STOCHASTIC PROCESSES OCCURRING IN THE THEORY OF QUEUES AND THEIR ANALYSIS BY THE METHOD OF THE IMBEDDED MARKOV CHAIN[1]

By David G. Kendall

Oxford University, England and Princeton University

1. Summary. The stochastic processes which occur in the theory of queues are in general not Markovian and special methods are required for their analysis. In many cases the problem can be greatly simplified by restricting attention to an imbedded Markov chain. In this paper some recent work on single-server queues is first reviewed from this standpoint, and the method is then applied to the analysis of the following many-server queuing-system:

Input: the inter-arrival times are independently and identically distributed in an arbitrary manner.

Queue-discipline: "first come, first served."

Service-mechanism: a general number, s, of servers; negative-exponential service-times.

If Q is the number of people waiting at an instant just preceding the arrival of a new customer, and if w is the waiting time of an arbitrary customer, then it will be shown that the equilibrium distribution of Q is a geometric series mixed with a concentration at $Q = 0$ and that the equilibrium distribution of w is a negative-exponential distribution mixed with a concentration at $w = 0$. (In the particular case of a single server this property of the waiting-time distribution was first discovered by W. L. Smith.)

The paper concludes with detailed formulae and numerical results for the following particular cases:

Numbers of servers: $s = 1, 2$ and 3.

Types of input: (i) Poissonian and (ii) regular.

2. Introduction. This paper follows an earlier one [8] in which the reader will find a detailed account of the history of the subject, the technological applications and the conventions of notation and terminology, but a thorough familiarity with the contents of [8] will not be assumed. A queuing-system of the type to be considered is specified when we know (i) the *input*, (ii) the *queue-discipline* and (iii) the *service-mechanism*. It will here be supposed that if the successive "customers" demand service at the epochs $\cdots, t_r, t_{r+1}, \cdots$, and if u_r denotes the inter-arrival time $t_{r+1} - t_r$, then the random variables $\cdots, u_r, u_{r+1}, \cdots$ are statistically independent and enjoy the same (arbitrary) distribution $dA(u)$ $(0 \leqq u < \infty)$. (Note that in some important applications this supposition of independence will not be admissible; for example, it cannot be made when the

Received 11/27/52.

[1] Work done partly under the sponsorship of the U. S. Office of Naval Research.

"customers" (which may be ships, or aircraft) are scheduled to arrive at specified times and are late or early by independent random time-errors.) When no further assumptions than these are made about the input I shall describe it as a *general independent* input and in the label denoting the system this state of affairs will be indicated by the letters *GI*. (Here I follow Lindley [3]; in [8] I called such an input "regenerative," but in the present paper I shall make no explicit use of the concept of a set of regeneration points (for this, see [8] and the subsequent discussion).) There are two important special types of input:

(1) D (deterministic, or regular):

$$A(u) \equiv 0(u < a); \qquad A(u) \equiv 1(u \geqq a),$$

(2) M ("random," or Poissonian):

$$A(u) \equiv 1 - e^{-u/a}.$$

With a D-input the customers arrive at regular intervals of time (the inter-arrival time being fixed and equal to a) while with an M-input the customers arrive "at random" (i.e., in a Poisson process). We may also mention an intermediate type of input:

(3) $$E_k \text{ (Erlangian)}: dA(u) \equiv \frac{(k/a)^k}{\Gamma(k)} e^{-ku/a} u^{k-1}\, du,$$

which coincides with M when $k = 1$ and which approaches D as k tends to infinity. The idea of bridging the gap between D and M in this way is derived from a similar device employed by Erlang (see [5]) in connection with the service-time distribution. (For a more detailed account of the method in relation to a problem in population mathematics see [7] and [9]. An extension due to Jensen and Palm is discussed in [5].) If the original input is Poissonian and if it is filtered in such a way that only every kth customer is admitted to the system then the net input will be of type E_k (the mean inter-arrival time being increased to ka); this remark is of interest in connection with a certain cyclic rule of queue-discipline different from the one to be considered here. In all three special cases it will be noted that $E(u) = a$; I shall give a this meaning in the general case also and I shall suppose throughout that $0 < a < \infty$.

So much for the input. Under the heading of queue-discipline it will be supposed that the customers form up into a single queue in the usual way and that the customer at the head of the queue is served as soon as a server is free to attend to him. In general there will be s servers and s will not necessarily be equal to unity.

The service-mechanism will be defined by the assertions that the service-times $\cdots, v_r, v_{r+1}, \cdots$ of the successive customers are statistically independent of one another and of the input (thus the presence of a long queue is here supposed to have no effect on the speed of service), and that for all customers (irrespective of the identity of the server) the service-time has the (arbitrary) distribution

$dB(v)(0 \leqq v < \infty)$. The distributions $dA(u)$ and $dB(v)$ will be classified in exactly the same way, with the aid of the symbols G, D, M and E_k, and it will be supposed throughout that $0 < E(v) \equiv b < \infty$. The symbol G denotes that no special assumption is made about $dB(v)$. With a D-distribution for the service-time each customer is served for exactly the same length of time b. With an M-distribution the service-times follow the negative-exponential law found to hold for unrestricted telephone conversations. Once again the E_k distribution is of an intermediate form.

With these conventions a particular type of queuing-system can be identified by giving it a label such as $D/G/3$ (regular arrivals; no special assumption about the service-time distribution; three servers). Table 1 summarizes the principal contributions to the subject and shows where accounts of the various types of queuing-system are to be found.

