J. Korean Math. Soc. **35** (1998), No. 3, pp. 503–525

RETRIAL QUEUES WITH A FINITE NUMBER OF SOURCES

J. R. ARTALEJO

ABSTRACT. In the theory of retrial queues it is usually assumed that the flow of primary customers is Poisson. This means that the number of independent sources, or potential customers, is infinite and each of them generates primary arrivals very seldom. We consider now retrial queueing systems with a homogeneous population, that is, we assume that a finite number K of identical sources generates the so called quasi-random input. We present a survey of the main results and mathematical tools for finite source retrial queues, concentrating on $M/G/1//K$ and $M/M/c//K$ systems with repeated attempts.

1. Introduction

The main characteristic of a retrial queue is that a customer who finds the service facility busy upon arrival is obliged to leave the service area, but some time later he comes back to re-initiate his demand. Between trials a customer is said to be "in orbit". Most papers assume that the population of potential customers is very large so the input stream is Poisson. In such a description, the probability of a new arrival during any interval of duration dt is given by $\lambda dt + o(dt)$ as $dt \to 0$, independently of the state of the system at time t. The analysis of retrial queueing systems in the case of non-exponentially distributed intervals between primary arrivals is still an open problem. Some recent contributions by Choi and Chang (1998) and Dudin and Klimenok (1998) consider arrival processes modelled as Batch Markovian Arrival Processes.

Received February 24, 1998.

1991 Mathematics Subject Classification: 60K25.

Key words and phrases: busy period, quasirandom input, retrial queues, steady state distribution, waiting time.

This is an invited paper to the International Conference on Probability Theory and its Applications.

The next extension concerns a limited number of sources K. Each source is either free or in the system (orbit and service facility) at any time. We consider that the input stream is the so called quasirandom input; that is, the probability that any particular source generates a request for service in any interval $(t, t+dt)$ is $\alpha dt + o(dt)$ as $dt \to 0$ if the source is idle at time t, and zero if the source is being served or in orbit at time t, independently of the behaviour of any other sources.

The main features of the theory of retrial queues can be found in Falin and Templeton (1997). On the other hand, systems with classical waiting lines and finite population have been reviewed in detail by Takagi (1993).

In this paper we give a complete survey of retrial queues with quasirandom input. The analysis of the basic models of type $M/G/1//K$ and $M/M/c//K$ in Kendall's notation is the subject matter of sections 2 and 3, respectively. In section 4, we give some examples of systems which can be modelled as retrial queues with a finite number of sources.

2. The $M/G/1//K$ retrial queue

2.1. Model description

We assume that the arrival process of primary calls is quasirandom with parameter α. If the server is free then the call is immediately served. When the server is busy, the source generates a Poisson flow of repeated attempts with parameter μ until it finds the server free. The service times have probability distribution function $B(x)$ with $B(0) = 0$. We denote its Laplace-Stieltjes transform as $\beta(s)$ and its nth moment as β_n. The input stream of primary arrivals, service times and intervals between successive repeated attempts are assumed to be mutually independent.

The state of the system can be described by means of the process $(C(t), N(t), \xi(t))$ where $C(t)$ is 0 or 1 according as the server is free or busy at time t, $N(t)$ is the number of sources in orbit and, if $C(t) = 1$, then $\xi(t)$ is the supplementary variable denoting the elapsed service time.

2.2. Joint distribution of the server state and the orbit length in steady state

There exists an abundance of techniques for handling the analysis of queue length distributions. Our first approach in this section will be

the method of supplementary variable combined with a discrete transformation. However, we will also discuss a second approach based on the theory of regenerative processes.

The first approach was introduced by Ohmura and Takahashi (1985). Later, Falin and Artalejo (1998) employ the same approach for improving the expressions of the main performance characteristics. Their results can be summarized as follows. We define the probabilities (densities)

$$(1) \qquad p_{0n} = \mathbf{P}\{C(t) = 0, N(t) = n\}, \qquad 0 \le n \le K-1,$$

$$(2) \quad p_{1n}(x) = \frac{d}{dx}\mathbf{P}\{C(t) = 1, \xi(t) \le x, N(t) = n\}, \qquad 0 \le n \le K-1,$$

$$(3) \quad p_{1n} = \mathbf{P}\{C(t) = 1, N(t) = n\} = \int_0^\infty p_{1n}(x)dx, \qquad 0 \le n \le K-1.$$

Then, following the method of supplementary variables, we find that the limiting probabilities as $t \to \infty$ satisfy the equations of statistical equilibrium:

$$(4) \qquad ((K-n)\alpha + n\mu)p_{0n} = \int_0^\infty p_{1n}(x)b(x)dx,$$

$$(5) \quad p'_{1n}(x) = -((K-n-1)\alpha + b(x))p_{1n}(x) + (K-n)\alpha p_{1,n-1}(x),$$

$$(6) \qquad p_{1n}(0) = (K-n)\alpha p_{0n} + (n+1)\mu p_{0,n+1},$$

where $b(x) = B'(x)/(1-B(x))$ is the hazard rate function of $B(x)$ and $p_{0K} = p_{1,-1} = 0$.

We can rewrite (5) with the help of so-called discrete transformations (Jaiswal (1968)). A discrete transformation is a specific linear replacement of variables where a set of unknown variables $p = (p_0, \dots, p_{K-1})$ is replaced by $q' = (q_0, \dots, q_{K-1})' = Ap'$, where A is a non-singular $K \times K$ matrix.

We introduce the transformation defined by

$$q_m = \sum_{n=0}^{K-1-m} \binom{K-1-n}{m} p_n, \tag{7}$$

so the inverse transformation is given by

$$p_n = \sum_{m=0}^{n} (-1)^m \binom{K-1-n+m}{m} q_{K-1-n+m}, \qquad 0 \le n \le K-1. \tag{8}$$

Let q_{0m}, $q_{1m}(x)$, q_{1m}, $0 \le m \le K-1$, be the images of sequences p_{0m}, $p_{1m}(x)$, p_{1m}, respectively, under discrete transformation (7).

