A $GI/M/1$ QUEUE WITH A MODIFIED SERVICE MECHANISM

A. G. PAKES

(Received Jan. 20, 1971; revised Sept. 20, 1971)

1. Introduction

In recent years there has appeared some work on queueing systems in which the service time distributions depend on the queue length. Such models may be expected to represent situations where the server's working rate is determined by his observations of the queue length. For example in most practical situations a server will work faster if the queue becomes long and conversely, he will be inclined to work rather more slowly if the queue is very short.

It appears that nearly all the work on systems of this nature has been devoted to the case of Poisson arrival processes. In this paper we shall consider the simplest case of a general input process. Specifically, we shall assume that customers arrive at a single counter according to a renewal process, and all service times are exponentially distributed, and service is in order of arrival. We consider two types of service discipline.

D_1: If a customer initiates a busy period his service rate is μ_1, and the service rate of all other customers in the busy period is μ.

D_2: If the queue length at the initiation of service of a customer $\leqq N$, a positive integer, the service rate is μ_1 and is μ otherwise.

In Section 2 of this paper we shall consider the $GI/M/1$ system with the discipline D_1. It is clear that the queue lengths at arrival epochs do not form a Markov chain since, in general, we lack knowledge of the server's working rate. We shall see that by introducing an appropriate supplementary variable a Markov chain is obtained, but to simplify the algebra we shall restrict ourselves to the calculation of limiting distributions. However, in Section 3 we shall briefly consider transient distributions and in Section 4 we show how the methods of Section 2 extend to cope with the discipline D_2. For the standard $GI/M/1$ system ($\mu=\mu_1$) our results yield some interesting facts about the limiting queue length distribution. In Section 5 we show that the methods of Section 2 yield results for the $GI/M/1$ version of Finch's model [2].

2. The discipline D_1

Assume that the customers arrive according to a renewal process where $A(\cdot)$ denotes the inter-arrival time D.F. of successive customers and T_n ($n=0, 1, \cdots$; $T_0=0$) denotes the epoch of arrival of the nth customer, C_n. Let Q_n be the queue length at epoch T_n-0. It is well known (e.g. [3]) that for the standard $GI/M/1$ system, $\{Q_n\}$ defines a Markov chain. However, for the discipline D_1 this is no longer true since it is not known at what rate the server works at (assuming he is).

Accordingly, when $Q_n>0$ define the random variable σ_n specifying the state of the server at T_n-0. Thus $\sigma_n=1$ if at T_n-0 service is in progress on the customer who initiated the current busy period, and $\sigma_n=2$ if any other customer is receiving service at T_n-0.

We consider a discrete time stochastic process which moves on the phase space $\bigcup_{i=1}^{\infty} \bigcup_{r=1}^{2} \{(i, r)\} \cup \{(0)\}$. State (i, r) is occupied at time n ($n=0, 1, \cdots$) if $Q_n=i$ and $\sigma_n=r$, and (0) is occupied at time n if $Q_n=0$. Assuming that the sequences of service times at rate μ_1 and service times at rate μ are both independent sequences of random variables, it is clear that the process defined above has the Markov property. In this section we shall obtain the limiting-stationary distribution of this Markov chain and thus the limiting waiting time distribution.

For $n=0, 1, \cdots$; $i, j=1, 2, \cdots$ and $r, s=1, 2$, let

$$p(i, r: j, s)=\Pr\{Q_{n+1}=j, \sigma_{n+1}=s \mid Q_n=i, \sigma_n=r\} ,$$

$$p(i, r: 0)=\Pr\{Q_{n+1}=0 \mid Q_n=i, \sigma_n=r\} ,$$

$$p(0: j, s)=\Pr\{Q_{n+1}=j, \sigma_{n+1}=s \mid Q_n=0\} ,$$

$$p(0: 0)=\Pr\{Q_{n+1}=0 \mid Q_n=0\} ,$$

$$\pi(j, s)=\lim_{n\to\infty} \Pr\{Q_n=j, \sigma_n=s\}$$

and

$$\pi(0)=\lim_{n\to\infty} \Pr\{Q_n=0\} .$$

Fig. 1 shows the allowed single step transitions.