TABLE 1

Analysis of the literature on the theory of queues[2]

Author (date)	Systems discussed	References
Erlang (1908–29).....	$M/M/s$, $M/D/s$, $M/E_k/1$	[5] (see also [11] and [13])
Pollaczek (1930).....	$E_k/G/1$	[14]
Khintchine (1932).....	$M/G/1$	[10]
Pollaczek (1934).....	$M/G/s$	[15]
Volberg (1939).....	$E_k/G/1$	[20]
Kendall (1951).....	$M/G/1$	[8]
Lindley (1952).....	$GI/G/1$	[12]
Pollaczek (1952).....	$GI/G/1$	[18]
Smith (1952).....	$GI/G/1$	[19]
Kendall (1952).....	$GI/M/s$	This paper.

Except in the case $M/M/s$ the stochastic processes associated with the fluctuations in queue-size are non-Markovian and special methods are required for their analysis. In [8] I examined the system $M/G/1$ by considering the behavior of a certain imbedded Markov chain and in this way obtained the distribution of queue-size in statistical equilibrium (a result originally found by Pollaczek [14] and Khintchine [10], each using quite different methods), and I discussed the ergodic properties of the system in relation to the value of the *relative traffic intensity* $\rho \equiv b/a$. ("Relative," because by a generally accepted convention it is measured in relation to the capacity of the system. When calculated in this way ρ is said to be expressed in erlangs, the erlang being the international unit of telephone traffic.)

[2] See also footnote 3.

Recently Lindley [12] has formulated and in several important cases solved an integral equation of the Wiener-Hopf type for the waiting-time distribution in statistical equilibrium when the queuing-system is of the more general type $GI/G/1$ (special attention being paid to the systems $D/E_k/1$ which Bailey and Welch [1], [2] have shown to be of importance in the design of appointment systems in hospital outpatient departments). (Lindley has given detailed solutions for the systems $M/G/1$, $E_k/G/1$ and $D/E_k/1$. For an independent treatment of the systems $GI/G/1$ see Pollaczek [18].) In continuation of Lindley's work W. L. Smith [19] has considered several other single-server systems, including those of the type $GI/M/1$, in similar detail.

I shall show here that the work of Lindley and Smith can also be regarded as an application of the method of the imbedded Markov chain, and I shall then apply the "imbedding" method to analyze the properties of the many-server system $GI/M/s$. It was observed by Smith in his study of $GI/M/1$ that the assumption of a negative-exponential service-time distribution leads to solutions of a very simple form, whatever the (general independent) input; it will be seen here that the same is true even when we allow a general number of servers.

In [8] I examined the ergodic behavior of the Markov chain imbedded in $M/G/1$ with the aid of Feller's theory of recurrent events; there are three quite different types of behavior when $\rho < 1$, when $\rho = 1$ and when $\rho > 1$. For the many-server system it has been observed by Pollaczek that the appropriate definition of the relative traffic-intensity is $\rho \equiv b/(sa)$. I shall assume here that Pollaczek's ρ is less than unity; with this assumption it will be shown that a stable equilibrium exists and the associated equilibrium distributions will be determined. (Dr. F. G. Foster has considered the dependence of the qualitative behavior of the system on the parameter ρ; his results are given elsewhere in this issue, (F. G. Foster, "On the stochastic matrices associated with certain queuing processes," *Ann. Math. Stat.*, Vol. 24 (1953)).)

3. The imbedded Markov chain. Let the state of a stochastic system at time t be denoted by $X(t)$, so that (in any actual realization) the history of the system can be represented as a function $X(\cdot)$ of the time with domain $(-\infty, \infty)$. (In the applications which follow it will be an integer-valued step-function defined to be continuous-to-the-right at its points of discontinuity.) Let Ω_t denote the set whose elements are the functions having as domain the time-interval $(-\infty, t]$ and having the same range as $X(\cdot)$. For each t in $(-\infty, \infty)$ let Θ_t be a specified subset of Ω_t, and corresponding to any actual realization of the process let Π be the set of those values of t in $(-\infty, \infty)$ for which Θ_t contains as an element the contraction of $X(\cdot)$ to the reduced domain $(-\infty, t]$. Let

$$Y(t) \equiv f_t\{X(\tau): \tau \leqq t\} \qquad \text{for } t \in \Pi,$$

where f_t is some specified functional with domain Θ_t. Now suppose that $\{\Theta_t, f_t : -\infty < t < \infty\}$ have been chosen in such a way that

(i) Π almost certainly has no finite point of accumulation. (We then write its members in increasing order as $\cdots, t_{n-1}, t_n, t_{n+1}, \cdots$.)

(ii) If Y_m denotes $Y(t_m)$ for each $t_m \varepsilon \Pi$, then distribution $\{Y_{n+1} \mid Y_n, Y_{n-1}, \cdots\} \equiv$ distribution $\{Y_{n+1} \mid Y_n\}$ for all n.

The variables $\cdots, Y_{n-1}, Y_n, Y_{n+1}, \cdots$ will then be said to constitute an *imbedded Markov chain*.

Such an imbedded chain can always be constructed, at least in a trivial way. Thus we could choose $\Theta_t = \Omega_t$ for each *integer* t, and require Θ_t to be void for all other values of t. Π would then be the set of integers, and by taking $f_t \equiv 1$ we would obtain an imbedded Markov chain. In practice, however, three conditions must be satisfied if the procedure is to be of any value. First, the system must be simple enough to permit a mathematical formulation of the present heuristics. (The abstract formulation employed in this and the preceding paragraph must not mislead the reader into thinking that it would be a simple matter to implement the program envisaged here in complete generality. Grave difficulties of definition would be encountered at the outset. The remarks in the present section of the paper are offered only as a guide to intuitive thinking.) Secondly, for Y to be useful as a reduced state-description, the functional f_t must be sufficiently and suitably sensitive to variations in its argument. Thirdly, the stochastic mechanism governing the transition from one instant in Π to the next must be simple enough to permit the calculation of the transition-probabilities associated with distribution $\{Y_{n+1} \mid Y_n\}$.

