ON TIME DEPENDENT QUEUING PROCESSES

By J. Keilson and A. Kooharian

Sylvania Applied Research Laboratory, Waltham, Massachusetts

1. Introduction. It is well known that the general class of stochastic processes with discrete states in continuous time arising in queuing theory, birth-death processes, etc., can be characterized as Markov processes provided the full set of random variables needed to specify the state of the process is employed. A detailed illustration of this approach is given by Cox in [1][1] for the case of a queue in equilibrium subject to a random (Poisson) arrival distribution and a general service time distribution. Our object in this paper is to initiate a systematic development of this approach in the theory of queues. It turns out that such a development for the time dependent version of the above described queuing problem requires analytical considerations not encountered in the equilibrium case. Similarly the systematic development of this approach for the queue with a general arrival distribution (as well as a general service time distribution) leads to a still different type of mathematical problem (simultaneous Wiener-Hopf integral equations with an analytic side condition) which we intend to report on elsewhere.

One final remark is in order concerning the formulation of the approach and derivation of the governing differential equations carried out in sections 2 and 3. While there is a general similarity between the arguments in these sections and those, for example, in [1], we prefer to give a self contained discussion in order to exhibit how the additional complications arising from the consideration of time dependence can be incorporated in the general approach.

2. Phase space. We assume a Poisson arrival distribution with a mean rate of arrival λ and that the service time x between an admission and completion is specified by an arbitrary probability density $D(x)$.

The state of the entire system (queue and service operation) at time t is specified by the number, m, of people in queue and the elapsed time, x, of the person currently in service. Our phase space Γ, accordingly, will be two dimensional with one discrete dimension consisting of the non-negative integers (queue lengths) and one continuous dimension consisting of the positive reals x. The state of the system is then characterized by a point in Γ. For completeness there should be a single additional point in Γ corresponding to the state of total vacancy of the system.

We can now introduce the probability density $W_m(x, t)$ on Γ for the probability that at time t the queue length, excluding servee, is m and the elapsed time in service is x. It is worth emphasizing that the characterization of the state of the system by means of the set of probability densities $W_m(x, t)$ is

Received July 6, 1959; revised August 27, 1959.

[1] We are indebted to the referee for bringing Cox's work to our attention.

well defined *independently* of the queuing discipline. Thus whereas the queue length m is essential in our characterization of the state of the system, nothing whatsoever is implied with respect to the discipline governing selection from the queue for servicing (e.g. first come first served, completely random, etc.). It can be shown, nevertheless, that queue-discipline dependent aspects of the system such as waiting time distributions are deducible from the $W_m(x, t)$ when the queuing discipline is of the first come first served or random selection type (see [1], for example).

3. Analysis. The derivation of the difference-differential equations for the $W_m(x, t)$ employs elementary continuity arguments concerning the motion of the system in Γ. Consideration of the continuity of the flow during a time interval $(t, t + \Delta)$ leads to the equations

$$\begin{aligned} &W_m(x + \Delta, t + \Delta) \\ &\quad = W_m(x, t)(1 - \lambda\Delta)(1 - \eta(x)\Delta) + W_{m-1}(x, t)\lambda\Delta, \quad m = 1, 2, \cdots, \end{aligned} \tag{3.1}$$

to first order terms in Δ. Equation (3.1) is the basic relationship connecting the state of the system at a time $t + \Delta$ to those at time t from which the present state is attainable in phase space by the occurrence or nonoccurrence of arrivals and departures in the interval Δ. The interpretation of $\eta(x)$ in (3.1), accordingly, is similar to that of λ; i.e., $\eta(x)\Delta$ is the first order probability that a service completion occurs in the interval $(x, x + \Delta)$ conditioned on the system having reached the state x. The relationship of $\eta(x)$ to $D(x)$ is given by

$$D(x) = \eta(x) \exp\left\{-\int_0^x \eta(y)\, dy\right\}. \tag{3.2}$$

Rearranging terms in (3.1), dividing by Δ and taking the limit as $\Delta \to 0$ in (3.1), we obtain, for $m = 1, 2, \cdots$,