Using the correspondence $(0)\leftrightarrow 0$, $(j, 1)\leftrightarrow 2j-1$ and $(j, 2)\leftrightarrow 2j$ ($j=1, 2, \cdots$), the states of the above Markov chain can be identified with those of a Markov chain having the non-negative integers as state-space. This latter Markov chain can be seen to be irreducible and aperiodic, hence if a stationary distribution can be shown to exist for the above Markov chain, it will be unique and given by the limits above, which are now seen to exist.

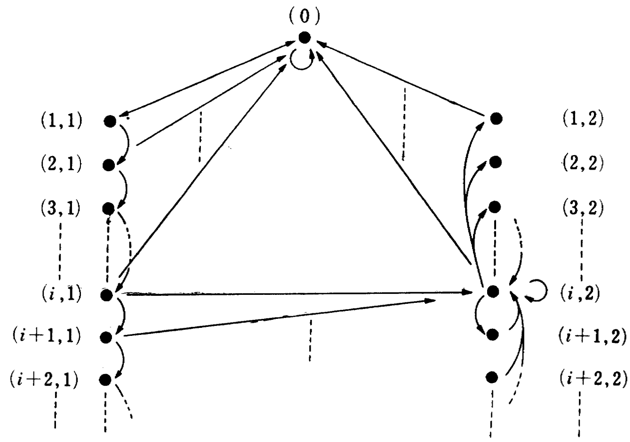

Fig. 1. Diagram of possible one step-transitions for the $GI/M/1$ system under discipline D_1.

Figure 1 shows that the stationary distribution must satisfy

(1) $$\pi(0)=\pi(0)p(0, 0)+\sum_{i,r\geqq 1}\pi(i, r)p(i, r: 0)$$

(2) $$\pi(j, 1)=\pi(j-1, 1)p(j-1, 1: j, 1) \qquad (j=2, 3,\cdots)$$

(3) $$\pi(1, 1)=\pi(0)p(0: 1, 1)$$

(4) $$\pi(j, 2)=\sum_{i=j}^{\infty}\pi(i, 1)p(i, 1: j, 2)+\sum_{i=j-1}^{\infty}\pi(i, 2)p(i, 2: j, 2) \qquad (j=2, 3,\cdots)$$

and

(5) $$\pi(1, 2)=\sum_{i=1}^{\infty}[\pi(i, 1)p(i, 1: 1, 2)+\pi(i, 2)p(i, 2: 1, 2)] .$$

It is easily seen that $p(0: 1, 1)$ and $p(j-1, 1: j, 1)$ $(j=2, 3,\cdots)$ equal the probability that no customer is dismissed during an inter-arrival period when the server is working at rate μ_1 and this is given by $a\equiv a(\mu_1)=\int_0^{\infty}e^{-\mu_1 x}dA(x)$ and (2) and (3) yield

(6) $$\pi(j, 1)=\pi(0)a^j \qquad j=1, 2,\cdots .$$

Clearly $p(i, 2: j, 2)$ $(i, j=1, 2,\cdots)$ is the probability that there are $i+1-j$ departures during an inter-arrival period when the server works at rate μ for all the services and so $p(i, 2: j, 2)=b_{i+1-j}$ $(j=1,\cdots, i+1;$ $i=1, 2,\cdots)$ and is zero otherwise, where

$$\beta(z)=\sum_{i=0}^{\infty}b_i z^i=a[\mu(1-z)] \qquad (|z|\leqq 1) .$$

Now, $p(i, 1; j, 2)$ is the probability that during an inter-arrival period

$i+1-j$ services are completed, the first at rate μ_1 and the remainder at rate μ. Define a modified renewal process by a sequence of independent events, the first having the lifetime D.F. $1-e^{-\mu_1 x}$ and the remainder having the lifetime D.F. $=1-e^{-\mu x}$. Letting $N(x)$ be the counting process for this renewal process, we have

$$p(i, 1: j, 2)=\int_0^\infty \Pr\{N(x)=i+1-j\}\, dA(x) \equiv c_{i+1-j}$$

for $j=1,\cdots, i+1$; $i=1, 2,\cdots$, and zero otherwise. On using equation (5) p. 38 of [1] and inverting the resulting Laplace transform we have

$$\gamma(z)=[(\mu-\mu_1)(1-z)a-\mu_1 z\beta(z)]/[\mu(1-z)-\mu_1] , \tag{7}$$

$$\gamma_1(z)=\mu_1 z(a-\beta(z))/[\mu(1-z)-\mu_1] \tag{8}$$

where $\gamma(z)=\gamma_1(z)+c_0=\sum_{j=0}^{\infty} c_j z^j$.