The stochastic processes with which we shall be concerned all have a Markovian origin and they are deprived of their Markovian character only because we are unwilling to work with a sufficiently comprehensive description of the present state. One way of remedying this difficulty would be to augment the description of the present state so as to imbed the given process in a more complicated one having the Markov property. To illustrate this procedure we might consider taking the state $Z(t)$ of the augmented process to be the whole past history of the given process:

$$Z(t) \equiv \text{the contraction of } X(\cdot) \text{ to the reduced domain } (-\infty, t].$$

However, certain difficulties are to be expected in defining the Z-process satisfactorily, and it is fortunate that such a drastic procedure is not necessary in queuing theory, where in the worst case ($GI/G/s$) it would be enough to replace the single initial state-variable (queue-size) by a vector variable of $s + 2$ components (the extra components specifying the expended service-times of the people being served and the expended inter-arrival time). This particular form of the "augmentation" technique can be carried through in some simple cases but only at the expense of very complicated calculations, and the method of contraction to an imbedded Markov chain is usually preferable even although by its very nature it must leave some of our questions unanswered.

I shall illustrate the "contraction" method by referring briefly to my earlier

treatment of $M/G/1$ and to Lindley's treatment of $GI/G/1$. I shall then apply it to obtain the equilibrium behavior of $GI/M/s$ (a queuing process which seems never to have been treated before except in some special cases).[3]

4. The system $M/G/1$. Let q be the number of people waiting or being served at time t; then $X(t) \equiv q$ does not constitute a Markov process (except in the special case $M/M/1$). The augmentation technique would not here be too difficult; it would suffice to take $Z(t) \equiv (q, v_0)$ where v_0 is the expended service-time of the person being served (v_0 being left undefined when $q = 0$). The contraction technique proceeds as follows (full details will be found in [8]; I quote here a few illustrative results only). We define $X(\cdot)$ to be continuous-to-the-right at its points of discontinuity, and we take the test for membership of the set Θ_t to be $X(t) = X(t - 0) - 1$, ("the value of q has just decreased by unity"); thus the set Π consists of *the epochs of departure*. When $t \in \Pi$, let $Y(t) = X(t) = q$, so that Y is *the number of persons left behind by a departing customer* (including the person, if any, whose service is just starting). The fact that the input is of the Poisson type (i.e., the M in $M/G/1$) ensures that the Y-chain is Markovian.

Let q' and q'' be the numbers of persons left behind by two consecutively departing customers and let

$$p_{ij} \equiv \text{pr}\{q'' = j \mid q' = i\}.$$

Then $\mathbf{P} \equiv \| p_{ij} \|$ is the transition-matrix for the imbedded Markov chain, and we can proceed with its analysis by the methods described in Feller's book [6]. It is found that the matrix $\mathbf{P}$ is

↗	0	1	2	3	·
0	k_0	k_1	k_2	k_3	·
1	k_0	k_1	k_2	k_3	·
2	0	k_0	k_1	k_2	·
3	0	0	k_0	k_1	·
·	·	·	·	·	·

where

$$k_j \equiv \frac{1}{j!} \int_0^\infty e^{-v/a} \left(\frac{v}{a}\right)^j dB(v) \qquad (j = 0, 1, 2, \cdots), \tag{4}$$

[3] On my communicating a MS. copy of the present paper to M. Pollaczek, he informed me that he has recently considered the system $GI/M/s$ from a different point of view, a number of his results being qualitatively equivalent with some of mine. His paper has since appeared [Pollaczek, 1953] and contains an attack on the problem of the more general system $GI/G/s$. It seems that manageable solutions may be expected whenever the Laplace transform of the v-distribution is rational.

and the chain is in all cases irreducible and aperiodic. It is ergodic when $\rho < 1$, recurrent-but-null when $\rho = 1$ and transient when $\rho > 1$, and in the ergodic case the limiting q-distribution is that generated by the function

$$H(z) \equiv (1 - \rho) \frac{(1 - z) K(z)}{K(z) - z}; \tag{5}$$

where $K(z) \equiv \sum_0^\infty k_j z^j$. From this the Pollaczek formula for the Laplace transform of the waiting-time distribution follows easily on noting that q is the number of arrivals during the sum of the departing customer's waiting-time and his service-time. In particular the probability of not having to wait is found to be $1 - \rho$, and the ratio of the mean waiting-time to the mean service-time is given by

$$\frac{E(w)}{E(v)} = \frac{\rho}{2(1 - \rho)} \left\{1 + \text{var}\left(\frac{v}{b}\right)\right\}, \tag{6}$$

a formula of great practical importance.