$$\frac{\partial W_m}{\partial t} + \frac{\partial W_m}{\partial x} + [\lambda + \eta(x)]W_m = \lambda W_{m-1} \tag{3.3}$$

as the governing partial differential equations in the interior of Γ. For $m = 0$ we similarly find

$$\frac{\partial W_0}{\partial t} + \frac{\partial W_0}{\partial x} + [\lambda + \eta(x)]W_0 = 0. \tag{3.4}$$

One additional probability, $E(t)$, that describing the completely vacant state of the system, must be considered. Again a continuity argument similar to that leading to (3.3) yields

$$\frac{dE}{dt} + \lambda E(t) = \int_0^\infty \eta(x)W_0(x, t)\, dx. \tag{3.5}$$

In order to complete our mathematical description, it is necessary to specify

(a) some initial state $\{W_m(x, 0); m = 0, 1, 2, \cdots\}$ and $E(0)$ from which the system starts at $t = 0$, and

(b) the boundary conditions on the boundaries of Γ. ((3.3) and (3.4) are *partial* differential equations).

As will become clear in the subsequent analysis, the solution of the set of equations (3.3), (3.4) together with appropriate boundary conditions will have a linear dependence on the initial conditions. It is no restriction, therefore, to consider the problem for initial states of the form

$$W_m(x, 0) = \delta_{mN}\delta(x - x_0), \qquad m = 0, 1, 2, \cdots, \tag{3.6}$$

where δ_{mN} is the ordinary Kronecker's delta and $\delta(x - x_0)$ the "delta function." In words (3.6) corresponds to starting the system with a specific queue length N and elapsed service time x_0.

The derivation of the boundary conditions, on the other hand, requires a consideration of the motion of the system in Γ when a completion occurs. If the system is in the state $(x, m + 1)$ and experiences a completion, it drops directly into the state $(0, m)$. Let us consider, therefore, the quantity

$$P_m(t) = \int_0^\Delta W_m(x, t)\, dx, \tag{3.7}$$

which is the probability that the system be located at time t in the set of states S_m: $(0, m)$ to (Δ, m). If we restrict ourselves to single transitions in the time interval $(t, t + \Delta)$, the following considerations determine $P_m(t + \Delta)$:

(a) A system in the state $(x, m + 1)$ at time t may experience a completion so that at $t + \Delta$ it lies in S_m,

(b) A system at (x, m) for $x > 0$ at time t, on the other hand, cannot lie in S_m at $t + \Delta$,

(c) Similarly systems at $(x, m - 1)$ at time t cannot lie in S_m at $t + \Delta$, since x is unaffected by arrivals.

Hence continuity requires that to the first order in Δ

$$P_m(t + \Delta) = \Delta \int_0^\infty W_{m+1}(x, t)\eta(x)\, dx, \qquad m = 1, 2, \cdots.$$

Expanding $P_m(t + \Delta)$ and keeping only first order terms in Δ, we obtain the boundary condition

$$W_m(0, t) = \int_0^\infty W_{m+1}(x, t)\eta(x)\, dx, \qquad m = 1, 2, \cdots. \tag{3.8a}$$

The previous argument must be modified for $m = 0$, since a system lying in the empty state which experiences an arrival in $(t, t + \Delta)$ also finds itself in S_0 at time $t + \Delta$. In this case we obtain

$$W_0(0, t) = \int_0^\infty W_1(x, t)\eta(x)\, dx + \lambda E(t). \tag{3.8b}$$

The set of equations (3.3–3.6) and conditions (3.8) then provide a complete description of the queuing problem posed above. It is to be observed that if the equations of motion (3.3–3.6) are integrated over all x and summed over m

taking account of the boundary conditions (3.8), it follows that

$$\frac{d}{dt}\left(E(t) + \sum_{m=0}^{\infty} \int_0^{\infty} W_m(x, t)\, dx\right) = 0,$$

which expresses the conservation of probability.

