Single Retrial Queues with Service Option on Arrival

K. FARAHMAND AND N. COOKE

Department of Mathematics, University of Ulster, Northern Ireland

Abstract. We analyze a queueing system in which customers can call in to request service. A proportion, say $1 - p$ of them on their arrival test the availability of the server. 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. The remaining p proportion of the initial customers enter the orbit directly, without examining the state of the server. 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.

Keywords: Single line queue, repeated demands, orbit size, waiting time, ergodic state, generating function.

1. Introduction

The retrial queues are characterized by waiting customers who, unlike ordinary queues, cannot be in continuous contact with the server, but can only call in to test the state of the server. If the server is free the service commences immediately. Otherwise, if the server is engaged, the customer produces a source of unsatisfied customers, call orbit customers, who may re-apply several times for service. We assume the customers requesting service from outside arrive according to the Poisson process with rate λ, join orbit directly with probability p and test the availability of the server with probability $1 - p$. If the server is available the arrival enters service, otherwise it has to join the orbit and reinitiate its request later. Previously, see for example Keilson and Kooharian [6], Keilson et al [5] and Falin [2], the case of $p = 0$ only has been considered and the aim of this paper is to give the arrival the option of joining the orbit or testing the availability of the server.

The customers in orbit seek service at subsequent epochs until they find the server free and then enter service. The intervals separating a customer "look-in" epoch will be assumed to be exponentially distributed with parameter ξ_n, when the orbit size is n. We consider two cases characterized by ξ_n, the orbit discipline, see also Conolly [1] for a similar work on state dependent arrivals. Firstly the case of fixed rate of re-applying for service, that is $\xi_n \equiv \xi$, which for $p = 0$ dominate previous works in the subject, see [3] for a survey and [8] for the latest development. Secondly,

the case of discouraged repeated demands, $\xi_n \equiv \xi/n$, in which the customers from orbit are relaxed, and are prepared to reduce their rate of service as more customers join the orbit. For $p = 0$ this case was studied in Farahmand [4] which contains the theoretical motivation of the subject as well as the computer representations and communication networks application of the above structure. In addition our models could arise naturally in the following circumstances. First our model can be seen in the context of ordinary M/G/1 queues in which the server takes a vacation upon each completion of service. The vacations are terminated according to an exponential distribution with parameter ξ_n, or interrupted by a proportion p of newly arrived customers. That is $(1 - p)$ of new arrivals do not wish or do not have the priority to interrupt the server's vacation and join the orbit directly on arrival. Also the case of discouraged repeated demand, $\xi_n = \xi/n$ could be looked upon as a FCFS or LCFS discipline in orbit. For this model the customers in orbit form a queue in which only one at the head of the queue could apply for service. It reinitiates its demand at intervals assumed to be exponentially distributed with parameter ξ, till it finds the server free. For this model, also the newly arrived customers have an option to join the head or end of the queue with probability p or examining the state of the server on his arrival.

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 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. Let $W_n(x, t)$ be the joint probability density that there are n customers in orbit at epoch t and a customer is present in service who has been there for time x. It is shown in [5] that $G(u, x, t) = \sum_{n=0}^{\infty} u^n W_n(x, t)$, the generating function of $W_n(x, t)$, satisfies the following relation

$$G(u, x, t) = G(u, 0, t - x) \exp\{\lambda ux - \lambda x - N(x)\}. \tag{1.1}$$

where

$$N(x) = \int_0^x \eta(y)\ dy = -\log\{1 - A(x)\}.$$

Therefore for $G_\infty(u, x) = \lim_{t\to\infty} G(u, x, t)$ from (1.1) we have

$$G_\infty(u, x) = G_\infty(u, 0) \exp\{\lambda ux - \lambda x - N(x)\}. \tag{1.2}$$

We shall now consider separately each of the two cases mentioned earlier.

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 orbit. It is easy to show that the equations governing this system are

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

(2.1)

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

(2.2)

and

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

Now assume the system is ergodic and let $\Pi(u,t) = \sum_{n=0}^\infty u^n p_n(t)$ be the generating function of $p_n(t)$, and $\Pi_\infty(u) = \lim_{t\to\infty} \Pi(u,t)$. ¿From (1.2), (2.1) and (2.2) we thus have

$$(\lambda - \lambda pu + \xi u \frac{d}{du})\Pi_\infty(u) = \int_0^\infty a(x)G_\infty(u,0)\exp\{-\lambda x(1-u)\} \; dx. \qquad (2.4)$$

