Journal of Applied Mathematics and Stochastic Analysis
9, Number 2, 1996, 221-228

RETRIAL QUEUES WITH RECURRENT DEMAND OPTION

K. FARAHMAND and N.H. SMITH
University of Ulster, Jordanstown
Co. Antrim BT37 0QB
United Kingdom

(Received January, 1996; Revised March, 1996)

ABSTRACT

The object of this paper is to analyze the model of a queueing system in which customers can call in only to request service: if the server is free, the customer enters service immediately. Otherwise, if the service system is occupied, the customer joins a source of unsatisfied customers called the orbit. On completion of each service the recipient of service has an option of leaving the system completely with probability $1-p$ or returning to the orbit with probability p. We consider two models characterized by the discipline governing the order of re-requests for service from the orbit. First, all the customers from the orbit apply at a fixed rate. Secondly, customers from the orbit are discouraged and reduce their rate of demand as more customers join the orbit. The arrival at and the demands from the orbit are both assumed to be according to the Poisson process. However, the service times for both primary customers and customers from the orbit are assumed to have a general distribution. We calculate several characteristic quantities of these queueing systems.

Key words: Single Line Queue, Repeated Demands, Orbit Size, Waiting Time, Ergodic State, Generating Function, Retrial Queue, Recurrent Customer.

AMS (MOS) subject classifications: 60M20, 60K25.

1. Introduction

The main significant difference between retrial queues and the usual queueing systems is that with retrial queues the server cannot be in continuous contact with the waiting customers, who can only call in to test the state of the server. If the server is free, service commences immediately. However, if the server is occupied, the customer must break contact and reinitiate his request later. These unsatisfied customers produce a source of customers called the *orbit*. Therefore, the server receives requests from arrivals from outside at a rate λ, and from customers in the orbit at a rate ξ_n, when the orbit size is n, both according to the Poisson process. Previously, the case of a fixed $\xi_n(\equiv \xi)$ has been studied, for example by Keilson and Kooharian [6], Keilson et al. [5] and Falin [2], and refers to telephone call problems. Farahmand [3] and [4] considered the case $\xi_n = \xi/n$ which can be looked upon as discouraged repeated demands, that is, when the customers reduce their rate of repeated demands as more customers join the orbit and, obviously, the competition to find the server idle is higher.

The aim of this paper is to give an option to the completed customer to rejoin the orbit and therefore remain unsatisfied or to leave the system entirely. It is natural for telephone callers to break contact when the line is engaged and reapply for connection later. Therefore, our structure occurs in many communication networks as well as in computer representations and it is of theoretical interest. We assume that upon completion of service a customer with probability p rejoins the orbit. In other words, only a proportion $1-p$ of customers leave the system on completion of service. In sections 2 and 3 we consider the cases of fixed and discouraged repeated demands, respectively. A similar problem in ordinary M/G/1 queues was studied by Boxma and Cohen [1] in which it was assumed that there was a fixed number of permanent customers present who rejoin the queue on their completion of service. This system with permanent customers in the retrial context was studied by Farahmand [4].

The service times x for both the primary customers and the customers from the orbit are assumed to be independent and to have a common probability distribution function $A(x)$. When $A(x)$ is absolutely continuous with probability density $a(x)$ then

$$a(x)=\eta(x)\exp\left\{-\int_0^x \eta(y)dy\right\},$$

where $\eta(x)$ is the conditional completion rate for service at time x.

In order to consider the changes that occur to the system during and after serving a customer, we condition on the event that the server is busy. Therefore, let $W_n(x,t)$ be the joint probability density that there are n customers in the orbit at epoch t and a customer is present in service who has been there for time x. Then, the following equations govern the system:

$$W_n(x+\Delta,t+\Delta)=W_n(x,t)(1-\lambda\Delta)(1-\eta(x)\Delta)$$
$$+W_{n-1}(x,t)\lambda\Delta \quad n\geq 1$$

and

$$W_0(x+\Delta,t+\Delta)=W_0(x,t)(1-\lambda\Delta)(1-\eta(x)\Delta).$$

Therefore, we can show that $G(u,x,t)=\sum_{n=0}^{\infty}u^nW_n(x,t)$, the generating function of $W_n(x,t)$, satisfies the following relation

$$\frac{\partial G}{\partial t}+\frac{\partial G}{\partial x}+\{\lambda+\eta(x)\}G=u\lambda G.$$

Therefore solving the above equation we obtain

$$G(u,x,t)=G(u,0,t-x)\exp\{\lambda ux-\lambda x-N(x)\},$$

where $N(x)=\int_0^x\eta(y)dy$. When the system is assumed to be ergodic with $G_{\infty}(u,x)=\lim_{t\to\infty}G(u,x,t)$ the above relation simplifies to

$$G_{\infty}(u,x)=G_{\infty}(u,0)\exp\{-\lambda x(1-u)-N(x)\}. \tag{1.1}$$

Since the above argument is independent of the discipline of repeated demands of customers from the orbit, we can use equation (1.1) for both our queueing models.