In order to facilitate the analysis of this system we introduce the generating function

$$G(s, x, t) = \sum_{m=0}^{\infty} s^m W_m(x, t). \tag{3.9}$$

In terms of $G(s, x, t)$, (3.3) and (3.4) condense into

$$\frac{\partial G}{\partial t} + \frac{\partial G}{\partial x} + [\lambda + \eta(x)]G = \lambda s G, \tag{3.10}$$

(3.5) becomes

$$\frac{dE}{dt} + \lambda E(t) = \int_0^{\infty} \eta(x) G(0, x, t)\, dx, \tag{3.11}$$

while the boundary conditions (3.8) combine into

$$sG(s, 0, t) = \int_0^{\infty} \eta(x) G(s, x, t)\, dx + \lambda s E(t) - \int_0^{\infty} \eta(x) G(0, x, t)\, dx. \tag{3.12}$$

Our first step in the analysis of the system (3.10)–(3.12) is to make the substitution

$$G(s, x, t) = H(s, x, t) e^{-N(x)} \tag{3.13}$$

in (3.10), where we define $N(x) = \int_0^x \eta(y)\, dy$. (3.10) then reduces to

$$\frac{\partial H}{\partial t} + \frac{\partial H}{\partial x} + \lambda(1 - s)H = 0, \tag{3.14}$$

which has the general solution

$$H(s, x, t) = H_0(s, t - x) e^{-\lambda(1-s)x}. \tag{3.15}$$

The addition of (3.11) and (3.12), and the use of (3.13) and (3.15) leads to

$$\frac{dE}{dt} + \lambda E(t) + s H_0(s, t) = \int_0^{\infty} D(x) H_0(s, t - x) e^{-\lambda(1-s)x}\, dx + \lambda s E(t), \tag{3.16}$$

where (3.2) has been used in the integrand. The problem has thus been reduced to the determination of $H_0(s, t)$ and $E(t)$ for $t > 0$ with only a single integro-differential equation, (3.16), available. Actually there is a second distinguishing fact about H_0 deriving from its relationship to G, (3.13), which is required to possess the analytical structure of a generating function. This latter fact leads to the analyticity condition discussed in Section 4.

The unknowns in (3.16), $H_0(s, t)$ and $E(t)$, differ significantly with respect

to their dependence on t; namely $E(t)$ has no meaning for $t < 0$ whereas $H_0(s, t)$ does. In fact from (3.13) and (3.15) we have

$$G(s, x, 0) = H_0(s, -x)e^{-N(x)}e^{-\lambda(1-s)x} \tag{3.17}$$

for $x \geqq 0$. Thus for negative values of t, $H_0(s, t)$ is known explicitly in terms of that part of the initial conditions corresponding to

$$G(s, x, 0) = \sum_{m=0}^{\infty} s^m W_m(x, 0).$$

Using the specific choice of $\{W_m(x, 0)\}$ in (3.6), we obtain for $x > 0$

$$H_0(s, -x) = e^{N(x)}e^{\lambda(1-s)x}s^N \delta(x - x_0). \tag{3.18}$$

In so far as the analysis of (3.16) is concerned, the decomposition of $H_0(s, t)$ into its known and unknown parts splits the integral into the sum of a known inhomogeneous term and a convolution integral involving $H_0(s, t)$ for $t > 0$ only.

