ELSEVIER

European Journal of Operational Research 135 (2001) 569–581

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

Analysis of multiserver queues with constant retrial rate

J.R. Artalejo [a,*], A. Gómez-Corral [a], M.F. Neuts [b]

[a] *Department of Statistics and Operations Research, Faculty of Mathematics, University Complutense of Madrid, Madrid 28040, Spain*
[b] *Department of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA*

Received 26 January 1999; accepted 20 November 2000

Abstract

We consider multiserver retrial queues in which the time between two successive repeated attempts is independent of the number of customers applying for service. We study a Markovian model where each arriving customer finding any free server either enters service or leaves the service area and joins a pool of unsatisfied customers called 'orbit'. This system is analyzed as a quasi-birth-and-death (QBD) process and its main performance characteristics are efficiently computed.

Keywords: Queueing; Quasi-birth-and-death processes; Performance analysis; Retrial queues

1. Introduction

We consider a multiserver queueing system in which primary customers arrive according to a Poisson stream of rate $\lambda > 0$. The service facility consists of c identical servers and customer service times are independent and exponentially distributed with mean $1/\mu$. An arriving customer finding one or more servers idle either obtains service immediately (with probability p) or joins the orbit (with probability $q = 1 - p$). Customers who find all servers busy go directly to the orbit. The recovery probability p allows us to consider simultaneously two possibilities described in the existing literature. Most papers [2,9,13,14,24,25] consider the case $p = 1$ where customers have direct access to the idle servers. On the other hand, Neuts and Ramalhoto [20] and Neuts and Rao [21] introduce retrial queues where $p = 0$, i.e., arriving customers always enter the orbit and start generating requests for service. Pearce [22] treats both models in a unified way by introducing the recovery probability p.

The pioneering studies of retrial queues [9,13,25] present the concept of 'retrial time' as an alternative to the classical models of telephone systems, queues with losses, that do not take repeated calls into account. In this context each blocked call generates a stream of repeated requests independently of the rest of customers in the retrial group. Thus, in the *classical retrial policy* the intervals between successive repeated

* Corresponding author. Fax: 34-913944607.
E-mail address: jesus_artalejo@mat.ucm.es (J.R. Artalejo).

0377-2217/01/$ - see front matter
PII: S0377-2217(00)00330-1

attempts are exponentially distributed with rate $\sigma_i = i\sigma$, when the orbit size is i. However, recent applications to communication protocols and local area networks show that there are queueing situations in which the retrial rate is independent of the number of customers (if any) in orbit, i.e., the retrial rate is $\sigma_i = \sigma(1 - \delta_{0i})$, where δ_{ab} denotes Kronecker delta. This *constant retrial policy* was introduced by Fayolle [14], who modeled a telephone exchange system. Later it was used for the stability of the *ALOHA* protocol [7] and unslotted *CSMA/CD* (Carrier Sense Multiple Access with Collision Detection) protocol [5] in communication systems. Artalejo and Gómez-Corral [3] combined both policies by defining a *linear retrial policy* with retrial rate $\sigma_i = \alpha(1 - \delta_{0i}) + i\sigma$.

Multiserver retrial queues have been analyzed under the classical retrial policy. The equilibrium distribution of the system state is expressed in contour integrals [9] or as limits of extended continued fractions [22]. From an analytical point of view, both solutions are significant attempts but practical implementation requires a variety of approximations and truncated methods. The multiserver retrial queue with retrial rate $\sigma_i = i\sigma$ can be viewed as an LDQBD process [4]. The main feature of its infinitesimal generator is the spatial heterogeneity caused by the transitions due to successful retrials. That lack of homogeneity explains the analytical difficulty of retrial queues. Several interesting papers are devoted to the approximation of the initial system. Wilkinson [25] truncates the capacity of the orbit at some value K. The resulting finite system can be solved to get the equilibrium distribution and the main performance measures. However, a direct truncation implies a large choice for K, when the level of congestion is high. This drawback can be avoided by using more sophisticated methods of truncation [24] or by imposing a simplifying assumption that yields an auxiliary queueing model with an infinite system state and a more appropriate infinitesimal generator [13,21]. For instance, Neuts and Rao [21] proposed an algorithmic solution based on an approximating multiserver queue with retrial rate $\sigma_i = \min(i, N)\sigma$ for any sufficiently large number N. This approximation leads to an infinitesimal generator which is homogeneous from the level N up. Then the queueing process is a level-independent quasi-birth-and-death (QBD) process with a large number of boundary states which the general theory of Neuts [18] can be applied.

Now we turn to multiserver retrial queues operating under the constant retrial policy. Since Fayolle [14], there has been a rapid growth in the literature [1,5–7,16]. This retrial policy is a useful device for modelling the retrial phenomenon in communication and computer networks where repeated attempts are made by processor units independently of the number of messages stored in each node of the network. An examination of the literature shows that only the case $c \leqslant 3$ and $p = 1$ has been studied. In the present paper, we use matrix-geometric methods for the $M/M/c$ retrial queue with retrial rate $\sigma_i = \sigma(1 - \delta_{0i})$ and service option upon arrival with probability p.

Some recent papers discuss related work. Falin and Artalejo [12] study a different multiserver retrial queue in which customers join a classical waiting line or the orbit depending on the number of customers in the queue. Choi et al. [6] exploit a retrial policy independent of the orbit size to consider an $M/M/1$ queue with general retrial times. Finally, there are a number of papers [8,10,11] devoted to algorithmic methods for retrial queues including the analysis of models with general interarrival and interrepetition times of the types BMAP, PH, etc.

The remainder of the paper is organized as follows. We summarize the main results for the case $c \leqslant 3$ and $p = 1$ in Section 2. In Section 3, we use results from the classical theory for QBD processes as a starting point for the performance analysis of the $M/M/c$ retrial queue with constant retrial discipline. The optimization of the retrial rate is investigated in Section 4. Section 5 deals with the case of direct access to the service facility in which some algorithmic simplifications are present. Finally, conclusions are given in Section 6.

2. The case $c \leqslant 3$ and $p = 1$

We briefly review the main results for Markovian retrial queues with direct access to the server facility and a small number of servers $c \leqslant 3$. We

focus on the existence and computation of the stationary distribution.

The system state at time t can be described by means of a bivariate process $X(t) = (N(t), C(t))$, where $N(t)$ is the number of customers in orbit and $C(t)$ is the number of busy servers. Note that $\{X(t);\ t \geqslant 0\}$ is an irreducible Markovian process with the lattice semi-strip $S = \mathbb{Z}_+ \times \{0, \ldots, c\}$ as the state space. The state space and transitions are shown in Fig. 1 for the case $c = 3$.

We define the limiting distribution

$$P_{ij} = \lim_{t\to\infty} P\{(N(t), C(t)) = (i,j)\}, \quad (i,j) \in S, \tag{2.1}$$

which for a standard Markov chain always exists.

