MULTISERVER QUEUEING SYSTEMS
WITH RETRIALS AND LOSSES

VYACHESLAV M. ABRAMOV[1]

Abstract

The interest in retrial queueing systems mainly lies in their application to telephone systems. This paper studies multiserver retrial queueing systems with n servers. The arrival process is a quite general point process. An arriving customer occupies one of the free servers. If upon arrival all servers are busy, then the customer waits for his service in orbit, and after a random time retries in order to occupy a server. The orbit has one waiting space only, and an arriving customer, who finds all servers busy and the waiting space occupied, is lost from the system. Time intervals between possible retrials are assumed to have arbitrary distribution (the retrial scheme is explained more precisely in the paper). The paper provides analysis of this system. Specifically the paper studies the optimal number of servers to decrease the loss proportion to a given value. The representation obtained for the loss proportion enables us to solve the problem numerically. The algorithm for numerical solution includes effective simulation, which meets the challenge of a rare events problem in simulation.

2000 *Mathematics subject classification*: primary 60K25; secondary 68M20, 60H30.
Keywords and phrases: multiuser queueing systems, retrials, loss systems, point processes, martingales and semimartingales, effective simulation.

1. Introduction

This paper studies multiserver queueing systems with retrials and losses. Retrial queueing systems are an object of investigation in numerous papers (see, for example, the reviews by Artalejo [2], Artalejo and Falin [3] and Falin [10]). The interest in these systems mainly lies in their application to telephone systems.

During the past few years there has been an especial interest in queueing systems with retrials and losses. Different queueing systems of this type have been studied in [6, 13, 17] and other papers.

[1]School of Mathematical Sciences, Monash University, Building 28M, Clayton Campus, Clayton, VIC 3800, Australia; e-mail: vyacheslav.abramov@sci.monash.edu.au.

The most well-known studies in this direction are devoted to analytical solution of one or more problems. Bocharov *et al.* [6] studied a multiclass single-server M/G/1 queueing system with a finite buffer and limited number of places in orbit. They deduced a relationship between the steady-state distributions of this system and a similar system with only one class of customers. Kovalenko [13] studied the loss probability in an M/G/n retrial queueing system with two customer classes. Mandelbaum *et al.* [17] studied queueing system with retrials and losses with time-dependent parameters within the framework of a Markovian service network (see [16]). They provide fluid and diffusion approximations to obtain numerical characteristics for queue length and virtual waiting time processes. Atencia and Phong [4] and Bocharov, Phong and Atencia [7] also studied characteristics of queueing systems with retrials and losses with the aid of analytic methods.

However, explicit representations can only be deduced for restricted classes of retrial queueing systems, and application of these results for real systems is very difficult.

Recently Abramov [1] studied the main multiserver retrial queueing system, where the input stream was a quite general point process (see [1] for more details). The martingale approach that has been used there can be adapted to systems with retrials and losses as well, and the present paper is devoted to analysis of these systems. The exact description of the system is as follows.

- The arrival process $A(t)$ is a point process (the assumptions on this process are given later).
- There are n servers, and an arriving customer occupies one of the free servers.
- If upon arrival all servers are busy, but the secondary queue, that is, orbit, which has only one space is free, then the customer occupies the orbit and retries in order to occupy a server.
- A customer, who upon arrival finds all servers busy and the orbit occupied, is lost from the system without delay.
- The service time of each customer is an exponentially distributed random variable with parameter μ.
- The time between retrials is arbitrarily distributed. It is generated by a point process $D(t)$ (the details on this assumption are given later). The processes $D(t)$ and $A(t)$ are assumed to be disjoint (that is, the probability of a simultaneous jump of these processes is equal to 0).

We observe that a significant difference between the main multiserver model of [1] and the model considered in the present paper is also that retrial times in the present model are generally distributed, while in [1] retrial times of each customer in orbit are exponentially distributed.

There are also differences in the assumptions. In the case of the main multiserver

retrial queueing system considered in [1], the point process $A(t)$ must satisfy the condition

$$\mathbf{P}\left\{\lim_{t\to\infty}\frac{A(t)}{t}=\lambda<n\mu\right\}=1, \tag{1.1}$$

leading to the stability of the system.

In the case of the system considered in the present paper the number of customers in the systems is always bounded, and the stability condition (1.1) is no longer required, and the following weaker conditions of convergence are used:

$$\mathbf{P}\left\{\lim_{t\to\infty}\frac{A(t)}{t}=\lambda\right\}=1 \tag{1.2}$$

and

$$\mathbf{P}\left\{\lim_{t\to\infty}\frac{D(t)}{t}=\Delta\right\}=1, \tag{1.3}$$

that is, λ, Δ, μ and n are not allied to one another.

Conditions (1.2) and (1.3) are technical conditions that are used in our analysis. (We often use the Lebesgue theorem on dominated convergence.)

The main aim of this paper is the numerical solution of a concrete problem associated with a specific performance measure. The main question that we want to answer in this paper is *under what number of servers will the loss proportion be less than a given small value?* The system that will be studied in this paper is described by quite general point processes, and finding an explicit analytic solution for this system in the general case is impossible. At the same time directly simulating this system in order to answer the aforementioned question is not possible either, since as the loss probability is chosen to be very small, we should deal with simulating a rare events problem. The simulating rare events problem is a well-known problem which has received considerable attention in the literature (see, for example, [11], [23] and many others). Then the main result of this paper is an answer to the question *how can we simulate this system?*

The main results of the present paper are represented by relation (3.8) and Theorem 4.1. These results enable us to *effectively simulate* models to study their most significant performance characteristic, the loss proportion.

