THE EFFECTIVE BANDWIDTH PROBLEM REVISITED

VYACHESLAV M. ABRAMOV

ABSTRACT. The paper studies a single-server queueing system with autonomous service and ℓ priority classes. Arrival and departure processes are governed by marked point processes. There are ℓ buffers corresponding to priority classes, and upon arrival a unit of the kth priority class occupies a place in the kth buffer. Let $N^{(k)}$, $k = 1, 2, \ldots, \ell$ denote the quota for the total kth buffer content. The values $N^{(k)}$ are assumed to be large, and queueing systems both with finite and infinite buffers are studied. In the case of a system with finite buffers, the values $N^{(k)}$ characterize buffer capacities. The paper discusses a circle of problems related to optimization of performance measures associated with overflowing the quota of buffer contents in particular buffers models. Our approach to this problem is new, and the presentation of our results is simple and clear for real applications.

1991 *Mathematics Subject Classification.* 60K25, 60K30, 90B18, 60H30, 41A58, 41A60, 40E05.

Key words and phrases. Autonomous queue, stochastic differential equation, martingales and semimartingales, point processes, loss systems, batch arrivals and services, loss probability, asymptotic analysis, mathematical programming, priority queues.

1. Introduction

1.1. **Approach.** During the last two decades there has been an increasing interest in the effective bandwidth problem for queueing systems with priorities. There are different classes of messages (units) arriving in telecommunication systems, and all of them are characterized by their quality of service requirements. In order to provide these quality of service guarantees and to allocate necessary network resources, different priority classes characterizing units arriving to that network are used.

There are a large number of papers related to this subject. A detailed review of the related literature (up to publication time) can be found in Berger and Whitt [9] (for further discussions see also [10]). For other relevant contributions to this subject see [11], [19], [22], [23], [34], [35], [48].

These papers all discuss approximations and suggest algorithms for optimal solutions for the allocation of resources or effective bandwidth problems. Most of these papers use large deviation techniques. For example, Elwalid and Mitra [21], [22] use Chernoff's inequality to approximate loss probabilities in finite buffer systems with large buffers. Berger and Whitt [9] also use exponential asymptotics [10], [45] for the workload high level crossing of the ith class priority unit. Other papers (e.g. [14], [32], [37], [48], [49]) also apply one or other techniques of the large deviation principle. Many of the aforementioned papers are aimed at solving concrete analytic problems, and their results are based on an analysis of analytic transformations (such as Laplace-Stieltjes or the z-transform) and their approximations. Many of these results are then applied to $M/G/1$-oriented queueing models or to models with more general arrival processes having a Markov structure.

The approach of the present paper substantially differs from these previous ones. The main focus of this paper is the solution of bandwidth problems for $GI/M/1$-related priority systems. To the knowledge of the author, such priority systems are not presented in the literature where the overwhelming majority of priority queueing systems studied are of $M/GI/1$ type. The innovations of the present paper are as follows.

1. We consider models of queues with an *autonomous service mechanism* (see e.g. [12], [13] as well as Section 1.3 of this paper). The main results of our analysis are based on stochastic equations, and our models are studied under a rather general setting and can be applied to a broad class of real telecommunication systems. The obtained stochastic equations are then used for analysis of particular systems with exponentially distributed service times, which are a subclass of queues with an autonomous service mechanism. Note that martingale techniques for priority queueing systems (different from the systems considered here) have been developed by Kella [30]. However, the approach of [30] differs from the present one. Specifically, [30] studies fluid networks of parallel queues with dependent Lévy inputs. It shows that the special construction given in the paper can be applied to the analysis of workload processes in $M/G/1$ queues with a preemptive resume

discipline. The paper of Kella [30] is based on an extension of the earlier results of Kella and Whitt [31]. In contrast, our approach is based on a direct construction of queues with autonomous service mechanisms, and can be applied both to $M/G/1$ and $GI/M/1$ oriented priority queueing systems. ($GI/M/1$ queues are precisely described in the paper.)

2. The buffer content process is described by the so-called *buffer type stochastic equation*. The buffer type stochastic differential equation is a special stochastic differential equation with discontinuous right-hand side (see Filippov [24]) and has already been used by Elwalid and Mitra [21], [22] to study the model with two priority classes. However [21] and [22] used the explicit forms of this equation related to low and high priority units. Analysis of these explicit equations is a hard problem. In contrast, our buffer type equation are represented in an (equivalent) integral form, and we discover a very simple representation for cumulative buffer contents, see Lemma 2.1 and Theorem 2.2. According to this representation, the system of equations for cumulative buffer content processes is the usual system of stochastic equations describing standard queue-length processes with an autonomous service mechanism. This finding essentially simplify the analysis, algorithms of solution and finally gives very simple approximation of the explicit solution. For example, it enables us to study the system with an arbitrary number of priorities.