First, we consider the case $c = 1$. Following Fayolle [14] we easily find that

$$P_{i0} = \frac{\lambda}{\lambda + \sigma(1-\delta_{0i})}(1-\rho_1)\rho_1^i, \quad i \geqslant 0, \tag{2.2}$$

$$P_{i1} = \frac{\lambda}{\mu}(1-\rho_1)\rho_1^i, \quad i \geqslant 0, \tag{2.3}$$

where $\rho_1 = \lambda(\lambda+\sigma)(\mu\sigma)^{-1}$. The system is stable if and only if $\rho_1 < 1$.

For the case $c = 2$, Artalejo [1] gives the following formulas:

$$P_{00} = \frac{2\mu\sigma(\lambda+\mu+\sigma) - \lambda(\lambda+\sigma)^2}{\sigma(2\mu(\lambda+\mu) + (\lambda+\sigma)(\lambda+2\mu))}, \tag{2.4}$$

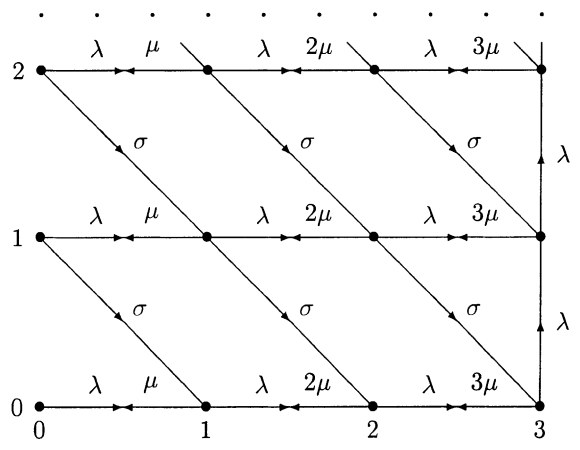

Fig. 1. State space and transitions.

$$P_{i0} = \frac{\lambda^2}{(\lambda+\sigma)^2 + \mu\sigma}\left(\frac{\lambda((\lambda+\sigma)^2+\mu\sigma)}{\mu\sigma(3\lambda+2(\mu+\sigma))}\right)^i P_{00}, \quad i \geqslant 1, \tag{2.5}$$

$$P_{i1} = \frac{\lambda + \sigma(1-\delta_{0i})}{\mu} P_{i0}, \quad i \geqslant 0, \tag{2.6}$$

$$P_{i2} = \frac{\lambda^2(\lambda+\mu+\sigma)}{\mu^2(3\lambda+2(\mu+\sigma))}\left(\frac{\lambda((\lambda+\sigma)^2+\mu\sigma)}{\mu\sigma(3\lambda+2(\mu+\sigma))}\right)^i P_{00}, \quad i \geqslant 0. \tag{2.7}$$

Now the necessary and sufficient condition for stability is $\rho_2 < 1$, where $\rho_2 = \lambda(\lambda+\sigma)^2(2\mu\sigma(\lambda+\mu+\sigma))^{-1}$. Note that when $c \leqslant 2$ the limiting probabilities are formulas of 'geometric' type. Thus, explicit expressions for the factorial moments of the number of customers in orbit can be easily given (see [1]).

If we compare these formulas with the corresponding expressions for the classical retrial policy, we observe that the constant retrial policy yields simpler analytical solutions. In fact, the solution for the classical case is given in terms of hypergeometric functions [13] instead of geometric progressions. However, the stability condition for the $M/M/c$ queue with $\sigma_i = i\sigma$ is $\lambda < c\mu$, i.e., the stability condition does not depend on the retrial parameter and agrees with the model without repeated attempts. That result is intuitive as the lengths of the idle server intervals tend to zero as $i \to \infty$. The stability conditions for models with constant retrial discipline are more interesting. All system parameters λ, μ, σ and c appear in the stability condition. In Section 3, we prove that the system with c channels is stable if and only if

$$\frac{\lambda+\sigma}{\sigma}\frac{1}{c!}\left(\frac{\lambda+\sigma}{\mu}\right)^c < \sum_{k=0}^{c}\frac{1}{k!}\left(\frac{\lambda+\sigma}{\mu}\right)^k. \tag{2.8}$$

For $c \leqslant 2$ the stationary distribution $\{P_{ij};\ (i,j) \in S\}$ is such that the partial sequences $\{P_{ij};\ i \geqslant 0\}$ satisfy a system of equations of 'birth-and-death' type. When $c > 2$, that birth-and-death structure is not preserved. Gómez-Corral and Ramalhoto [16] develop a recursive procedure for

the case $c = 3$. Under the stability condition (2.8), the stationary distribution is reduced to finding the probabilities P_{00} and P_{10} satisfying

$$P_{10} = \lim_{i\to\infty} \widetilde{\xi}_i P_{00}/\xi_i, \tag{2.9}$$

$$P_{i0} = \xi_i P_{10} - \widetilde{\xi}_i P_{00}, \quad i \geqslant 2, \tag{2.10}$$

where the coefficients $(\xi_i, \widetilde{\xi}_i)$ can be numerically computed.

3. The $M/M/c$ queue with constant retrial rate and recovery factor p

We study the stationary characteristics of the retrial queue of type $M/M/c$ with retrial rate $\sigma_i = \sigma(1 - \delta_{0i})$ and recovery factor p, described in Section 1. This model is a QBD process on the state space $S = \mathbb{Z}_+ \times \{0, \ldots, c\}$. The system state at time t consists of the number of customers in orbit and the number of busy servers as we defined in Section 2. Its infinitesimal generator Q is of the form

$$Q = \begin{pmatrix} B_0 & A_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots \\ A_2 & A_1 & A_0 & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & A_2 & A_1 & A_0 & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & A_2 & A_1 & A_0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}, \tag{3.1}$$

where all blocks are square matrices of order $c + 1$. In the case $c = 3$, the matrices A_0, A_1 and A_2 are given by

$$A_0 = \begin{pmatrix} \lambda q & 0 & 0 & 0 \\ 0 & \lambda q & 0 & 0 \\ 0 & 0 & \lambda q & 0 \\ 0 & 0 & 0 & \lambda \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0 & \sigma & 0 & 0 \\ 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & \sigma \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} -(\lambda+\sigma) & \lambda p & 0 & 0 \\ \mu & -(\lambda+\mu+\sigma) & \lambda p & 0 \\ 0 & 2\mu & -(\lambda+2\mu+\sigma) & \lambda p \\ 0 & 0 & 3\mu & -(\lambda+3\mu) \end{pmatrix}. \tag{3.2}$$

The matrix B_0 is similar to A_1 but with the retrial rate σ omitted. Note that the matrix $A = A_0 + A_1 + A_2$ is the generator of the $M/M/c/c$ loss system with arrival rate $\lambda p + \sigma$ and service rate μ. The stationary probability vector of A is given by

$$\pi = \left(\sum_{k=0}^{c} \frac{1}{k!} \left(\frac{\lambda p + \sigma}{\mu} \right)^k \right)^{-1} \left[1, \frac{\lambda p + \sigma}{\mu}, \ldots, \frac{1}{c!} \left(\frac{\lambda p + \sigma}{\mu} \right)^c \right]. \tag{3.3}$$

The general theory in [18, Theorem 3.1.1, pp. 82], states that:

(i) $\pi A_2 \mathbf{e} > \pi A_0 \mathbf{e}$ is the necessary and sufficient condition for stability. $\mathbf{e}$ denotes a column vector with all its elements equal to 1.

