REVISTA MATEMÁTICA COMPLUTENSE
(2002) vol. XV, num. 1, 101-129

ISSN 1139-1138

STANDARD AND RETRIAL QUEUEING SYSTEMS: A COMPARATIVE ANALYSIS

Jesús ARTALEJO and Gennadi FALIN

Abstract

We describe main models and results of a new branch of the queueing theory, theory of retrial queues, which is characterized by the following basic assumption: a customer who cannot get service (due to finite capacity of the system, balking, impatience, etc.) leaves the service area, but after some random delay returns to the system again. Emphasis is done on comparison with standard queues with waiting line and queues with losses. We give a survey of main results for both single server $M/G/1$ type and multiserver $M/M/c$ type retrial queues and discuss similarities and differences between the retrial queues and their standard counterparts. We demonstrate that although retrial queues are closely connected with these standard queueing models they, however, possess unique distinguished features. We also mention some open problems.

1 Introduction

1.1 Motivating examples

In classical queueing theory it is usually assumed that a customer who cannot get service immediately after arrival either joins the waiting line (and then is served according to some queueing discipline) or leaves the system forever. Sometimes impatient customers leave the queue, but it is also assumed that they are leaving the system forever. However as a matter of fact the assumption about loss of customers which elected to leave the system is just a first order approximation to a real situation. Usually such a customer after some random period of time returns to the system and tries to get service again.

The following are just a few examples which explain this general remark in more detail.

2000 Mathematics Subject Classification: 60K25.
Servicio de Publicaciones. Universidad Complutense. Madrid, 2002

A) Telephone systems

Everybody knows from his/her own experience that a telephone subscriber who obtains a busy signal repeats the call until the required connection is made. As a result, the flow of calls circulating in a telephone network consists of two parts: the flow of primary calls, which reflects the real wishes of the telephone subscribers, and the flow of repeated calls, which is the consequence of the lack of success of previous attempts.

B) Retail shopping queue

In a shop a customer who finds that a queue is too long may wish to do something else and return later on with the hope that the queue dissolves. Similar behavior may demonstrate some impatient customers who entered the waiting line but then discovered that the residual waiting time is too long.

C) Random access protocols in digital communication networks

Consider a communication line with slotted time which is shared by several stations. The duration of the slot equals the transmission time of a single packet of data. If two or more stations are transmitting packets simultaneously then a collision takes places, i.e. all packets are destroyed and must be retransmitted. If the stations involved in the conflict would try to retransmit destroyed packets in the nearest slot, then a collision occurs with certainty. To avoid this, each station independently of other stations, transmits the packet with probability p and delays actions until the next slot with probability $1-p$, or equivalently, each station introduces a random delay before next attempt to transmit the packet.

1.2 General structure of retrial queues

The standard queueing models do not take into account the phenomenon of retrials and therefore cannot be applied in solving a number of practically important problems. L. Kosten [33, p.33] notes that "*any theoretical result that does not take into consideration this repetition effect should be considered suspect*". Retrial queues (or queues with returning customers, repeated orders, etc.) have been introduced to solve this deficiency. The general structure of a retrial queue is shown in Figure 1.

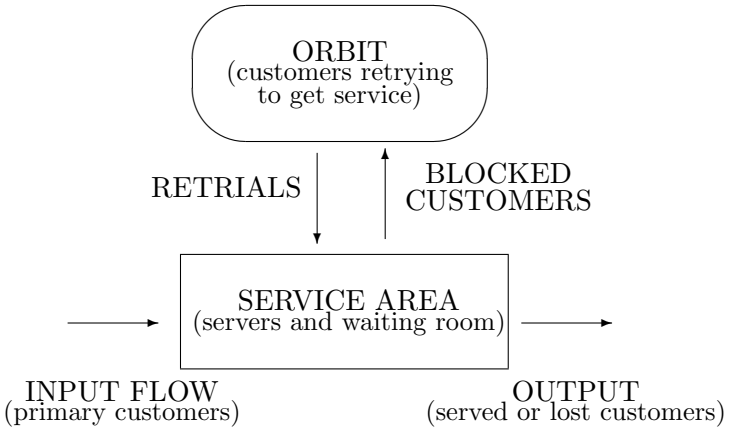

Figure 1. General structure of a retrial queue

It is clear from this picture that retrial queues can also be regarded as a special type of queueing networks. In the most general form these networks contain two nodes: the main node where blocking is possible and a delay node for repeated trials. To describe specific retrial queues with a certain structure and queueing discipline more nodes have to be introduced.

1.3 Bibliographical remarks

The early work of Kosten [32], Wilkinson [48] and Cohen [15] shows that retrial queues are suitable mathematical models for the modelling of subscribers' behavior in telephone networks. Since the pioneering works published in the 50's about 400 papers have been published in mathematical and statistical journals such as *Journal of Applied Probability, Advances in Applied Probability, Journal of the Royal Statistical Society,* etc., operations research journals such as *Queueing Systems, European Journal of Operational Research, Operations Research,* etc., telecommunication journals such as *The Bell System Technical Journal, IEEE Journal on Selected Areas in Communications,* etc. Several textbooks [34, 42, 44, 49] include sections or chapters devoted to retrial queues and a specific monograph was recently published by Falin and Templeton [26]. For a comprehensive review of the main results and

literature the reader is referred to the papers [5, 7, 23, 35, 50]. Some journals devoted special issues to the theory of retrial queues [4, 6, 46]. International Teletraffic Congresses and Seminars had sections devoted to retrial queues and recently a series of international workshops on retrial queues started in Madrid (September 1998) and continued in Minsk (June 1999) and Amsterdam (March 2000). The next one will be held in Cochin (December 2002).

2 The mathematical formalism

We consider a multiserver queueing system in which primary customers arrive according to a Poisson flow of rate λ. The service facility consists of c identical servers, and service times are exponentially distributed with parameter ν. If a primary customer finds some server free he automatically occupies a server and leaves the system after service. On the other hand, any customer who finds all servers busy upon arrival is obliged to leave the service area but he repeats his demand after an exponential time with parameter μ, i.e. we are assuming that the repeated attempts follow the classical retrial policy described later on in Subsection 3.2. We also assume that the interarrival periods, service times and retrial times are mutually independent.

The system state at time t can be described by means of a bivariate process $X = \{(C(t), N(t))\,;\, t \geq 0\}$, where $C(t)$ is the number of busy servers and $N(t)$ is the number of customers in orbit. Under the above assumptions the process X is a regular continuous time Markov chain with the lattice semi-strip $S = \{0, ..., c\} \times \mathbb{Z}_+$ as the state space.

