1$ no such stability of behaviour is to be expected. I shall now establish the truth of these conjectures (when suitably fomulated), and as a secondary consequance of the investigation I shall be able to describe the curious behaviour of the system when $\rho = 1$.

The non-stationary behaviour of the congestion process has been studied by Volberg (1939a, b). In the one paper of his which I have seen he considered the general problem of a many-counter queue governed by the rule of queue-discipline associated with (3), and his analysis is based on an integral equation technique. It seems possible that his second paper may have rather more in common with the methods used here, but I have had no opportunity of verifying this. The ergodic properties of the Markovian queueing systems studies by Erlang have been investigated by Jensen (1948), and another (related) congestion problem has been considered from the present point of view of M. H. Johnson (1950) in a privately circulated note. I am very much indebted to Mr. Johnson for the opportunity of seeing this account of his work, and I hope that we shall be privileged to hear him describe it to us later to-day.

I shall start from the fact that in a single-counter system with a Poissonian input the epochs of departure are points of regeneration; this makes it possible to reduce the problem to one concerning a Markov chain in "discrete time", and this is despite the fact that the congestion process itself is not Markovian. As before, let q and q' be the queues left behind by two consecutively departing customers, and let p_{ij} be the probability that $q' = j$

when it is given that $q = i$. Then, if one agrees to consider the history of the system as it would be observed at these epochs of departure, one is concerned with the ultimate behvaiour of a Markov chain having an enumerable infinity of states, and for which the matrix of one-step transition-probabilities is $\{p_{ij}\}$. This is a stochastic matrix; that is to say, it has non-negative elements and unit row/sums.* The analysis of such chains was begun by Kolmogorov (1936, 1937), and important simplifications and continuations of his work were later published by Doeblin (1939), Yosida and Kakutani (1940), and Doob (1942, 1945). A complete and highly original presentation of the subject has recently been given by Feller (1950) with the aid of his general theory of recurrent events, and it is precisely this work of Feller which makes it possible to investigate the ergodic properties of the congestion process in a relatively elementary manner.

For the purposes of the present paper some of the general properties of Markov chains will be required and these I shall now summarise.

(i) Let p^n_{ij} be the probability of an n-step transition from the ith to the jth state. As n tends to infinity, the p^n_{ij} need not converge to limiting values in the ordinary sense, but they always do so in the generalized ($C, 1$) sense (to limits π_{ij}, say).

(ii) The row-sums of the matrix $\{\pi_{ij}$ satisfy the inequality

$$\sum_{j=0}^{\infty} \pi_{ij} \leq 1 \qquad (i = 0, 1, 2, \ldots), \tag{19}$$

and the sign of equality need not hold. (If the sign of equality holds for every i, I shall follow Foster (1951) and say that the system is non-dissipative.)

(iii) The rows of the π-matrix need not be identical, and so the ultimate behaviour of the system may depend on its initial state. In any case, however, one will have

$$\sum_{a=0}^{\infty} \pi_{ia} p_{aj} = \sum_{a=0}^{\infty} p_{ia} \pi_{aj} = \sum_{a=0}^{\infty} \pi_{ia} \pi_{aj} = \pi_{ij} \quad \ldots \tag{20}$$

for every i and j. It will be noticed that the π-matrix is always idempotent.

Elegant and surprisingly simple proofs of the results (i), (ii) and (iii) will be found in the paper of Yosida and Kakutani which has just been cited. In the present problem all the extremes of pathological behaviour are not realized, and to establish this fact two further results will be needed.

[*] Note that the second suffix (denoting the column) identifies the new state, and the first suffix (denoting the row) identifies the old state. The reverse convention is sometimes adopted.

(i)' The transition-probabilities $p_{ij}^{(n)}$ converge to the limits π_{ij} in the ordinary sense if it happens that all the diagonal elements of the p-matrix are positive.

(ii)' The system will be non-dissipative (so that the π-matrix will have unit row-sums) if

$$\sum_{j=0}^{\infty} jp_{ij} \leq i \quad . \quad . \quad . \quad . \quad . \quad . \qquad (21)$$

for every i .