Similarly from (2.3) we obtain

$$G_\infty(u,0) = (\lambda - p\lambda + \xi \frac{d}{du})\Pi(u). \qquad (2.5)$$

Let $\alpha(s) = \int_0^\infty a(x)\exp(-sx) \; dx$. Then (2.4) and (2.5) give

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

Consequently

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

where $\Pi_\infty(1)$ is the probability of an idle server at the steady state. This constant $\Pi_\infty(1)$ can now be determined from normalization. Let

$$p_n^*(t) = \int_0^\infty W_n(x,t) \; dx$$

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

$$\Pi_\infty^*(u) = \sum_{n=0}^\infty u^n p_n^*(\infty).$$

Then since

$$\int_0^\infty \exp(-sx - N(x)) \; dx = \frac{1 - \alpha(s)}{s} \qquad (2.7)$$

we obtain
$$\Pi_\infty^*(u) = G_\infty(u,0)\{1 - \alpha(\lambda - \lambda u)\}/(\lambda - \lambda u).$$

Substituting for $G_\infty(u,0)$ from (2.5) and rearranging gives

$$\Pi_\infty^*(u) = \Pi_\infty(u)\{1 - \alpha(\lambda - \lambda u)\}/\{\alpha(\lambda - \lambda u) - u\}. \tag{2.8}$$

The normalization requirement $\Pi_\infty(1) + \Pi_\infty^*(1) = 1$ and L'Hospital's rule applied to (2.8) then gives

$$\Pi_\infty(1) = 1 - \lambda\overline{T} \tag{2.9}$$

and

$$\Pi_\infty^*(1) = \lambda\overline{T} \tag{2.10}$$

where $\overline{T} = \int_0^\infty x a(x)\ dx = -(d/ds)\alpha(s)|_{s=0}$ is the expected service time. It is interesting to note that $\Pi_\infty(1)$ and $\Pi_\infty^*(1)$ are only functions of λ and $\overline{T}$ and not p and ξ. However, since all the customers are eventually served, it soon becomes obvious that the proportion of time that the server is busy or idle, $\Pi_\infty(1)$ and $\Pi_\infty^*(1)$, should be the same no matter what values p and ξ have.

2.1. Measures of effectiveness

(a) The average orbit size

The mean number of customer in orbit is given by $\overline{n} = \Pi_\infty'(1) + \Pi_\infty^{*'}(1)$. First using (2.6), applying L'Hospital's rule and since

$$\frac{d}{du}\alpha'(\lambda - \lambda u)|_{u=1} = -\lambda \int_0^\infty x^2 a(x)\ dx = -\lambda(\sigma^2 + \overline{T}^2)$$

where σ^2 is the variance of the service time, we obtain

$$\Pi_\infty'(1) = (\lambda/\xi)\{\lambda\overline{T}(1 - p) + p\} \tag{2.11}$$

and in a similar way

$$\Pi_\infty^{*'}(1) = \Pi_\infty'(1)\lambda\overline{T}/(1 - \lambda\overline{T}) + \Pi_\infty(1)\{\lambda^2(\sigma^2 + \overline{T}^2)/2(1 - \lambda\overline{T})^2\}. \tag{2.12}$$

Hence from (2.11) and (2.12) we can obtain

$$\overline{n} = \left[(\lambda/\xi)\{\lambda\overline{T}(1 - p) + p\} + (\lambda^2/2)(\sigma^2 + \overline{T}^2)\right]/(1 - \lambda\overline{T}). \tag{2.13}$$

(b) The average waiting time of a customer

To calculate the mean waiting time $\overline{\tau}$ of a customer we use a method applied by Keilson et al [5] or Little [7]. Let us consider a particular system sample observed over a long time interval $(0, t)$. Let t_n be the total time interval length in which n customers wait. During that period nt_n units of time are expended waiting. Hence

for sufficiently large t, the average time spent waiting by the customers who arrive is

$$\overline{\tau} \;=\; \sum_{n=0}^{\infty} nt_n/\{\text{number of arrivals in}(0,t)\}$$

$$=\; \sum_{n=0}^{\infty} nt_n/\lambda t$$

$$=\; \overline{n}/\lambda.$$

Using (2.13) this gives

$$\overline{\tau} = \left[\{\lambda\overline{T}(1-p)+p\}/\xi + (\lambda/2)(\sigma^2+\overline{T}^2)\right]/(1-\lambda\overline{T}). \tag{2.14}$$

(c) The average number of "look-ins" per customer

To calculate the mean number of "look-ins" per customer, ϕ, for each completed service we shall use an argument similar to that for (2.14). In the total interval t_n an average of $n\xi t_n$, $n\geq 1$ "look-ins" from orbit occur and $\lambda(1-p)t_n$, $n\geq 0$, new customers apply for service. Hence for sufficiently large t

$$\phi \;=\; \sum_{n=0}^{\infty} \lambda(1-p)t_n/\lambda t + \sum_{n=1}^{\infty} n\xi t_n/\lambda t$$

$$=\; \{\lambda(1-p)+\overline{n}\xi\}/\lambda$$

$$=\; 1-p+\xi\overline{\tau}. \tag{2.15}$$

The value of ϕ in (2.15) is a natural generalization of the expected orbit size to reflect the fact that an arrival may forego the first look-in with probability p. When we put $p=0$ in these last three results (2.13)-(2.15) they correspond to those of Keilson et al [5].