3. Some papers (e.g. [21], [22]) assume that buffers have large capacities and discuss the probabilities of buffer overflow. They use general estimates given by large deviation theory, and particularly, by Chernoff's inequality. Being well-motivated theoretically, these estimates do not properly solve real practical problems. There is an example in [18] showing that inequalities based on exponential bounds can give unrealistic results.

We offer a *unified approach* to systems with finite and infinite buffers. Large parameters $N^{(k)}$, $k = 1, 2, \ldots, \ell$, that are used in the sequel, are referred to as *quota* for buffer content and are related to finite and infinite buffers systems. In the case of finite buffers models with recurrent input and exponentially distributed service times of batches, we develop the known asymptotic results on losses in $GI/M/1/n$ queues as $n \to \infty$ [2] to the case of $GI/M^{Y=C}/1/N_k$ queues ($k = 1, 2, \ldots, \ell$) with large buffers N_k (the second position $M^{Y=C}$ of the notation $GI/M^{Y=C}/1/N_k$ means that the service time of units is exponentially distributed, and batch size is equal to C) and then adapt the obtained asymptotic result to estimate the loss probability in systems with large finite buffers. The asymptotic representation of this paper, that is used for the loss probability in $GI/M^{Y=C}/1/N_k$ queues and then for the probability of buffer overflow, is preferable to general type estimates such as Cramer or Chernoff inequalities. The asymptotic results for the loss probability in $GI/M^{Y=C}/1/N_k$ queues are expressed via the roots of the appropriate functional equations (see Sections 5 and 6). They are also useful in studying the behaviour of losses in the case of heavy load conditions. Cramer and Chernoff inequalities are rougher, but their advantage is that they are explicit. However there are exact estimates in the form of explicit inequalities in the literature for the stationary probabilities of $GI/M/1/N$ large buffer queueing systems as well (see, [16]), and they can be easily adapted to the loss probabilities of the standard $GI/M/1/n$ queueing system (see the discussion section in [2]) with application to models such as the $GI/M^{Y=C}/1/N_k$ queues considered in the paper. We however are not going so far.

The results of this paper can be also applied to $M/G^{Y=C}/1/N$ oriented large buffers models. However, in this case a special asymptotic analysis similar to that given in Abramov [4] is necessary. This asymptotic analysis is routine and not provided in the paper.

1.2. **Convention on the notation.** For any increasing random sequence of points t_1, t_2,..., the associated point process $Z(t)=\sum_{n=1}^{\infty}\mathbf{I}\{t_n\leq t\}$ is always denoted by a capital Latin letter. If ζ_1, ζ_2, ... is a sequence of marks, then the associated marked point process $\mathcal{Z}(t)=\sum_{j=1}^{Z(t)}\zeta_j$ is always denoted by calligraphic letters. All processes considered in the paper are assumed to be right-continuous having left-limits and starting at zero. Exceptions from this rule are especially mentioned in the text (e.g. Remark 1.1). For an arbitrary point process $\mathcal{Z}(t)$, its jump in point t is denoted $\triangle\mathcal{Z}(t)=\mathcal{Z}(t)-\mathcal{Z}(t-)$, where $\mathcal{Z}(t-)$ is the left-limit of the process in point t. For arrival processes we use letters A and $\mathcal{A}$ with sub- or super-script (the notation is given in Section 1.3), and for departure process we use letters D and $\mathcal{D}$. The buffer processes describing the buffer contents will be denoted by calligraphic letter $\mathcal{Q}$ with sub- or super-script (the notation is in Section 1.3). All processes of this paper are assumed to be given on a common filtered probability space $\{\Omega,\mathcal{F},\mathbf{F}=(\mathcal{F}_t)_{t\geq0},\mathbb{P}\}$.

1.3. **Description of the system.** The paper is concerned with priority queueing system having ℓ buffers. Units arrive at the kth buffer at random time instants $t_1^{(k)}=\tau_1^{(k)}$, $t_2^{(k)}=\tau_1^{(k)}+\tau_2^{(k)}$, ..., and the nth unit arriving at the kth buffer has a positive integer random length $\vartheta_n^{(k)}$. (In telecommunication systems length can represent required memory for the message.) Denote $A^{(k)}(t)=\sum_{n=1}^{\infty}\mathbf{I}\left\{t_n^{(k)}\leq t\right\}$, where $\mathbf{I}\{\cdot\}$ denotes an indicator of the event, and $\mathcal{A}^{(k)}(t)=\sum_{j=1}^{A^{(k)}(t)}\vartheta_j^{(k)}$.

The departure process $\mathcal{D}(t)$ is assumed to be a point process with constant positive integer jumps C. Let χ_1, χ_2,... denote times between departures, and let $x_n=\sum_{i=1}^{n}\chi_i$ denote the nth departure moment. Then $\mathcal{D}(t)=C\sum_{n=1}^{\infty}\mathbf{I}\{x_n\leq t\}$. The constant C is called *depletion rate*.