Denoting the first sum on the right of (4) and (5) by χ_j ($j=1, 2,\cdots$), equation (6) yields

$$\chi_j=\pi(0)a^{j-1}\gamma_1(a) .$$

Let $1/\rho=-\mu a'(0)$. It is well known (e.g. [3]) that if $\rho<1$ there exists $0<\delta<1$ uniquely satisfying $z=\beta(z)$ in the open unit disc. We shall now show that a solution of the form

$$\pi(j, 2)=\pi(0)(Ca^j+D\delta^j)$$

satisfies equations (1)-(5). Substituting this expression into (4) and (5) yields

$$Ca+D\delta=\gamma_1(a)+C(\beta(a)-b_0)+D(\beta(\delta)-b_0)$$

and

$$Ca^j+D\delta^j=a^{j-1}\gamma_1(a)+Ca^{j-1}\beta(a)+D\delta^j \qquad (j=2, 3,\cdots) \tag{9}$$

which combine to give

$$C=\mu_1 a/[\mu(1-a)-\mu_1] , \tag{10}$$

which, with (9) and on noting that $b_0>0$, yields $D=-C$. Note that these manipulations are only valid if $a\neq\delta$, that is, if $\mu(1-a)-\mu_1\neq 0$. This exceptional case will be dealt with below.

Hence equations (2)-(5) are satisfied by solutions of the form

$$\pi(j, 1)=\pi(0)a^j \quad \text{and} \quad \pi(j, 2)=\pi(0)C(a^j-\delta^j) \qquad (j=1, 2,\cdots;\ a\neq\delta) . \tag{11}$$

Noting that the probabilities must sum to unity, we obtain

(12) $$\pi(0)=(1-\delta)[\mu(1-a)-\mu_1]/[(1-\delta)(\mu-\mu_1)-\mu_1 a] \qquad (a\neq\delta)\,.$$

On observing that $p(0:0)=1-a$, $p(i,1:0)=1-\sum_{k=0}^{i} c_k$ and $p(i,2:0)=1-\sum_{k=0}^{i} b_k$, it is easy to check that (1) is satisfied by (11) and (12).

It only remains to prove the positivity of $\pi(0)$ and $\pi(i,r)$. To show that this inequality holds when $r=2$, consider the denominator of the right hand side of (10) as a function, $f(\cdot)$, of μ_1. We have

$$f(\mu_1)=\mu\left[1-\int_0^\infty e^{-\mu_1 x}dA(x)\right]-\mu_1\,.$$

Clearly $f(0)=0$ and $f(\mu(1-\delta))=0$, since in this latter case $a=\delta$. On noting that $f'(0+)=\rho^{-1}>0$, $f'(\mu_1)\to -1$ $(\mu_1\to\infty)$ and $f''(\mu_1)>0$ $(\mu_1>0)$ we have $f(\mu_1)>0$ (resp. <0) if $a>\delta$ (resp. $a<\delta$). Hence $\pi(0)$ and $\pi(j,r)$ are either all positive or all negative and since their sum is unity, the former is the case.

Let $\pi(j)=\pi(j,1)+\pi(j,2)$, the limiting probability of an arriving customer finding j customers ahead of him. Then

$$\pi(j)=\frac{(1-\delta)(\mu_1-\mu)(1-a)a^j+a\mu_1(1-\delta)\delta^j}{(1-\delta)(\mu_1-\mu)+\mu_1 a}\,.$$

When $\mu_1>\mu$ the queue length distribution is a mixture of two geometric distributions.