In our case, we discuss three relations for loss proportions: (3.8), (3.9) and (3.10). Relations (3.9) and (3.10) are not effective. For example, (3.10) requires us to simulate the events $\{Q_1(s-)=n\}$ and $\{Q_2(s-)=1\}$ directly. These events become rare as n increases indefinitely. By contrast, relation (3.8) is effective. Its expression contains $\lim_{t\to\infty}t^{-1}\mathbf{E}\int_0^t Q(s)\,ds$, which according to Theorem 4.1 can be easily approximated, so that the probabilities of rare events presented in the calculation of this expression can be removed. Specifically, the Poisson measure in (3.2) can be used for the calculation

of (3.8), that is, one Poisson measure can be replaced by another and we can then use the Radon-Nikodim derivative. In other words, the Poisson process under which the required characteristic of the system has a small probability, is replaced by another Poisson process under which the usual simulation becomes available, and then the required small probability is recalculated using the Radon-Nikodim derivative (see, for example, [18] or [22]). Recall that increments of a Poisson process are exponentially distributed, and for two exponentially distributed random variables with parameters ϑ_1 and ϑ_2, the Radon-Nikodim derivative is

$$\frac{d\mathbf{P}_{\vartheta_1}(x)}{d\mathbf{P}_{\vartheta_2}(x)} = \frac{\vartheta_1}{\vartheta_2} e^{(\vartheta_2 - \vartheta_1)x},$$

where $\mathbf{P}_{\vartheta_k}(x)$, $k = 1, 2$, are the corresponding exponential probability distributions.

Relations (3.9) and (3.10) are derived from (2.2) and (2.3) respectively. The presence of these equations is illustrative in showing the difference between direct simulation and simulation based on (3.8) and Theorem 4.1.

The paper is organized as follows. In Section 2 we discuss the basic equations giving us then the three aforementioned relations (3.8), (3.9) and (3.10) for loss proportions. In Section 3 we use semimartingale decomposition and the Lenglart-Rebolledo inequality to derive the crucial relation (3.8). In Section 4 we prove Theorem 4.1 and the corollary from this theorem related to the particular case where the point processes $A(t)$ and $D(t)$ are Poisson. Specifically, Theorem 4.1 establishes relationships for the proportions

$$\lim_{t \to \infty} t^{-1} \int_0^t \mathbf{P}\{Q_1(s) = i, Q_2(s) = j\}\, ds, \quad i = 0, \ldots, n-1; \; j = 0, 1.$$

Section 5 is devoted to analysis of the performance measure based on the loss proportion. We formulate the proposed performance measure as an optimization problem and describe the algorithms for its solution. Section 6 provides two numerical examples. The concluding remarks are in Section 7.

2. Basic equations

All point processes considered in this paper are assumed to be right-continuous having left-side limits.

The following notation is used. The arrival process is denoted by $A(t)$. The queueing process (the number of occupied servers process) is denoted by $Q_1(t)$. The orbit state is denoted by $Q_2(t)$. It can take only the values 0 and 1 corresponding to the cases of empty and occupied orbit space. The loss process is denoted by $Q_3(t)$. Specifically, $Q_3(t)$ is the cumulated number of losses up to time t. The moments of

retrials are governed by the point process $D(t)$, that is, the retrial moments are $t_1, t_2, \ldots$ and $D(t) = \max\{k : t_k \leq t\}$. The sequence of independent Poisson processes all with the same rate μ is denoted by $\pi_l(t)$.

Then we have the following equations ($I\{\mathscr{A}\}$ denotes the indicator of event $\mathscr{A}$):

$$Q_1(t) + Q_2(t) + Q_3(t) = A(t) - \int_0^t \sum_{l=1}^n I\{Q_1(s-) \geq l\}\, d\pi_l(s), \tag{2.1}$$

$$Q_2(t) + Q_3(t) = \int_0^t I\{Q_1(s-) = n\}\, dA(s)$$

$$- \int_0^t I\{Q_1(s-) \neq n\}I\{Q_2(s-) = 1\}\, dD(s), \tag{2.2}$$

$$Q_3(t) = \int_0^t I\{Q_1(s-) = n\}I\{Q_2(s-) = 1\}\, dA(s). \tag{2.3}$$

Clearly, the left-hand side of relation (2.1) is the total number of customers in the system including customers in service, one in orbit (in the case of all servers being occupied) as well as all losses up to time t. Observe that the right-hand side of (2.1) is the same as the right-hand side of (2.1) in [1]. The first term of the right-hand side of (2.2) represents the total number of arrivals to the busy system where all servers are occupied. The second term of the right-hand side of (2.2) is the number of successful retrial customers up to time t subtracted from the total number of arrivals to the busy system during the same time t. Successful retrials occur in the cases where immediately before retrial instants there is at least one busy server. Notice that relation (2.2) is structured similarly to the corresponding relation (2.2) of [1]. Relation (2.3) characterizes the number of losses. Losses occur in the cases where immediately before arrival all servers are busy and the place in orbit is occupied. This is indicated by the right-hand side of (2.3).

3. Semimartingale decompositions and normalization

In this section we discuss the normalized processes by dividing these processes by t. As t increases to infinity, the behaviour of the process $Q_3(t)$ divided by t is the significant performance characteristic characterising losses per time unit.

We will need normalized characteristics for the semimartingale decomposition of the processes (see, for example, [12] and [14]).