(ii) the stationary probability vector $\mathbf{x}$, partitioned as $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \ldots)$, of Q is given by

$$\begin{aligned} &\mathbf{x}_0(B_0 + RA_2) = \mathbf{0}, \\ &\mathbf{x}_i = \mathbf{x}_0 R^i, \quad i \geqslant 1, \\ &\mathbf{x}_0 (I - R)^{-1} \mathbf{e} = 1, \end{aligned} \tag{3.4}$$

where R is the minimal non-negative solution to the matrix equation $R^2 A_2 + RA_1 + A_0 = \mathbf{0}$ and $\mathbf{x}_i$, $i \geqslant 0$, are row vectors of dimension $c + 1$.

In practice, we would like to understand the influence of the system parameters on the main performance characteristics. For example, to plot any system descriptor versus σ we need to know the range of σ which satisfies the stability condition. This basic question cannot be solved from the general inequality $\pi A_2 \mathbf{e} > \pi A_0 \mathbf{e}$. Alternatively, after some algebra, we reexpress the stability condition as

$$\frac{\lambda p + \sigma}{c!} \left(\frac{\lambda p + \sigma}{\mu} \right)^c < (\sigma - \lambda q) \sum_{k=0}^{c} \frac{1}{k!} \left(\frac{\lambda p + \sigma}{\mu} \right)^k. \tag{3.5}$$

We may always normalize by setting $c\mu = 1$. λ is then the traffic intensity of the classical $M/M/c$.

We now fix λ. Then, as intuition tells us, there should be a stability abscissa $\sigma^*(\lambda,c,p)$ such that the stability condition is fulfilled if and only if $\sigma > \sigma^*(\lambda,c,p)$.

After some elementary algebra Eq. (3.5) reduces to finding the unique root $u^*(\lambda,c,p)$ of the polynomial $f(u) = \sum_{k=0}^{c}(c^{k-1}/k!)(k-\lambda c)u^k$ in the interval $(\lambda p, \infty)$. The coefficients of $f(u)$ have only one variation of sign so the uniqueness of $u^*(\lambda,c,p)$ follows trivially. Finally, the critical σ value is given by $\sigma^*(\lambda,c,p) = u^*(\lambda,c,p) - \lambda p$.

On the other hand, by fixing σ we expect to find a value $\lambda^*(\sigma,c,p)$ such that condition (3.5) is equivalent to $\lambda < \lambda^*(\sigma,c,p)$. By similar arguments we may show that

$$\lambda^*(\sigma,c,0) = \sum_{k=0}^{c-1}\frac{c^k}{k!}\sigma^{k+1}\left(\sum_{k=0}^{c}\frac{c^k}{k!}\sigma^k\right)^{-1} \tag{3.6}$$

when $p = 0$. If $p \neq 0$, we first compute the unique root $u^*(\sigma,c,p)$ in the interval (σ,∞) of the polynomial $g(u) = \sum_{k=0}^{c}(c^{k-1}/k!)(c\sigma - kq)u^k - (c^c/c!)u^{c+1}$. Then we have $\lambda^*(\sigma,c,p) = (u^*(\sigma,c,p) - \sigma)/p$.

The stability abscissa $\sigma^*(\lambda,c,p)$ (respectively, $\lambda^*(\sigma,c,p)$) determines the domain of σ (respectively, λ) when the rest of parameters are fixed. Thus, the computation of the stability abscissa is the first step of any numerical investigation.

Figs. 2–7 illustrate the effect on $\sigma^*(\lambda,c,p)$ and $\lambda^*(\sigma,c,p)$ of varying system parameters. In Figs. 2, 4 and 6 the value $\sigma^*(\lambda,c,p)$ is plotted versus the arrival rate λ. We have presented three curves in each figure which correspond to $c = 5$, 15 and 30. In addition, the three figures correspond to the recovery probabilities $p = 0$, 0.5 and 1, respectively. We notice that, $\lambda < c\mu$ is a necessary condition for the stability and $c\mu = 1$, so the parameter λ lies in $(0,1)$. The curves, which in decreasing order correspond to higher number of servers, show that, as is to be expected, $\sigma^*(\lambda,c,p)$ increases with increasing arrival rate λ and decreases with increasing c. A comparison among the

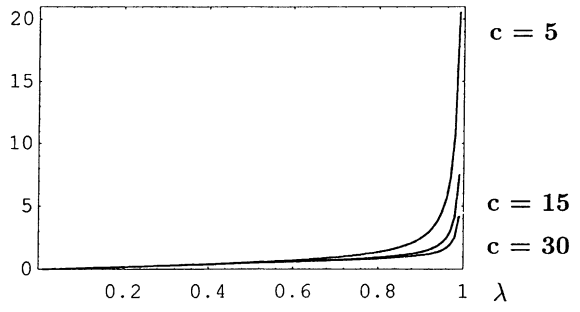

Fig. 2. $\sigma^*(\lambda,c,p)$ versus λ. Case $p = 0$, $c = 5$, 15, 30.

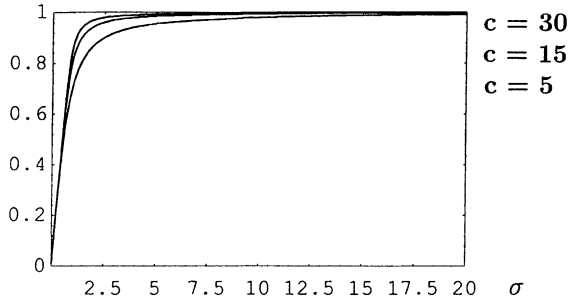

Fig. 3. $\lambda^*(\sigma,c,p)$ versus σ. Case $p = 0$, $c = 5$, 15, 30.

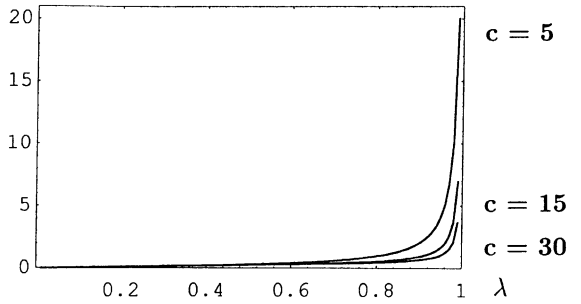

Fig. 4. $\sigma^*(\lambda,c,p)$ versus λ. Case $p = 0.5$, $c = 5$, 15, 30.