Putting $\mu_1=\mu$ in (11) and (12) yields $\pi(j,1)=(1-\delta)a^j$ and $\pi(j,2)=(1-\delta)(\delta^j-a^j)$ which for a standard $GI/M/1$ system gives, for example, the limiting probability that an arriving customer finds j others in front of him and that the customer receiving service began the busy period. If π_r $(r=1,2)$ is the limiting probability that $\sigma_n=r$, we have

$$\pi_1=a\pi(0)/(1-a)\,,\quad =a(1-\delta)/(1-a) \qquad \text{if } \mu=\mu_1$$

$$\pi_2=C\pi(0)(a-\delta)/[(1-a)(1-\delta)]\,,\quad =(\delta-a)/(1-a) \qquad \text{if } \mu=\mu_1\,.$$

Let $\pi(j\,|\,r)$ $(j=1,2,\cdots;\ r=1,2)$ be the limiting probability of an arrival finding j customers ahead of him conditioned on the state of the server. We have

$$\pi(j\,|\,1)=(1-a)a^{j-1}\,,\quad \pi(j\,|\,2)=(1-a)(1-\delta)(a^j-\delta^j)/(a-\delta) \qquad (a\neq\delta)\,.$$

Thus we see that given that the customer receiving service began the current busy period, the queue lengths have a geometric distribution on $\{j=1,2,\cdots\}$ and given that the customer being served did not begin the busy period, the queue length distribution is a convolution of two geometric distributions.

Still assuming stationarity, we have

$$E(Q_n)=\pi(0)[a+C(a-\delta)(1-a\delta)/(1-\delta)^2]/(1-a)^2 .$$

If $a=\delta$, use of L'Hospital's rule on the expressions above, or direct substitution, shows that

$$\pi(0)=(1-\delta)(1-\beta'(\delta))/(1-(1-\delta)\beta'(\delta)) , \qquad \pi(j, 1)=\pi(0)\delta^j ,$$

$$\pi(j, 2)=(1-\delta)^2\beta'(\delta)j\delta^{j-1}/(1-(1-\delta)\beta'(\delta)) ,$$

$$\pi(j, 1)=(1-\delta)\delta^{j-1} , \qquad \pi(j, 2)=(1-\delta)^2 j\delta^{j-1}$$

and

$$E(Q_n)=[\delta+\beta'(\delta)]/(1-\delta)(1-(1-\delta)\beta'(\delta)) .$$

If $g(\theta)$ is the L.S.T. of the limiting waiting time D.F. of an arriving customer, then a standard argument shows that (e.g. [3], p. 121)

$$g(\theta)=\begin{cases} \pi(0)\left[1+\dfrac{a\mu_1[\mu_1-\mu(1-\delta)](\mu+\theta)}{(\mu_1-\mu(1-a))(\mu(1-\delta)+\theta)(\mu_1+\theta)}\right] & (a\neq\delta) \\ \pi(0)\left[1+\dfrac{\mu\delta(1-\delta)(\mu+\theta)}{(1-\beta'(\delta))(\mu(1-\delta)+\theta)^2}\right] & (a=\delta) . \end{cases}$$

3. The transient distribution

For the Markov chain discussed in the last section denote the n-step transition probability from state $(\cdot)$ to state $(*)$ by $p^{(n)}(\cdot:*)$. The task of finding generating functions of these quantities is tedious and the resulting expressions are complicated. However, for the sake of completeness we shall outline the derivation of the generating function for $p^{(n)}(\cdot:0)$.

The transition diagram (Fig. 1) shows that the relevant backwards Chapman-Kolmogorov equations take the form

$$p^{(n+1)}(0:0)=ap^{(n)}(1, 1:0)+(1-a)p^{(n)}(0:0) ,$$

$$p^{(n+1)}(i, 1:0)=ap^{(n)}(i+1, 1:0)+\sum_{k=1}^{i+1} c_{i+1-k}p^{(n)}(k, 2:0) + \left(1-\sum_{k=0}^{i} c_k\right)p^{(n)}(0:0)$$

and

$$p^{(n+1)}(i, 2:0)=\sum_{k=1}^{i+1} b_{i+1-k}p^{(n)}(k, 2:0)+\left(1-\sum_{k=0}^{i} b_k\right)p^{(n)}(0:0) .$$