The variables q_{0m}, $0 \le m \le K-1$, can be determined with the help of the following recursion

$$q_{0m} = C_m q_{0,K-1}, \qquad 0 \le m \le K-2, \tag{9}$$

$$q_{0,K-1} = \nu((\nu + \alpha + (K-1)\mu)C_0 + (\alpha - \mu)C_1)^{-1}, \tag{10}$$

where $\nu = 1/\beta_1$ and the coefficients C_m can be recursively computed by putting $q_{0,K-1} = 1$ in the equation

$$\begin{aligned} &(((K-m-1)\mu + (m+1)\alpha)(1-\beta(m\alpha)) \\ &+m\mu\beta(m\alpha))q_{0m} - (K-m)\mu\beta(m\alpha)q_{0,m-1} \\ &+(m+1)(\alpha-\mu)(1-\beta(m\alpha))q_{0,m+1} = 0, \qquad 1 \le m \le K-1, \end{aligned} \tag{11}$$

where we have assumed that $q_{0K} = q_{0,-1} = 0$.

Then, $q_{1m}(0)$ and $q_{1m}(x)$ are given by

$$\begin{aligned} &\beta(m\alpha)q_{1m}(0) \\ &= ((K-m-1)\mu + (m+1)\alpha)q_{0m} + (m+1)(\alpha-\mu)q_{0,m+1}, \end{aligned} \tag{12}$$

$$q_{1m}(x) = q_{1m}(0)(1 - B(x)) \exp\{-m\alpha x\}. \tag{13}$$

Obviously, q_{1m} follows by integrating $q_{1m}(x)$.

Once we have obtained the limiting probabilities and probability densities, we can derive formulae for the main system performance measures.

It should be pointed out that all main characteristics can be expressed in terms of the server utilization $p_1 = \mathbf{P}\{C(t) = 1\}$. Namely,

1. *The server utilization*

$$p_1 = 1 - q_{00}. \tag{14}$$

2. *The mean number of sources in orbit*

$$N = \mathbf{E}[N(t)] = K - \alpha^{-1}(\alpha + \nu)p_1. \tag{15}$$

3. *The mean rate of generation of primary arrivals*

$$\overline{\lambda} = \alpha \mathbf{E}[K - C(t) - N(t)] = \nu p_1. \tag{16}$$

4. *The mean waiting time*

$$\mathbf{E}[W] = (\overline{\lambda})^{-1} N = (\nu p_1)^{-1} K - \alpha^{-1} - \nu^{-1}. \tag{17}$$

5. *The variance of the orbit length*

$$\begin{aligned} VarN(t) = & - p_1((K-1)\alpha\mu(2-\beta(\alpha)) \\ & + (\alpha+\nu)((3\mu-\alpha)(1-\beta(\alpha)) + \mu\beta(\alpha))) \\ & \times (\alpha(1-\beta(\alpha))(\alpha-\mu))^{-1} \\ & + K\mu(2-\beta(\alpha))((1-\beta(\alpha))(\alpha-\mu))^{-1} \\ & - (\alpha^{-1}(\alpha+\nu))^2 p_1^2, \qquad \text{if} \quad \alpha \neq \mu, \end{aligned} \tag{18}$$

$$\begin{aligned} \mathbf{Var}N(t) = & \quad \nu(K-1)(\nu+\alpha K)^{-1} \\ & \times ((K-2)\beta(\alpha)(K-(K-1)\beta(\alpha))^{-1} \\ & + ((\alpha-\nu)K + 2\nu)(\nu+\alpha K)^{-1}), \qquad \text{if} \quad \alpha = \mu. \end{aligned} \tag{19}$$

We next discuss a second approach for computing the limiting probabilities p_{0n} and p_{1n}. De Kok (1984) and Schellhaas (1986) have considered single server retrial queues in which the arrival process is modelled as a state dependent Markov process with parameter λ_{in} when $(C(t), N(t))$ is in state (i, n). The $M/G/1//K$ retrial queue can be obtained as a special case when $\lambda_{in} = \alpha(K-i-n)$, $i \in \{0,1\}$, $0 \leq n \leq K-1$. The advantage of this approach is that it holds valid in the context of complex systems

such as models operating under the simultaneous presence of repeated attempts and negative arrivals (see Artalejo and Gomez-Corral (1996)).

Following the methodology under the assumptions considered by De Kok (1984), we find that the limiting probabilities satisfy the following equations

$$n\mu p_{0n} = \lambda_{1,n-1} p_{1,n-1}, \qquad n \geq 1, \tag{20}$$

$$p_{1n} = \lambda_{00} A_{0n} p_{00} + \sum_{k=1}^{n+1} \frac{\lambda_{0k} + k\mu}{k\mu} \lambda_{1,k-1} A_{kn} p_{1,k-1}, \qquad n \geq 0, \tag{21}$$

$$\sum_{i=0}^{1} \sum_{n=0}^{\infty} p_{in} = 1, \tag{22}$$

where

$$A_{n+1,n} = \frac{(n+1)\mu}{\lambda_{0,n+1} + (n+1)\mu} B_{nn}, \qquad n \geq 0, \tag{23}$$

$$A_{kn} = \frac{k\mu}{\lambda_{0k} + k\mu} B_{k-1,n} + \frac{\lambda_{0k}}{\lambda_{0k} + k\mu} B_{kn}, \qquad 0 \leq k \leq n. \tag{24}$$

The quantities B_{kn} might be determined for each specific model. For the $M/G/1//K$, we have

$$B_{kn} = \int_0^{\infty} \binom{K-k-1}{n-k} (1 - e^{-\alpha t})^{n-k} (e^{-\alpha t})^{K-n-1} (1 - B(t)) dt, \qquad 0 \leq k \leq n \leq K-1. \tag{25}$$

The above expressions provide a stable recursive scheme for computing $\{p_{0n}\}_{n=1}^{\infty}$ and $\{p_{1n}\}_{n=0}^{\infty}$ in terms of p_{00}. Then, we find p_{00} by using the normalization equation (22).