By ordering the states as $S = \{(0,0), ..., (c,0), (0,1), ..., (c,1), ...\}$ we can express the infinitesimal generator Q of the process X in the following matrix-block form:

$$Q = \begin{pmatrix} A_0^{(0)} & A_0^{(+1)} & 0 & 0 & ... \\ A_1^{(-1)} & A_1^{(0)} & A_1^{(+1)} & 0 & ... \\ 0 & A_2^{(-1)} & A_2^{(0)} & A_2^{(+1)} & ... \\ 0 & 0 & A_3^{(-1)} & A_3^{(0)} & ... \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \end{pmatrix},$$

where $A_j^{(-1)}$, $A_j^{(0)}$ and $A_j^{(+1)}$ are the following $(c+1)\times(c+1)$ matrices:

$$A_j^{(-1)} = \begin{pmatrix} 0 & j\mu & 0 & ... & 0 \\ 0 & 0 & j\mu & ... & 0 \\ .. & .. & .. & ... & .. \\ 0 & .. & .. & ... & j\mu \\ 0 & .. & .. & ... & 0 \end{pmatrix}, \quad A_j^{(+1)} = \begin{pmatrix} 0 & ... & ... & 0 & 0 \\ 0 & ... & ... & 0 & 0 \\ ... & ... & ... & .. & .. \\ 0 & ... & ... & 0 & 0 \\ 0 & ... & ... & 0 & \lambda \end{pmatrix},$$

$$A_j^{(0)} = \begin{pmatrix} -(\lambda+j\mu) & \lambda & 0 & 0 & ... & 0 \\ \nu & -(\lambda+\nu+j\mu) & \lambda & 0 & ... & 0 \\ 0 & 2\nu & -(\lambda+2\nu+j\mu) & \lambda & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & c\nu & -(\lambda+c\nu) \end{pmatrix}.$$

Geometrically stochastic behavior of the process X can be represented with the help of the transition diagram shown in Figure 2, for the case $c = 3$.

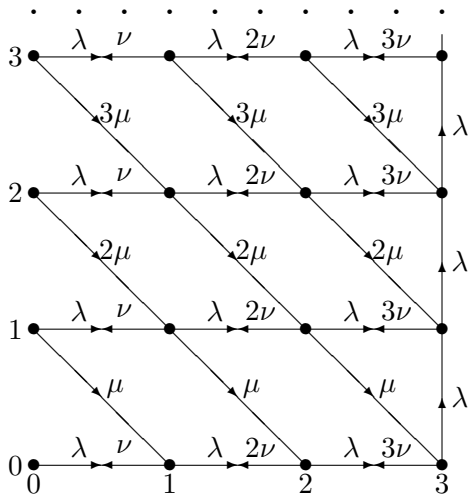

Figure 2. State space and transitions

The above mathematical description corresponds to the main retrial queue of type $M/M/c$. Many variants differ only in details but not in essential ideas. For instance, the consideration of a system with c servers and a waiting room with finite capacity d arises naturally in

modern telephone systems. This model can be obtained by considering service rates depending on the system state in the way $\nu_{ij} = \min(i, c)\nu$, for $0 \leq i \leq c + d$. In Section 3, we describe other interesting variants.

It should be noted that random walks on the product of a finite set and the set of non-negative integers (i.e. on a lattice semi-strip) arise in many applications. The best-know family of such walks was introduced by Neuts [38] and Malyshev [37]. The main assumption of their theories is the following condition of limited spacial homogeneity

$$A_j^{(k)} \equiv A^{(k)}, \text{ if } j \geq j^*,$$

for all k and some positive integer j^*.

This assumption allows extensive mathematical analysis of both stationary and transient behavior of the process. In contrast to this, retrial models operating under the classical retrial policy have transitions from states (i, j) which depend on the second coordinate. The main analytical difficulties and the most interesting properties of retrial queues are connected with this fact. To show the nature of the difficulties in more detail, we now consider the simplest problem: the calculation of the stationary distribution $\mathbf{p} = (p_{ij})_{(i,j)\in S}$ of the process X. This is usually done with the help of Kolmogorov equations $\mathbf{p}Q = 0$. Partitioning the stationary probability vector $\mathbf{p}$ as $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, ...)$, where $\mathbf{p}_j = (p_{0j}, ..., p_{cj})$, we can write Kolmogorov equations in the following matrix form

$$\mathbf{p}_{j-1}A_{j-1}^{(+1)} + \mathbf{p}_j A_j^{(0)} + \mathbf{p}_{j+1}A_{j+1}^{(-1)} = 0, \quad j = 0, 1, .. \tag{2.1}$$

where $\mathbf{p}_{-1}$ and $A_{-1}^{(+1)}$ are defined to be zero.

Alternatively, we may introduce partial generating functions

$$p_i(z) = \sum_{j=0}^{\infty} z^j p_{ij},\ 0 \leq i \leq c,$$

and transform the Kolmogorov equations into the following set of differential equations

$$\mu\mathbf{p}'(z)A(z) = \mathbf{p}(z)B(z), \tag{2.2}$$

where $\mathbf{p}(z) = (\mathbf{p}_0(z), ..., \mathbf{p}_c(z))$, $\mathbf{p}'(z) = (\mathbf{p}_0'(z), ..., \mathbf{p}_c'(z))$, and $A(z)$ and $B(z)$ are the following $(c + 1) \times (c + 1)$ matrices:

$$A(z) = \begin{pmatrix} z & -1 & 0 & ... & 0 & 0 \\ 0 & z & -1 & ... & 0 & 0 \\ .. & .. & .. & ... & .. & .. \\ 0 & 0 & 0 & ... & z & -1 \\ 0 & 0 & 0 & ... & 0 & 0 \end{pmatrix},$$

$$B(z) = \begin{pmatrix} -\lambda & \lambda & 0 & ... & 0 & 0 \\ \nu & -(\lambda+\nu) & \lambda & ... & 0 & 0 \\ 0 & 2\nu & -(\lambda+2\nu) & ... & 0 & 0 \\ .. & .. & .. & ... & .. & .. \\ 0 & 0 & 0 & ... & -(\lambda+(c-1)\nu) & \lambda \\ 0 & 0 & 0 & ... & c\nu & -(\lambda(1-z)+c\nu) \end{pmatrix}.$$

In the case $c > 2$ both infinite set of linear equations (2.1) and set of differential equations (2.2) do not have a closed form solution. Based on these equations some theoretical approaches provide solutions in terms of contour integrals [15] or as limit of extended continued fractions [41]. However, from a practical point of view, the stationary probabilities p_{ij} cannot be expressed in a tractable form and do not lend to a direct recursive computation. For $c \leq 2$, the probabilities p_{ij} satisfy a set of equations of birth-and-death type so explicit solutions in terms of special functions are available. Some information about explicit expressions (case $c \leq 2$) and approximations and numerical methods (case $c > 2$) is given in Subsection 4.1. Equation (2.2) is also the key to get the following stationary mean values [26, Section 2.3.3]:

$$E[C] = \frac{\lambda}{\nu}, \tag{2.3}$$

$$E[N] = \frac{(\nu+\mu)(\lambda - \nu Var(C))}{\mu(c\nu - \lambda)}, \tag{2.4}$$

where $E[C]$, $E[N]$ and $Var(C)$ are defined as the mean values and variance of $C(t)$ and $N(t)$ as $t \to \infty$.

Formula (2.3) can be thought as a variant of Little's formula. On the other hand, the use of formula (2.4) reduces the calculation of the mean number of customers in orbit to the variance of the number of busy servers, which is a simpler problem.