2. Fixed Rate of Repeated Demands

First we consider a case that allows each customer in the orbit to apply for service with a constant rate ξ. Let $p_n(t)$ be the probability that at epoch t, the server is idle and n customers are in the orbit. Since upon completion of service the customer with probability p chooses to rejoin the orbit, the equations governing the system are

$$\frac{dp_n(t)}{dt} = -(\lambda + n\xi)p_n(t) + (1-p)\int_0^\infty W_n(x,t)\eta(x)dx$$
$$+ p\int_0^\infty W_{n-1}(x,t)\eta(x)dx \qquad n \geq 1 \tag{2.1}$$

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + (1-p)\int_0^\infty W_0(x,t)\eta(x)dx \tag{2.2}$$

and

$$W_n(0,t) = \lambda p_n(t) + (n+1)\xi p_{n+1}(t) \qquad n \geq 0. \tag{2.3}$$

Let $\Pi(u,t) = \sum_{n=0}^\infty u^n p_n(t)$ be the generating function of $p_n(t)$. Assuming that the system is ergodic, multiplying (2.1) and (2.2) by u^n and summing it up over n, we obtain

$$(\lambda + \xi u\frac{d}{du})\Pi_\infty(u) = (1-p+pu)\int_0^\infty \eta(x)G_\infty(u,x)dx$$
$$= (1-p+pu)\int_0^\infty \eta(x)G_\infty(u,0)\exp\{-\lambda x(1-u) - N(x)\}dx, \tag{2.4}$$

where $\Pi_\infty(u) = \lim_{t\to\infty}\Pi(u,t)$. The value of $G_\infty(u,0)$ in (2.4) can be obtained from (2.3) as

$$G_\infty(u,0) = (\lambda + \xi\frac{d}{du})\Pi_\infty(u), \tag{2.5}$$

which together with (2.4) gives

$$(1-p+pu)\alpha(\lambda - \lambda u)(\lambda - \xi\frac{d}{du})\Pi(u) = (\lambda - \xi u\frac{d}{du})\Pi(u),$$

where $\alpha(s) = \int_0^\infty a(x)\exp(-sx)dx$. Therefore, we can obtain $\Pi_\infty(u)$ in terms of $\Pi_\infty(1)$, the probability of an idle server at the steady state,

$$\Pi_\infty(u) = \Pi_\infty(1)\exp\left\{\frac{\lambda}{\xi}\int_1^u \frac{1-(1-p+p\omega)\alpha(\lambda-\lambda\omega)}{(1-p+p\omega)\alpha(\lambda-\lambda\omega)-\omega}\right\}d\omega. \tag{2.6}$$

In order to obtain $\Pi_\infty(1)$, and therefore a closed formula for $\Pi_\infty(u)$, let

$$q_n(t) = \int_0^\infty W_n(x,t)dx$$

be the probability that the server is busy and n customers are in the orbit at time t. Let $Q_\infty(u)$ be the generating function of this probability in the steady state defined as

$$Q_{\infty}(u) = \sum_{n=0}^{\infty} u^n q_n(\infty).$$

Then, since by integrating by parts, we can show

$$\alpha(s) = 1 - \int_0^{\infty} \exp\{-sx - N(x)\}dx,$$

from (2.5) we have,

$$Q_{\infty}(u) = \{1 - \alpha(\lambda - \lambda u)\}\{\lambda + \xi\frac{d}{du}\}\Pi_{\infty}(u)/(\lambda - \lambda u). \tag{2.7}$$

Therefore, $\Pi_{\infty}(1)$ can be found from the normalized equation $\Pi_{\infty}(1) + Q_{\infty}(1) = 1$. To this end, from (2.6) we obtain

$$\frac{d\Pi_{\infty}}{du}(1) = \frac{\lambda(p + \lambda\bar{T})\Pi_{\infty}(1)}{\xi(1 - p - \lambda\bar{T})}, \tag{2.8}$$

where $\bar{T} = \int_0^{\infty} xa(x)\ dx = -(d/ds)\alpha(s)\mid_{s=0}$ is the expected service time. Therefore (2.7) and (2.8) yield

$$Q_{\infty}(1) = \frac{\lambda\bar{T}\Pi_{\infty}(1)}{1 - p - \lambda\bar{T}}. \tag{2.9}$$