5. The system *GI/G/1*. Let q and $X(t)$ be defined as in Section 4 but now let the test for membership of the set Θ_t be the requirement that *either*

$$X(t - 0) = 0 \quad \text{and} \quad X(t) = 1$$

or

$$X(t - 0) \geqq 2 \quad \text{and} \quad X(t) = X(t - 0) - 1$$

("the service of a customer has just commenced"); thus the set Π will consist of *the epochs of commencement of service.* If $t \in \Pi$ then we take $Y(t)$ to be *the waiting-time w of the customer whose service has just commenced* (that is, the time since his arrival); of course $Y(t)$ may be zero. The value of $Y(t)$ can be found by examining the graph of q against t, starting at some epoch when q was equal to zero; thus $Y(t)$ is a rather complicated functional of the contraction of $X(\cdot)$ to the domain $(-\infty, t]$, and it is this functional which is f_t. A little intuitive consideration will show that the Y-chain is Markovian, although (in contrast to the previous example) it has a continuum of states. (If w_r is given, then the prediction of w_{r+1} is equivalent to the prediction of $v_r - u_r$ (formula (7) below). Now information about $w_{r-1}, w_{r-2}, \cdots$ would be of no assistance in making this prediction.) It should be emphasized that the Y-chain would not be Markovian if the input were not of the special type indicated by the symbol *GI*. The determination of distr $\{Y_{n+1} \mid Y_n\}$ here depends on the fact that

$$w_{r+1} = \max\{w_r + v_r - u_r, 0\}, \tag{7}$$

where w_r is the waiting-time of the customer whose service-time is v_r, and u_r is the time which elapses between the arrival of this and of the next customer. The simple matrix relations of Section 4 are here replaced by an integral equation of the Wiener-Hopf type due to D. V. Lindley [12], (see also Smith [19]).

6. The many-server system *GI/M/s*. Here q is the number of persons waiting in the queue or being served at one of the s service-points at the time t, and $X(t) = q$. The system is Markovian only in the special case $M/M/s$ treated by Erlang [5], Molina [13] and Kolmogorov [11]. Let the test for membership of the set Θ_t be $X(t) = X(t - 0) + 1$, ("a customer has just arrived"); then the set Π will consist of *the epochs of arrival.* If $t \in \Pi$ let $Y(t) = X(t - 0)$, so that Y is *the number of persons found to be ahead of him (waiting, or being served) by the newly arrived customer.* As before, a little consideration will show that (because of the negative-exponential service-times) the Y-chain is Markovian. It will be convenient to use q for the value of Y but it should be borne in mind that it is not q but $Q \equiv \max(q - s, 0)$ which is the length of the queue in the ordinary sense of the word; if $0 \leqq q \leqq s$ then q of the service-points will be occupied and $s - q$ will be free and no one will be waiting.

We commence with an examination of the general form of the matrix $\mathbf{P}$ whose (i, j)th element is $p_{ij} \equiv \mathrm{pr}\{q'' = j \mid q' = i\}$, where q' and q'' refer to two consecutive epochs of arrival. The best way to describe the matrix $\mathbf{P}$ is first to partition it as follows:

$$\mathbf{P} \equiv \left[\begin{array}{c} \mathbf{A} \\ \hdashline \mathbf{B} \end{array} \vdots \mathbf{C}\right]. \tag{8}$$

Here $\mathbf{A}$ is a square matrix of s rows and s columns; in describing the elements of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ they will always be given the labels which they bear in virtue of position in $\mathbf{P}$; thus the top left-hand element of $\mathbf{C}$ will be called $c_{0,s}$ and not $c_{1,1}$.

Consider the (i, j)th element of the matrix $\mathbf{A}$. We must have $i \leqq s - 1$, and so the newly arriving customer will find $s - 1$ or fewer customers ahead of him, all of whom are being served. There will therefore be at least one server free and so the service-time of the new customer can commence immediately. We have to account for the events during the period of time u which elapses between his own arrival and the arrival of the next customer. Suppose that n customers conclude their service during this period, and let us write $[n \mid m; u]$ for the conditional probability associated with the stated value of n when $m(\equiv i + 1)$ customers are being served at the commencement of the period. We can think of these m customers as the members of a colony which is subject to a randomly-operating death-rate of amount $1/b$ per head per unit of time. The theory of the simple death-process (see, for example, [6]) then gives

$$[n \mid m; u] = \binom{m}{n}(1 - e^{-u/b})^n e^{-(m-n)u/b}, \tag{9}$$

(a result which can also be obtained directly without much difficulty). If we put

$$[n \mid m] \equiv \int_0^\infty [n \mid m; u]\, dA(u) \tag{10}$$

then we shall have

$$p_{ij} = [i + 1 - j \mid i + 1] \qquad \text{when} \qquad j \leqq i + 1$$

and

$$p_{ij} = 0 \text{ elsewhere,}$$

provided that $(i, j) \varepsilon \mathbf{A}$. Here is the form of $\mathbf{A}$ when $s = 4$:

$$\begin{bmatrix} [1 \mid 1] & [0 \mid 1] & 0 & 0 \\ [2 \mid 2] & [1 \mid 2] & [0 \mid 2] & 0 \\ [3 \mid 3] & [2 \mid 3] & [1 \mid 3] & [0 \mid 3] \\ [4 \mid 4] & [3 \mid 4] & [2 \mid 4] & [1 \mid 4] \end{bmatrix}$$

It is important to notice that the elements of $\mathbf{A}$ are positive when $j \leqq i + 1$.

The elements of $\mathbf{B}$ are less simple in form. Here $i \geqq s$, and so immediately after the commencement of the inter-arrival time u there will be $m \equiv i - s + 1$ persons waiting *in addition* to the s persons who are being served. Also $j < s$, and so at the end of the interval no one can be waiting and $n \equiv s - j$ servers must be free. We require the conditional probability $\{n \mid s; m; u\}$ associated with the stated value of n when m and u are given. Suppose that the last of the m waiting customers is received at a service-point after the lapse of a time U. The distribution of U can be written down at once because $sU/b = \frac{1}{2}\chi^2_{2m}$, and so

$$\{n \mid s; m; u\} = \frac{1}{(m-1)!}\left(\frac{s}{b}\right)^m \int_0^u e^{-sU/b} U^{m-1} [n \mid s; u - U]\, dU. \tag{11}$$

If now we put

$$\{n \mid s; m\} \equiv \int_0^\infty \{n \mid s; m; u\}\, dA(u) \tag{12}$$

then we shall have

$$p_{ij} = \{s - j \mid s; i - s + 1\} \qquad \text{when} \qquad (i, j) \varepsilon \mathbf{B};$$

it is to be noted that all these probabilities will be positive.