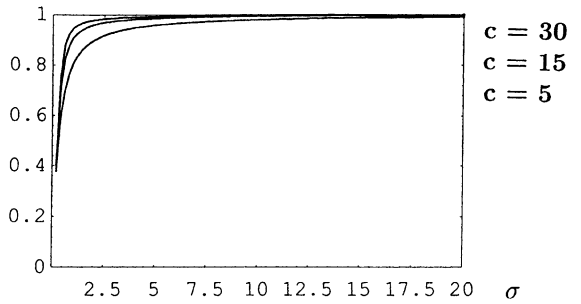

Fig. 5. $\lambda^*(\sigma,c,p)$ versus σ. Case $p = 0.5$, $c = 5$, 15, 30.

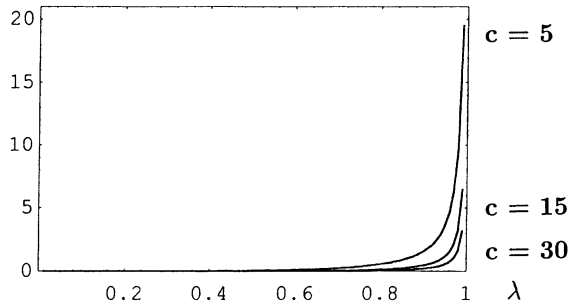

Fig. 6. $\sigma^*(\lambda, c, p)$ versus λ. Case $p = 1$, $c = 5$, 15, 30.

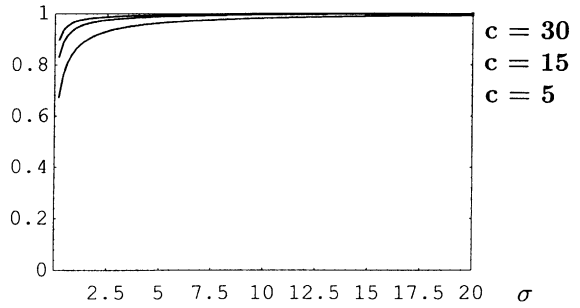

Fig. 7. $\lambda^*(\sigma, c, p)$ versus σ. Case $p = 1$, $c = 5$, 15, 30.

three figures also shows that $\sigma^*(\lambda, c, p)$ is decreasing as a function of p.

On the other hand, Figs. 3, 5 and 7 show the influence of σ, c and p on $\lambda^*(\sigma, c, p)$. Now the lowest curve of each figure corresponds to the case $c = 5$. We observe that $\lambda^*(\sigma, c, p)$ increases with increasing parameters σ, c and p.

We are now ready to compute the probability vector $\mathbf{x}$. There is extensive research on algorithms for computing the equilibrium distribution $\mathbf{x}$. In particular, Latouche and Ramaswani [17] developed an iterative algorithm with a logarithmic reduction over the number of iterations of earlier algorithms. These authors mainly focused on discrete time but they also showed how the algorithm can be adapted for continuous time. Recently Bright and Taylor [4] extended the algorithm to cover the case of level-dependent QBD processes. For the sake of completeness, we next adapt the algorithm given in [4] to calculate the matrix R corresponding to a QBD of form (3.1).

Algorithm.

$i := 0;$
$U := A_0(-A_1)^{-1};$
$D := A_2(-A_1)^{-1};$
$S := U;$
$\Pi := I;$

repeat
$i := i + 1;$
$\Pi := D\Pi;$
$Q_0 := U^2;$
$Q_1 := UD + DU;$
$Q_2 := D^2;$
$U := Q_0(I - Q_1)^{-1};$
$D := Q_2(I - Q_1)^{-1};$
$S := S + U\Pi$

until $(U\Pi)_{\max} < \varepsilon;$
$R := S.$

That algorithm provides a stable recursive method for computing the matrix R. Once R is known, the stationary distribution $\mathbf{x}$ is readily obtained from (3.4).

It should be pointed out that the case of direct access to the service facility leads to a special matrix structure which will be exploited in detail in Section 5 to get an alternative way for the computation of R.

Once the vector $\mathbf{x}$ is computed, a variety of other performance characteristics may be routinely obtained. Some of these are:

1. *The overall rate of retrials*

$$\sigma_1^* = \sigma \sum_{i=1}^{\infty} \sum_{j=0}^{c} P_{ij} = \sigma(1 - \mathbf{x}_0\mathbf{e}). \tag{3.7}$$

2. *The rate of retrials that are successful*

$$\begin{aligned}\sigma_2^* &= \sigma \sum_{i=1}^{\infty} \sum_{j=0}^{c-1} P_{ij} \\ &= \sigma_1^* - \sigma\left(\mathbf{x}_0 R(I - R)^{-1}\right)_c \\ &= \lambda q + \lambda p\left(\mathbf{x}_0(I - R)^{-1}\right)_c.\end{aligned} \tag{3.8}$$

3. *The fraction of retrials that are successful*

$$f = \frac{\sigma_2^*}{\sigma_1^*}. \tag{3.9}$$

4. *The blocking probability*

$$P_{\cdot c} = \sum_{i=0}^{\infty} P_{ic} = \left(\mathbf{x}_0 (I - R)^{-1}\right)_c. \tag{3.10}$$

5. *The mean number of busy servers*

$$Y = \sum_{i=0}^{\infty} \sum_{j=1}^{c} j P_{ij} = \mathbf{x}_0 (I - R)^{-1} \mathbf{a} = \frac{\lambda}{\mu}, \tag{3.11}$$

where $\mathbf{a} = (0, 1, \ldots, c)'$.

6. *The factorial moments of the number of customers in orbit*

$$N^k = k!, \mathbf{x}_0 R^k (I - R)^{-1-k} \mathbf{e}, \quad k \geqslant 1. \tag{3.12}$$

7. *The mean busy period*

$$T = \lambda^{-1}\left((\mathbf{x}_0)_0^{-1} - 1\right). \tag{3.13}$$

The point is that all main performance measures can be directly expressed in terms of $\mathbf{x}_0$ and R. Since the stationary analysis of our queueing system with constant retrial policy does not require the truncation of the orbit, the above formulas can be considered as closed form expressions.

Next, we present some numerical results that illustrate the effect of the parameters on the performance measures. To that end, we show in Table 1 the influence of λ and p on the mean number of customers in orbit N, and the expected amount of time in a cycle during which $C(t) = c$, $E[T_c]$. From the theory of regenerative processes, $E[T_c]$ is related to the other system parameters by

$$E[T_c] = P_{\cdot c}\left(\lambda^{-1} + T\right) = \frac{P_{\cdot c}}{\lambda P_{00}}. \tag{3.14}$$

For $(\mu, \sigma, c) = (1/c, 1, 5)$ and various choices of p, we give the values of N and $E[T_c]$. Note that the stability abscissa $\lambda^*(\sigma, c, p)$ determines the domain of the arrival rate λ. We can observe that both measures increase for increasing values of λ and the increase is more apparent as λ tends to $\lambda^*(\sigma, c, p)$. It is also clear that N and $E[T_c]$ are strongly affected by the recovery factor p. For instance, to take $p = 1$ implies a fast reduction of N and $E[T_c]$.