The derivation of the above methodology is based on a well-known property of regenerative processes. Let us define a regeneration cycle T as the time interval between two successive visits of process $(C(t), N(t))$ to state $(0,0)$; further, define T_{in} as the amount of time in T during which $(C(t), N(t)) = (i,n)$. Then, we have $p_{in} = \mathbf{E}[T_{in}] / \mathbf{E}[T]$. An appeal to PASTA property and Wald's identity is also needed.

2.3. The busy period

Let assume that all sources are free at time $t = 0$ and one of them just generates a request for service. Then, a busy period L starts. The busy period concludes at the first service completion epoch at which $(C(t), N(t))$ returns to the state $(0, 0)$. The busy period of the $M/G/1//K$ retrial queue was studied by Falin and Artalejo (1998). Their results can be summarized as follows.

We consider the following transient taboo probabilities and density probabilities

$$P_{0n}(t) = \mathbf{P}\{L > t, C(t) = 0, N(t) = n\}, \qquad 1 \leq n \leq K-1, \tag{26}$$

$$P_{1n}(t,x) = \frac{d}{dx}\mathbf{P}\{L > t, C(t) = 1, \xi(t) \leq x, N(t) = n\}, \quad 0 \leq n \leq K-1, \tag{27}$$

$$P_{1n}(t) = \int_0^\infty P_{1n}(t,x)dx, \qquad 0 \leq n \leq K-1. \tag{28}$$

By writing the differential equations that govern the motion of these taboo probabilities and using again the transformation (7), we can find the following explicit expressions for the Laplace transform $\mathbf{E}\left[e^{-sL}\right]$ and the first moments of L:

$$\begin{aligned} \mathbf{E}\left[e^{-sL}\right] &= (\beta(s) - (s + ((K-1)\mu + \alpha)(1 - \beta(s)))B_0(s) \\ &- (\alpha - \mu)(1 - \beta(s))B_1(s)) \\ &\times (1 + (s + ((K-1)\mu + \alpha)(1 - \beta(s)))A_0(s) \\ &+ (\alpha - \mu)(1 - \beta(s))A_1(s))^{-1}. \end{aligned} \tag{29}$$

The coefficients $A_i(s)$ and $B_i(s)$, $i \in \{0, 1\}$, can be determined from the recursive equations:

$$A_{K-1}(s) = 0, \qquad B_{K-1}(s) = 0,$$
$$A_{K-2}(s) = (\mu\beta(s + (K-1)\alpha))^{-1}, \qquad B_{K-2}(s) = -\mu^{-1},$$
$$u_m(s)A_m(s) + v_m(s)A_{m+1}(s) - w_m(s)A_{m-1}(s) = -\binom{K-1}{m},$$
$$m = K-2, \ldots, 1 \tag{30}$$
$$u_m(s)B_m(s) + v_m(s)B_{m+1}(s) - w_m(s)B_{m-1}(s) = \binom{K-1}{m}\beta(s + m\alpha),$$
$$m = K-2, \ldots, 1$$

where

$$\begin{aligned} u_m(s) &= s + ((K-m-1)\mu + (m+1)\alpha)(1 - \beta(s+m\alpha)) \\ &+ m\mu\beta(s+m\alpha), \qquad (31) \\ v_m(s) &= (m+1)(\alpha - \mu)(1 - \beta(s+m\alpha)), \\ w_m(s) &= (K-m)\mu\beta(s+m\alpha). \end{aligned}$$

$$\begin{aligned} \mathbf{E}[L] &= \beta_1 + (1 + \beta_1((K-1)\mu + \alpha))(A_0(0) + B_0(0)) \\ &+ (\alpha - \mu)\beta_1(A_1(0) + B_1(0)), \qquad (32) \\ \mathbf{E}[L^2] &= \beta_2 + 2\mathbf{E}[L]\,((1 + \beta_1((K-1)\mu + \alpha))A_0(0) \\ &+ (\alpha - \mu)\beta_1 A_1(0)) \\ &+ \beta_2((K-1)\mu + \alpha)(A_0(0) + B_0(0)) \\ &+ (\alpha - \mu)\beta_2(A_1(0) + B_1(0)) \\ &- 2(1 + \beta_1((K-1)\mu + \alpha))(A_0'(0) + B_0'(0)) \\ &- 2(\alpha - \mu)\beta_1(A_1'(0) + B_1'(0)). \qquad (33) \end{aligned}$$

The quantities $A_i(0)$ and $B_i(0)$, $i \in \{0, 1\}$, follow from equations (30) by putting $s = 0$. Finally, $A_i'(0)$ and $B_i'(0)$, $i \in \{0, 1\}$, can be found with the help of the following system of recursive equations

$$
\begin{aligned}
& A'_{K-1}(0) = 0, \qquad B'_{K-1}(0) = 0, \\
(34)\quad & A'_{K-2}(0) = -\beta'((K-1)\alpha)(\mu\beta((K-1)\alpha))^{-1}, \qquad B'_{K-2}(0) = 0, \\
& u_m(0)A'_m(0) + v_m(0)A'_{m+1}(0) - w_m(0)A'_{m-1}(0) \\
= \;& (m+1)(\alpha-\mu)\beta'(m\alpha)A_{m+1}(0) \\
- \;& (1-\beta'(m\alpha)((K-1-2m)\mu + (m+1)\alpha))A_m(0) \\
+ \;& (K-m)\mu\beta'(m\alpha)A_{m-1}(0), \quad m = K-2,\dots,1 \\
& u_m(0)B'_m(0) + v_m(0)B'_{m+1}(0) - w_m(0)B'_{m-1}(0) \\
= \;& (m+1)(\alpha-\mu)\beta'(m\alpha)B_{m+1}(0) \\
- \;& (1-\beta'(m\alpha)((K-1-2m)\mu + (m+1)\alpha))B_m(0) \\
+ \;& (K-m)\mu\beta'(m\alpha)B_{m-1}(0) \\
+ \;& \binom{K-1}{m}\beta'(m\alpha), \qquad m = K-2,\dots,1.
\end{aligned}
$$