Finally we require the elements of $\mathbf{C}$, and here $j \geqq s$. Obviously we shall have $p_{ij} = 0$ if $j > i + 1$, because no one can enter the system during the inter-arrival interval. Let us write $n \equiv i + 1 - j$ for the number of customers whose service is concluded during the inter-arrival time, and then seek the conditional probability $(n \mid s; u)$ associated with the stated value of n when u is given. (The notation implies that this probability is independent of the value of i (which is also to be supposed given). That this is so will be the principal result of the argument which follows.) Because $j \geqq s$ for the elements of $\mathbf{C}$ we need only consider the values of n such that $0 \leqq n \leqq i + 1 - s$, and if n has one of these values then *there will be no service-point unoccupied during or at the end of the inter-arrival time* u. Thus $(n \mid s; u)$ is equal to the probability that, in time u, n incidents will be registered in a Poisson process for which the expected incident-rate is s/b per unit of time. That is,

$$(n \mid s; u) = e^{-su/b} \frac{(su/b)^n}{n!}, \tag{13}$$

and so if we write

$$(n \mid s) \equiv \int_0^\infty (n \mid s; u)\, dA(u) \tag{14}$$

then we shall have

$$p_{ij} = (i + 1 - j \mid s) \qquad \text{if} \qquad i \leqq i + 1,$$

and

$$p_{ij} = 0 \text{ elsewhere,}$$

provided that $(i, j) \varepsilon \mathbf{C}$. Here is the form of $\mathbf{C}$ when $S = 4$:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & \cdot \\ 0 & 0 & 0 & 0 & \cdot \\ 0 & 0 & 0 & 0 & \cdot \\ (0 \mid 4) & 0 & 0 & 0 & \cdot \\ (1 \mid 4) & (0 \mid 4) & 0 & 0 & \cdot \\ (2 \mid 4) & (1 \mid 4) & (0 \mid 4) & 0 & \cdot \\ (3 \mid 4) & (2 \mid 4) & (1 \mid 4) & (0 \mid 4) & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

It is to be noted that the diagonal and super-diagonal elements of $\mathbf{P}$ which lie in $\mathbf{C}$ are respectively equal to $(1 \mid s)$ and to $(0 \mid s)$ and that both these quantities are positive.

7. The many-server system *GI/M/s* (continued). Now that we know the form of the matrix $\mathbf{P}$ we can apply Feller's treatment of denumerable Markov chains to obtain the principal properties of the imbedded Markov chain in the practically important case when the relative traffic-intensity, $\rho \equiv b/(sa)$, is less than unity. Familiarity with chapter 15 of Feller's book [6] will be assumed.

In the first place, the chain is *irreducible*. This can either be seen analytically, or made self-evident by the following intuitive considerations.

(a) The transition $i \to 0$ always has a positive probability because it can happen that all the $i + 1$ customers will be served and will leave the system during the inter-arrival period.

(b) The transition $i \to i + 1$ always has a positive probability because it can happen that no customer will leave the system during the inter-arrival period. Thus it is possible in a suitable (finite) number of steps for the system to move from any given state to the zero state and thence to any other given state.

Accordingly, every state is of the same "type." That they are all *aperiodic* follows from Feller's theory and the fact that the diagonal elements of $\mathbf{P}$ are positive. We shall now show that with the given restriction on the relative traffic-

intensity ($\rho < 1$), the states are *ergodic*. This will imply that the probability p_{ij}^n (that the system will be in state j after n steps) converges in the ordinary sense (as n tends to infinity) towards the jth element, π_j, of a limiting distribution which is independent of the initial state, i.

From Feller's theorems 1 and 2 we know that *either* (i) every state is transient (or every state is recurrent-but-null) and $p_{ij}^n \to 0$ as $n \to \infty$ for all i and j; *or* (ii) every state is ergodic and $p_{ij}^n \to \pi_j$ as $n \to \infty$ for all i and j, where the π's are positive and sum to unity. Suppose that the matrix $\mathbf{P}$ could be shown to possess a nonnull invariant row-vector $\mathbf{x}$, the components of which form the terms in an absolutely convergent series. Then $\mathbf{x} = \mathbf{xP} = \mathbf{xP}^n$, and so in case (i) we should have

$$x_j = \sum_{\alpha=0}^{\infty} x_\alpha p_{\alpha j}^n \to 0 \qquad \text{as } n \to \infty,$$

and in case (ii) we should have

$$x_j = \sum_{\alpha=0}^{\infty} x_\alpha p_{\alpha j}^n \to (\sum x_\alpha)\pi_j \qquad \text{as } n \to \infty,$$

and because $\mathbf{x}$ is supposed not to be null it would then follow that the system must be ergodic, that $\sum x_\alpha \neq 0$ and that the vector $\mathbf{x}$ could be normalized to give the limiting distribution, $\boldsymbol{\pi}$. Thus, in order to establish ergodicity and determine the limiting distribution it will be sufficient to construct a vector $\mathbf{x}$ having the stated properties. (It is important to note that we do not need to worry about the signs of the components of $\mathbf{x}$.)