For understanding the physical behavior of the $M/M/c$ retrial queue, it is convenient to analyze the system state at service completion epochs. It should be noted that a server in a standard queueing system is rendering service in a continuous manner until the queue becomes empty. In contrast, in a retrial queue a server remains unavailable for the system over some interval of time after each service completion. For ease of notation, let us describe this situation for the single server case $c = 1$. At epoch η_{i-1}, the $(i-1)$th customer completes his service and the server becomes free. The next customer enters service after some random time R_i, during which the server is free although there may be customers in the orbit. If the number of customers in orbit at time η_{i-1}, N_{i-1}, is equal to j, then R_i is exponentially distributed with rate $\lambda + j\mu$. Observe that the type (i.e. primary or orbiting customer) of the ith service time is determined by a competition between two exponential laws of rates λ and $j\mu$, respectively. Note also that repeated attempts that occur during the service time S_i that starts at epoch $\xi_i = \eta_{i-1} + R_i$ do not modify the state of the system. At epoch $\eta_i = \xi_i + S_i$ the server becomes idle again. Thus, the evolution of a retrial queue is described in terms of an alternating sequence $\{(R_i, S_i);\ i \geq 0\}$ of idle and busy periods for the server (see Figure 3). This alternating structure is a root of the stochastic decomposition property described in Subsection 4.2.

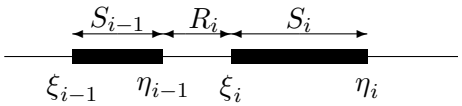

Figure 3. Description of the system behavior

Based on the above figure, the following equation which describes the dynamic of the orbit can be derived:

$$N_i = N_{i-1} - B_i + V_i,$$

where V_i is the number of customers arriving during the ith service time and $B_i = 1$ if the ith customer in service proceeds from the orbit and

$B_i = 0$ otherwise. The random vector (R_i, B_i) depends on the history of the system prior to time η_{i-1} only through N_{i-1} and

$$P\left(R_i > t, B_i = 1 \mid \ N_{i-1} = j\right) = \frac{j\mu}{\lambda + j\mu} e^{-(\lambda + j\mu)t}.$$

On the other hand, the pair (V_i, S_i) does not depend on the history of the system before ξ_i and

$$P\left(S_i \in (t, t+dt),\ V_i = n\right) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \nu e^{-\nu t} dt.$$

3 Other queueing systems with retrials

In complex models of computer and communication systems the repeated attempts are combined with a variety of queueing phenomena leading to a large number of variants and generalizations of the main retrial queue described in Section 2. In investigating variants, it should be distinguished between basic structural properties which are extended analogues of the derivations for the main models of type $M/M/c$ and $M/G/1$, and in particular significant properties of each specific model. For purposes of illustration, we now briefly describe a few of such variants.

3.1 The single server model of type M/G/1

The consideration of the single server case has intrinsic interest for the stochastic modelling of communication protocols arising from local area networks [30]. If $c = 1$, the service distribution can be generalized to follow a general law with probability distribution function $B(x)$ $(B(0) = 0)$, Laplace-Stieltjes transform $\beta(s)$ and first moments β_k. Now the mathematical model can be viewed as a Markov regenerative process. The main $M/G/1$ retrial queue and many of its variants can be studied by using a variety of methodologies including Markov renewal theory, supplementary variable analysis, embedded Markov chains, etc. For a systematic account of the fundamental methods and results, we refer to the reader to the monograph by Falin and Templeton [26, Chapter 1].

3.2 Generalized retrial policies

In many applications to telephony a call receiving a busy signal is not allowed to await for the termination of the busy condition. In this context, each blocked call generates a source of repeated requests for service independently of the rest of calls in the orbit. Thus, the classical retrial policy assumes that the probability of a repeated attempt during the interval $(t, t+dt)$, given that j calls are in orbit at time t, is $j\mu dt + o(dt)$. In contrast to this, some applications to computer and communication networks are based on the feature that the time between two successive repeated attempts is controlled by an electronic device and, consequently, is independent of the number of units applying for service, so the probability of a repeated attempt during $(t, t+dt)$, given that the orbit is no empty, is $\alpha dt + o(dt)$. This second type of discipline is called constant retrial policy. Artalejo and Gomez-Corral [3] treat both models in a unified way by defining the linear retrial policy.

3.3 Retrials due to balking and impatience

Most queueing systems with retrials are motivated by computer and telecommunication applications where a repeated attempt appears due to blocking in a system with limited service capacity. However, the existence of retrials can be due to another reason. For instance, Fayolle and Brun [27] study a single server system with retrials where the repeated attempts are due to impatience of the queueing customers. A second possibility is provided by the consideration of mixed models with waiting line and orbit [25], where a customer finding a long queue upon arrival may decide to attend another secondary job and come back later hoping to find a shorter queue.

3.4 Models with nonpersistent subscribers

Suppose that a calling subscriber after some unsuccessful retrials decides to abandon the system. This practical variant can be modelled with the help of the so-called persistence function $\{H_j;\ j \geq 0\}$, where H_j represents the probability that after the jth attempt fails, a subscriber will make the $(j+1)$ th one [23, Section 13].

3.5 Multiclass retrial queues

In the main model it is assumed that the input process is homogeneous from the point of view of such characteristics as the service time and the interretrial time distributions. In practice, however, these characteristics may differ widely for different subscriber groups. This leads us to multiclass retrial queues [23, Section 12]. Multiclass models are far more difficult for mathematical analysis than single class models because now the joint queue length process is a random walk on the multidimensional integer lattice $\mathbb{Z}_+^n$ rather than on $\mathbb{Z}_+$.

3.6 Batch arrival retrial queues

It is very common the consideration of communication systems at which units arrive in batches. In batch arrival retrial queues it is assumed that at every arrival epoch a batch of k primary units arrives with probability c_k. If $c = 1$ and the channel is busy at the arrival epoch, then the whole group joins the orbit, whereas if the channel is free, then one of the arriving units starts its service and the rest form sources of repeated calls. The consideration of multiserver retrial queues with batch arrivals leads to an infinitesimal generator Q of $M/G/1$ type. Although the methods required to investigate this type of multiserver models seem standard, we do not know any work dealing with the algorithmic analysis of these systems.

3.7 Retrial queues with a finite number of sources

It is usually assumed that the flow of primary arrivals is Poisson. Usually this means that primary arrivals are generated by a very large number of sources and each of them generates primary calls very seldom. From this point of view a model with Poisson input flow is a model with an infinite number of sources. However, in many practical situations it is important to take into account the fact that the rate of generation of new primary calls decreases as the number of customers in the system increases. Examples of this behavior arise from the performance analysis of hybrid fiber-coax, cellular networks and star-like local area networks with collision avoidance circuits [29, 30, 47]. This can be done with the help of finite source models where each individual source generates its own flow of primary demands.

3.8 Other advanced retrial queues

The retrial literature is vast and rich so it is possible to find a great number of variants and generalizations including systems with priorities [12], models with negative arrivals and disasters [8], polling systems [36], overloading systems [1], etc. The interested reader may find useful material about the above variants and many other retrial models in the monograph [26], the papers [5, 7] and the references therein. At this moment, it should be pointed out the impossibility of getting analytical solutions for retrial systems in the case of non-exponentially distributed intervals between primary arrivals and interretrial periods. It means a significant difference with standard queues which admit closed form expressions for the model $G/M/c$ and its variants [31]. As an alternative, many efforts have addressed during the last decade to the numerical investigation of complex retrial queues. In this sense, we especially mention the use of matrix-analytical methods [13, 17, 18] for the investigation of versatile retrial models with interarrival and interrepetition distributions of type PH, MAP, etc.