Hence, it is easy to show

$$\Pi_{\infty}(1) = \frac{1 - p - \lambda\bar{T}}{1 - p}. \tag{2.10}$$

This shows that the condition for ergodicity is $\lambda\bar{T} < 1 - p$ which is indeed necessary to make (2.9) meaningful as well. It is interesting to note that the probability of the server being busy, $Q_{\infty}(1) = \lambda\bar{T}/(1 - p)$, is independent of the rate of repeated demands of service by customers from orbit. This independence can be justified by noting that, if the customers increase their rate of demand from orbit, they are more likely to find the server free and, with probability $1 - p$, depart from the system after being served. Therefore, on average, there will be a smaller orbit size and the increase in repeated demands will be compensated by the decrease in the number of customers.

2.1 The measure of effectiveness

The expected number of customers in the orbit at ergodicity is given by $\bar{n} = \Pi'_{\infty}(1) + Q'_{\infty}(1)$. Now (2.8) and (2.10) give

$$\Pi'_{\infty}(1) = \lambda(p + \lambda\bar{T})/\xi(1 - p).$$

This together with (2.7) gives

$$Q'_{\infty}(1) = \frac{2\lambda^2 p\xi\bar{T}^2 + \lambda^2\xi(1 - p)(\sigma^2 + \bar{T}^2) + 2\lambda^2\bar{T}(p + \lambda\bar{T})}{2\xi(1 - p)(1 - p - \lambda\bar{T})}.$$

By algebraic transformation we have

$$\bar{n} = \lambda^2\frac{2p\bar{T}/(1 - p) + 2(p + \lambda\bar{T})/\lambda\xi + \sigma^2 + \bar{T}^2}{2(1 - p - \lambda\bar{T})}. \tag{2.11}$$

This, for $p = 0$, corresponds to the expected value of orbit size obtained in [4]. Note also that when $\xi \rightarrow \infty$, that is, when there is constant surveillance of the service by customers from orbit, the system behavior reduces to that of ordinary queues with Poisson arrivals and a general service time, in which a proportion p of customers choose to rejoin the queue on completion of service.

The expected waiting time $\bar{\tau}$ *of a customer* is obtained using a method similar to Keilson et al. [5] or Little [7]. Let t be a long time interval in which a particular system sample is observed and t_n be the total length of sub-intervals of t in which n customers wait. Obviously, during t_n a total of nt_n customers unit time is spent waiting and therefore, for this system sample, the expected time spent waiting by an arrived customer is

$$\bar{\tau} = \sum_{n \geq 0} nt_n/\lambda t = \bar{n}/\lambda \quad \text{as } t \to \infty. \tag{2.12}$$

The denominator of the middle term in (2.12) is the number of arrivals in the time interval of length t, which is λt. Hence, from (2.11) and (2.12) we find that

$$\bar{\tau} = \lambda \frac{2p\bar{T}/(1-p) + 2(p+\lambda\bar{T})/\lambda\xi + \sigma^2 + \bar{T}^2}{2(1-p-\lambda\bar{T})}.$$

The average number of "look ins" per customer for each completed call Γ is calculated using the above sample system. During the time interval t_n, an average of $n\xi t_n$ "look ins" occur from the orbit and λt_n new customers apply for service. Hence, for sufficiently large t,

$$\Gamma = \sum_{n \geq 0} (\lambda t_n + n\xi t_n)/\lambda t = 1 + \xi\bar{n}/\lambda = 1 + \xi\bar{\tau}. \tag{2.13}$$

This together with (2.11) gives

$$\Gamma = \frac{1 + \lambda\xi p\bar{T}/(1-p) - \lambda\xi(\sigma^2 + \bar{T}^2)/2}{1 - p - \lambda\bar{T}}.$$

3. Discouraged Repeated Demands

Here the customers in the orbit seek service at subsequent epochs with a rate which is a decreasing function of the orbit size. We assume that $\xi_n = \xi/n$ when the orbit size is n. This is indeed the same model as that with the assumption that the customers in the orbit form a queue such that only one (which is more natural if it is the first) can reapply for service with rate ξ. Also, our analysis covers the queueing model in which the server upon completion of service takes a vacation, whose duration is exponentially distributed with parameter ξ. A vacation may be interrupted and terminated if an arrival occurs before its normal termination. In this, the service of the arriving customer starts immediately. Otherwise, if the vacation terminates normally, the server will serve the first customer, if any, in the orbit.