3. Case of discouraged repeated demands

The equations governing this system are formed in the normal way and using a similar method as §2. We obtain

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

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

and

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

Again we assume the system is ergodic, and using (3.1) and (3.2) gives

$$(\lambda-\lambda pu+\xi)\Pi_{\infty}(u) - \xi p_0 \;=\; \int_0^{\infty} G_{\infty}(u,0)a(x)\exp\{-\lambda x(1-u)\}\,dx \tag{3.4}$$

Similarly from (3.3) we can obtain

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

Using (3.4) and(3.5) we therefore have

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

Applying L'Hospital's rule to (3.6) gives

$$\Pi_\infty(1) = \frac{\xi p_0(1 - \lambda\overline{T})}{\xi - \lambda\overline{T}(\xi + \lambda) + \lambda p(\lambda\overline{T} - 1)}. \tag{3.7}$$

Also from (3.5) we have

$$\begin{aligned}
\Pi_\infty^*(u) &= \int_0^\infty G_\infty(u, x) \, dx \\
&= \frac{\{1 - \alpha(\lambda - \lambda u)\}\{(\lambda - \lambda p + \xi/u)\Pi_\infty(u) - p_0\xi/u\}}{\lambda - \lambda u}.
\end{aligned} \tag{3.8}$$

Applying L'Hospital's rule to (3.8) gives

$$\Pi_\infty^*(1) = \frac{\lambda\xi p_0\overline{T}}{\xi - \lambda\overline{T}(\xi + \lambda) + \lambda p(\lambda\overline{T} - 1)}. \tag{3.9}$$

Now it is required that $\Pi_\infty(1) + \Pi_\infty^*(1) = 1$, thus giving

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

Back substituting for p_0 in (3.7) and (3.9) gives respectively

$$\Pi_\infty(1) = 1 - \lambda\overline{T} \tag{3.11}$$

and

$$\Pi_\infty^*(1) = \lambda\overline{T}. \tag{3.12}$$

3.1. Measures of effectiveness

(a) The average orbit size

Like before the mean number of customers in orbit is given by $\overline{n} = \Pi_\infty'(1) + \Pi_\infty^{*'}(1)$.
Using (3.6.), (3.8) and applying L'Hospital's rule we obtain

$$\Pi_\infty'(1) = \frac{(\lambda^2/2)(2\overline{T} - \lambda\overline{T} + \lambda\sigma^2) + \lambda p(1 - \lambda\overline{T})^2}{\xi - \lambda^2\overline{T} - \lambda\overline{T}\xi - \lambda p(1 - \lambda\overline{T})} \tag{3.13}$$

and

$$\Pi_{\infty}^{*'}(1) = \frac{\lambda^2 \overline{T}\{\lambda \overline{T}(1-p)+p\}+(\lambda^2/2)(\sigma^2+\overline{T}^2)(\xi-\lambda p)}{\xi-\lambda^2\overline{T}-\lambda\overline{T}\xi-\lambda p(1-\lambda\overline{T})}. \tag{3.14}$$

Thus

$$\overline{n} = \frac{\lambda^2\overline{T}(1-p)+\lambda p+(\lambda^2/2)(\sigma^2+\overline{T}^2)\{\lambda(1-p)+\xi\}}{\xi-\lambda^2\overline{T}-\lambda\overline{T}\xi-\lambda p(1-\lambda\overline{T})}. \tag{3.15}$$

As expected, for $p = 0$ this result is consistent with that found in Farahmand [4].

(b) The average waiting time of a customer

The mean waiting time $\overline{\tau}$ of a customer is given as previously by $\overline{\tau} = \overline{n}/\lambda$. Therefore from this and (3.15) the required result is obtained.

(c) The average number of "look-ins" per customer.