For simplicity we shall continue the analysis for the special case of the system starting from the completely unoccupied state, i.e., $G(s, x, 0) = 0$ and $E(0) = 1$. In this case the inhomogeneous term vanishes so that (3.16) becomes

$$\frac{dE}{dt} + \lambda(1 - s)E(t) + sH_0(s, t) = \int_0^t D(x)H_0(s, t - x)e^{-\lambda(1-s)x}\,dx. \tag{3.19}$$

If we take the Laplace transform of (3.19), and adopt the notation of using lower case letters for the Laplace transforms of capital lettered functions, we find

$$[p + \lambda\{1 - s\}]e(p) = [d(p + \lambda\{1 - s\}) - s]h_0(s, p) + 1, \tag{3.20}$$

or equivalently,

$$h_0(s, p) = \frac{[p + \lambda\{1 - s\}]e(p) - 1}{d(p + \lambda\{1 - s\}) - s}. \tag{3.21}$$

4. The analyticity condition. Since we may properly restrict ourselves to that class of possible solutions $G(s, x, t)$ which are $L_1(0, \infty)$ in x for $0 \leqq s \leqq 1$ and $t \geqq 0$, it follows that $h_0(s, p)$ must be analytic in the right half plane $\mathrm{Re}(p) > 0$. If we consider the possible singularities in h_0 arising from the roots of the denominator in (3.21), i.e. the set of points p_s satisfying

$$s = d(p_s + \lambda\{1 - s\}), \qquad 0 \leqq s \leqq 1, \tag{4.1}$$

it is possible to show that there is a continuum of roots—the positive real p axis—in the right half plane. In view of the preceding remark, therefore, $e(p)$ must be chosen so that the numerator of (3.21) cancels these roots of denominator. This argument specifies the values of $e(p)$ on the positive real axis which together with the fact that $e(p)$ must be analytic in $\mathrm{Re}(p) > 0$ serves to uniquely determine e by analytic continuation in $\mathrm{Re}(p) > 0$.

In order to give an explicit representation of $e(p)$, it is necessary to determine under what conditions the function p_s, defining the locus of roots of (4.1) as s runs from 0 to 1, has an inverse s_p. That is we seek the conditions under which there exists a solution s_p of the functional equation

$$s_p = d(p + \lambda\{1 - s_p\}). \tag{4.2}$$

This equation, interestingly enough, has previously arisen in the study of a rather distinct problem in the *equilibrium* theory of queues; namely, the study of the distribution of occupation times of the server ([2], [3]). The function s_p, when it exists, is actually the Laplace transform of this distribution. We give an analysis of the significance of this identification with respect to the time dependent theory in the appendix. In [2] theorem 6, it is shown that there is a unique analytic solution of (4.2) under the condition $\lambda/\eta < 1$, where

$$\eta = 1/\int_0^\infty xD(x)\,dx$$

is the mean rate of service. This is the familar stability condition in queuing theory. Under this condition $e(p)$ can be explicitly given as

$$e(p) = \frac{1}{p + \lambda\{1 - s_p\}}. \tag{4.3}$$

In general s_p and, therefore, $e(p)$ will require branch cuts in the p plane in order to be well defined for $\mathrm{Re}(p) < 0$. We shall illustrate the nature of the situation by considering a specific case in Section 6.

An alternative integral expression for $E(t)$ may be obtained by utilizing the transformation of variables suggested by (4.2) itself. Indeed using the transformation

$$u = p + \lambda(1 - s_p) \tag{4.4}$$

in the usual inversion formula for $e(p)$ yields

$$E(t) = \frac{1}{2\pi i}\int \frac{e^{[\lambda d(u)+u-\lambda]t}(\lambda d'(u) + 1)}{u}\,du, \tag{4.5}$$

where it is easily shown that the contour in the u plane may be taken to be the imaginary u-axis indented to the right of the origin.