2.4. The model with server vacations

In this section we consider the modification of the main model in which the server operates according to a "starting vacation" strategy; i.e., when the server is idle at the arriving epoch of a primary or returning source, it either starts its service time (with probability α_k) or takes a vacation (with probability $\overline{\alpha}_k = 1 - \alpha_k$).The recovery probabilities α_k depend on the number of sources in the orbit at arrival time and excludes to the arriving source if it comes from the orbit. The system description given in section 2.1 must be modified as follows. Now $C(t)$ takes values on $\{0,1,2\}$. The case $C(t) = 2$ means that the server is on vacation at time t. Successive server vacation times are independent random variables with probability distribution function $V(x)$ and first moment v. The independence among the vacation times and the rest of system components is assumed.

This model was considered by Li and Yang (1995) who investigated the limiting distribution following the method of supplementary variables and the method of characteristics for solving quasi-linear partial differential equations. They showed that the limiting probabilities $\{p_{0n}\}_{n=0}^{K}$ and probability densities $\{p_{1n}(x)\}_{n=0}^{K-1}$ and $\{p_{2n}(x)\}_{n=1}^{K}$ (the continuous parameter x denotes either the elapsed time of the source being

served (if $C(t) = 1$) or the elapsed time of the vacation period in progress at time t (if $C(t) = 2$)) are given by

(35)
$$p_{00} = \left(\sum_{n=0}^{K} (1 + \alpha(K-n)(\alpha_n \beta_1 + \overline{\alpha}_n \upsilon) + n\mu(\alpha_{n-1}\beta_1 + \overline{\alpha}_{n-1}\upsilon)) Q_n \right)^{-1},$$

(36)
$$p_{0n} = Q_n p_{00}, \qquad 1 \leq n \leq K,$$

(37)
$$\begin{aligned} p_{1n}(x) &= (1 - B(x)) \sum_{k=0}^{n} \left(\binom{K-k-1}{K-n-1} e^{-(K-n-1)\alpha x} (1 - e^{-\alpha x})^{n-k} \right. \\ &\quad \left. \times (\alpha(K-k)Q_k + (k+1)\mu\, Q_{k+1})\alpha_k \right) p_{00}, \\ &\quad 0 \leq n \leq K-1, \end{aligned}$$

(38)
$$\begin{aligned} p_{2n}(x) &= (1 - V(x)) \sum_{k=0}^{n} \left(\binom{K-k}{K-n} e^{-(K-n)\alpha x} (1 - e^{-\alpha x})^{n-k} \right. \\ &\quad \left. \times (\alpha(K-k+1)Q_{k-1} + k\mu\, Q_k)\overline{\alpha}_{k-1} \right) p_{00}, \\ &\quad 1 \leq n \leq K, \end{aligned}$$

where $\{Q_i\}_{i=0}^{K}$ satisfies the following recursion

(39)
$$\begin{aligned} Q_{i+1} &= \frac{1}{q_{i+1,i}} \sum_{k=0}^{i} \sum_{j=i+1}^{K} q_{kj} Q_k, \qquad 0 \leq i \leq K-1, \qquad Q_0 = 1, \\ q_{ki} &= a_{ki}\alpha_k(K-k)\alpha + a_{k-1,i}\alpha_{k-1}k\mu + b_{k+1,i}\overline{\alpha}_k(K-k)\alpha + b_{ki}\overline{\alpha}_{k-1}k\mu, \\ &\quad 0 \leq k \leq i+1, \\ a_{ki} &= \begin{cases} \binom{K-k-1}{K-i-1} \int_0^\infty e^{-(K-i-1)\alpha x}(1 - e^{-\alpha x})^{i-k} dB(x) & \text{if } 0 \leq k \leq i < K, \\ 0, & \text{otherwise,} \end{cases} \\ b_{ki} &= \begin{cases} \binom{K-k}{K-i} \int_0^\infty e^{-(K-i)\alpha x}(1 - e^{-\alpha x})^{i-k} dV(x) & \text{if } 0 \leq k \leq i \leq K, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

The major performance characteristics can be expressed as follows:

1. *The server utilization*

$$p_1 = \beta_1 p_{00} \sum_{n=0}^{K} (\alpha_n \alpha(K-n) + \alpha_{n-1} n\mu) Q_n. \tag{40}$$

2. *The mean number of sources in the orbit*

$$\mathbf{E}[N(t)] + p_1 = K - \alpha^{-1} \nu p_1. \tag{41}$$

3. *The mean total time spent in the system*

$$\mathbf{E}[W] + \beta_1 = (\nu p_1)^{-1} K - \alpha^{-1}. \tag{42}$$

2.5. Approximations and numerical results

It is clear that real teletraffic and computer systems often do not satisfy assumptions made by classical queueing models. As a consequence, a possible alternative based on information theoretic techniques has been successfully developed since the mid-60s. The principles of maximum entropy and minimum cross-entropy provide a new and powerful framework for the approximate analysis of queueing systems.

Artalejo and Gomez-Corral (1995) applied the maximum entropy principle to $M/G/1//K$ retrial queues. Two different retrial policies were considered. In Model I any source finding the server busy joins the orbit. After an exponentially distributed time with rate μ, and independently of each other source, the marked one tries his luck again. Thus this policy agrees with the classical description given in section 2.1. The description of the repeated attempts in Model II is collective. If $N(t) > 0$ then the next attempt to get service is exponentially distributed with rate μ independently of the orbit size. This second retrial policy arises in computer networks where the server must check if the transmission medium is or not available, or in models where the server is required to search for customers.