The effect of the retrial rate σ and the recovery probability p on N is shown in Table 2. The numerical results show that, as is to be expected, N decreases, as $\sigma \to \infty$, to the mean number of waiting customers in the classical $M/M/c$ queue

$$\lim_{\sigma \to \infty} N = \sum_{i=1}^{\infty} i Q_{i+c} = \frac{c^c}{c!} \frac{\rho_c^{c+1}}{(1 - \rho_c)^2} Q_0, \tag{3.15}$$

Table 1
N and $E[T_c]$ versus λ and p, $(\mu, \sigma, c) = (1/c, 1.0, 5)$

λ	$p = 0.0$		$p = 0.5$		$p = 1.0$	
	N	$E[T_c]$	N	$E[T_c]$	N	$E[T_c]$
0.05	0.0526	0.0001	0.0256	0.0001	0.68×10^{-6}	0.0001
0.10	0.1111	0.0030	0.0526	0.0029	0.37×10^{-4}	0.0027
0.15	0.1768	0.0169	0.0814	0.0156	0.0003	0.0145
0.20	0.2521	0.0590	0.1130	0.0528	0.0017	0.0478
0.25	0.3409	0.1597	0.1495	0.1385	0.0060	0.1224
0.30	0.4499	0.3710	0.1944	0.3107	0.0158	0.2677
0.35	0.5902	0.7805	0.2537	0.6279	0.0360	0.5270
0.40	0.7803	1.5386	0.3365	1.1819	0.0732	0.9644
0.45	1.0533	2.9183	0.4575	2.1192	0.1379	1.6766
0.50	1.4726	5.4560	0.6409	3.6860	0.2460	2.8154
0.55	2.1782	10.349	0.9294	6.3240	0.4236	4.6307
0.60	3.5529	20.892	1.4073	10.907	0.7162	7.5645
0.65	7.1530	50.521	2.2655	19.434	1.2123	12.478
0.70	35.095	290.36	4.0551	37.685	2.1092	21.305
0.75	–	–	9.2976	92.128	3.9596	39.467
0.80	–	–	106.44	1111.6	9.1348	90.244
0.85	–	–	–	–	64.201	630.88

Table 2
N versus σ and p, $(\lambda, \mu, c) = (0.5, 1/c, 5)$

σ	$p = 0.0$	$p = 0.5$	$p = 1.0$
0.05	–	–	77.514
0.1	–	–	2.1822
0.2	–	–	0.8332
0.3	–	262.19	0.5568
0.4	–	3.6364	0.4372
0.5	–	1.8981	0.3704
0.6	11.906	1.3132	0.3277
0.7	4.1268	1.0197	0.2979
0.8	2.5379	0.8431	0.2761
0.9	1.8536	0.7252	0.2593
1	1.4726	0.6409	0.2460
2	0.5495	0.3420	0.1875
3	0.3791	0.2640	0.1683
4	0.3072	0.2280	0.1588
5	0.2676	0.2073	0.1531
10	0.1951	0.1677	0.1417
50	0.1427	0.1376	0.1326
100	0.1365	0.1340	0.1315
500	0.1316	0.1310	0.1305
1000	0.1309	0.1307	0.1304

where

$$Q_0 = \left(\sum_{i=0}^{c-1} \frac{(\lambda/\mu)^i}{i!} + \frac{c^c}{c!} \frac{\rho_c^c}{1-\rho_c} \right)^{-1}, \quad \rho_c = \frac{\lambda}{c\mu}, \tag{3.16}$$

and the sequence $\{Q_i;\ i \geqslant 0\}$ is the stationary distribution of the $M/M/c$ queue with parameters λ and μ.

Finally, Fig. 8 shows the influence of σ and p on $E[T_c]$. The curves, which in decreasing order correspond to lower recovery probability, are consistent with the following limiting result:

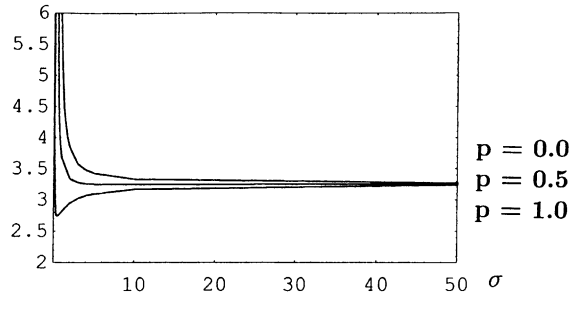

Fig. 8. $E[T_c]$ versus σ and p, $(\lambda, \mu, c) = (0.5, 1/c, 5)$.

$$\lim_{\sigma \to \infty} P_{\cdot c} = \sum_{i=c}^{\infty} Q_i = \frac{c^c}{c!} \frac{\rho_c^c}{1-\rho_c} Q_0. \tag{3.17}$$

It should be noted that the highest curve, where $p = 0$, is monotone; but the cases $p = 0.5$ and 1 start decreasing to a minimum and then $E[T_c]$ converges to its asymptotic value.

4. Optimization of the retrial rate

We now turn our attention to the optimization of the retrial rate σ.

Our numerical experience indicates that a naive optimization of the classical performance characteristics, such as those given in Eqs. (3.7)–(3.13), leads to improper solutions, i.e., the retrial parameter optima σ_{opt} equals $\sigma^*(\lambda, c, p)$ or ∞. We could specify a cost structure on the retrials and completed jobs and find the value of σ that optimizes the resulting objective function. However, we prefer to concentrate probabilistic criteria that are independent of costs.

The following are some specific probabilistic descriptors of the system that attain their optima at proper values $\sigma_{\text{opt}} \in (\sigma^*(\lambda, c, p), \infty)$.

Ideal retrials: We would like to choose the retrial rate so that many retrials result in rendering all c servers busy. We shall call a retrial *ideal* if it results in a transfer of a job from the orbit to a unique free server. Following an ideal retrial, all servers are busy.

After an ideal retrial all servers are rendering work and, in addition, we avoid the possibility of an unsuccessful repeated attempt. Thus, from a social point of view, the ideal retrial represents the best possible choice for a repeated attempt.