The processes $A(t)$, $D(t)$ and $\pi_l(t)$, $l = 1, 2, \ldots, m$, are all assumed to be given on the common filtered probability space $(\Omega, \mathscr{F}, \mathbf{F} = (\mathscr{F}_t)_{t \geq 0}, \mathbf{P})$. All of these processes are semimartingales. The semimartingale decomposition for an arbitrary point process $X(t)$ is written as $X(t) = \widehat{X}(t) + M_X(t)$, where $\widehat{X}(t)$ is a compensator

and $M_X(t)$ is a local square integrable martingale. For example: $A(t)=\widehat{A}(t)+M_A(t)$. In some cases for local square integrable martingales we use the notation M with indexes such as $M_{i,j,k}(t)$. Such types of notation will be especially explained.

We rewrite (2.1) as

$$Q_1(t) + Q_2(t) + Q_3(t) = A(t) - C(t), \tag{3.1}$$

$$C(t) = \int_0^t \sum_{l=1}^n \mathbf{I}\{Q_1(s-) \geq l\}\, \mathrm{d}\pi_l(s). \tag{3.2}$$

The semimartingale $C(t)$ admits a Doob-Meyer decomposition

$$C(t) = \widehat{C}(t) + M_C(t). \tag{3.3}$$

In turn, the compensator $\widehat{C}(t)$ admits the representation

$$\widehat{C}(t) = \mu \int_0^t Q_1(s)\, \mathrm{d}s \tag{3.4}$$

(for details see [9], [14] and [15, Theorem 1.6.1]). Therefore, in view of (3.2)–(3.4), one can rewrite (3.1) as follows:

$$Q_1(t) + Q_2(t) + Q_3(t) = A(t) - \mu \int_0^t Q_1(s)\, \mathrm{d}s - M_C(t). \tag{3.5}$$

Let us pass now to normalized processes. For an arbitrary process $X(t)$, for $t > 0$ its normalization is denoted by a lower case, so $x(t) = X(t)/t$. Therefore, passing to normalized processes in (3.5), we write

$$q_1(t) + q_2(t) + q_3(t) = a(t) - \frac{\mu}{t} \int_0^t Q_1(s)\, \mathrm{d}s - m_C(t). \tag{3.6}$$

Assume now that t increases to infinity. Then, because of the stochastic boundedness of $Q_1(t)$ and $Q_2(t)$, the terms $q_1(t)$ and $q_2(t)$ vanish with probability 1, and the left-hand side of (3.6) in limit becomes $\mathbf{P}^- \lim_{t\to\infty} q_3(t)$.

Next, according to (1.2), as t increases to infinity, $a(t)$ converges with probability 1 to λ. The term $m_C(t)$ vanishes in probability, since according to the Lenglart-Rebolledo inequality we have:

$$\mathbf{P}\{|m_C(t)| > \delta\} \leq \mathbf{P}\left\{\sup_{0<s\leq t}\left|\frac{s m_C(s)}{t}\right| > \delta\right\} = \mathbf{P}\left\{\sup_{0<s\leq t} |M_C(s)| > \delta t\right\}$$

$$\leq \frac{\epsilon}{\delta^2} + \mathbf{P}\left\{\sum_{l=1}^n \pi_l(t) > \epsilon t^2\right\} = \frac{\epsilon}{\delta^2} + \mathbf{P}\left\{\frac{1}{t}\sum_{l=1}^n \pi_l(t) > \epsilon t\right\}, \tag{3.7}$$

and because of the arbitrariness of ϵ in (3.7) the statement is right.

Therefore, applying the Lebesgue theorem on dominated convergence

$$\mathbf{P}^{-}\lim_{t\to\infty} q_3(t) = \lambda - \mu\left(\mathbf{P}^{-}\lim_{t\to\infty}\frac{1}{t}\int_0^t Q_1(s)\mathrm{d}s\right)$$
$$= \lambda - \mu\left(\lim_{t\to\infty}\frac{1}{t}\mathbf{E}\int_0^t Q_1(s)\,\mathrm{d}s\right). \tag{3.8}$$

The meaning of the term of the right-hand side of (3.8), taken in brackets, is the expected number of occupied servers during the long-run history, or the expected stationary number of occupied servers.

Relations (2.2) and (2.3) provide additional information on the behaviour of the process $q_3(t)$. Specifically, we have

$$\mathbf{P}^{-}\lim_{t\to\infty} q_3(t) = \lim_{t\to\infty}\frac{1}{t}\mathbf{E}\int_0^t \mathbf{I}\{Q_1(s-) = n\}\,\mathrm{d}A(s)$$
$$- \lim_{t\to\infty}\frac{1}{t}\mathbf{E}\int_0^t \mathbf{I}\{Q_1(s-) \neq n\}\mathbf{I}\{Q_2(s-) = 1\}\,\mathrm{d}D(s) \tag{3.9}$$

and

$$\mathbf{P}^{-}\lim_{t\to\infty} q_3(t) = \lim_{t\to\infty}\frac{1}{t}\mathbf{E}\int_0^t \mathbf{I}\{Q_1(s-) = n\}\mathbf{I}\{Q_2(s-) = 1\}\,\mathrm{d}A(s). \tag{3.10}$$

In the particular case where the process $A(t)$ is Poisson, from (3.10) we obtain

$$\mathbf{P}^{-}\lim_{t\to\infty} q_3(t) = \lambda\lim_{t\to\infty}\frac{1}{t}\mathbf{E}\int_0^t \mathbf{I}\{Q_1(s) = n\}\mathbf{I}\{Q_2(s) = 1\}\,\mathrm{d}s$$
$$= \lambda\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{P}\{Q_1(s) = n, Q_2(s) = 1\}\,\mathrm{d}s, \tag{3.11}$$

and from (3.9) we obtain

$$\mathbf{P}^{-}\lim_{t\to\infty} q_3(t) = \lambda\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{P}\{Q_1(s) = n\}\,\mathrm{d}s$$
$$- \lim_{t\to\infty}\frac{1}{t}\mathbf{E}\int_0^t \mathbf{I}\{Q_1(s-) \neq n\}\mathbf{I}\{Q_2(s-) = 1\}\,\mathrm{d}D(s). \tag{3.12}$$