5. Steady state limit. The state densities for the queue under discussion in the equilibrium case are well known [1], [2]. Our object here is to show how easily these results follow from the above expressions for the time dependent solution. Indeed, by standard Tauberian arguments

$$\lim_{t\to\infty} E(t) = \lim_{p\to 0+} pe(p) = \lim_{p\to 0+} \frac{p}{p + \lambda\{1 - s_p\}} = \frac{1}{1 - \lambda s_0'(0)} = 1 - \frac{\lambda}{\eta}, \tag{5.1}$$

and

$$\lim_{t\to\infty} H_0(s, t) = \lim_{p\to 0+} ph_0(s, p) = \frac{\lambda(1 - s)\left(1 - \frac{\lambda}{\eta}\right)}{d(\lambda\{1 - s\}) - s}, \tag{5.2}$$

where

$$s_p'(0) \equiv \left(\frac{ds_p}{dp}\right)_{p=0} = \frac{1}{\lambda - \eta}.$$

6. The Poisson/Poisson time dependent queue. When the service time distribution is also exponential with mean service time η then

$$D(x) = \eta e^{-\eta x}, \tag{6.1}$$

so that

$$d(p) = \frac{\eta}{\eta + p}. \tag{6.2}$$

The functional equation (4.2), accordingly, becomes

$$\frac{\eta}{\eta + p + \lambda(1 - s_p)} = s_p. \tag{6.3}$$

Solving for s_p yields

$$s_p = \frac{(p + \lambda + \eta) \pm [(p + \lambda + \eta)^2 - 4\lambda\eta]^{1/2}}{2\lambda}. \tag{6.4}$$

Rewriting the expression within the square brackets in the form

$$\{(p + \lambda + \eta) - 2(\lambda\eta)^{1/2}\}\{(p + \lambda + \eta) + 2(\lambda\eta)^{1/2}\} \tag{6.5}$$

shows that s_p has branch points at

$$p = -(\lambda + \eta) \pm 2(\lambda\eta)^{1/2} = -(\sqrt{\lambda} \pm \sqrt{\eta})^2. \tag{6.6}$$

The branch points are thus seen to lie on the negative p axis. By (4.3) and (6.4)

$$e(p) = \frac{-(p + \lambda - \eta) + [\{p + (\sqrt{\lambda} - \sqrt{\eta})^2\}\{p + (\sqrt{\lambda} + \sqrt{\eta})^2\}]^{1/2}}{2\eta p}. \tag{6.7}$$

The branch of $e(p)$ corresponding to the choice of $+$ sign in (6.7) is required to insure the vanishing of $e(p)$ as $p \to +\infty$.

The inversion of $e(p)$ can now be carried out. The contour we choose is indicated in Fig. 1 below where a finite branch cut has been made between the branch points p_1, p_2 as given by (6.6). Taking account of the simple pole at $p = 0$ as well as the branch cut, we obtain

$$E(t) = 1 - \frac{\lambda}{\eta} - \frac{1}{2\pi i}\int_{c_2} e(p)e^{pt}dp. \tag{6.8}$$

We point out that for $p \in C_2$, $\mathrm{Re}(p) < 0$ so that the second term in (6.8) represents transient behavior. The integral appearing in (6.8) can be simplified leading to

$$E(t) = 1 - \frac{\lambda}{\eta} + \frac{1}{\pi}\int_{u_1}^{u_2} \frac{[(u - u_1)(u_2 - u)]^{1/2}e^{-ut}}{2\eta u}\,du, \tag{6.9}$$