We define $P^{(i)}$ as the steady-state fraction of ideal retrials. Alternatively, $P^{(i)}$ is the steady-state conditional probability that a retrial is ideal given that a retrial occurs. We have that

$$P^{(i)} = \frac{\sum_{i=1}^{\infty} P_{i,c-1}}{\sum_{i=1}^{\infty} \sum_{j=0}^{c} P_{ij}} = \frac{\left(\mathbf{x}_0 R (I-R)^{-1} \right)_{c-1}}{1 - \mathbf{x}_0 \mathbf{e}}. \tag{4.1}$$

In Fig. 9, we plot $P^{(i)}$ versus σ and several values of p. We consider that $\lambda = 0.5$, $c = 5$ and

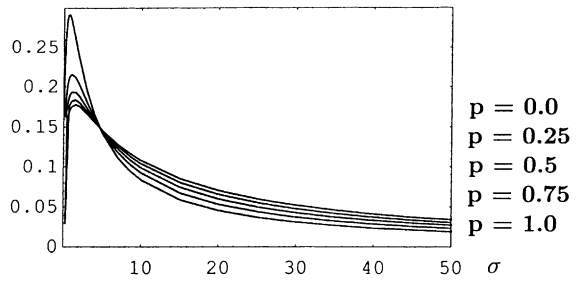

Fig. 9. $P^{(i)}$ versus σ and p.

$c\mu = 1$. $P^{(i)}$ decreases to 0, as $\sigma \to \infty$. Note that a proper value of σ maximizing the probability $P^{(i)}$ always exists.

Successive ideal retrials: A more stringent, yet still tractable criterion is to maximize the steady-state fraction of retrials that are ideal and that are followed by another ideal retrial. Let $P^{(i,i)}$ denote that fraction.

If $p = 0$, we can easily find that

$$P^{(i,i)} = \frac{P_{1,c-1}}{\left(\mathbf{x}_0 R(I-R)^{-1}\right)_{c-1}} \frac{c\lambda\mu\sigma}{(\lambda + c\mu)((c-1)\mu + \sigma)} \times \left(\frac{1}{\lambda + (c-1)\mu} + \frac{1}{c\mu + \sigma}\right) + \frac{\left(\mathbf{x}_0 R^2 (I-R)^{-1}\right)_{c-1}}{\left(\mathbf{x}_0 R(I-R)^{-1}\right)_{c-1}} \times \frac{c\mu}{c\mu + \sigma} \frac{\sigma}{(c-1)\mu + \sigma}. \tag{4.2}$$

For the case of a general recovery probability p, $P^{(i,i)}$ is given by

$$P^{(i,i)} = \frac{P_{1,c-1}}{\left(\mathbf{x}_0 R(I-R)^{-1}\right)_{c-1}} \times \left(\frac{\lambda}{\lambda + c\mu} \frac{c\mu}{c\mu + \sigma} A_{c-1} + \frac{c\mu}{\lambda + c\mu} B_{c-1}\right) + \frac{\left(\mathbf{x}_0 R^2 (I-R)^{-1}\right)_{c-1}}{\left(\mathbf{x}_0 R(I-R)^{-1}\right)_{c-1}} \frac{c\mu}{c\mu + \sigma} A_{c-1}, \tag{4.3}$$

where the constants $\{(A_j, B_j);\ 0 \leqslant j \leqslant c\}$ satisfy the following recursion:

$$A_c = \frac{\lambda}{\lambda + c\mu + \sigma} A_c + \frac{c\mu}{\lambda + c\mu + \sigma} A_{c-1},$$

$$B_c = \frac{\lambda}{\lambda + c\mu} A_c + \frac{c\mu}{\lambda + c\mu} B_{c-1},$$

$$A_j = \frac{\lambda p}{\lambda + j\mu + \sigma} A_{j+1} + \frac{\lambda q}{\lambda + j\mu + \sigma} A_j + \frac{j\mu}{\lambda + j\mu + \sigma} A_{j-1} + \frac{\sigma \delta_{j,c-1}}{\lambda + j\mu + \sigma}, \quad 0 \leqslant j \leqslant c-1,$$

$$B_j = \frac{\lambda p}{\lambda + j\mu} B_{j+1} + \frac{\lambda q}{\lambda + j\mu} A_j + \frac{j\mu}{\lambda + j\mu} B_{j-1}, \quad 0 \leqslant j \leqslant c-1. \tag{4.4}$$

Eqs. (4.2)–(4.4) are proved from the first principles. A sketch of the proof is as follows. First, we condition on the system state just before the first ideal retrial occurs. Then we need to distinguish several cases but the general objective is to guarantee that the minimum between the next service completion and the next retrial ends in completion of any service time. After this service completion time, we consider a first step analysis so we condition on the system state just after the next event (primary arrival, service completion, retrial) occurs. Now A_j (respectively, B_j) is the probability that the next retrial is ideal given that the number of busy servers is j and the orbit is non-empty (respectively, empty).

In Fig. 10, we again consider that $\lambda = 0.5$, $c = 5$ and $c\mu = 1$. The influence of σ and p is now essentially different but $P^{(i,i)}$ still has its optima at a proper value of σ. Figs. 9 and 10 also show that σ_{opt} decreases with increasing values of p.

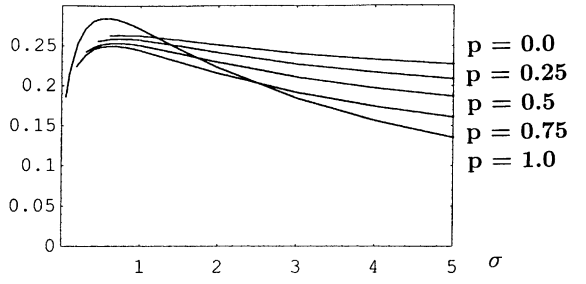

Fig. 10. $P^{(i,i)}$ versus σ and p.

Vain retrials: Non-ideal retrials either transfer a customer from the orbit to one of several idle servers or they accomplish nothing at all because when the retrial occurs, all servers are occupied. Let us call these last retrials *vain*.

The steady-state fraction of vain retrials is given by

$$P^{(v)} = \frac{\sum_{i=1}^{\infty} P_{ic}}{\sum_{i=1}^{\infty} \sum_{j=0}^{c} P_{ij}} = \frac{\left(\mathbf{x}_0 R(I-R)^{-1}\right)_c}{1 - \mathbf{x}_0 \mathbf{e}}, \qquad (4.5)$$

and we can choose σ so as to *minimize* that quantity.

5. The model with direct access to the server facility

We now take advantage of the special matrix structure associated with the case $p = 1$. We note that A_0 can be written as

$$A_0 = \lambda \mathbf{e}_c \mathbf{e}_c', \qquad (5.1)$$

where $\mathbf{e}_c$ is a column vector of dimension $c+1$ such that all its elements are equal to 0 except for the last one which is equal to 1.