4 Comparing standard and retrial queueing systems

It is now clear that there exists a rich variety of different single server and multiserver queueing systems with retrials. Although the study of some of them implies a special insight of their peculiarities, in many other cases an extended investigation based on the methods developed for the $M/M/c$ and the $M/G/1$ retrial queue may be carried out for structural complex retrial models too. Therefore, in this section, we concentrate on the main models of type $M/M/c$ and $M/G/1$ and establish a comparative analysis of the standard models versus their retrial counterparts.

4.1 The main M/M/c model

In addition to process $X = \{(C(t), N(t));\ t \geq 0\}$ which describes the system state for the retrial queue, we now consider a second process $Y = \{Q(t);\ t \geq 0\}$ which indicates the number of customers in the system for the standard $M/M/c$ queue without repeated attempts. In fact, the

process Y is a simple birth-and-death process with birth (arrival) rates $\lambda_i = \lambda$, $i \geq 0$, and death (service) rates $\nu_i = \min(i, c)\nu$, $i \geq 1$.

As usual, the first question to be investigated is the positive recurrence of X and Y. It can be shown that both processes are positive recurrent if and only if

$$\rho = \frac{\lambda}{c\nu} < 1. \tag{4.1}$$

In the case of process Y, the proof follows from the classical results for the classification of states in birth-and-death processes [31]. The proof for the retrial process X uses Foster's criterion based on mean drifts [26, Section 2.2]. Essentially more interesting is the fact that $\rho = 1$ provides a necessary and sufficient condition for the null recurrence of Y, whereas the behavior of X in the case $\rho = 1$ depends on the retrial rate. For instance, if $c = 1$ and $\rho = 1$, then X is null recurrent if and only if $\mu \geq \nu$ [26, Section 1.3.1].

It is interesting to observe that the positive recurrence condition of process X is independent of the retrial rate μ. An intuitive explanation follows by assuming a very congested orbit, then the idle time R_i converges to zero and the system performs like the standard queue with random order discipline.

If $\rho < 1$, the steady state exists. In the case of the standard $M/M/c$ queue [31] the stationary distribution of the number of customers in the system is given by

$$p_j = \begin{cases} p_0 \left(\frac{\lambda}{\nu}\right)^j \frac{1}{j!}, & \text{if } 0 \leq j \leq c, \\ p_0 \rho^j \frac{c^c}{c!}, & \text{if } j > c, \end{cases}$$

where

$$p_0 = \left(\frac{c^c \rho^{c+1}}{c!(1-\rho)} + \sum_{j=0}^{c} \left(\frac{\lambda}{\nu}\right)^j \frac{1}{j!} \right)^{-1}.$$

In particular, in the single server case we have a geometric distribution with parameter ρ:

$$p_j = (1-\rho)\rho^j, \ j \geq 0.$$

The stationary distribution of the system state for the $M/M/1$ retrial queue is as follows [26, Section 1.2]:

$$p_{0j} = \frac{\rho^j}{j!\mu^j}(1-\rho)^{1+\frac{\lambda}{\mu}} \prod_{k=0}^{j-1}(\lambda + k\mu),\ j \geq 0,$$

$$p_{1j} = \frac{\rho^{j+1}}{j!\mu^j}(1-\rho)^{1+\frac{\lambda}{\mu}} \prod_{k=1}^{j}(\lambda + k\mu),\ j \geq 0.$$

Thus, we also have the following expression for the stationary distribution of the total number of customers in the system

$$p_j = \frac{\rho^j}{j!\mu^j}(1-\rho)^{1+\frac{\lambda}{\mu}} \prod_{k=1}^{j}(\lambda + k\mu),\ j \geq 0,$$

which can be identified as negative binomial distribution.

The model with $c = 2$ can be reduced to hypergeometric expressions [26, Section 2.3]. However, the consideration of more than two servers complicates the transitions among states and, as consequence, the underlying structure of birth-and-death type is not preserved. As we mention earlier, indeed the recursive computation of p_{ij} cannot be performed. Thus, the analysis of numerical methods of calculation of p_{ij} has been the subject matter of many papers [21, 40, 43, 48]. Among them, the so-called generalized truncated models propose to approximate an infinite system (which cannot be solved directly) with the help of another infinite calculable system. The fact that we approximate the original infinite system by another infinite system implies better accuracy than direct methods [48] based on finite truncation of the state space. We next describe briefly two of such generalized truncation methods.

Falin [21] introduces a simple model which may be described as follows. Assume that the retrial rate becomes equal to infinite when the number of customers in orbit exceeds a level M. It means that, from the level M up, the system performs as an ordinary $M/M/1$ queue with arrival rate λ and service rate $c\nu$. Let $\widetilde{X} = \{(\widetilde{C}(t), \widetilde{N}(t));\ t \geq 0\}$ be the process denoting the system state in the approximate model. If we denote by μ_j the retrial rate given that there are j customers in orbit, then the generalized truncated model $\widetilde{X}$ corresponds to the case

$$\mu_j = \begin{cases} j\mu, & \text{if } 0 \leq j \leq M, \\ \infty, & \text{if } j \geq M+1. \end{cases}$$

It is easy to see that $\widetilde{X}$ is a Markov chain with state space $\widetilde{S} = \{0, ..., c-1\} \times \{0, ..., M\} \cup \{c\} \times \mathbb{Z}_+$. The condition (4.1) is again necessary and sufficient for the positive recurrence. Although the stationary probabilities $\widetilde{p}_{ij}$ for $0 \leq i \leq c$, $0 \leq j \leq M$, coincide up to a normalizing constant with the corresponding stationary probabilities of the finite truncated model obtained by placing a fictitious limit in the orbit capacity [48], the numerical results show that a significant reduction in the value of M is obtained when we use the formulation $\widetilde{X}$.

Neuts and Rao [40] investigate a second possibility to get a numerically tractable approximation. They consider that the number of customers in orbit who are allowed to conduct retrials is restricted to an appropriate number N, so the retrial rate is $\mu_j = \min(j, N)\mu$, $j \geq 0$. It yields an approximate process $\widehat{X} = \{(\widehat{C}(t), \widehat{N}(t));\ t \geq 0\}$ which can be reformulated as a quasi-birth-and-death (QBD) process with a large number of boundary states. QBD processes have been dealt widely in the queueing literature [39], so the methods for determining the ergodicity condition and for computing the stationary distribution $\widehat{p}_{ij}$ are well investigated. Note that the state space is $\widehat{S} = \{0, ..., c\} \times \mathbb{Z}_+$, so it agrees with the initial state space S. However the condition (4.1) does not hold for $\widehat{X}$ because the retrial rate is constant from the level N up. Following the general theory for QBD processes, it can be proved that the process $\widehat{X}$ is positive recurrent if and only if

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

We now turn our attention to other important performance characteristics such as the busy period and the waiting time. We now assume that a busy period is defined as the period starting with the arrival of a customer who finds the system empty and ends at the first completion epoch at which the system becomes empty again. The busy period analysis is important from the server's point of view and is also helpful in the efficient planning of the system resources. We first analyze the

standard $M/M/c$ queue and denote its busy period as L_∞. The existing studies (see [10, 31] and their references) mainly deal with the existence of closed form solutions for the case $c \leq 2$. The solution is given in terms of the Bessel function of the first kind of order r, $I_r(x)$, defined as

$$I_r(x) = \sum_{n=0}^{\infty} \frac{(x/2)^{r+2n}}{(n+r)!n!}.$$

In the case $c = 1$, the probability density function, $g(x)$, of L_∞ is

$$g(x) = \frac{e^{-(\lambda+\nu)x}}{x\rho^{1/2}} I_1\left(2x\sqrt{\lambda\nu}\right), \ x > 0.$$

The first moments of L_∞ can be obtained as a particular case of the corresponding formulas given in Subsection 4.2 for the standard $M/G/1$ queue.