Once again, in order to calculate the mean number of "look ins" per customer for each completed service, we use a similar argument to that in the previous corresponding section. In the total time t_n in which n customers are waiting, an average of ξt_n, $n \geq 1$, "look ins" from orbit occur and $\lambda(1-p)t_n$, $n \geq 0$, new customers apply for service. Hence for sufficiently large t

$$\phi = \sum_{n=0}^{\infty}\lambda(1-p)t_n/\lambda t+\sum_{n=1}^{\infty}\xi t_n/\lambda t$$
$$= \{\lambda(1-p)+\xi\}/\lambda-\xi t_0/\lambda t. \tag{3.16}$$

Now t_0/t is the probability that there are no customers in orbit. Therefore by rearranging (3.8) we can write

$$t_0/t = p_0+\Pi_{\infty}^{*}(0)$$
$$= p_0+p_0\{1-\alpha(\lambda)\}/\alpha(\lambda),$$

where $\alpha(\lambda) = \int_0^{\infty} a(x)\exp(-\lambda x)\ dx$. Substituting for p_0 from (3.10) and using (3.16) we have

$$\phi = \{\lambda(1-p)+\xi\}/\lambda-\{\lambda p(\lambda\overline{T}-1)-\lambda^2\overline{T}-\lambda\overline{T}\xi+\xi\}/\lambda\alpha(\lambda)$$

which for $p = 0$ corresponds to the known result, see [4].

4. Numerical observations

The following graphs shows the relations between the probability p and the expected orbit sizes for both cases of $\xi_n \equiv \xi$ and $\xi_n = \xi/n$. This expected orbit size is, obviously, smaller for the case of fixed rate of repeated demands than for the discouraged case. Indeed, in both cases these expected orbit sizes are an increasing functions of p. However, somehow this increase is more significant in the discouraged case.
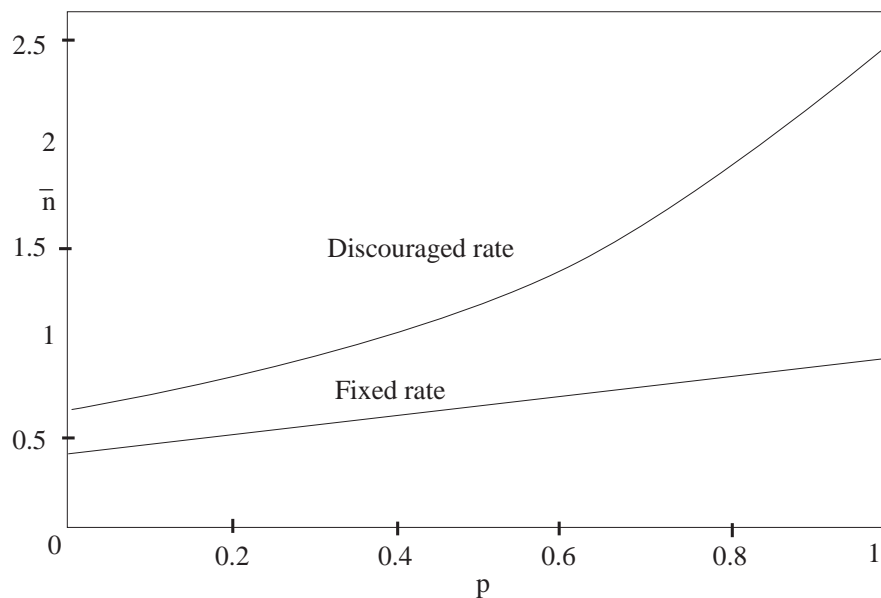
Figure 1. Graphs of the Expected Orbit Sizes

Acknowledgments

References

1. B.W. Conolly. Generalized state-dependent Erlangian queues: Speculations about calculating measures of effectiveness. *J. Appl. Prob.*, 23:358–363, 1975.
2. G.I. Falin. On the waiting-time process in a single line queue with repeated calls. *J. Appl. Prob.*, 23:185–192, 1986.
3. G.I. Falin. A survey of retrial queues. *Queueing Systems*, 7:127–167, 1990.
4. K. Farahmand. Single line queue with repeated demands. *Queueing Systems*, 6:223–228, 1990.
5. J. Keilson, J. Cozzolino, and H. Young. A service system with unfilled requests repeated. *Oper. Res.*, 16:1126–1132, 1968.
6. J. Keilson and A. Kooharian. On time dependent queueing processes. *Ann. Math. Stat.*, 31:104–112, 1960.
7. J.D.C. Little. A proof for queueing formula $L = \lambda W$. *Oper. Res.*, 9:383–387, 1961.
8. M. Martine and J.R. Artalejo. Analysis of an M/G/1 queue with two types of impatient units. *Adv. Appl. Prob.*, 27:840–861, 1995.