Suppose that the available information about the queueing system places a number of constraints. Then the principle of maximum entropy gives a method for computing a unique estimate for the unknown probability distribution $\{p_n / \ n \in S\}$. To that end, the principle states that, of all distributions satisfying the constraints, the minimally prejudiced distribution is the one that maximizes Shannon's entropy

$$H(p) = -\sum_{n \in S} p_n \ln p_n. \tag{43}$$

It is assumed that the constraints can be expressed in terms of mean value formulae of the form

$$\sum_{n \in S} p_n = 1, \tag{44}$$

$$\sum_{n \in S} f_k(n) p_n = F_k, \qquad 1 \leq k \leq m. \tag{45}$$

The maximization of Shannon's functional can be carried out using Lagrange's method of undetermined multipliers leading to the solution

$$\begin{aligned} \widehat{p}_n &= \exp\left\{-\beta_0 - \sum_{k=1}^{m} f_k(n)\beta_k\right\}, \qquad n \in S, \\ \exp\{\beta_0\} &= \sum_{n \in S} \exp\left\{-\sum_{k=1}^{m} f_k(n)\beta_k\right\}, \end{aligned} \tag{46}$$

where β_k, $1 \leq k \leq m$, are the Lagrangian multipliers corresponding to the set of mean value constraints (45), and β_0 is associated to the normalization constraint (44).

Artalejo and Gomez-Corral (1995) discussed the maximum entropy solution of $M/G/1//K$ retrial queues when the first moments of the stationary distribution and the equality relation

$$n\mu p_{0n} = \alpha(K-n)p_{1,n-1}, \qquad 1 \leq n \leq K-1, \tag{47}$$

are simultaneously considered as possible constraints.

The accuracy of the maximum entropy approach was tested for several service time distributions and compared with the results obtained from the classical queueing methodology. The use of the first two moments of sequences $\{p_{in}\}_{n=0}^{K-1}$, for $i \in \{0,1\}$, and the auxiliary constraint (47) leads to very good levels of accuracy.

Numerical results showing how the system performance measures are affected by the system parameters are given in the papers by Ohmura and Takahashi (1985) and Falin and Artalejo (1997). In these papers, the

interested reader may found a variety of figures and tables illustrating the effect of the system parameters (α, μ, β_1 and K) on the mean waiting time, the server utilization, the mean rate of generation of primary arrivals and other performance characteristics. In addition, Ohmura and Takahashi (1985) developed some simulation results for analyzing the relation between $\mathbf{E}[W]$and $\overline{\lambda}\beta_1$ when the retrial times are not exponentially distributed.

3. The $M/M/c//K$ retrial queue

3.1. Model description

We consider a retrial queueing system with c servers where primary calls are generated by $K > c$ sources according to a quasirandom input with rate α. If all servers are occupied at time of arrival of a primary call, then the source joins the orbit with probability H_1 and produces a Poisson flow of repeated attempts with rate μ. If all servers are busy at time of a retrial time completion then the source decides to retry again with probability H_2, otherwise it leaves the system. The service times are exponentially distributed with rate ν (without loss of generality we may assume $\nu = 1$). The independence among primary calls, retrial and service times is assumed.

The bivariate process $(C(t), N(t))$ describing the system state is now Markovian with state space $S = \{0, \ldots, c\} \times \{0, \ldots, M\}$, where $M = K - c$. The elements of its infinitesimal generator are given by

(48)
$$q_{(i,j)(n,m)} = \begin{cases} (K-i-j)\alpha, & \text{if } (n,m) = (i+1, j), \\ i, & \text{if } (n,m) = (i-1, j), \\ j\mu, & \text{if } (n,m) = (i+1, j-1), \\ -((K-i-j)\alpha + i + j\mu), & \text{if } (n,m) = (i,j), \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{if } 0 \le i \le c-1,$$

(49)
$$q_{(c,j)(n,m)} = \begin{cases} (M-j)\alpha H_1, & \text{if } (n,m) = (c, j+1), \\ c, & \text{if } (n,m) = (c-1, j), \\ j\mu(1-H_2), & \text{if } (n,m) = (c, j-1), \\ -((M-j)\alpha H_1 + c + j\mu(1-H_2)), & \text{if } (n,m) = (c,j), \\ 0, & \text{otherwise.} \end{cases}$$

3.2. Joint distribution of the server state and the orbit length in steady state

The $M/M/c//K$ retrial queue with persistent subscribers (when $H_1 = H_2 = 1$) was introduced by Kornyshev (1969) who investigated the limiting probabilities and obtained some useful relationships among the system performance measures. Later Falin and Templeton (1997) extended the analysis and studied the waiting time. The model under consideration including non-persistent sources was investigated by Falin (1998).

The limiting probabilities satisfy the following set of equations (below p_{ij} equals 0 if $(i,j) \notin S$):

$$\begin{aligned} & ((K-i-j)\alpha + i + j\mu)p_{ij} \\ (50) \quad = \; & (K-i+1-j)\alpha p_{i-1,j} + (j+1)\mu p_{i-1,j+1} + (i+1)p_{i+1,j}, \\ & 0 \le i \le c-1, \end{aligned}$$

$$\begin{aligned} & ((M-j)\alpha H_1 + c + j\mu(1-H_2))p_{cj} \\ (51) \quad = \; & (M+1-j)\alpha p_{c-1,j} + (j+1)\mu p_{c-1,j+1} \\ + \; & (M-j+1)\alpha H_1 p_{c,j-1} + (j+1)\mu(1-H_2)p_{c,j+1}. \end{aligned}$$

Falin (1998) proposed an algorithm for computing recursively the probabilities p_{ij}. First, we introduce new unknowns $r_{ij} = p_{ij}/p_{0M}$, for $(i,j) \in S$. If we could find r_{ij} then p_{ij} could be computed by means of $p_{ij} = r_{ij}\left(\sum_{i=0}^{c}\sum_{j=0}^{M} r_{ij}\right)^{-1}$. The problem is reduced to find the solution of a tridiagonal set of linear equations. It can be solved by using the method of "forward elimination, backward substitution" which reduces

the original system to a triangular one. Finally, we avoid subtractions by introducing appropriate new variables.