In the case where both $A(t)$ and $D(t)$ are Poisson, $q_3(t)$ converges to its limit with probability 1 as t increases to infinity, because both $m_A(t)$ and $m_D(t)$ vanish with probability 1. Specifically, in this case we obtain

$$\mathbf{P}\left\{\lim_{t\to\infty} q_3(t) = \lambda\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{P}\{Q_1(s) = n\}\,\mathrm{d}s - \Delta\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{P}\{Q_1(s) \neq n, Q_2(s) = 1\}\,\mathrm{d}s\right\}$$
$$= \mathbf{P}\left\{\lim_{t\to\infty} q_3(t) = \lambda\lim_{t\to\infty}\mathbf{P}\{Q_1(t) = n\} - \Delta\lim_{t\to\infty}\mathbf{P}\{Q_1(t) \neq n, Q_2(t) = 1\}\right\}$$
$$= 1. \tag{3.13}$$

We use the existence of the limiting stationary probabilities in (3.13) as $t \to \infty$, as well as the equalities

$$\lim_{t \to \infty} \mathbf{P}\{Q_1(t) = n\} = \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = n\} \, ds,$$

$$\lim_{t \to \infty} \mathbf{P}\{Q_1(t) \neq n, \, Q_2(t) = 1\} = \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) \neq n, \, Q_2(s) = 1\} \, ds.$$

4. Limiting frequencies and distributions

The results obtained in the previous section do not provide complete information about losses in this system. We have no information about the quantity

$$\lim_{t \to \infty} t^{-1} \int_0^t Q_1(s) \, ds$$

entering the right-hand side of (3.8). Information about the right-hand sides of (3.9) and (3.10) is unknown as well.

Therefore in this section we derive equations for the following frequencies:

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = i, \, Q_2(s) = j\} \, ds, \quad i = 0, 1, \ldots, n; \; j = 0, 1. \tag{4.1}$$

For this purpose we introduce the processes

$$I_{i,j}(t) = \mathbf{I}\{Q_1(t) = i \cap Q_2(t) = j\}, \quad i = 0, 1, \ldots, n; \; j = 0, 1, \tag{4.2}$$

taking the value 0 if at least one of the indices i or j is negative.

The jump of a point process is denoted by adding $\triangle$.[1] For example, $\triangle A(t)$ is a jump of $A(t)$, $\triangle \pi_l(t)$ is a jump for $\pi_l(t)$ and so on.

Denote $\Pi_i(t) = \sum_{l=1}^i \pi_l(t)$. Then we have the following equations:

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \cap Q_2(t) = Q_2(t-) = j\}$$
$$= I_{i-1,j}(t-)\triangle A(t) + I_{i+1,j}(t-)\triangle \Pi_{i+1}(t)$$
$$+ I_{i,j}(t-)[1 - \triangle A(t)][1 - \triangle \Pi_i(t)][1 - j\triangle D(t)],$$
$$i = 0, \ldots, n - 1; \; j = 0, 1, \tag{4.3}$$

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \cap Q_2(t-) = 1 \cap Q_2(t) = 0\}$$
$$= I_{i-1,1}(t-)\triangle D(t), \quad i = 0, \ldots, n - 1, \tag{4.4}$$

---

[1]This is the triangle sign. We hope that this sign will not be mixed with the parameter (capital Greek letter) $\triangle$.

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \cap Q_2(t) = Q_2(t-) = 0\}$$
$$= I_{n-1,0}(t-)\triangle A(t) + I_{n,0}(t-)[1 - \triangle\Pi_n(t)][1 - \triangle A(t)], \quad (4.5)$$

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \cap Q_2(t-) = 0 \cap Q_2(t) = 1\}$$
$$= I_{n,0}(t-)\triangle A(t), \quad (4.6)$$

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \cap Q_2(t) = Q_2(t-) = 1\}$$
$$= I_{n-1,1}(t-)\triangle A(t)$$
$$+ I_{n,1}(t-)\triangle A(t) + I_{n,1}(t-)[1 - \triangle A(t)][1 - \triangle\Pi_n], \quad (4.7)$$

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \cap Q_2(t-) = 1 \cap Q_2(t) = 0\}$$
$$= I_{n-1,1}(t-)\triangle D(t). \quad (4.8)$$

Then, for $i = 0, 1, \ldots, n$ and $j = 0, 1$,

$$\triangle I_{i,j}(t) = \mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \cap Q_2(t) = Q_2(t-) = j\}$$
$$+ \mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \cap Q_2(t-) \neq Q_2(t) = j\} - I_{i,j}(t-). \quad (4.9)$$

Since $\sum_{s \leq t} \triangle I_{i,j}(s) = I_{i,j}(t) - I_{i,j}(0)$, then taking into account (4.3)–(4.9) we have the following: For $i = 0, 1, \ldots, n - 1$,

$$I_{i,j}(t) = I_{i,j}(0) + \int_0^t [I_{i-1,j}(s-) - I_{i,j}(s-)]\,\mathrm{d}A(s)$$
$$- \int_0^t I_{i,j}(s-)\,\mathrm{d}\Pi_i(s) - j\int_0^t I_{i,j}(s-)\,\mathrm{d}D(s)$$
$$+ \int_0^t I_{i+1,j}(s-)\,\mathrm{d}\Pi_{i+1}(s) + (1 - j)\int_0^t I_{i-1,1}(s-)\,\mathrm{d}D(s). \quad (4.10)$$