The equations corresponding to (2.1)-(2.3) which govern the above systems are

$$\frac{dp_n(t)}{dt} = -(\lambda+\xi)p_n(t) + (1-p)\int_0^\infty W_n(x,t)\eta(x)dx + p\int_0^\infty W_{n-1}(x,t)\eta(x)dx \qquad n \geq 1 \tag{3.1}$$

and

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + (1-p)\int_0^\infty W_0(x,t)\eta(x)dx \tag{3.2}$$

$$W_n(0,t) = \lambda p_n(t) + \xi p_{n+1}(t). \tag{3.3}$$

As in the case of a fixed rate of repeated demands in section 2, we can obtain the relevant generating functions at ergodicity. A little algebra together with (3.1) and (3.2) yields

$$(\lambda+\xi)\Pi_{\infty}(u)-\xi p_0=(1-p+pu)\int_0^{\infty}\eta(x)G_{\infty}(u,x)dx. \tag{3.4}$$

Also from (3.3) we can easily show

$$G_{\infty}(u,0)=(\lambda+\xi/u)\Pi_{\infty}(u)-\xi p_0/u. \tag{3.5}$$

Since the value of $G_{\infty}(u,x)$ in (3.4) relates to $G_{\infty}(u,0)$ by (1.2), we can substitute (3.5) in (3.4) to obtain

$$\Pi_{\infty}(u)=\xi p_0\frac{\{(1-p+up)/u\}\alpha(\lambda-\lambda u)-1}{-\lambda-\xi+(1-p+pu)(\lambda+\xi/u)\alpha(\lambda-\lambda u)}. \tag{3.6}$$

In order to obtain p_0 with the same definitions already developed in section 2, as in (2.7), we can show that

$$Q_{\infty}(u)=\int_0^{\infty}G_{\infty}(u,x)dx=\frac{\{1-\alpha(\lambda-\lambda u)\}\{(\lambda+\xi/u)\Pi_{\infty}(u)-p_0\xi/u\}}{\lambda-\lambda u}. \tag{3.7}$$

Now we can use the normalized relation $\Pi_{\infty}(1)+Q_{\infty}(1)=1$ to evaluate p_0. To this end, (3.6) and (3.7) yield

$$\Pi_{\infty}(1)=\frac{\xi p_0(1-p-\lambda\bar{T})}{\xi-(\lambda+\xi)(p+\lambda\bar{T})} \tag{3.8}$$

and

$$Q_{\infty}(1)=\frac{\lambda\xi p_0\bar{T}}{\xi-(\lambda+\xi)(p+\lambda\bar{T})}. \tag{3.9}$$

Therefore, from (3.8) and (3.9) we can easily evaluate

$$p_0=\frac{\xi-(\lambda+\xi)(p+\lambda\bar{T})}{\xi(1-p)}. \tag{3.10}$$

To make (3.10) meaningful, and therefore to find the conditions for the existence of the ergodicity, we require $\lambda\bar{T}(1+\lambda/\xi)<1$ and $(\lambda/\xi+1)(p+\lambda\bar{T})<1$, which are also required in (3.8) and (3.9). The first condition is that also required in [3] and is independent of p.

3.1 The measure of effectiveness

In order to obtain *the expected number of customers in the orbit* we need to evaluate $\Pi'_{\infty}(1)$ and $Q'_{\infty}(1)$. To this end, (3.6) and (3.7) yield

$$\Pi'_{\infty}(1)=\lambda\frac{\lambda\bar{T}+p+\lambda^2(\sigma^2-\bar{T}^2)/2(1-p)}{\xi-(\lambda+\xi)(p+\lambda\bar{T})}$$

and

$$Q'_{\infty}(1)=\lambda^2\frac{2\bar{T}(p+\lambda\bar{T})+\sigma^2(\xi-p\xi-\lambda p)+\bar{T}^2(\xi+\lambda p+\xi p)}{2(1-p)\{\xi-(\lambda+\xi)(p+\lambda+\bar{T})\}}. \tag{3.11}$$

Now we are in a position to find the expected orbit size $\bar{n}=\Pi'_{\infty}(1)+Q'_{\infty}(1)$ as

$$\bar{n}=\lambda\frac{\lambda\sigma^2(\lambda+\xi)+2p+\lambda\bar{T}\{2+\bar{T}(1+p)(\lambda+\xi)\}/(1-p)}{2\{\xi-(\lambda+\xi)(p+\lambda\bar{T})\}}. \tag{3.12}$$

For $p = 0$ the value of $\bar{n}$ corresponds to that in [3].

The expected waiting time can be obtained as in section 2 to be $\bar{\tau} = \bar{n}/\lambda$ and therefore the result is easily derived from (3.11).