Then, the system quality measures can be expressed in terms of the limiting probabilities as follows:

1. *The mean number of sources of repeated calls*

$$N = \mathbf{E}[N(t)] = \sum_{i=0}^{c} \sum_{j=0}^{M} j p_{ij}. \tag{52}$$

2. *The probability of all servers busy*

$$p_c = \mathbf{P}\{C(t) = c\} = \sum_{j=0}^{M} p_{cj}. \tag{53}$$

3. *The mean number of busy servers*

$$Y = \mathbf{E}[C(t)] = \sum_{i=0}^{c} \sum_{j=0}^{M} i p_{ij}. \tag{54}$$

4. *The mean rate of generation of primary calls*

$$\overline{\lambda} = \alpha \mathbf{E}[K - C(t) - N(t)] = \alpha(K - Y - N). \tag{55}$$

5. *The probability of losing a primary call*

$$L = 1 - Y/\overline{\lambda}. \tag{56}$$

6. *The blocking fraction of primary calls*

$$B_A = \frac{Mp_c - N_c}{K - Y - N}, \qquad \text{where } N_c = \sum_{j=0}^{M} j p_{cj}. \tag{57}$$

7. *The blocking fraction of repeated attempts*

$$B_R = N_c/N. \tag{58}$$

8. *The global blocking probability*

$$B = \frac{\alpha M p_c + (\mu - \alpha) N_c}{\alpha(K - Y - N) + \mu N}. \tag{59}$$

The particular case $H_1 = H_2 = 1$ was investigated by Kornyshev (1969) by using a different methodology based on the following series of transformations of the limiting probabilities:

$$p_{i,j-1}(q) = j p_{ij}(q-1) \left(\sum_{i=0}^{c} \sum_{j=0}^{M-q} j p_{ij}(q-1) \right)^{-1}, \tag{60}$$

$$0 \le i \le c,\ 0 \le j \le M - q,\ 1 \le q \le M.$$

Kornyshev (1969) also showed how the major performance characteristics (mean number of sources in orbit, mean waiting time, the probability of all servers busy, the blocking fractions of primary calls and repeated attempts) can be computed from transforms (60).

In the case $c = 1$ and $\alpha \neq \mu$ Falin (1998) found the following explicit expressions for the joint distribution of the pair $(C(t), N(t))$:

$$p_{0n} = \frac{\gamma^n}{n!} \prod_{i=0}^{n-1} \frac{(a+i)(b+i)}{c+i} \left(F(a,b;c;\gamma) + K\alpha F(a+1,b;c;\gamma) \right)^{-1}, \tag{61}$$

$$\begin{aligned} p_{1n} &= \frac{\gamma^n}{n!} \prod_{i=0}^{n-1} \frac{(a+1+i)(b+i)}{c+i} K\alpha (F(a,b;c;\gamma) \\ &+ K\alpha F(a+1,b;c;\gamma))^{-1}, \end{aligned} \tag{62}$$

where

$$a = \frac{K\alpha}{\mu - \alpha}, \qquad b = -K + 1,$$

$$c = \frac{1 + (1 - H_2)((K-1)\alpha + \mu)}{(1 - H_2)(\mu - \alpha)}, \qquad \gamma = -\frac{\alpha H_1}{\mu(1 - H_2)}$$

and $F(a,b;c;x)$ is the hypergeometric function defined by

$$F(a,b;c;x) = \sum_{j=0}^{\infty} \frac{x^j}{j!} \prod_{i=0}^{j-1} \frac{(a+i)(b+i)}{c+i}.$$

3.3. Waiting time

Assume that some fixed source i_0 places a primary request for service at time $t = 0$. Then its virtual waiting time W just starts and ends at the time at which the source starts to be served or decides to leave the system. Falin (1998) and Falin and Templeton (1997) gave a method for deriving the distribution of W and its expected value.

Suppose that at time $t = 0$ there are j sources of repeated calls and i sources are receiving service. Then we mark one of the sources in orbit and denote by τ_{ij} its residual waiting time. Let

$$f_{ij}(t) = \mathbf{P}\{\tau_{ij} < t, \mathcal{F} = S\}, \qquad g_{ij}(t) = \mathbf{P}\{\tau_{ij} < t, \mathcal{F} = L\}, \tag{63}$$

where the event $\{\mathcal{F} = S\}$ (respectively $\{\mathcal{F} = L\}$) denotes that the marked source is accepted for service (respectively is not served).

Then, the waiting time distribution function is given by

$$\mathbf{P}\{W < t, \mathcal{F} = S\} = \sum_{i=0}^{c-1}\sum_{j=0}^{M}\pi_{ij} + \sum_{j=0}^{M-1}\pi_{cj}H_1 f_{c,j+1}(t), \tag{64}$$

$$\mathbf{P}\{W < t, \mathcal{F} = L\} = \sum_{j=0}^{M-1}\pi_{cj}((1 - H_1) + H_1 g_{c,j+1}(t)), \tag{65}$$

where π_{ij} denotes the probability that the marked source finds the system in the state (i, j) upon arrival. With the help of PASTA property π_{ij} can be reduced to the limiting distribution as follows

$$\pi_{ij} = \frac{\alpha(K - i - j)}{\overline{\lambda}} p_{ij}. \tag{66}$$

The analysis of probabilities $f_{ij}(t)$ and $g_{ij}(t)$ can be done by introducing an auxiliary Markov process $Z(t)$ with state space $S \cup \{\mathcal{S}\} \cup \{\mathcal{L}\}$. The special states $\mathcal{S}$ and $\mathcal{L}$ are absorbing states. A transition to them means that the event $\{\mathcal{F} = S\}$ or $\{\mathcal{F} = L\}$ has occurred. The analysis of Kolmogorov's backward equations for the Markov chain $Z(t)$ leads to the following solution

$$\mathbf{P}\{W < t, \mathcal{F} = S\} = 1 - L - \frac{1}{\bar{\lambda}} \sum_{i=0}^{c} \sum_{j=1}^{M} j p_{ij} f'_{ij}(t), \tag{67}$$

$$\mathbf{P}\{W < t, \mathcal{F} = L\} = L - \frac{1}{\bar{\lambda}} \sum_{i=0}^{c} \sum_{j=1}^{M} j p_{ij} g'_{ij}(t). \tag{68}$$