Since A_0 is of form (5.1), the matrix R can be explicitly determined, once its spectral radius η is known (see [23, Theorem 4]). To be precise, R is of the form

$$R = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{u} \end{pmatrix}, \qquad (5.2)$$

and the problem reduces to determine the row vector $\mathbf{u}$. A simple proof follows taking into account that the matrix R is given by $\lim_{n\to\infty} R_n$, where $\{R_n;\ n \geqslant 0\}$ is a sequence of matrices defined by

$$\begin{aligned} &R_0 = \mathbf{0}, \\ &R_n = -\left(A_0 A_1^{-1} + R_{n-1}^2 A_2 A_1^{-1}\right), \quad n \geqslant 1. \end{aligned} \qquad (5.3)$$

Thus, we note that a zero row in A_0 produces a zero row in R_1 and, consequently, in all successive iterates.

From (5.2) we find that

$$\mathbf{u}R = u_c \mathbf{u}, \qquad (5.4)$$

$$\eta = u_c, \qquad (5.5)$$

where u_c denotes the last element of $\mathbf{u}$ and η is the spectral radius of R. Then multiplying the fundamental equation $R^2 A_2 + R A_1 + A_0 = \mathbf{0}$ by $\mathbf{u}$ and using (5.4) and (5.5), we obtain that

$$\det\left(\eta^2 A_2 + \eta A_1 + A_0\right) = 0. \qquad (5.6)$$

In fact, the general theory [15] establishes that η is the unique root in $(0,1)$ of Eq. (5.6). The equation $\mathbf{u}(\eta^2 A_2 + \eta A_1 + A_0) = \mathbf{0}$ is useful to determine the vector $\mathbf{u}$ up to a constant. Furthermore, the general equation $R A_2 \mathbf{e} = A_0 \mathbf{e}$ reduces the computation of $\mathbf{u}$, to the solution of the system

$$\begin{aligned} &\mathbf{u}\left(\eta^2 A_2 + \eta A_1 + A_0\right) = \mathbf{0}, \\ &\mathbf{u}\mathbf{e} = \frac{\lambda}{\sigma} + \eta. \end{aligned} \qquad (5.7)$$

Together, Eqs. (5.6), (5.7) and (5.2) give an alternate method for computing R.

One procedure for computing the spectral radius η uses Elsner's algorithm and a bisection method, see [18]. An alternative scheme, appropriate to the present example, is to solve directly Eq. (5.6) which leads to the computation of the unique root in $(0,1)$ of a polynomial equation. When $c = 2k$, for $k \geqslant 1$, the polynomial has degree $k+1$. If $c = 2k+1$, for $k \geqslant 0$, the degree is $k+2$. The polynomial can be recursively obtained from the following equations:

$$\begin{aligned} \det\left(\eta^2 A_2 + \eta A_1 + A_0\right) &= \eta^c((\lambda - (\lambda + c\mu)\eta) Z_{c-1}(\eta) \\ &\quad - c\mu\eta(\lambda + \sigma\eta) Z_{c-2}(\eta)), \end{aligned} \qquad (5.8)$$

$$\begin{aligned} &Z_0(\eta) = -(\lambda + \sigma), \\ &Z_1(\eta) = (\lambda + \sigma)^2 + \mu\sigma(1 - \eta), \end{aligned} \qquad (5.9)$$

$$\begin{aligned} Z_n(\eta) &= -(\lambda + n\mu + \sigma) Z_{n-1}(\eta) \\ &\quad - n\mu(\lambda + \sigma\eta) Z_{n-2}(\eta), \quad 2 \leqslant n \leqslant c-1. \end{aligned} \qquad (5.10)$$

Finally, we present some numerical results to understand how η varies with λ, and σ. First, in Fig. 11, we plot η versus λ for various values of c. We consider that $c\mu = 1$ and $\sigma = 1$, then λ takes values on the interval $(0, \lambda^*(\sigma, c, 1))$. The highest

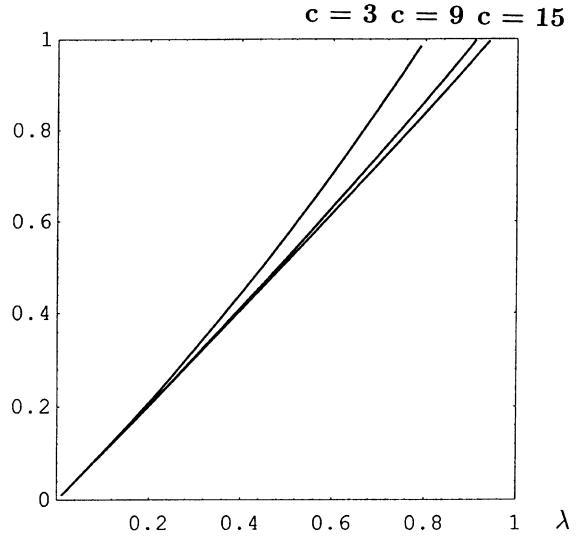

Fig. 11. η versus λ. $(\sigma,\mu)=(1,1/c)$, $c\in\{3,9,15\}$.

curve in the figure corresponds to the lowest value of c. The calculus involved in the recursive solution of Eqs. (5.8)–(5.10) is easily done by using MATHEMATICA (see [26]). Observe that, as it must, η converges to 1, as λ tends to $\lambda^*(\sigma,c,1)$.

The influence of the retrial parameter σ on η, is shown in Fig. 12. We again assume the normalization $c\mu=1$ and consider $c=3$, 9 and 15, and $\lambda=0.5$. The highest curve again corresponds to the case $c=3$. The spectral radius converges to

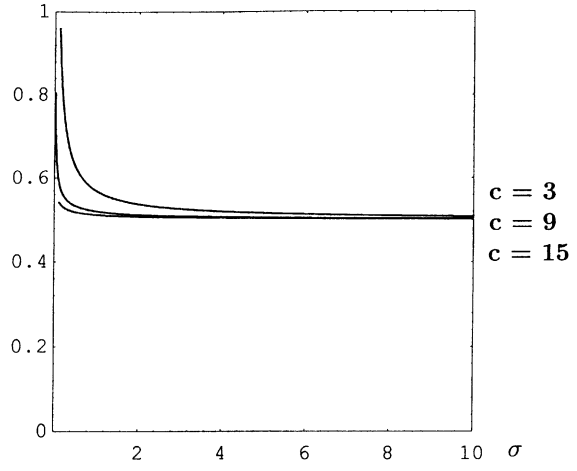

Fig. 12. η versus σ. $(\lambda,\mu)=(0.5,1/c)$, $c\in\{3,9,15\}$.

$\eta_\infty=\lambda$, as one expects because $\rho_c=\lambda(c\mu)^{-1}$ is the spectral radius for the classical $M/M/c$ queue.