For $|z|, |w|<1$ let

$$P_r(z, w)=\sum_{n=0}^{\infty}\sum_{i=1}^{\infty} p^{(n)}(i, r:0)z^i w^n \qquad (r=1, 2) ,$$

$$U_0(w)=\sum_{n=0}^{\infty} p^{(n)}(0:0)w^n \quad \text{and} \quad U_r(w)=\sum_{n=0}^{\infty} p^{(n)}(1, r:0)w^n \qquad (r=1, 2) .$$

Noting that $p^{(0)}(0:0)=1$, the above equations yield

(13) $U_0(w)=[1+awU_1(w)]/[1-w(1-a)]$,

(14) $P_1(z, w)=w[\beta(z)P_2(z, w)-azU_1(w)-c_0zU_2(w)+zA(z)U_0(w)]/(z-aw)$

and

(15) $P_2(z, w)=zw[L(z)U_0(w)-b_0U_2(w)]/(z-w\beta(z))$

where $A(z)=c_0+(z-\gamma(z))/(1-z)$ and $L(z)=b_0+(z-\beta(z))/(1-z)$. Thus $P_r(z, w)$ is expressible in terms of the three functions $U_i(w)$ $(i=0, 1, 2)$. Equation (13) gives one relation between these functions and another two may be found by observing that regularity of $P_r(z, w)$ in $|z|<1$ for each w in $|w|<1$ implies that the numerator of (14) must have a zero at $z=aw$ and that of (15) has a zero at $\delta(w)$, the unique zero in $|z|<1$ of $z-w\beta(z)$ (e.g. see [3]). Thus we can evaluate $P_r(z, w)$.

4. The discipline D_2

Assuming that service times with parameter μ_1 are mutually independent and those with parameter μ are mutually independent for the $GI/M/1$ queue under the discipline D_2, we see that the following stochastic process is a Markov chain which will enable us to determine the probabilities of queue lengths at customer arrival epochs. The process is in state (i) $(i=0, 1, \cdots, N)$ at time n $(n=0, 1, \cdots)$ if C_n arrives to find i customers ahead of him, and it is in state (i, r) $(i=N+1, N+2, \cdots; r=1, 2)$ at time n if C_n arrives to find i customers adead of him and the server is working at rate μ_1 $(r=1)$ or μ $(r=2)$.

This Markov chain may be dealt with in much the same manner as that in Section 2. Letting $\pi(j)$ $(j=0, 1, \cdots, N)$ and $\pi(j, r)$ $(j=N+1, \cdots; r=1, 2)$ be the limiting-stationary probabilities, it can be shown that

$$\pi(j, 1)=\pi(N)a^{j-N} \qquad \text{and} \qquad \pi(j, 2)=\pi(N)C[a^{j-N}-\delta^{j-N}]$$

where a, δ and C are as in Section 2. By working recursively $\pi(j)$ can be found in terms of $\pi(j+1), \cdots, \pi(N)$ $(j=0, \cdots, N-1)$ from the equations

$$\pi(j)=\sum_{i=j-1}^{N} a_{i+1-j}\pi(i)+\sum_{i=N+1}^{\infty} \pi(i, 1)d_{i-N, N+1-j}+\sum_{i=N+1}^{\infty} \pi(i, 2)d_{i+1-N, N-j} \qquad (j=1, \cdots, N)$$

where $a_i=\int_0^{\infty} e^{-\mu_1 x}(\mu_1 x)^i/i!dA(x)$ and d_{ij} is the probability of $i+1$ events in the random interval $(0, T)$ of a modified renewal process where the lifetime distribution of the first event is Erlangian with parameters μ_1 and j and succeeding lifetime distributions are exponential with parameter μ, and D.F. of T is $A(\cdot)$.