It should be noted that the above expressions reduce the calculation of the nth moment of W to find the $(n-1)$th moments of the conditional waiting times. This fact was exploded by Falin and Artalejo (1998) for the single server queue with persistent subscribers. They obtained the following results:

$$\mathbf{E}[W^n] = \delta_{n0} + \frac{n}{\bar{\lambda}} \sum_{i=0}^{1} \sum_{j=0}^{K-1} j p_{ij} \mathbf{E}\left[\tau_{ij}^{n-1}\right], \qquad n \geq 0. \tag{69}$$

We now denote $\mathbf{E}\left[\tau_{0j}^n\right]$ by $a_j^{(n)}$ and $\mathbf{E}\left[\tau_{1j}^n\right]$ by $b_j^{(n)}$. These moments satisfy the following set of equations for $n \geq 1$ and $1 \leq j \leq K-1$:

$$-((K-j)\alpha + j\mu)a_j^{(n)} + (j-1)\mu b_{j-1}^{(n)} + (K-j)\alpha b_j^{(n)} = -n a_j^{(n-1)}, \tag{70}$$

$$-((K-j-1)\alpha + 1)b_j^{(n)} + a_j^{(n)} + (K-j-1)\alpha b_{j+1}^{(n)} = -n b_j^{(n-1)}. \tag{71}$$

Eliminating $a_j^{(n)}$ from these relations we get

$$\begin{aligned} & n(a_j^{(n-1)} + ((K-j)\alpha + j\mu)b_j^{(n-1)}) \\ = \; & -((K-j)\alpha + j\mu)(K-j-1)\alpha b_{j+1}^{(n)} \\ & -(j-1)\mu b_{j-1}^{(n)} + (((K-j)\alpha + j\mu)(K-j-1)\alpha + j\mu)b_j^{(n)}, \\ & n \geq 1, \qquad 1 \leq j \leq K-1. \end{aligned} \tag{72}$$

The above set of equations has the form

$$-\alpha_{j-1}\chi_{j-1} + \beta_j \chi_j - \gamma_j \chi_{j+1} = \delta_j, \qquad 1 \leq j \leq K-1, \tag{73}$$

where

$$
\begin{aligned}
&\chi_0 = \chi_K = 0, \qquad \alpha_j = j\mu, \\
&\beta_j = ((K-j)\alpha + j\mu)(K-j-1)\alpha + j\mu, \\
&\gamma_j = ((K-j)\alpha + j\mu)(K-j-1)\alpha, \\
&\delta_j = n(a_j^{(n-1)} + ((K-j)\alpha + j\mu)b_j^{(n-1)}).
\end{aligned} \tag{74}
$$

The method of "forward elimination, backward substitution" is again the key for solving the set of linear equations (73). Following this method we first calculate variables B_j and D_j according to the recursive formulae:

$$
\begin{aligned}
B_1 &= \beta_1, \quad B_j = \beta_j - (B_{j-1})^{-1}\alpha_{j-1}\gamma_{j-1}, \quad 2 \le j \le K-1, \\
D_1 &= \delta_1, \quad D_j = \delta_j + (B_{j-1})^{-1}\alpha_{j-1}D_{j-1}, \quad 2 \le j \le K-1.
\end{aligned} \tag{75}
$$

χ_j can be recursively computed in reverse order. Note that $\beta_j = \alpha_j + \gamma_j$ and the sequence α_j is increasing, so it is convenient to introduce an auxiliary variable $Z_j = B_j - \gamma_j$.

Then, we obtain

$$
\begin{aligned}
&Z_1 = \alpha_1, \quad D_1 = \delta_1, \\
&Z_j = (Z_{j-1} + \gamma_{j-1})^{-1}(\alpha_j Z_{j-1} + (\alpha_j - \alpha_{j-1})\gamma_{j-1}), \quad 2 \le j \le K-1, \\
&D_j = \delta_j + (Z_{j-1} + \gamma_{j-1})^{-1}\alpha_{j-1}D_{j-1}, \qquad 2 \le j \le K-1, \\
&\chi_{K-1} = (Z_{K-1} + \gamma_{K-1})^{-1}D_{K-1}, \\
&\chi_j = (Z_j + \gamma_j)^{-1}(D_j + \gamma_j\chi_{j+1}), \qquad j = K-2, \ldots, 1
\end{aligned} \tag{76}
$$

so we only deal with positive numbers.

3.4. Miscellaneous

Numerical examples showing the influence of the system parameters (α, μ) on the main performance measures $(\overline{\lambda}, B_A, B_R, L, \mathbf{E}[W])$ can be found in Kornyshev (1969), Falin (1998) and Falin and Artalejo (1998).

The waiting time process was also studied by Dragieva (1994). However, only expected characteristics were investigated and some results seem to be incorrect.

Recently, Artalejo et al. (1997) described a versatile finite retrial queue in a Markovian environment that covers as special cases a wide variety of queueing phenomena. Its infinitesimal generator can be reduced to a finite block-tridiagonal one so well-known matrix methods

are the key to investigate the limiting distribution and the first passage times. Models with quasirandom input are also included in this framework.

4. Applications

4.1. Analysis of subscribers' behaviour in telephone networks

The pioneering paper by Kornyshev (1969) was motivated by the analysis of the behaviour of subscribers in real telephone networks. It was evident that the classical models of telephone systems (waiting lines with infinite capacity and queues with losses) did not take into account the existence of a real flow of repeated calls. The main reason for getting a blocked signal when a subscriber calls is to find all trunks busy. In this sense, the systems with repeated attempts provided an elegant alternative to understand the retrial phenomena. In this context, Kornyshev's paper was the first attempt for investigating the performance characteristics of servicing systems with an accessible bunch of c lines and a finite population of subscribers.