For $i = n$, we have the following pair of equations. In the case $j = 0$ we have

$$I_{n,0}(t) = I_{n,0}(0) + \int_0^t [I_{n-1,0}(s-) - I_{n,0}(s-)]\,\mathrm{d}A(s)$$
$$- \int_0^t I_{n,0}(s-)\,\mathrm{d}\Pi_n(s) + \int_0^t I_{n-1,1}(s-)\,\mathrm{d}D(s). \quad (4.11)$$

In turn, in the case $j = 1$ we have

$$I_{n,1}(t) = I_{n,1}(0) + \int_0^t I_{n-1,1}(s-)\,\mathrm{d}A(s)$$
$$+ \int_0^t I_{n,0}(s-)\,\mathrm{d}A(s) - \int_0^t I_{n,1}(s-)\,\mathrm{d}\Pi_n(s). \quad (4.12)$$

By semimartingale decomposition, from (4.10)–(4.12), we obtain the following equations. In the cases $i = 0, 1, \ldots, n - 1$, $j = 0, 1$, we have

$$I_{i,j}(t) = I_{i,j}(0) + \int_0^t [I_{i-1,j}(s-) - I_{i,j}(s-)]\,\mathrm{d}A(s) - i\mu\int_0^t I_{i,j}(s)\,\mathrm{d}s$$

$$+ (i + 1)\mu \int_0^t I_{i+1,j}(s) \, ds - j \int_0^t I_{i,1}(s-) \, dD(s)$$

$$+ (1 - j) \int_0^t I_{i-1,1}(s-) \, dD(s) + M_{i,j}(t), \tag{4.13}$$

with the martingale

$$M_{i,j}(t) = - \int_0^t I_{i,j}(s-) \, dM_{\Pi_i}(s) + \int_0^t I_{i+1,j}(s-) \, dM_{\Pi_{i+1}}(s). \tag{4.14}$$

For $i = n$ we have the following pair of equations. In the case $j = 0$ we have

$$I_{n,0}(t) = I_{n,0}(0) + \int_0^t [I_{n-1,0}(s-) - I_{n,0}(s-)] dA(s)$$

$$- n\mu \int_0^t I_{n,0}(s) \, ds + \int_0^t I_{n-1,1}(s-) \, dD(s) + M_{n,0}(t), \tag{4.15}$$

with the martingale

$$M_{n,0}(t) = - \int_0^t I_{n,0}(s-) \, dM_{\Pi_n}(s). \tag{4.16}$$

In turn, in the case $j = 1$ we have

$$I_{n,1}(t) = I_{n,1}(0) + \int_0^t I_{n-1,1}(s-) \, dA(s) - n\mu \int_0^t I_{n,1}(s) \, ds$$

$$+ \int_0^t I_{n,0}(s-) \, dA(s) + M_{n,1}(t), \tag{4.17}$$

with the martingale

$$M_{n,1}(t) = - \int_0^t I_{n,1}(s-) \, dM_{\Pi_n}(s). \tag{4.18}$$

THEOREM 4.1. *For the limiting state frequencies of the processes $Q_1(t)$ and $Q_2(t)$ we have the following equations.*

In the boundary cases $i = 0$ for $j = 0, 1$ we have

$$\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = 1, Q_2(s) = j\} \, ds$$

$$= \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = 0, Q_2(s-) = j\} \, dA(s)$$

$$+ j \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = 0, Q_2(s-) = j\} \, dD(s). \tag{4.19}$$

In the cases $i = 1, 2, \ldots, n - 1$, $j = 0, 1$, we have

$$i\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = i, Q_2(s) = j\} \, ds$$

$$- (i + 1)\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = i + 1, Q_2(s) = j\} \, ds$$

$$= \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \left[\mathbf{I}\{Q_1(s-) = i - 1, Q_2(s-) = j\} \right.$$

$$- \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\} \big] \, dA(s)$$

$$+ (1 - j) \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i - 1, Q_2(s-) = 1\} \, dD(s)$$

$$- j \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = 1\} \, dD(s). \tag{4.20}$$

In the case $i = n$ and $j = 0$ we have

$$n\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = n, Q_2(s) = 0\} \, ds$$

$$= \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \left[\mathbf{I}\{Q_1(s-) = n - 1, Q_2(s-) = 0\} \right.$$

$$- \mathbf{I}\{Q_1(s-) = n, Q_2(s-) = 0\} \big] \, dA(s)$$

$$+ \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = n, Q_2(s-) = 1\} \, dD(s). \tag{4.21}$$

In the case $i = n$ and $j = 1$ we have

$$n\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = n, Q_2(s) = 1\} \, ds$$

$$= \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = n - 1, Q_2(s-) = 1\} \, dA(s)$$

$$+ \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = n, Q_2(s-) = 0\} \, dA(s). \tag{4.22}$$

PROOF. The proof of this theorem easily follows from the representations (4.13), (4.15) and (4.17). We divide the left- and right-hand sides by t, and pass to the limits in probability as t increases to infinity.