In order to implement this program, let us write

$$\mathbf{x} \equiv [\mu_0, \mu_1, \cdots, \mu_{s-2}, 1, \lambda, \lambda^2, \lambda^3, \cdots] \tag{15}$$

(the μ-terms being absent when $s = 1$). I shall show that λ and the μ's can be chosen so as to give this vector the required properties (it will be enough if it is invariant and if $|\lambda| < 1$). The invariance of $\mathbf{x}$ requires that

$$x_j = \sum_{\alpha=0}^{\infty} x_\alpha p_{\alpha j} \qquad (0 \leqq j < \infty)$$

and when $j \geqq s$ each of these equations will be found to be equivalent to the following equation for λ:

$$F(\lambda) = \lambda \qquad (0 < \lambda < 1) \tag{16}$$

where

$$F(\lambda) \equiv \int_0^\infty e^{-(1-\lambda)su/b}\, dA(u). \tag{17}$$

There is a momentary advantage in writing the λ-equation in the form, $\sum_{n=0}^{\infty} (n \mid s)\lambda^n = \lambda$, and then noting that the coefficients $(n \mid s)$ are all positive and are the terms in a probability distribution whose mean is equal to $1/\rho$ (and so is

greater than unity). It is now an immediate consequence of the fundamental lemma of branching-process theory (see, e.g., [6]) that the λ-equation has a unique root in the interval $0 < \lambda < 1$. In what follows the symbol λ will always denote this root.

The equations $x_j = \sum x_\alpha p_{\alpha j}$ $(1 \leqq j \leqq s-1)$ can now be used for the successive determination of the μ's (this depends on the fact that $[0 \mid 1]$, $[0 \mid 2]$, $\cdots$, $[0 \mid s-1]$ are all positive). Thus we have only to verify that the vector $\mathbf{x}$ so constructed satisfies the last of the invariance conditions, $x_0 = \sum x_\alpha p_{\alpha 0}$. But this is an immediate consequence of the fact that the row-sums of the matrix $\mathbf{P}$ are all equal to unity (the intuitive meaning of the row-sum condition is obvious; its analytical verification is tedious but elementary). The following statement summarizes the results so far obtained:

I: *when the relative traffic-intensity is less than unity, the Markov chain imbedded in $GI/M/s$ is irreducible and ergodic. The limiting distribution is a geometric series save for modifications to its first $(s-1)$ terms, the common ratio being the unique root of the equation*

$$F(\lambda) = \lambda \qquad (0 < \lambda < 1).$$

Now i, the number of persons waiting or being served in the system, is not in practice so interesting as the "true" queue-size encountered by the newly-arriving customer:

$$Q \equiv \max(i - s, 0). \tag{18}$$

A random variable of equal importance is the waiting-time of the new customer, w. If $k \equiv \max(i - s + 1, 0)$ then w will be the sum of k independent variables each distributed like $\frac{1}{2}b\chi_2^2/s$, and so the Q- and w-distributions are quite easy to find once the i-distribution is known. Thus, in the statistical equilibrium which is ultimately attained when the relative traffic-intensity ρ is less than unity, we shall have:

II: *the probability that Q is zero is*

$$\alpha \equiv \frac{\sum\mu + 1 + \lambda}{\sum\mu + 1/(1-\lambda)}, \tag{19}$$

and the probability that w is zero is

$$\beta \equiv \frac{\sum\mu + 1}{\sum\mu + 1/(1-\lambda)}; \tag{20}$$

III: *the mean value of Q is given by*

$$E(Q) = \frac{1-\alpha}{1-\lambda}, \tag{21}$$

and the mean value of w is given by

$$\frac{E(w)}{E(v)} = \frac{1-\beta}{s(1-\lambda)}; \tag{22}$$

IV: *when Q is known to be positive, its conditional distribution is*

$$(1 - \lambda)\lambda^{Q-1} \qquad (Q = 1, 2, 3, \cdots), \tag{23}$$

and when w is known to be positive, its conditional distribution is

$$e^{-w/c}\, dw/c \qquad (0 < w < \infty), \tag{24}$$

where

$$c \equiv \frac{b}{s(1 - \lambda)}. \tag{25}$$

The second part of Theorem IV is a generalization of a remarkable result recently discovered by W. L. Smith [19]. He observed that for a single server and with certain restrictions on the form of the distribution $dA(u)$ defining the general independent input, a negative-exponential distribution of service-times produces a negative-exponential distribution of waiting-time (apart from a probability-concentration at the origin). We now see that the restrictions on the form of the distribution $dA(u)$ are unnecessary and that the result is true whatever the number of servers. Moreover, this simple property of the waiting-time distribution is associated with an equally simple property of the queue-size distribution, which we have shown to be of the geometric-series form apart from a probability-concentration at $Q = 0$.

8. Detailed results for $GI/M/s$ when $s \leqq 3$. Suppose first that the inter-arrival time u has a distribution $dA(u) \equiv dA_1(u/a)$, where $dA_1(u)$ is a fixed distribution with a mean equal to unity, so that the average inter-arrival time, a, enters $dA(u)$ only as a scale-parameter. Then the (λ, ρ)-relation can be written

$$\lambda = \int_0^\infty e^{-(1-\lambda)x/\rho}\, dA_1(x) \qquad (0 < \lambda < 1), \tag{26}$$

and so it is independent of a and of the number of servers, s. Two special cases are of interest.

Poissonian input (system $M/M/s$). Here $dA_1(u) \equiv e^{-u}du$, and the (λ, ρ)-equation is

$$\lambda^2 - (1 + \rho)\lambda + \rho. = 0. \tag{27}$$

The root in the interval (0, 1) is $\lambda = \rho$. (The results for the system $M/M/s$ are well known and are to be found in references [5], [13], [11] and [6].)