4.2. Magnetic disk memory systems

Ohmura and Takahashi (1985) described an application of the $M/G/1//K$ retrial queue to the waiting time analysis of magnetic disk memory systems. They consider a memory system where K disk units share a disk controller (server) and transmit information when they find the controller idle. Unsatisfied requests are repeated after a disk's rotation which can be modelled as a constant repetition interval. But, as simulation shows, an exponential approximation of rotation interval is suitable in practice.

4.3. Local area networks with CSMA/CD protocol

Li and Yang (1995) applied the $M/G/1//K$ with retrials and server vacations to the study of a local area network operating under the communication protocol "Carrier Sense Multiple Access with Collision Detection" (CSMA/CD). In these networks, a finite number K of users are connected by a bus (server). The rules governing the transmission of information in the system, in principle, match the $M/G/1//K$ description. In addition, a phenomenon called "collision" is also considered. Suppose that an user senses the channel and finds a free signal. Then

the transmission starts but, due to non-zero propagation delay, during a certain amount of time $\mathcal{T}$ any other users may sense the channel and transmit their messages. In such a case a collision occurs. The user which transmission time is in progress joins the retrial group and the rest of users involved in the collision remain in their previous states as the collision had not occurred. When a collision occurs a recovery time is needed by the channel to be free again. Li and Yang (1995) modelled the recovery time as a deterministic vacation period $3\mathcal{T}$, where $\mathcal{T}$ is the time required by a signal to travel from one extreme of the channel to the other.

The interested reader is also referred to Khomichkov (1993, 1995) where the stationary characteristics of local area networks with protocol CSMA/CD are investigated. These papers study interesting models where more complex assumptions with regard to the time of occupation, the recovery time and other system devices have been considered.

4.4. Collision avoidance local area networks

An usual feature in local area networks is that several stations use a common medium for transmission so collisions among messages occur. These collisions imply the destruction of information and consequently the performance quality decreases. To avoid this problem a number of collision-avoidance local area networks have been developed (bus topology, star topology, etc.). Janssens (1997) described a local area network consisting of K network access controllers and the hub (server). To solve the problem of possible collisions among packets coming from different controllers, the network is modelled as the $M/G/1//K$ retrial queue and the analysis is based on the regenerative approach introduced by De Kok (1984).

ACKNOWLEDGEMENTS. The work of the author is supported by the DGICYT PB95-0416 and by the European Commission under INTAS 96-0828.

References

[1] Artalejo, J. R. and Gomez-Corral, A., *Information theoretic analysis for queueing systems with quasi-random input*, Mathematical and Computer Modelling **22** (1995), 65-76.

[2] ——, *A. On the $M/G/1$ queue with negative arrivals and request repeated*, Dept. de Estadistica e I. O., UCM, Working Paper, 1996.

[3] Artalejo, J. R., Rajagopal, V. and Sivasamy, R., *A note on finite Markovian queues with repeated attempts*, Dept. de Estadistica e I. O., UCM, Working Paper, 1997.

[4] Choi, B. D. and Chang, Y. $MAP_1, MAP_2/M/c$ *retrial queue with the retrial group of finite capacity and geometric loss*, Mathematical and Computer Modelling, (to appear).

[5] Dragieva, V. I., *Single-line queue with finite source and repeated calls*, Problems of Information Transmission **30** (1994), 283-289.

[6] Dudin, A. N. and Klimenok, V. *Queueing system* $BMAP/G/1$ *with repeated calls*, Mathematical and Computer Modelling, (to appear).

[7] Falin, G. I. and Artalejo, J. R., *A finite source retrial queue*, European Journal of Operational Research, **108** (1998), 409-424.

[8] Falin, G. I. and Templeton, J. G. C., *Retrial Queues*, Chapman and Hall, London, 1997.

[9] Falin, G. I. *A multiserver retrial queue with finite number of sources of primary calls*, Mathematical and Computer Modelling, (to appear).

[10] Jaiswal, N. K., *Priority Queues*, Academic Press, New York, 1968.

[11] Janssens, G. K., *The quasi-random input queueing system with repeated attempts as a model for a collision-avoidance star local area network*, IEEE Transactions on Communications **45** (1997), 360-364.

[12] Khomichkov, I. I., *Study of models of local area networks with multiple access protocols*, Automation and Remote Control **54** (1993), 1801-1811.

[13] ——, *Calculation of the characteristics of local area network with p-persistent protocol of multiple random access*, Automation and Remote Control **56** (1995), 208-218.

[14] Kok, A. G. de, *Algorithmic methods for single server systems with repeated attempts*, Statistica Neerlandica **38** (1984), 23-32.

[15] Kornyshev, Y. N., *Design of a fully accessible switching system with repeated calls*, Telecommunications **23** (1969), 46-52.

[16] Li, H. and Yang, T., *A single-server retrial queue with server vacations and a finite number of input sources*, European Journal of Operational Research **85** (1995), 149-160.

[17] Schellhaas, H., *Computation of the state probabilities in a class of semi-regenerative queueing models*, in: Semi-Markov Models: Theory and Applications (Jansen, J. ed.), Plenum Press, pp. 111-130, 1986.

[18] Ohmura, H. and Takahashi, Y., *An analysis of repeated call model with a finite number of sources*, Electronics and Communications in Japan **68** (1985), 112-121.

[19] Takagi, H., *Queueing Analysis*, vol. 2, *Finite Systems*, North-Holland, Amsterdam, 1993.

Departamento de Estadistica e I. O.
Facultad de Matematicas
Universidad Complutense de Madrid
28040 Madrid, Spain
E-mail: jesus_artalejo@mat.ucm.es