Specifically, (4.19) and (4.20) follow from (4.13) and from the fact that the martingales $m_{i,j}(t)$ vanish with probability 1 as $t \to \infty$. The last fact is the consequence of vanishing $m_{\Pi}(t)$ and $m_{\Pi_{i+1}}(t)$ with probability 1 as $t \to \infty$. (The right-hand side of (4.14) is divided by t, and then we pass to the limit with probability 1 as $t \to \infty$.) Relations (4.21) and (4.22) follow similarly from the corresponding relations (4.15) and

(4.17) and vanishing the martingales $m_{n,0}(t)$ and $m_{n,1}(t)$ given by (4.16) and (4.18). We also use the Lebesgue dominated convergence theorem to replace $\mathbf{P}^- \lim_{t\to\infty}$ by $\lim_{t\to\infty} \mathbf{E}$ in the places where it is required. $\qquad\square$

The standard particular cases from Theorem 4.1 are the cases where $A(t)$ is Poisson, $D(t)$ is Poisson, and both $A(t)$ and $D(t)$ are Poisson. We do not consider the first two cases, and only the last of these three cases is formulated and proved below. In this case the representation is deduced explicitly and has the form of limiting distributions rather than frequencies.

COROLLARY 4.2. *In the case where $A(t)$ is Poisson with rate λ, and $D(t)$ is Poisson with rate Δ, denote $P_{i,j} = \lim_{t\to\infty} \mathbf{P}\{Q_1(t) = i, Q_2(t) = j\}$. We have the following. In the boundary cases $i = 0$ for $j = 0, 1$, we have*

$$\mu P_{1,j} = \lambda P_{0,j} + j\Delta P_{0,j}. \tag{4.23}$$

In the cases $i = 1, 2, \ldots, n - 1$, $j = 0, 1$, we have

$$i\mu P_{i,j} - (i + 1)\mu P_{i+1,j} = \lambda P_{i-1,j} - \lambda P_{i,j} + (1 - j)\Delta P_{i-1,1} - j P_{i,1}. \tag{4.24}$$

In the case $i = n$ and $j = 0$ we have

$$n\mu P_{n,0} = \lambda P_{n-1,0} - \lambda P_{n,0} + \Delta P_{n,1}. \tag{4.25}$$

In the case $i = n$ and $j = 1$ we have

$$n\mu P_{n,1} = \lambda P_{n-1,1} + \lambda P_{n,0}. \tag{4.26}$$

PROOF. In the case where $A(t)$ and $D(t)$ are Poisson, the proof of this corollary is based on the following two relations for $i = 0, 1, \ldots, n$, $j = 0, 1$:

$$\lim_{t\to\infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\} \, dA(s) = \lambda P_{i,j} \tag{4.27}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\} \, dD(s) = \Delta P_{i,j}. \tag{4.28}$$

Then inserting (4.27) and (4.28) into (4.19)–(4.22) leads to the statement of the corollary. Thus we now only need to prove (4.27) and (4.28).

Prove (4.27). By semimartingale decomposition of the Poisson process we have

$$A(t) = \lambda t + M_A(t).$$

Substituting this representation for (4.27) we obtain for $i = 0, 1, \ldots, n$, $j = 0, 1$,

$$
\lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, \, Q_2(s-) = j\} \, dA(s)
$$
$$
= \lim_{t \to \infty} \frac{\lambda}{t} \int_0^t \mathbf{P}\{Q_1(s-) = i, \, Q_2(s-) = j\} \, ds
$$
$$
+ \lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, \, Q_2(s-) = j\} \, dM_A(s). \qquad (4.29)
$$

The first term of the right-hand side is equal to $\lambda P_{i,j}$, and the second term is equal to 0. Therefore (4.27) is proved.

The proof of (4.28) is analogous. The corollary is proved. $\qquad\qquad\qquad\square$

5. Performance analysis and algorithms

One of the most important performance characteristics is the loss proportion per arriving customer. Specifically, denoting the loss proportion by f, we write

$$
f = \frac{\mathbf{P}^- \lim_{t \to \infty} q_3(t)}{\mathbf{P}^- \lim_{t \to \infty} a(t)} = \frac{1}{\lambda} \mathbf{P}^- \lim_{t \to \infty} q_3(t). \qquad (5.1)
$$

The parameters λ, μ and Δ are assumed to be given. So, the problem is to find the appropriate number of servers n such that $f \le \alpha$. More accurately, the problem is to minimize n subject to $f \le \alpha$.

In the general case, we have no explicit representation for the processes, and so simulation techniques are used. There are three equivalent relations obtained for $\mathbf{P}^- \lim_{t \to \infty} q_3(t)$ in Section 2: (3.8), (3.9) and (3.10).

The simplest of them looks to be (3.10). According to (3.10) the appropriate limit $\mathbf{P}^- \lim_{t \to \infty} q_3(t)$ can be obtained straightforwardly. Immediately before each arrival (one occurring say at instant s) we only check the event

$$
\{Q_1(s-) = n \cap Q_2(s-) = 1\}.
$$

Such types of simulation do not require any theory. However, in the case of small α this simulation is not effective. As the number of servers is chosen to be large, and the loss proportion should be small, this can result in an erroneous conclusion.

The other relation is given by (3.9). For a small α it is also based on rare events, and its application is impossible.

In this general case only (3.8) is available. It is based on the average number of busy servers $\lim_{t \to \infty} t^{-1} \mathbf{E} \int_0^t Q_1(s) \, ds$ which can be estimated by simulation. More

specifically, the simulation is based on application of Theorem 4.1, helping us to obtain the frequencies ($i = 0, 1, \ldots, n$)

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = i\} \, ds = \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = i, Q_2(s) = 0\} \, ds$$
$$+ \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbf{P}\{Q_1(s) = i, Q_2(s) = 1\} \, ds. \qquad (5.2)$$

The first and second terms of the right-hand sides are obtained from Theorem 4.1. They are expressed via the frequencies

$$\lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\} \, dA(s),$$
$$\lim_{t \to \infty} \frac{1}{t} \mathbf{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\} \, dD(s),$$

$i = 0, 1, \ldots, n$, $j = 0, 1$, which in turn are easily evaluated by simulation.