The first of these results is relatively a "deep" one; it follows eqsily from the theory given in Feller's book on noting that (because of the stated condition) every state in the Markov chain is necessarily "aperiodic"*. (I owe this helpful remark to Mr. F. G. Foster.) The useful criterion (ii)' for non-dissipation is due to Foster (1951); in words, his condition requires that the expected increase in state-label is never to be positive. In the same paper he has given one or two other conditions sufficient for non-dissipation, and I have generalized (ii)' to give a criteron invariant under a change of state-lebelling (Kendall, 1951). It should be observed that neither of (i)' and (ii)' makes any statement about necessary conditions.

The matrix $\{p_{ij}\}$ defining the Markov chain associated with the congestion process has a particularly simple form #; it is

$$\begin{array}{c|ccccc}
 & 0 & 1 & 2 & 3 & 4 \\
\hline
0 & k_0 & k_1 & k_2 & k_3 & k_4 \\
1 & k_0 & k_1 & k_2 & k_3 & k_4 \\
2 & \cdot & k_0 & k_1 & k_2 & k_3 \\
3 & \cdot & \cdot & k_0 & k_1 & k_2 \\
4 & \cdot & \cdot & \cdot & k_0 & k_1
\end{array}$$

where

$$k_r = \frac{1}{r!}\int_0^{\infty} e^{-v/a}\left(\frac{v}{a}\right)^r dB(v) \quad (r=0,1,2,\ldots)\ldots(22)$$

[*] The period λ of the jth state is defined by Feller to be the greatest common divisor of those values of n for which $p_{jj}^{(n)} > 0$. When $\lambda = 1$ the jth state is said to be aperiodic, and Feller has shown that π_{ij} then exists as an ordinary limit for each value of i.

[#] The occurrence in congestion theory of stochastic matrices having this particular form was first noticed by M. H. Johnson; they play an important role in his own investigation of non-stationary queues with a constant service-time and a general input.

i.e., k_r is the probability that there will be precisely r new arrivals during a single service-time. None of the k's can vanish, and so the diagonal elements of the p-matrix are all positive. The following result is therefore an immediate consequence of (i)'.

Theorem 1. - For the congestion process * the n-step transition-probabilities p^n_{ij} converge to limits π_{ij} as $n \to \infty$. There is another simple consequence of the form of the p-matrix; it can be seen at a glance that there is always a positive probability k_0 that the system will move in a single transition from a given (non-zero) state into the next lowest state, and also that any state can be reached in a single transition from the zero-state ($q = 0$). Accordingly it is always possible to move from one to another of a given ordered pair of states in some finite number of steps with positive probability. In the terminology of Feller's book the chain is irreducible (and aperiodic).

It has just been shown that the limits π_{ij} , exist in the ordinary sense; I shall next show that either all are zero, or else all are positive. For let $\pi_{ab} = 0$. Then since

$$p^{n+M+N}{}_{ab} \geq p^M{}_{ai}\, p^n_{ij}\, p^N{}_{jb} \quad \cdots\cdots \quad (23)$$

it follows that

$$\pi_{ab} \geq p^M_{ai}\, \pi_{ij}\, p^N_{jb} \quad \cdots\cdots \quad (24)$$

for every M and N . Now for given a, b, i and j it is possible to choose M and N so that the p-factors in (24) are both positive, and thus the vanishing of π_{ab} must imply the vanishing of π_{ij} for every i and j .

Because of (19) and because the k's form a probability distribution with a finite mean value ρ , the three power-series

$$K(z) \equiv \sum_{r=0}^{\infty} k_r z^r$$

$$L(z) \equiv \frac{1-K(z)}{1-z} \equiv \sum_{r=0}^{\infty} z^r \sum_{s=r+1}^{\infty} k_s$$

and

$$\pi_i(z) \equiv \sum_{=0}^{\infty} z^j \pi_{ij} \qquad (i = 0, 1, 2, \cdots) \cdots (25)$$

all converge absolutely when $|z| \leq 1$. It follows from Abel's continuity theorem that

[*] It is to be understood that the congestion process has here been redefined in terms of the Markov chain having the one-step transition-probabilities . A complete discussion of the original process in "continuous time" would be much less elementary.

$$\lim_{z\to 1-0} K(z) = 1 ; \qquad \lim_{z\to 1-0} L(z) = \rho$$

and

$$\lim_{z\to 1-0} \Pi_i(z) = \Pi_i(1) \leq 1 \qquad (26)$$

when z approaches the unit point along the real axis from the left. It is to be noted that the generating function $K(z)$ is related to the Laplace transform $\beta(s)$ of the service-time distribution $dB(v)$ by the identity

$$K(z) = \beta\left(\frac{1-z}{a}\right) \quad (|z| \leq 1) \cdots \qquad (27)$$

As soon as $dB(v)$ has been specified, therefore, the functions $K(z)$ and $L(z)$ and the sequence of k's can be determined.