If $c = 2$, the density function, $h(x)$, of a busy period is given by

$$h(x) = \frac{e^{-(\lambda+2\nu)x}}{x} \sum_{r=0}^{\infty} (r+1) \left(\frac{\nu}{2\lambda}\right)^{(r+1)/2} I_{r+1}(2x\sqrt{2\lambda\nu}), \ x > 0.$$

The first moments of L_∞ are

$$E[L_\infty] = \frac{1}{\nu(1-\rho)}, \quad E[L_\infty^2] = \frac{2-\rho}{\nu^2(1-\rho)^3}.$$

In the case $c > 2$, we do not have explicit expressions. However, from the theory of regenerative processes we know that the stationary probability of an empty system, p_0, is equal to $(1 + \lambda E[L_\infty])^{-1}$. In [10] the investigation of the Laplace-Stieltjes transform of L_∞ and the computation of its moments are reduced to the recursive solution of some simple systems of linear equations.

In comparison with the above results, the study of the busy period, L_μ, of the $M/M/c$ retrial queue can be considered as an open problem. Both explicit results for the density function and transform solutions are unknown. In the case $c = 1$, the moments of L_μ can be recursively computed following the method described in [14].

We now consider the virtual waiting time, $W(t)$, of a customer who arrives to the system at time t. According to this definition, $W(t)$ means

the time that the customer spends waiting at the queue (standard model) or at the orbit (retrial model) excluding the service time. We deal with the system at steady state so we simply denote $W(t)$ by W. Clearly, the distribution of W depends on the service discipline. Retrial queues arising from teletraffic applications operate under a random order access discipline. This means that all calls waiting in the orbit have an equal chance of being allocated to free servers when these become available. The case in which customers are served in order of arrivals is of minor interest and its solution is simpler. Thus, in what follows, we assume the random access discipline. Random servicing is complicate due to the overtaking phenomena, i.e. it is necessary to consider not only the number of customers present in the system at the time of arrival of a customer whose delay is to be investigated, but also the possibility that customers arriving at later time will compete for free servers. In what follows we denote the waiting time for standard and retrial queues respectively as W_∞ and W_μ.

The complementary distribution function of the waiting time in the standard $M/M/c$ queue can be expressed in the form of an integral [44, Section 9.1.3] which can be expanded using Lagrange orthogonal polynomials or Mclaurin-series methods [16, Section 5.15]. The analysis of W_μ in the $M/M/c$ retrial queue is a work that remains to be done.

To finish this section, we now consider the limit behavior of the $M/M/c$ retrial queue under high and low rate of retrials. This is an important feature due to lack of analytical formulas for the main performance characteristics, since limit theorems allow us to understand the influence of the repeated attempts in some domains of the system parameters. Besides limit results provide insight into correspondence between retrial queues and the classical queueing models with waiting line and losses.

First, we consider the case of high rate of retrials. As $\mu \to \infty$ (i.e. the intervals between two successive retrials tend to zero) the $M/M/c$ retrial queue can be viewed as the standard one with waiting line. This general heuristic observation can be transformed into a rigorous mathematical result in several ways. In this sense, it is very interesting to obtain asymptotic expansions for stationary performance characteristics in a power series in the mean time between successive retrials $1/\mu$. The first term of such an expansion is the corresponding performance charac-

teristic for the standard $M/M/c$ queue, and the second term describes the influence of retrials and hence is of special interest. To illustrate this, we consider the stationary blocking probability B_μ of the retrial queue, i.e. the probability that all servers are busy. Then, we have [26, Section 2.7.1]

$$B_\mu = B_\infty + \frac{(c-1)\nu - \lambda + \nu B_\infty}{\mu} \ln(1-\rho) B_\infty + o\left(\frac{1}{\mu}\right),$$

where B_∞ is the blocking probability in the standard $M/M/c$ given by

$$B_\infty = \frac{\left(\frac{\lambda}{\nu}\right)^c \frac{1}{(c-1)!}}{\left(\frac{\lambda}{\nu}\right)^c \frac{1}{(c-1)!} + \left(c - \frac{\lambda}{\nu}\right) \sum_{i=0}^{c-1} \left(\frac{\lambda}{\nu}\right)^i \frac{1}{i!}}.$$

On the other hand, the limit behavior of retrial queues as $\mu \to 0$ is of interest on account of the weak dependence of the stationary distribution $\{p_{i.}(\mu);\ 0 \le i \le c\}$ of the number of busy servers upon μ. This fact is numerically illustrated in [26, section 2.6.4]. Because for complex retrial queues $\lim_{\mu\to 0} p_{i.}(\mu)$ can be found more simply than $\lim_{\mu\to\infty} p_{i.}(\mu)$, it is reasonable to use this limit as an approximation of $p_{i.}(\mu)$ for all $\mu > 0$. For the $M/M/c$ retrial queue in steady state, as $\mu \to 0$, the distribution of the number of busy servers converges to the corresponding distribution for the standard Erlang loss system $M/M/c/0$ (which is a truncated Poisson distribution), but with increased arrival rate $\Lambda = \lambda + r$, where r is the unique positive root of the polynomial equation [26, Section 2.7.2]

$$r \sum_{i=0}^{c-1} \left(\frac{\lambda + r}{\nu}\right)^i \frac{1}{i!} = \lambda \left(\frac{\lambda + r}{\nu}\right)^c \frac{1}{c!}.$$

The additional arrival rate $r = \lim_{\mu\to 0} \mu E[N]$ and can be thought of as a load formed by sources of repeated calls. This result shows that it is important to distinguish between the cases $\mu = 0$ and $\mu \to 0$. If $\mu = 0$, then the blocked customers do not send repeated attempts at all. Thus, the retrial queue becomes the standard Erlang loss system with the same arrival rate λ with stationary distribution

$$p_{i.}(0) = \frac{\left(\frac{\lambda}{\nu}\right)^i \frac{1}{i!}}{\sum_{i=0}^{c} \left(\frac{\lambda}{\nu}\right)^i \frac{1}{i!}},\ 0 \le i \le c.$$

Obviously, $N(t)$ converges to ∞, as $t \to \infty$. In contrast, if $\mu \to 0$, then the retrial queue in steady state can be viewed as the standard Erlang loss system but with the increased arrival rate $\Lambda = \lambda + r$.

4.2 The main M/G/1 model

We now consider the single server case so the service time distribution can be extended to follow a general law. The subsequent necessary notation was introduced in Subsection 3.1. We assume that $\rho = \lambda\beta_1 < 1$ so our queueing models are stable and the limiting probabilities

$$p_j = \lim_{t\to\infty} P\{Q(t) = j\},\ j \in \mathbb{Z}_+,$$

$$p_{ij} = \lim_{t\to\infty} P\{C(t) = i, N(t) = j\},\ (i,j) \in \{0,1\} \times \mathbb{Z}_+,$$

exist and are positive.