It is worth noting that direct simulation of

$$\lim_{t \to \infty} t^{-1} \int_0^t \mathbf{P}\{Q_1(s) = i, Q_2(s) = j\} \, ds \quad (i = 0, 1, \ldots, n, j = 0, 1)$$

and consequently $\lim_{t \to \infty} t^{-1} \mathbf{E} \int_0^t Q_1(s) \, ds$ is not possible. (See [1, Section 9] for a detailed discussion of this issue.)

Thus, for simulation purposes we use (3.8) and the relations of Theorem 4.1.

Recall that the problem is to minimize m subject to $f \leq \alpha$. The upper bound for n is unknown, and a special search procedure is necessary. A relevant search procedure is known due to Rubalskii [21], who proposed the search algorithm for minimization of a unimodal function on an unbounded set. The optimal algorithm is an extension of the standard Fibonacci procedure.

Thus, solution of the problem is based on two main steps:

- Simulation
- Search step

These procedures are repeated until the optimal solution is found.

In the particular case where both $A(t)$ and $D(t)$ are Poisson, for stationary probabilities $P_{i,j}$, $i = 0, 1, \ldots, n$; $j = 0, 1$, we have a system of algebraic equations. Furthermore, the upper bound for the value n can be evaluated as well. In order to evaluate the upper bound for n, one can imagine that the arrival rate is $\lambda + \Delta$, that is, retrials occur all the time continuously and the retrial space is always occupied. In this case the loss probability is greater than the original, and consequently the corresponding value n is greater than that required. This upper bound n can be easily

calculated by Erlang's loss formula. Then, having the upper and lower bounds, we obtain the optimal number of servers more easily.

A similar approach can be used for a non-Markovian system. In the most practical cases where the input stream is recurrent the upper bound for n can be easily evaluated by the known formulae for the loss systems.

However in general we have to use a special algorithm, say the method of [21], in order to find the optimal value of n.

6. Numerical examples

In this section we provide numerical solutions for two systems. The first system is the Markovian system with $\lambda = 10$, $\Delta = 2$, $\mu = 1$ and $\alpha = 0.0001$. In the second example we assume that the increments of the processes $A(t)$ and $D(t)$ are deterministic with the same parameters.

Example 1. Assuming that the lower bound for m is equal to 1, let us find the upper bound. For this purpose we consider the M/M/n/0 queueing system with input rate $12 = \lambda + \Delta$ and $\mu = 1$. According to Erlang's loss formula, for the loss probability we have

$$p_n = \frac{\rho^n/n!}{\sum_{i=0}^{n}(\rho^i/i!)},$$

where $\rho = \lambda/\mu = 12$. Assuming that the loss probability is equal to $\alpha = 0.0001$, let us find the value n.

For $n = 26$ the loss probability is 0.000174, and for $n = 27$ the loss probability is 0.000078. The first of these probabilities is greater than 0.0001, and the second one is smaller than 0.0001. Therefore, the upper bound for n is 27.

Let us now find the optimal value of n. The solution of this problem can be achieved as follows. We find a new value of $n = 1 + (27 - 1)/2 = 14$. Now we solve the system (4.23)–(4.26) given by Corollary 4.2. We also use (5.1) and (3.13). We obtain a loss probability of 0.064841. Therefore in the next step we find a new value n by $14 + (27 - 14)/2 = 20.5$ taking then the integer part. With the new value $n = 20$ we solve the system again, and for the loss probability we obtain 0.001226. This value is greater than α, and therefore the new value of n is the integer part of $20 + (27 - 20)/2$ which is equal to 23. With this value of n we have 0.000110 for the loss probability. Thus the last step is anticipated with $n = 24$. (Formally we must analyze first the value $n = 25$ and only then accept $n = 24$.) For this value of n the loss probability is 0.000045. Thus the problem is solved, and the value $n = 24$ is the optimal number of servers decreasing the loss probability to the required value.

Example 2. Assume that all parameters are the same as in Example 1, but increments of the processes $A(t)$ and $D(t)$ are deterministic. In this case the scheme of calculation includes simulation, as explained in Section 5. For this concrete system the algorithm can be simplified. Specifically, as in Example 1 we first evaluate the upper bound for n. For our approximation of the upper bound of n we consider the D/M/n/0 queueing system, the interarrival times of which all are equal to $1/(\lambda + \Delta)$, that is, in our case $1/12$. Notice that under the assumption that customers constantly arrive to the main queue by $A(t)$ and $D(t)$, the process is not a D/M/n/0 queue exactly. Although the input stream is based on two sources with deterministic interarrival times, the structure of the overall input stream is complicated. The arguments for such approximating by the D/M/n/0 queueing system are heuristic. Nevertheless, such approximation is available. This is supported by computation as well.

The loss probability for GI/M/n/0 loss systems is well known (see, for example, [8, 19, 20, 24] as well as [5]):

$$p_n = \left[\sum_{i=0}^{n} \binom{n}{i} \prod_{j=1}^{i} \frac{1 - r_j}{r_j} \right]^{-1}, \tag{6.1}$$

where $r_j = \int_0^\infty e^{-j\mu x} \, dR(x)$, and $R(x)$ is the probability distribution function of an interarrival time. In our case interarrival times are deterministic and of length $1/12$ with $\mu = 1$. Therefore, in (6.1), $r_j = e^{-j/12}$.

According to our calculations the upper bound for n is 22. The value of the loss probability is 0.000036. Let us note that for $n = 21$ the value of the loss probability is 0.000137, that is, it is slightly greater than $\alpha = 0.0001$.