Regular input (system $D/M/s$). (This system has only been studied before in the case $s = 1$.) Here the (λ, ρ)-equation can most conveniently be put in the form

$$1 - e^{-X} = \rho X, \quad \text{where} \quad X \equiv (1 - \lambda)/\rho, \quad \text{and} \quad 0 < \lambda < 1. \tag{28}$$

A few corresponding values of λ and ρ are given in Table 2.

In order to complete the study of any particular case when $s \geqq 2$ we need the values of the μ's. When $s = 2$ or 3 the procedure is as follows; it will be obvious how this is to be extended when $s \geqq 4$.

TABLE 2

The (λ, ρ)*-relation for the system* $D/M/s$

ρ	λ	ρ	λ
0.0	0	0.6	0.3242
0.1	0.0000 4542	0.7	0.4670
0.2	0.0069 77	0.8	0.6286
0.3	0.0408 8	0.9	0.8069
0.4	0.1073 6	1.0	1
0.5	0.2032		

Determination of μ_0 *when* $s = 2$. To simplify the formulae, I shall suppose that $a = 1$; this will not result in any loss of generality. The μ_0-equation is

$$1 = [0 \mid 1]\mu_0 + [1 \mid 2] + \sum_{\alpha=1}^{\infty} \{1 \mid 2; \alpha\}\lambda^\alpha, \tag{29}$$

and after a few transformations this becomes

$$\left(\frac{2}{2\lambda - 1} - \mu_0\right) \int_0^\infty e^{-u/2\rho}\, dA(u) = \frac{1}{2\lambda - 1}. \tag{30}$$

Thus we have:

$$\text{System } M/M/2: \qquad \mu_0 = 1/(2\rho). \tag{31}$$

$$\text{System } D/M/2: \qquad \mu_0 = \frac{2 - e^{1/(2\rho)}}{2\lambda - 1} \tag{32}$$

Determination of μ_0 *and* μ_1 *when* $s = 3$. The equations determining the μ's are

$$1 = [0 \mid 2]\mu_1 + [1 \mid 3] + \sum_{\alpha=1}^{\infty} \{1 \mid 3; \alpha\}\lambda^\alpha, \tag{33}$$

and

$$\mu_1 = [0 \mid 1]\mu_0 + [1 \mid 2]\mu_1 + [2 \mid 3] + \sum_{\alpha=1}^{\infty} \{2 \mid 3; \alpha\}\lambda^\alpha, \tag{34}$$

and after some transformations these become

$$\left(\frac{3}{3\lambda - 1} - \mu_1\right) \int_0^\infty e^{-2u/3\rho}\, dA(u) = \frac{1}{3\lambda - 1} \tag{35}$$

and

$$\left(\frac{6}{3\lambda - 2} - \mu_0 - 2\mu_1\right)\int_0^\infty e^{-u/3\rho}\, dA(u) = \frac{4}{3\lambda - 2} - \mu_1. \tag{36}$$

It is now quite a simple matter to find μ_0 and μ_1 when $dA(u)$ is given. For example:

$$\text{System } M \mid M \mid 3: \qquad \mu_1 = \frac{2}{3\rho} \quad \text{and} \quad \mu_0 = \frac{2}{(3\rho)^2}. \tag{37}$$

$$\text{System } D \mid M \mid 3: \qquad \text{if } x \text{ is the positive root of } x^{3\rho} = e,$$

then

$$\mu_1 = \frac{3 - x^2}{3\lambda - 1}$$

and

$$\mu_0 = 2\left(\frac{3 - 2x}{3\lambda - 2}\right) - (2 - x)\mu_1. \tag{38}$$

The above formulae have been used in the construction of Table 3, which shows the effect of varying the quality and intensity of the input and the number of servers when the service-time has a negative-exponential distribution. *For a specified relative traffic-intensity, ρ, Table 3 gives*

(a) the probability of not having to wait, and

(b) the ratio of the mean waiting-time to the mean service-time. "Random" arrivals and "Regular" arrivals refer to the systems $M/M/s$ and $D/M/s$, respectively.

TABLE 3

The many-server queuing-system with a negative-exponential service-time

Relative traffic-intensity ρ	One server				Two servers				Three servers			
	Random arrivals		Regular arrivals		Random arrivals		Regular arrivals		Random arrivals		Regular arrivals	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
0.0	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
0.1	0.90	0.11	1.00	0.00	0.98	0.01	1.00	0.00	1.00	0.00	1.00	0.00
0.2	0.80	0.25	0.99	0.01	0.93	0.04	1.00	0.00	0.98	0.01	1.00	0.00
0.3	0.70	0.43	0.96	0.04	0.86	0.10	0.99	0.00	0.93	0.03	1.00	0.00
0.4	0.60	0.67	0.89	0.12	0.77	0.19	0.96	0.02	0.86	0.08	0.98	0.01
0.5	0.50	1.00	0.80	0.26	0.67	0.33	0.90	0.06	0.76	0.16	0.94	0.02
0.6	0.40	1.50	0.68	0.48	0.55	0.56	0.79	0.15	0.65	0.30	0.86	0.07
0.7	0.30	2.33	0.53	0.88	0.42	0.96	0.65	0.33	0.51	0.55	0.73	0.17
0.8	0.20	4.00	0.37	1.69	0.29	1.78	0.47	0.71	0.35	1.08	0.54	0.41
0.9	0.10	9.00	0.19	4.18	0.15	4.26	0.25	1.93	0.18	2.72	0.30	1.21
1.0	0.00	∞	0.00	∞	0.00	∞	0.00	∞	0.00	∞	0.00	∞

I should like to express my gratitude to Mr. W. L. Smith for his kindly allowing me to see a copy of his paper in advance of publication. I am also indebted to

Professor J. W. Tukey and Mr. D. M. G. Wishart for their helpful comments, and to the referee for several useful suggestions.