Both sequences $\{p_j\}$ and $\{p_{ij}\}$ can be computed recursively with the help of the following equations [26, 34]:

$$p_0 = 1 - \rho,\ p_1 = \frac{1-a_0}{a_0}p_0,\ p_2 = \frac{1-a_0-a_1}{a_0}(p_0+p_1),$$

$$p_{j+1} = \frac{1-\sum_{i=0}^{j} a_i}{a_0}\sum_{i=0}^{j} p_i + \sum_{i=2}^{j} p_i \sum_{k=j-i+2}^{j} \frac{a_k}{a_0},\ j \geq 2,$$

$$p_{0j} = \frac{\lambda}{\lambda + j\mu}\pi_j,\ p_{1j} = \frac{(j+1)\mu}{\lambda}p_{0,j+1},\ j \geq 0,$$

$$\pi_j = \sum_{i=0}^{j} \pi_i \frac{\lambda}{\lambda + i\mu} a_{j-i} + \sum_{i=1}^{j+1} \pi_i \frac{i\mu}{\lambda + i\mu} a_{j-i+1},\ j \geq 0,$$

$$\pi_0 = (1-\rho)\exp\left\{-\frac{\lambda}{\mu}\int_0^1 \frac{1-\beta(\lambda-\lambda u)}{\beta(\lambda-\lambda u)-u}du\right\},$$

where

$$a_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dB(t),\ j \geq 0,$$

is the probability that exactly j customers arrive during the service time.

In fact, the sequence $\{\pi_j;\ j \geq 0\}$ corresponds to the distribution of the embedded Markov chain at service completion epochs.

An alternative solution [26, 31, 34] in terms of the generating functions

$$p(z) = \sum_{j=0}^{\infty} z^j p_j,\ p_i(z) = \sum_{j=0}^{\infty} z^j p_{ij},\ i \in \{0,1\},$$

is given by

$$p(z) = \frac{(1-\rho)(1-z)\beta(\lambda - \lambda z)}{\beta(\lambda - \lambda z) - z}, \tag{4.2}$$

$$p_0(z) = (1-\rho) \exp\left\{ -\frac{\lambda}{\mu} \int_z^1 \frac{1-\beta(\lambda-\lambda u)}{\beta(\lambda-\lambda u)-u} du \right\}, \tag{4.3}$$

$$p_1(z) = \frac{1-\beta(\lambda-\lambda z)}{\beta(\lambda-\lambda z)-z} p_0(z). \tag{4.4}$$

Note that the solution for both standard and retrial queues are given in terms of the Laplace-Stieltjes transform of the service times but the retrial model exhibits a more complex expression mainly due to the integral arising in the right-hand side of (4.3).

In particular, the corresponding expectations are given by

$$E[Q] = \rho + \frac{\lambda^2 \beta_2}{2(1-\rho)},$$

$$E[C] = \rho,\ E[N] = \frac{\lambda^2}{1-\rho} \left(\frac{\beta_1}{\mu} + \frac{\beta_2}{2} \right).$$

In Section 2, we remarked the existence of a sequence of idle periods in which the server is unavailable for the system. Due to this fact, a retrial queue can be considered as a special type of vacation model in

which the vacation begins at the end of each service time [2]. To exploit this fact, we now denote the process (C, N) as (C_μ, N_μ) to remark the dependence on the retrial rate μ. Let (C_∞, N_∞) be the corresponding formulation for the standard $M/G/1$. This vector represents the server state and the number of customers in queue at steady state, so that $Q = C_\infty + N_\infty$.

From equations (4.2)-(4.4) we observe that the vector (C_μ, N_μ) can be represented as a sum of two independent random vectors as follows

$$(C_\mu, N_\mu) = (C_\infty, N_\infty) + (0, R_\mu),$$

where R_μ represents the number of customers in orbit given that the server is free so its generating function is $R_\mu(z) = p_0(z)/(1-\rho)$.

An application of the stochastic decomposition property yields explicit relationships among the factorial moments of the number of customers in orbit in the $M/G/1$ retrial queue and the factorial moments in the standard $M/G/1$ queue [2, Section 4.1]. Moreover, we can estimate the following measure of proximity between the stationary distributions of the $M/G/1$ queues with and without retrials

$$D = \sum_{i=0}^{1} \sum_{j=0}^{\infty} \left| p_{ij}(\mu) - p_{ij}(\infty) \right|,$$

as follows

$$2(1-\rho-\pi_0) < D < 2\left(1 - \frac{\pi_0}{1-\rho}\right).$$

Several methodologies can be used to analyze the busy period of the standard $M/G/1$ queue. In particular, the length of L_∞ is independent of the queueing discipline so we choose to permute the order in which customers are served and create a last-come-first-served discipline. In this way, the analysis of L_∞ is connected with a branching process in which each customer arriving during the first service time generates a sub-busy period distributed as the initial busy period under study [31]. This yields the following equation for the Laplace-Stieltjes transform of L_∞:

$$L^*_\infty(s) = \beta(s + \lambda - \lambda L^*_\infty(s)). \tag{4.5}$$

This result gives the Laplace-Stieltjes transform for the busy period expressed as a functional equation which can be easily differentiated to find the first moments of L_∞. In particular, we have

$$E[L_\infty] = \frac{\beta_1}{1-\rho}, \quad E[L_\infty^2] = \frac{\beta_2}{(1-\rho)^3}.$$

On the other hand, the structure of the busy period, L_μ, of the $M/G/1$ retrial queue and its analysis in terms of Laplace transforms have been investigated by several methods [19, 26]. Thus, we have

$$L_\mu^*(s) = \frac{\int_0^{L_\infty^*(s)} \frac{\beta(s+\lambda-\lambda u)}{e(s,u)(\beta(s+\lambda-\lambda u)-u)} du}{\int_0^{L_\infty^*(s)} \frac{du}{e(s,u)(\beta(s+\lambda-\lambda u)-u)}}, \ s > 0,$$

where $L_\infty^*(s)$ is the Laplace-Stieltjes transform for the busy period in the standard queue without retrials given in (4.5) and $e(s,u)$ is

$$e(s,u) = \exp\left\{ \frac{1}{\mu} \int_0^u \frac{s+\lambda-\lambda\beta(s+\lambda-\lambda v)}{\beta(s+\lambda-\lambda v)-v} dv \right\}, \ 0 \le u < L_\infty^*(s).$$

The above expression provides a theoretical solution but it has serious limitations in practice. For instance, we cannot compute the first moments of L_μ by direct differentiation. The expectation follows easily from the theory of regenerative processes and is equal to $E[L_\mu] = \lambda^{-1}\left(p_{00}^{-1} - 1\right)$. A direct method of calculation for the second moment [9] yields

$$E[L_\mu^2] = \frac{1}{\pi_0}\left(\frac{1}{(1-\rho)^2}\left(\frac{2\rho\beta_1}{\mu} + \beta_2 \right) - \int_0^1 \frac{2}{\lambda\mu(\beta(\lambda-\lambda t)-t)} \right.$$

$$\left. \times \left(1 - \frac{\lambda(1-t)\beta'(\lambda-\lambda t)}{\beta(\lambda-\lambda t)-t} - \frac{1}{1-\rho} \exp\left\{ \frac{\lambda}{\mu} \int_t^1 \frac{1-\beta(\lambda-\lambda u)}{\beta(\lambda-\lambda u)-u} du \right\} \right) dt \right).$$

By assuming exponential service times, the above expression reduces to a simpler form. As an alternative, it can be numerically evaluated. The busy period of the single server retrial queue can also be connected with branching processes but with more complex structure [28].