On multiplying the j^{th} of the equations

$$\sum_{a=0}^{\infty} \pi_{ia} p_{aj} = \pi_{ij}$$

by z^j (where $|z| < 1$) and summing over all the values of j (for a fixed value of i) one obtains the identity

$$\Pi_i(z)\{1 - L(z)\} = \pi_{i0} K(z) \qquad (i = 0, 1, 2, \cdot\cdot) \cdot\cdot \quad (28)$$

from which (on letting $z \to 1-0$) it follows that

$$(1-\rho)\Pi_i(1) = \pi_{i0} \quad \cdots\cdots \qquad (29)$$

Now π_{i0} and

$$\Pi_i(1) = \sum_{j=0}^{\infty} \pi_{ij} \leq 1$$

are both finite and non-negative, and so π_{i0} must vanish if $\rho \geq 1$. This completes the proof of

Theorem 2. - If the traffic intensity ρ is greater than or equal to one erlang then the congestion process is completely dissipative, the limits π_{ij} being equal to zero for every i and j .

The practical interpretation of this result is as follows: if $\rho \geq 1$, then as n tends to infinity there is a vanishingly small probability that the n^{th} departing customer will leave behind him a queue of fewer than Q waiting customers.)Here Q is fixed, but it may be arbitrarily lagge .) The most interesting feature of the result is the "unstable" character of the system when the input is equal to the capacity ($\rho = 1$).

A more detailed analysis of the behaviour of the system when $\rho > 1$ will be given in § 4, and the remainder of the present section will be devoted to a discussion of the situation when $\rho < 1$. It is convenient to write

$$H(z) = \sum_{q=0}^{\infty} h_q z^q \equiv (1-\rho)\frac{(1-z)K(z)}{K(z)-z}$$
$$\equiv (1-\rho)\frac{K(z)}{1-L(z)} \qquad . \quad . \quad . \quad . \quad . \qquad (30)$$

(this is identical with the function $H(z)$ which occurred in § 2), and (28) then becomes

$$\pi_i(z) = \pi_{i0} H(z)/(1-\rho) \qquad (|z| < 1) \quad . \quad . \quad . \quad (31)$$

My aim will now be to establish

Theorem 3. - If the traffic intensity ρ is less than one erlang then $\pi_{ij} = h_j$ for every i and j.

This means that when $\rho < 1$, the probability that the n^{th} departing customer leaves behind him a queue of size q approaches (as n tends to infinity) a non-zero limit h_q which is independent of the initial state of the system. The limiting distribution $\{h_q\}$ is identical with that found by Pollaczek and Khintchine in their analysis of the "equilibrium" problem, and in any particular case it can be found by expanding the generating function $H(z)$ in powers of z (it will be recalled that $H(z)$ is itself determined as soon as the service-time distribution $dB(v)$ has been specified). For example, if the service-time distribution has the negative-exponental form (5), then $\beta(s) = 1/(1+bs)$ and so

$$K(z) = \frac{a}{a+b-bz}, \quad L(z) = \frac{b}{a+b-bz}$$

and

$$H(z) = \frac{1-\rho}{1-\rho z},$$

and the size of a queue left by a departing customer is then distributed according to the geometric law

$$(1-\rho)\rho^q \qquad (q = 0, 1, 2, \ldots)$$

(the "equilibrium" solution found by Erlang).

In order examples the calculation of the h's will be more tedious, but it is always in principle a straightforward problem of power-series expansion. It is interesting to notice (and this in fact constitutes an important step in the proof of Theorem 3) that $H(z)$ does always generate a genuine probability distribution when $\rho < 1$. This follows from the expansion

$$H(z) = (1-\rho)K(z)\sum_{r=0}^{\infty}[L(z)]^r$$

(the double series being absolutely convergent if $\rho < 1$ and $|z| \leq 1$), which displays the non-negative character of the h's , and from (30) and (26) which together imply that