REFERENCES

*[1] N. T. J. Bailey and J. D. Welch, "Appointment systems in hospital outpatient departments," *The Lancet*, (1952), p. 1105.

*[2] N. T. J. Bailey, "Study of queues and appointment systems in outpatient departments, with special reference to waiting-times," *J. Roy. Stat. Soc.* (B) Vol. 14 (1952), pp. 185–199.

*[3] G. E. Bell, D. V. Lindley, K. D. Tocher and J. D. Welch, "Marshalling and queuing," *Operational Research Quarterly*, Vol. 3 (1952), pp. 4–13.

*[4] G. S. Berkeley, "Traffic and delay formulae," *Post Office Electrical Engineers' J.*, Vol. 29.

[5] E. Brockmeyer, H. L. Halstrøm and A. Jensen, *The Life and Works of A. K. Erlang* (Copenhagen, 1948). (English translations of all of Erlang's papers will be found in this volume, together with a number of valuable memoirs by the three editors.)

[6] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley and Sons, 1950.

[7] D. G. Kendall, "On the role of variable generation time in the development of a stochastic birth process," *Biometrika*, Vol. 35 (1948), pp. 316–330.

*[8] D. G. Kendall, "Some problems in the theory of queues," *J. Roy. Stat. Soc.* (B) Vol. 13 (1951), pp. 151–173 and pp. 184–185. (This contains a general review of the subject and a detailed bibliography. Since the latter was compiled, a number of other papers have been noted, these have been added to the present list and are indicated by an asterisk.)

*[9] D. G. Kendall, "Les processus stochastiques de croissance en biologie," *Ann. Inst. H. Poincaré.*, Vol. 13 (1952), pp. 43–108.

[10] A. Khintchine, "Mathematisches über die Erwartung vor einem öffentlichen Schalter," *Rec. Math.*, Vol. 39 (1932), pp. 73–84. (Russian, with German summary.)

[11] A. Kolmogorov, "Sur le problème d'attente," *Rec. Math.*, Vol. 38 (1931), pp. 101–106.

*[12] D. V. Lindley, "The theory of queues with a single server," *Proc. Cambridge Philos. Soc.*, Vol. 48 (1952), pp. 277–289.

[13] E. C. Molina, "Application of the theory of probability to telephone trunking problems," *Bell System Tech. J.*, Vol. 6 (1927), pp. 461–494.

[14] F. Pollaczek, "Uber eine Aufgabe der Wahrscheinlichkeitstheorie," *Math. Zeit.*, Vol. 32 (1930), pp. 64–100 and pp. 729–750.

[15] F. Pollaczek, "Uber das Warteproblem," *Math Zeit.*, Vol. 38 (1934), pp. 492–537.

*[16] F. Pollaczek, "Répartition des délais d'attente des avions arrivant à un aéroport qui possède s pistes d'atterrissages," *C. R. Acad. Sci. Paris*, Vol. 232 (1951), pp. 1901–1903 and pp. 2286–2288.

*[17] F. Pollaczek, "Sur la répartition des périodes d'occupation ininterrompue d'un guichet," *C. R. Acad. Sci. Paris*, Vol. 234 (1952), pp. 2042–2044.

*[18] F. Pollaczek, "Fonctions caractéristiques de certaines répartitions définies au moyen de la notion d'ordre," *C. R. Acad. Sci. Paris*, Vol. 234 (1952), pp. 2334–2336.

*[19] W. L. Smith, "On the distribution of queuing times," (to be published).

[20] O. Volberg, "Problème de la queue stationnaire et nonstationnaire," *Doklady Akad. Nauk SSSR*, Vol. 24 (1939), pp. 657–661.

The following additional references were noted after the present paper was written.

*F. Benson, "Further notes on the productivity of machines requiring attention at random intervals," *J. Roy. Stat. Soc.* (B), Vol. 14 (1952), pp. 200–210.

*F. G. Foster, "On the stochastic matrices associated with certain queuing processes," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 355–360.

*G. R. Frost, W. Keister and A. E. Ritchie, "A throwdown machine for telephone traffic studies," *Bell System Tech. J.*, Vol. 32 (1953), pp. 292–359.

*B. D. Greenshields and F. M. Weida, *Statistics, with applications to Highway Traffic Analyses*, Eno Foundation for Traffic Control, Saugatuck, Conn., 1952.

*L. Kosten, "On the accuracy of measurements of probabilities of delay in telecommunication systems," *Appl. Sci. Research (B)*, Vol. 2 (1951–1952), pp. 108–130 and pp. 401–415.

*F. Pollaczek, "Délais d'attente des avions atterrissant selon leur ordre d'arrivée sur un aéroport à s pistes," *C. R. Acad. Sci. Paris*, Vol. 234 (1952), pp. 1246–1248.

*F. Pollaczek, "Sur une généralisation de la théorie des attentes," *C. R. Acad. Sci. Paris*, Vol. 236 (1953), pp. 578–580.

*J. Riordan, "Telephone traffic time-averages," *Bell System Tech. J.*, Vol. 30 (1951), pp. 1129–1144.

*J. Riordan, "Delay curves for calls served at random," *Bell System Tech. J.*, Vol. 32 (1953), pp. 100–119.

*R. I. Wilkinson, "Working curves for delayed exponential calls served in random order," *Bell System Tech. J.*, Vol. 32 (1953), pp. 360–383.