The analysis for the waiting time of the standard $M/G/1$ queue with random order of service [45] leads to the following expression for the Laplace-Stieltjes transform of W_∞:

$$W^*_\infty(s) = 1-\rho+\frac{\lambda(1-\rho)}{s}\int_{L^*_\infty(s)}^1 \frac{(1-u)(\beta(\lambda-\lambda u)-\beta(s+\lambda-\lambda u))}{(\beta(\lambda-\lambda u)-u)(u-\beta(s+\lambda-\lambda u))} \times \exp\left\{-\int_u^1 \frac{dv}{v-\beta(s+\lambda-\lambda v)}\right\} du.$$

The mean and second moments are given by

$$E[W_\infty] = \frac{\lambda\beta_2}{2(1-\rho)},$$

$$E[W^2_\infty] = \frac{2\lambda\beta_3}{3(1-\rho)(2-\rho)} + \frac{\lambda^2\beta_2^2}{(1-\rho)^2(2-\rho)}.$$

The formula for the Laplace-Stieltjes transform of W_μ in the $M/G/1$ retrial queue [24] is still more formidable:

$$W^*_\mu(s) = 1-\rho+\frac{\lambda(1-\rho)}{s}\int_{L^*_\infty(s)}^1 \frac{(1-u)(\beta(\lambda-\lambda u)-\beta(s+\lambda-\lambda u))}{(\beta(\lambda-\lambda u)-u)(u-\beta(s+\lambda-\lambda u))}$$

$$\times \exp\left\{\int_u^1 \frac{s+\mu+\lambda-\lambda v}{\mu(\beta(s+\lambda-\lambda v)-v)}dv\right\} \exp\left\{\int_1^u \frac{\lambda-\lambda v}{\mu(\beta(\lambda-\lambda v)-v)}dv\right\} du.$$

The mean waiting time can be easily obtained with the help of Little's formula:

$$E[W_\mu] = \frac{\lambda\beta_2}{2(1-\rho)} + \frac{\rho}{\mu(1-\rho)}.$$

Recently, Artalejo et al. [11] have obtained the following formula for the second moment of W_μ

$$E[W^2_\mu] = \frac{2\lambda\beta_3}{3(1-\rho)(2-\rho)} + \frac{\lambda^2\beta_2^2}{(1-\rho)^2(2-\rho)} + \frac{\lambda\beta_2}{\mu}\left(\frac{2}{(1-\rho)^2(2-\rho)} + \frac{\rho}{(1-\rho)^2}\right) + \frac{2\rho}{\mu^2(1-\rho)^2}.$$

Of course, if $\mu \to \infty$, the above formulas for the retrial queue agree with the corresponding results for the standard $M/G/1$ queue. Before

dealing in more detail with limit theorems for the $M/G/1$ retrial queue, we mention the possibility of studying discrete processes closely related to L_μ and W_μ. In the case of L_μ the study can be extended to the number of customers arriving to the system during the length of a busy period [26]. It is also natural to measure the waiting time by the number of retrials $Z_\mu(t)$ make by a primary customer entering the system at time t, before he starts service [22].

Even for the standard $M/G/1$ queue the type of distribution of the queue length in steady state is an unknown; the Pollaczek-Khinchin equation (4.2) gives only the generating function of the distribution in terms of $\beta(s)$. However, under the heavy traffic analysis (which is of special practical interest), the queue length distribution can be approximated by an exponential law. To be more exact, as λ varies in such a way that $\rho \to 1-$, then the distribution of the scaled queue length, $(1-\rho)Q(t)$, weakly converges to an exponential distribution with mean $\beta_2/2\beta_1^2$. A similar result holds for the $M/G/1$ retrial queue, but now the limiting distribution of the scaled number of customers in orbit [26, Section 1.4.1] is Gamma with mean

$$\frac{\beta_2}{2\beta_1^2} + \frac{1}{\mu\beta_1},$$

and variance

$$\frac{\beta_2^2}{4\beta_1^4} + \frac{\beta_2}{2\mu\beta_1^3}.$$

Since a retrial queueing model involves one additional parameter, μ, another interesting limit situations are $\mu \to \infty$ (short intervals between retrials) and $\mu \to 0$ (long intervals between retrials). Of course, the corresponding counterparts for the standard queue do not exist.

In the case $\mu \to \infty$, it follows easily from (4.2)-(4.4) that the steady state distribution of the number of customers in orbit converges to the steady state distribution of the number of customers in the standard queue $M/G/1$.

As $\mu \to 0$, the number of customers in orbit [26, Section 1.4.2] is asymptotically Gaussian with mean

$$\frac{\lambda\rho}{\mu(1-\rho)},$$

and variance

$$\frac{2\lambda\rho(1-\rho)+\lambda^3\beta_2}{2\mu(1-\rho)^2}.$$

Finally, we consider the departure process [19] which is defined as the sequence of times $\{\eta_i;\ i \geq 0\}$ at which customers complete service and leave the system. Equivalently, we consider the sequence of inter-departure times $T_i = \eta_i - \eta_{i-1}$. Taking into account that the departure process from one queueing system can be the input process for another queueing system, it is important to find conditions under which the departure process is Poisson, or, at least, it is a renewal process. For the standard $M/G/1$ queue in steady state the departure process is a renewal one if and only if the service time distribution is exponential, in which case the process in fact is Poisson. The departure process for the $M/G/1$ retrial queue is never a renewal process, except in the trivial case of instantaneous service times.

The first two moments of the departure process in the standard queue [45] are

$$E[T_i] = \frac{1}{\lambda},$$

$$Var(T_i) = \frac{1}{\lambda^2} + \beta_2 - 2\beta_1^2.$$

In the case of the $M/G/1$ retrial queue [20], we have

$$E[T_i] = \frac{1}{\lambda},$$

$$\begin{aligned} Var(T_i) &= \frac{1}{\lambda^2} + \beta_2 - 2\beta_1^2 - \frac{2(1-\rho)}{\lambda^2} \\ &\quad + \frac{2(1-\rho)}{\lambda\mu} \int_0^1 u^{\frac{\lambda}{\mu}-1} \exp\left\{ \frac{\lambda}{\mu} \int_1^u \frac{1-\beta(\lambda-\lambda v)}{\beta(\lambda-\lambda v)-v} dv \right\} du. \end{aligned}$$

Note that $E[T_i]$ does not depend on μ. It is consequence of the fact that in steady state the rate of the departure flow must be equal to the rate of the input flow.

Acknowledgements

The authors thank the referee for his/her comments on an earlier version of this paper. It was finished during a visit of G.I. Falin to Madrid which was organized according the cooperation framework between Complutense University and Moscow State University. J.R. Artalejo thanks the support received from DGES grant PB98-0837.

References

[1] V.V. Anisimov, Averaging methods for transient regimes in overloading retrial queueing systems, *Mathematical and Computer Modelling* **30** (1999), 65-78.