However, in order to obtain *the expected number of "look ins"* we need to modify arguments used in section 2. The corresponding formula in (2.12) for Γ for the discouraged case becomes

$$\Gamma = \sum_{n \geq 0} (\lambda t_n + \xi t_n)/\lambda t - \xi t_0/\lambda t. \tag{3.13}$$

The additional term in (3.12) compared with (2.13) appears to disable the effect of ξ when the orbit becomes idle. Hence here

$$\Gamma = 1 + \xi/\lambda - \xi t_0/\lambda t. \tag{3.14}$$

Notice that t_0/t, for sufficiently large t, is the probability of an empty orbit and hence is equal to $p_0 + q_0(\infty) = p_0 + Q_\infty(0)$. From (3.7) we obtain

$$Q_\infty(0) = p_0 \frac{1 - \alpha(\lambda)}{(1-p)\alpha(\lambda)}$$

and therefore we can show that

$$\lim_{t \to \infty} \frac{t_0}{t} = p_0 \frac{1 - p\alpha(\lambda)}{(1-p)\alpha(\lambda)}$$

which together with (3.10) and (3.13) evaluates Γ.

4. Numerical Comparison of the Orbit Sizes

For arbitrary values of $\lambda = \xi = \sigma^2 = 1$ which satisfy the required ergodicity conditions, we present the graphs of $\bar{n}$ against p for values of $\bar{T} = 1/3$ and $\bar{T} = 1/6$ for the two cases of fixed and discouraged rate of repeated demands.

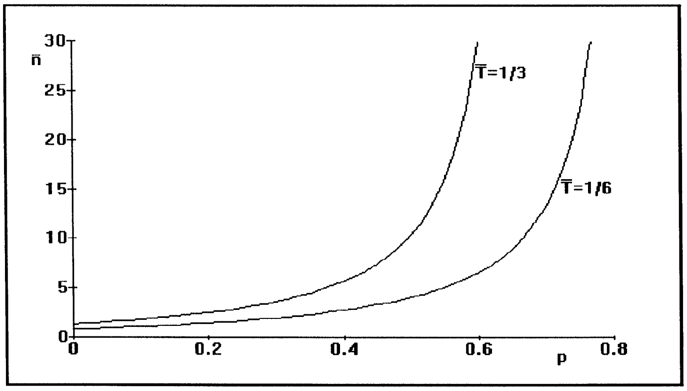

Figure 1. Fixed rate of repeated demands.

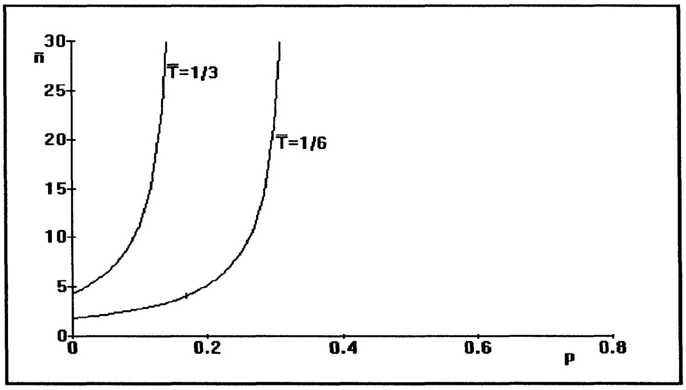

Figure 2. Discouraged repeated demands.

As is expected for the discouraged case, Figure 2, $\bar{n}$ increases much faster than that in the case of a fixed rate of repeated demands, Figure 1. However, this rate of increase turns out to be surprisingly fast. Each curve in Figures 1 and 2 behaves asymptotically as a rectangular hyperbola. At $p=0$, the results corresponds to $\bar{n}$ obtained in [5] and [3].

References

[1] Boxma, O.J. and Cohen, J.W., The M/G/1 queue with permanent customers, *IEEE J. on Selected Areas in Communications* **9** (1991), 179-184.

[2] Falin, G.I., On the waiting time process in a single line queue with repeated calls, *J. Appl. Prob.* **23** (1986), 185-192.

[3] Farahmand, K., Single line queue with repeated demands, *Queueing Sys.* **6** (1990), 223-228.

[4] Farahmand, K., Single line queue with recurrent repeated demands, *Queueing Sys.*, in press.

[5] Keilson, J., Cozzolino, J. and Young, H., A service system with unfilled requests repeated, *Oper. Res.* **16** (1968), 1126-1132.

[6] Keilson, J. and Kooharian, A., On time dependent queueing processes, *Ann. Math. Stat.* **31** (1960), 104-112.

[7] Little, J.D.C., A proof for queueing formula $L=\lambda W$, *Oper. Res.* **9** (1961), 383-387.