If $Q_n=j$ then C_{n-j} is the customer receiving service on C_n's arrival. We have for $j=N+1, N+2, \cdots$

$$P_r(Q_n=j \mid \partial_{n-j} \leqq N)=(1-a)a^{j-N-1}$$

$$P_r(Q_n=j \mid \partial_{n-j} \geqq N+1)=(1-a)(1-\delta)(a^{j-N}-\delta^{j-N})/(a-\delta)$$

where ∂_n is the queue length at C_n's service inception and assuming that the stationary distribution obtains. The first distribution is geometric on $\{N+1, N+2, \cdots\}$ and the second is a convolution of two geometric distributions on this set. When $\mu_1=\mu$ we obtain the following results for the standard $GI/M/1$ system

$$P_r\{Q_n=j,\, \partial_{n-j} \leqq i\} = (1-\delta)\delta^i a^{j-i}$$

$$P_r\{Q_n=j,\, \partial_{n-j} > i\} = (1-\delta)(\delta^j-\delta^i a^{j-i}) \qquad (j=i+1, i+2, \cdots;\ i=1, 2, \cdots)\,.$$

These results complement those near the end of Section 2.

5. A queueing model of Finch

Finch [2] considered a system in which customers arrive according to a renewal process and were served in order of arrival with identically and independently distributed service times. Those customers who arrive to find the counter vacant are delayed before their service begins. These delay periods are identically and independently distributed.

In this section we consider the case of exponentially distributed service and delay times having parameters μ and μ_1, respectively. Finch has shown that a limiting waiting time D.F. exists when $1<\mu \int_0^\infty x dA(x) <\infty$ where $A(\cdot)$ is the D.F. of interarrival times. We obtain the L.S.T. of the limiting waiting time D.F. by first finding the limiting queue size distribution.

This latter task is carried out in substantially the same manner as that of Section 2. We consider a discrete time process having the state space as in Section 2. State (0) is occupied at time n if C_n arrives to find the counter vacant and (j, r) is occupied at n if C_n arrives to find j customers ahead of him and the server idle $(r=1)$ or serving $(r=2)$. Fig. 1 is applicable here, but with the following additional transitions: $(0)\to(1, 2)$ with probability c_1 (notation as in Section 2) and $(j-1, 1)\to (j, 2)$ $(j=2, 3, \cdots)$. The transitions $(i, 1)\to(j, 2)$ have probability c_{i-j+2} $(i=j-1, j, \cdots;\ j=2, 3, \cdots$, and $i=1, 2, \cdots$ when $j=1)$. All other transitions into (j, r) have the same probabilities as in Section 2. Letting $\pi(0)$ and $\pi(j, r)$ be the limiting probabilities, it is not difficult to show that if $a \neq \delta$

$$\pi(0)=(1-\delta)(\mu(1-a)-\mu_1)/(\mu(1-\delta)-\mu_1)$$

$$\pi(j, 1)=\pi(0)a^j \qquad \text{and} \qquad \pi(j, 2)=\pi(0)E(a^j-\delta^j)$$

where $E=\mu_1/(\mu(1-a)-\mu_1)$. These results may be used to calculate quantities of interest as in Section 2. One measure of customer dissatisfaction in this system that may be of interest is the probability, P, that a customer arrives to find others ahead of him and the server idle. This is given by

$$\begin{aligned} P &= a[(1-\delta)/(1-a)][(\mu(1-a)-\mu_1)/(\mu(1-\delta)-\mu_1)] \\ &\cong a(1-\delta) \qquad \text{if } \mu_1 \text{ is large but not } \mu \\ &\cong a \qquad \text{if } \mu \text{ is large but not } \mu_1 . \end{aligned}$$

Thus it would appear that shortening the delay period may please customers more than having shorter service times, although, other factors must of course be considered.

The L.S.T., $g(\theta)$, of the limiting waiting time D.F. may be found in the usual way and is given by

$$g(\theta)=\pi(0)\mu_1(\mu+\theta)/(\mu(1-\delta)+\theta)(\mu_1+\theta) \qquad (a \neq \delta) .$$

MONASH UNIVERSITY, CLAYTON, VICTORIA, AUSTRALIA

REFERENCES

[1] Cox, D. R. (1962). *Renewal Theory*, Methuen, London.

[2] Finch, P. D. (1959). A probability limit theorem with application to a generalisation of queueing theory, *Acta. Math. Acad. Sci. Hung.*, **10**, 317-325.

[3] Takacs, L. (1962). *Introduction to the Theory of Queues*, O.U.P., New York.