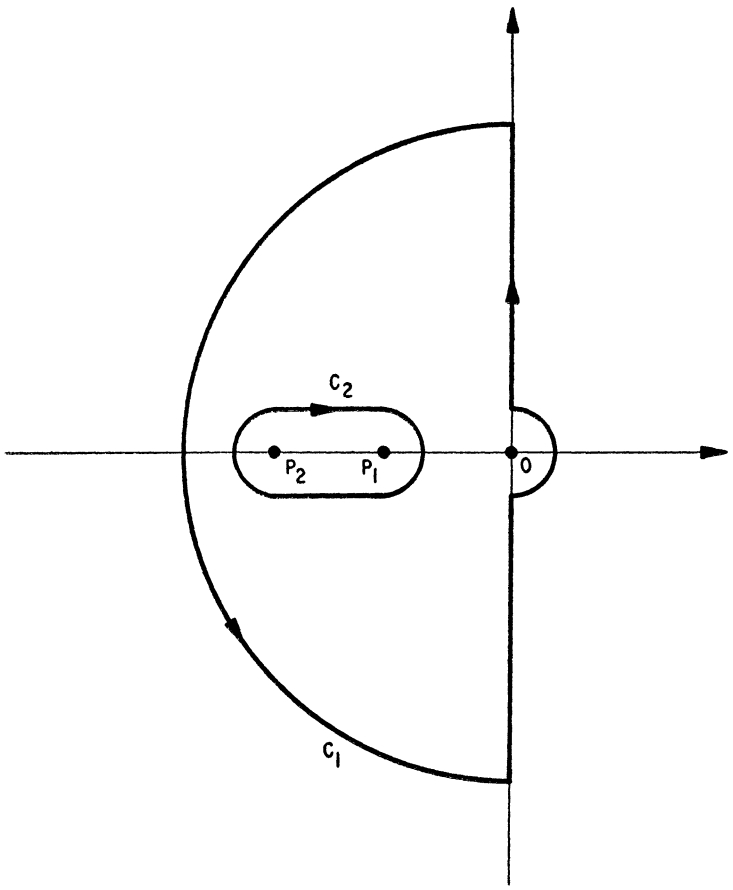

FIG. 1

where $u_1 = -p_1$, $u_2 = -p_2$. We do not pursue the details of this solution further since this case has already been discussed by Morse [5] using a completely different approach. It is straightforward to bring (6.9) into the form obtained by Morse.

APPENDIX

By occupation times we mean the time intervals between the state of complete vacancy. The distribution of these occupation times is well known [2]. It may, however, be obtained independently from the time dependent formalism developed in the text in the following way. At $t = 0$, the queue is started in the state $m = 0$, $x = 0$. Let $J(t)$ be the probability at time t that the system has emptied. Then $dJ/dt = W_J(t)$ is the probability density function for the occupation times. Let $U_m(x, t)$ be the p.d.f. for the state (m, x) at time t conditioned on the system's *not* having emptied, and let $G(s, x, t) = \sum_0 s^m U_m(x, t)$.

It is clear that

$$W_J(t) = \frac{dJ}{dt} = \int_0^\infty \eta(x) G(0, x, t)\, dx. \tag{1}$$

The boundary conditions $U_m(0, t) = \int_0^\infty \eta(x) U_{m+1}(x, t)\, dx$ for all m imply

$$G(s, 0, t) = \int \eta(x) \frac{G(s, x, t) - G(0, x, t)}{s}\, dx, \tag{2}$$

and, as before, $G(s, 0, t - x)$ obeys Eq. (3.10). Thus

$$\overset{\circ}{G}(s, x, t) = G(s, 0, t - x) \exp\{-\lambda(1 - s)x - N(x)\} + \delta(x - t) \exp\{-\lambda(1 - s)x - N(x)\}, \tag{3}$$

where $\delta(x)$ is the delta function, and $G(s, 0, t)$ is zero for negative t. If one substitutes (3) into (2) and takes the Laplace transform, $\tilde{G}(s, 0, p)$ is determined by the analyticity condition of Section 4.

From (1) and (3) we then have

$$\int_0^\infty e^{-pt} W_J(t)\, dt = s_p, \tag{4}$$

where s_p is defined by Eq. (4.2).

REFERENCES

[1] D. R. Cox, "The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables," *Proc. Camb. Phil. Soc.*, Vol. 51 (1955), pp. 433–441.

[2] L. Takacs, "Investigation of waiting time problems by reduction to Markov processes," *Acta Math. Acad. Sci. Hung.*, Vol. 6 (1955), pp. 101–129.

[3] D. G. Kendall, "Some problems in the theory of queues," *J. Roy. Stat. Soc., Ser. B*, Vol. 13 (1951), pp. 151–185.

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

[5] P. M. Morse, "Stochastic properties of waiting lines," *J. Opns. Res. Soc. Amer.* Vol. 3 (1955), pp. 255–261.