[2] J.R. Artalejo and G.I. Falin, Stochastic decomposition for retrial queues, *Top* **2** (1994), 329–342.

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

[4] J.R. Artalejo (Ed.), *Retrial Queuing Systems, Mathematical and Computer Modelling* **30**, No. 3-4 (1999), 1-228.

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

[6] J.R. Artalejo (Ed.), *1^{st} International Workshop on Retrial Queues, Top* **7**, No. 2 (1999), 169-353.

[7] J.R. Artalejo, A classified bibliography of research on retrial queues: Progress in 1990-1999, *Top* **7** (1999), 187-211.

[8] J.R. Artalejo and A. Gomez-Corral, On a single server queue with negative arrivals and request repeated, *Journal of Applied Probability* **36** (1999), 907-918.

[9] J.R. Artalejo and M.J. Lopez-Herrero, On the busy period of the M/G/1 retrial queue, *Naval Research Logistics* **47** (2000), 115-127.

[10] J.R. Artalejo and M.J. Lopez-Herrero, Analysis of the busy period for the M/M/c queue: an algorithmic approach, *Journal of Applied Probability* **38** (2001), 209-222.

[11] J.R. Artalejo, G.I. Falin and M.J. Lopez-Herrero, A second order analysis of the waiting time in the M/G/1 retrial queue, *Asia-Pacific Journal of Operational Research* **19** (2002) (to appear).

[12] B.D. Choi, Y. Chang, Single server retrial queues with priority calls, *Mathematical and Computer Modelling* **30** (1999), 7-32.

[13] B.D. Choi, Y. Chang and B. Kim, MAP_1,MAP_2/M/c retrial queue with guard channels and its application to cellular networks, *Top* **7** (1999), 231-248.

[14] Q.H. Choo and B. Conolly, New results in the theory of repeated orders queueing systems, *Journal of Applied Probability* **16** (1979), 631-640.

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

[16] R.B. Cooper, *Introduction to Queueing Theory,* Edward Arnold, (1981).

[17] J.E. Diamond and A.S. Alfa, Matrix analytical methods for multi-server retrial queues with buffers, *Top* **7** (1999), 249-266.

[18] A.N. Dudin and V.I. Klimenok, A retrial BMAP/SM/1 system with linear repeated requests, *Queueing Systems* **34** (2000), 47-66.

[19] G.I. Falin, A single-line system with secondary orders, *Engineering Cybernetics* **17** (1979), 76-83.

[20] G.I. Falin, Effect of the recurrent calls on output flow of a single channel system of mass service, *Engineering Cybernetics* **17** (1979), 99-102.

[21] G.I. Falin, Calculation of probability characteristics of a multiline system with repeat calls, *Moscow University Computational Mathematics and Cybernetics* **1** (1983), 43-49.

[22] G.I. Falin, On the waiting time process in single-line queue with repeated calls, *Journal of Applied Probability* **23** (1986), 185-192.

[23] G.I. Falin, A survey of retrial queues, *Queueing Systems* **7** (1990), 127-167.

[24] G.I. Falin and C. Fricker, On the virtual waiting time in an M/G/1 retrial queue, *Journal of Applied Probability* **28** (1991), 446-460.

[25] G.I. Falin and J.R. Artalejo, Approximations for multiserver queues with balking/retrial discipline, *OR Spektrum* **17** (1995), 239-244.

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

[27] G. Fayolle and M.A. Brun, On a system with impatience and repeated calls. In: *Queueing Systems and its Applications. Liber Amicorum for J.W. Cohen* (Eds. O.J. Boxma and R. Syski), pp. 283-305, North-Holland, Amsterdam (1988).

[28] S.A. Grishechkin, Multiclass batch arrival retrial queue analyzed as branching processes with immigration, *Queueing Systems* **11** (1992), 395-418.

[29] D.J. Houck and W.S. Lai, Traffic modeling and analysis of hybrid fiber-coax systems, *Computer Networks and ISDN Systems* **30** (1998), 821-834.

[30] G.K. Janssens, 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.

[31] L. Kleinrock, *Queueing Systems, Volume 1: Theory*, Wiley, New York (1975).

[32] L. Kosten, On the influence of repeated calls in the theory of probabilities of blocking, *De Ingenieur* **59** (1947), 1-25.

[33] L. Kosten, *Stochastic Theory of Service Systems*, Pergamon Press, Oxford (1973).

[34] V.G. Kulkarni, *Modeling and Analysis of Stochastic Systems*, Chapman and Hall, London (1995).

[35] V.G. Kulkarni and H.M. Liang, Retrial queues revisited. In: *Frontiers in Queueing* (Ed. J.H. Dshalalow), pp. 19-34, CRC Press, Boca Raton (1997).

[36] C. Langaris, Gated polling systems with customers in orbit, *Mathematical and Computer Modelling* **30** (1999), 171-187.

[37] V.A. Malyshev, Homogeneous random walks on the product of a finite set and a half-line. In: *Probabilistic Methods of Research,* Moscow State University (in Russian), pp. 5-13 (1972).

[38] M.F. Neuts, Markov chains with applications in queueing theory, which have a matrix-geometric invariant probability vector, *Advances in Applied Probability* **10** (1978), 185-212.

[39] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach,* The Johns Hopkins University Press, Baltimore (1981).

[40] M.F. Neuts and B.M. Rao, Numerical investigation of a multiserver retrial model, *Queueing Systems* **7** (1990), 169-190.

[41] C.E.M. Pearce, Extended continued fractions, recurrence relations and two-dimensional Markov processes, *Advances in Applied Probability* **21** (1989), 357-375.

[42] J. Riordan, *Stochastic Service Systems*, Wiley, New York (1962).

[43] S.N. Stepanov, Markov models with retrials: the calculation of stationary performance measures based on the concept of truncation, *Mathematical and Computer Modeling* **30** (1999), 207-228.

[44] R. Syski, *Introduction to Congestion in Telephone Systems*, Elsevier Science Publishers, Amsterdam (1986).

[45] H. Takagi, *Queueing Analysis, Volume 1: Vacation and Priority Systems*, North-Holland, Amsterdam (1991).

[46] J.G.C. Templeton (Ed.), *Retrial Queues, Queueing Systems* **7**, No. 2 (1990), 125-228.

[47] P. Tran-Gia and M. Mandjes, Modeling of customer retrial phenomenon in cellular networks, *IEEE Journal on Selected Areas in Communications* **15** (1997), 1406-1414.

[48] R.I. Wilkinson, Theories for toll traffic engineering in the USA, *The Bell System Technical Journal* **35** (1956), 421-514.

[49] R.W. Wolff, *Stochastic Modeling and the Theory of Queues*, Pretince-Hall, New Jersey (1989).

[50] T. Yang and J.G.C. Templeton, A survey on retrial queues, *Queueing Systems* **2** (1987), 201–233.

Department of Statistics and O.R.
Faculty of Mathematics
Complutense University of Madrid
Madrid 28040
Spain
E-mail: jesus_artalejo@mat.ucm.es

Department of Probability Theory
Moscow State University
Moscow 119899, Russia
E-mail: falin@mech.math.msu.su

Recibido: 21 de Junio de 2001
Revisado: 27 de Noviembre de 2001