In the next step we choose then the integer part of $1 + (22 - 1)/2$, and the value of n is 11. Simulating with this value of n yields the value of the loss probability 0.125444. Therefore in the next step we choose the integer part of $11 + (22 - 11)/2$, that is, $n = 16$. With this value of n by simulating we have a value of loss probability of 0.002784. In the next step $n = 16 + (22 - 16)/2 = 19$. With this value of n we correspondingly obtain a loss probability of 0.000060. In the next step we choose the integer part of $19 - (19 - 16)/2$ for n, that is, $n = 17$. The loss probability in this case is 0.000786. There is one value, $n = 18$, which must be checked. For this value the loss probability is 0.000214. Thus our conclusion is $n = 17$. Recall that the loss probability in this case is 0.000060.

7. Concluding remarks

In this paper we studied a multiserver retrial queueing system having only one space in orbit. For this queueing system we derived analytic results permitting us to provide

performance analysis. Specifically we solved the problem to find the minimal number of servers required to decrease the loss probability to a given small value. In the general case the algorithm of the solution requires effective simulation, meeting the challenge of a rare events problem. Numerical examples were provided for two cases. In the first case all processes were assumed to be Poisson, and the system Markovian. For the Markovian system the calculations were based on analytic results. In the second case the processes $A(t)$ and $D(t)$ had deterministic increments. The analysis in this case was based on effective simulation. Comparison of these results showed that in the case of the Markovian system the number of servers required to make the loss proportions smaller than a particular given value is relatively greater than that in the system having the processes $A(t)$ and $D(t)$ with deterministic increments.

Acknowledgement

References

[1] V. M. Abramov, "Analysis of multiserver retrial queueing system: A martingale approach and an algorithm of solution", *Ann. Oper. Res.* **141** (2006) 19–50.

[2] J. R. Artalejo, "Accessible bibliography on retrial queues", *Math. Comput. Model.* **30** (1999) 1–6.

[3] J. R. Artalejo and G. I. Falin, "Standard and retrial queueing systems. A comparative analysis", *Mat. Complut.* **15** (2002) 101–129.

[4] I. Atencia and N. H. Phong, "A queueing system under LCFS PR discipline with Markovian arrival process and general times of searching for service", *Investig. Oper.* **25** (2004) 293–298.

[5] A. T. Bharucha-Reid, *Elements of the theory of Markov processes and their application* (McGraw-Hill, New York, 1960).

[6] P. P. Bocharov, C. D'Apice, R. Manzo and N. H. Phong, "On retrial single-server queueing system with finite buffer and multivariate Poisson flow", *Prob. Inf. Transm.* **37** (2001) 397–406.

[7] P. P. Bocharov, N. H. Phong and I. Atencia, "Retrial queueing systems with several input flows", *Investig. Oper.* **22** (2001) 135–143.

[8] J. W. Cohen, "The full availability group of trunks with an arbitrary distribution of interarrival times and negative exponential holding time distribution", *Bull. Belg. Math. Soc. Simon Stevin* **31** (1957) 169–181.

[9] C. Dellacherie, *Capacites et Processus Stochastiques* (Springer, Berlin, 1972).

[10] G. I. Falin, "A survey on retrial queues", *Queueing Syst.* **7** (1990) 127–168.

[11] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models", *ACM Trans. Model. Comput. Sim.* **5** (1995) 43–85.

[12] J. Jacod and A. N. Shiryayev, *Limit theorems for stochastic processes* (Springer, Berlin, 1987).

[13] I. N. Kovalenko, "The loss probability in M/G/m queueing systems with T-retrials in light traffic", *Dopovidi NAN Ukrainy (Ukrainian Academy of Sciences), Ser. A* **5** (2002) 77–80.

[14] R. S. Liptser and A. N. Shiryayev, *Statistics of random processes.* Vols I, II. (Springer, Berlin, 1977/1978).

[15] R. S. Liptser and A. N. Shiryayev, *Theory of martingales* (Kluwer, Dordrecht, 1989).

[16] A. Mandelbaum, W. A. Massey and M. I. Reiman, "Strong approximation for Markovian service network", *Queueing Syst.* **30** (1998) 149–201.

[17] A. Mandelbaum, W. A. Massey, M. I. Reiman, A. Stolyar and B. Rider, "Queue-lengths and waiting times for multiserver queues with abandonment and retrials", *Telecommunication Syst.* **21** (2002) 149–171.

[18] B. Melamed and R. Y. Rubinstein, *Modern simulation and modelling* (John Wiley, Chichester, 1998).

[19] C. Palm, "Intensitätschwankungen im Fernsprechverkehr", *Ericsson Technics* **44** (1943) 1–189.

[20] F. Pollaczek, "Generalisation de la théorie probabiliste des systèmes telephoniques sans dispositif d'attente", *C.R. Math. Acad. Sci. Paris* **236** (1953) 1469–1470.

[21] G. B. Rubalskii, "The search of an extremum of unimodal function of one variable in an unbounded set", *Comput. Math. Math. Phys.* **22** (1982) 8–15. Transl. from Russian: *Zh. Vychisl. Mat. Mat. Fiz.* **22** (1982), 10–16, 251.

[22] R. Y. Rubinstein and A. Shapiro, *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method* (John Wiley, Chichester, 1993).

[23] P. Shahabuddin, "Rare event simulation in stochastic models", in *Proceedings of the 1995 Winter Simulation Conference* (eds. C. Alexopoulos, K. Kang, W. R. Lilegdon and D. Goldsman), (IEEE Press, 1995) 178–185.

[24] L. Takács, "On a probability problem concerning telephone traffic", *Acta Math. Hungar.* **8** (1957) 319–324.