6. Concluding remarks

This paper deals with the numerical investigation of the $M/M/c$ retrial queue operating under the so-called constant retrial policy. Although our approach proceeds along the matrix-geometric formalism, the study is significant due to the interest of the system itself and its specific analysis. First, it is showed that the queueing model under consideration may be presented within the simpler framework of QBD processes [17,18,23]. The advantage of working with QBD processes is that one may present the basic features of the $M/M/c$ queue with constant retrial rate by thinking in terms of block states and transition submatrices. Second, our study provides a new insight showing that the queue under the constant retrial policy is well suited for numerical purposes in comparison with the difficulties presented by the queue with classical retrial policy. Third, we obtain a successful solution to the problem of finding specific descriptors of the queue that attain their optima at proper values of the retrial parameter. Fourth, we show what analytical simplifications occur for the case of direct access to the service facility.

The current study enriches the existing literature on queueing systems with constant retrial rate which, until now, deals either with Markovian models with fewer than three servers or with simple variants of the $M/G/1$ retrial queue.

Our work can be generalized in several directions. A first possibility is to introduce a dynamical control of the number of active servers depending on the number of customers in orbit. Suppose that at time $t=0$ the system is empty and only one server is active. We assume that the system evolves as an $M/M/1$ queue with constant retrial rate and $p=1$ until the first epoch at which a primary arrival finds the server busy and a critical number, we say M, of customers in orbit. Then a second server is switched on and the last arrival automatically enters service. This dynamic control is iterated by switching on the servers when the number of customers in orbit crosses up the

critical levels $(M, 2M, \ldots, (c-1)M)$. For fixed (λ, μ, σ), the number of servers should be chosen as the first non-negative integer for which the stability condition holds. From a mathematical point of view this model can be thought as a QBD process with a large number of boundary states.

A second generalization consists in assuming the following full access rule: when there are $i \geqslant 1$ customers in orbit, a signal is sent out in accordance with an exponential law of rate σ and the number k of idle servers is reported back; then a number $\min(i,k)$ of customers in orbit are taken into service. It is clear that the interest of this new retrial rule is connected with a better use of the system resources. The analysis is now based on the general theory for Markov processes of $GI/M/1$ type [18].

As a last generalization, we mention a multiserver queue with general retrial times, i.e., every time that the orbit is non-empty, we start a renewal process of repeated attempts. This description generalizes the single server model investigated in [6]. Now the approach is based on the theory for Markov chains of $M/G/1$ type [19].

The above generalizations suggest natural ways to continue working on the applicability of matrix-geometric methods to multiserver queues with homogeneous repeated attempts. It is our hope to develop some of these models in any future work.

Acknowledgements

The research of J.R. Artalejo and A. Gómez-Corral was supported by the European Commission through INTAS project 96-0828 and the DGES through project PB98-0837. M.F. Neuts acknowledges the support of NSF Grant Nr. DMI-9306828 and a Senior Fulbright Award to Spain during five months of 1998.

References

[1] J.R. Artalejo, Stationary analysis of the characteristics of the $M/M/2$ queue with constant repeated attempts, Opsearch 33 (1996) 83–95.

[2] J.R. Artalejo, Accessible bibliography on retrial queues, Mathematical and Computer Modelling 30 (1999) 1–6.

[3] J.R. Artalejo, A. Gómez-Corral, Steady state solution of a single-server queue with linear repeated requests, Journal of Applied Probability 34 (1997) 223–233.

[4] L. Bright, P.G. Taylor, Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes, Stochastic Models 11 (1995) 497–525.

[5] B.D. Choi, Y.W. Shin, W.C. Ahn, Retrial queues with collision arising from unslotted CSMA/CD protocol, Queueing Systems 11 (1992) 335–356.

[6] B.D. Choi, K.K. Park, C.E.M. Pearce, An $M/M/1$ retrial queue with control policy and general retrial times, Queueing Systems 14 (1993) 275–292.

[7] B.D. Choi, K.H. Rhee, K.K. Park, The $M/G/1$ retrial queue with retrial rate control policy, Probability in the Engineering and Informational Sciences 7 (1993) 29–46.

[8] B.D. Choi, Y. Chang, $MAP_1/MAP_2/M/c$ with the retrial group of finite capacity and geometric loss, Mathematical and Computer Modelling 30 (1999) 99–114.

[9] J.W. Cohen, Basic problems of telephone traffic theory and the influence of repeated calls, Philips Telecommunication Review 18 (1957) 49–100.

[10] J.E. Diamond, A.S. Alfa, The $MAP/PH/1$ retrial queue, Stochastic Models 14 (1998) 1151–1177.

[11] A.N. Dudin, V.I. Klimenok, Queueing system $BMAP/G/1$ with repeated calls, Mathematical and Computer Modelling 30 (1999) 115–128.

[12] J.R. Falin, J.R. Artalejo, Approximations for multiserver queues with balking/retrial discipline, OR Spektrum 17 (1995) 239–244.

[13] G.I. Falin, J.G.C. Templeton, Retrial Queues, Chapman & Hall, London, 1997.

[14] G. Fayolle, A simple telephone exchange with delayed feedbacks, in: O.J. Boxma, J.W. Cohen, H.C. Tijms (Eds.), in: Teletraffic Analysis and Computer Performance Evaluation, Elsevier, Amsterdam, 1986.

[15] H.R. Gail, S.L. Hantler, B.A. Taylor, Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains, Advances in Applied Probability 28 (1996) 114–165.

[16] A. Gómez-Corral, M.F. Ramalhoto, On the stationary distribution of a Markovian process arising in the theory of multiserver retrial queueing systems, Mathematical and Computer Modelling 30 (1999) 141–158.

[17] G. Latouche, V. Ramaswani, A logarithmic reduction algorithm for quasi-birth–death processes, Journal of Applied Probability 30 (1993) 650–674.

[18] M.F. Neuts, Matrix-geometric Solutions in Stochastic Models – An Algorithmic Approach, Johns Hopkins University Press, Baltimore, MD, 1981.

[19] M.F. Neuts, Structured Stochastic Matrices of $M/G/1$ Type and Their Applications, Marcel Dekker, New York, 1989.

[20] M.F. Neuts, M.F. Ramalhoto, A service model in which the server is required to search for customers, Journal of Applied Probability 21 (1984) 157–166.

[21] M.F. Neuts, B.M. Rao, Numerical investigation of a multiserver retrial model, Queueing Systems 7 (1990) 169–190.
[22] C.E.M. Pearce, Extended continued fractions, recurrence relations and two-dimensional Markov processes, Advances in Applied Probability 21 (1989) 357–375.
[23] V. Ramaswami, G. Latouche, A general class of Markov processes with explicit matrix-geometric solutions, OR Spektrum 8 (1986) 209–218.
[24] S.N. Stepanov, Markov models with retrials: The calculation of stationary performance measures based on the concept of truncation, Mathematical and Computer Modelling 30 (1999) 207–228.
[25] R.I. Wilkinson, Theories for toll traffic engineering in the USA, Bell System Technical Journal 35 (1956) 421–514.
[26] S. Wolfram, The MATHEMATICA Book, Cambridge University Press, New York, 1996.