The buffers are numbered 1,2,...,ℓ, and the buffer with lower order number has higher priority. Assume that the buffers are infinite. Then the equation for the first buffer content (*highest* priority buffer) is

$$\mathcal{Q}^{(1)}(t) = \max\left\{0, \mathcal{Q}^{(1)}(t-) + \triangle\mathcal{A}^{(1)}(t) - \triangle\mathcal{D}(t)\right\}. \tag{1.1}$$

According to (1.1), the buffer content $\mathcal{Q}^{(1)}(t)$ is governed by the processes $\mathcal{A}^{(1)}(t)$ and $\mathcal{D}(t)$ and is referred to as a queueing process with an *autonomous service mechanism*. Queues with autonomous service mechanism were introduced and originally studied by Borovkov [12], [13]. For different applications see [1], [3], [5], [7], [25], [26] and [27]. The term $\triangle\mathcal{A}^{(1)}(t)$ is called the *arrival jump* at time t, and the term $\triangle\mathcal{D}(t)$ is called the *possible departure jump* at time t. The prefix *possible* underlines the fact that departures can occur only if the system is not empty. For further simplifications, throughout the paper we assume that arrival and departure processes are *disjoint*, i.e. the probability of simultaneous arrival and departure is 0.

If t is a jump point of the process $\mathcal{D}(t)$, then the *real departure jump* at time t is $\min\{\mathcal{Q}^{(1)}(t-), C\}$. Thus, if $\mathcal{Q}^{(1)}(t-)$=0, then there is no departure jump.

$\mathcal{Q}^{(2)}(t)$ is the second buffer content, the priority of which is lower than that of the first buffer. $\mathcal{Q}^{(2)}(t)$ satisfies the equation:

$$\begin{aligned}\mathcal{Q}^{(2)}(t) = \max\Big\{0, \mathcal{Q}^{(2)}(t-) + \triangle\mathcal{A}^{(2)}(t) \\ - \Big[\triangle\mathcal{D}(t) - \mathcal{Q}^{(1)}(t-)\Big]\mathbf{I}\{\mathcal{Q}^{(1)}(t) = 0\}\Big\}.\end{aligned} \tag{1.2}$$

Despite the fact that equation (1.2) has a more complicated form than equation (1.1), both of these equations are of the same type. The term $\triangle\mathcal{A}^{(2)}(t)$ is an arrival jump at time t. The structure of the departure jump is more difficult. For simplicity we discuss the case $\ell = 2$ below. If t is at a departure jump and $\{\mathcal{Q}^{(1)}(t) > 0\}$, then $\{\mathcal{Q}^{(1)}(t-) > C\}$, and the jump is related to the first buffer only. Otherwise, if $\{\mathcal{Q}^{(1)}(t) = 0\}$, then the following two cases are possible:

$$\left\{0 < \mathcal{Q}^{(1)}(t-) \leq C\right\}, \tag{i}$$

$$\left\{\mathcal{Q}^{(1)}(t-) = 0\right\}. \tag{ii}$$

In case (i) departures occur from the first buffer, the first buffer is completely emptied, and if the second buffer is not empty, then in the case $\{\mathcal{Q}^{(1)}(t-) < C\}$ departures occur also from the second buffer. In case (ii) departures occur merely from the second buffer, provided that this buffer is not empty. Thus the real departure jump in this case is

$$\min\left\{\mathcal{Q}^{(1)}(t-) + \mathcal{Q}^{(2)}(t-), C\right\}.$$

Equation (1.2) is easily extended to the kth buffer content for any k $=1,2,\ldots,\ell$. Indeed, denoting

$$\mathcal{Q}_k(t) = \mathcal{Q}^{(1)}(t)+\mathcal{Q}^{(2)}(t)+\ldots+\mathcal{Q}^{(k)}(t), \tag{1.3}$$

$$\mathcal{A}_k(t) = \mathcal{A}^{(1)}(t)+\mathcal{A}^{(2)}(t)+\ldots+\mathcal{A}^{(k)}(t), \tag{1.4}$$

we have the following equation ($k=1,2,\ldots,\ell-1$):

$$\mathcal{Q}^{(k+1)}(t) = \max\Big\{0, \mathcal{Q}^{(k+1)}(t-)+\triangle\mathcal{A}^{(k+1)}(t) - \Big[\triangle\mathcal{D}(t)-\mathcal{Q}_k(t-)\Big]\mathbf{I}\{\mathcal{Q}_k(t)=0\}\Big\}. \tag{1.5}$$

The extension of (1.2) given by (1.5) is quite clear. The term $\triangle\mathcal{A}^{(k+1)}(t)$ is an arrival jump at time t (if any) to the buffer content $\mathcal{Q}^{(k+1)}(t-)$. The other term of (1.5)

$$\Big[\triangle\mathcal{D}(t)-\mathcal{Q}_k(t-)\Big]\mathbf{I}\{\mathcal{Q}_k(t)=0\}$$

is also similar to the corresponding term of (1.2). If t is a jump point, then the meaning of $\mathcal{Q}_k(t-)$ is the total content of all buffers, the priority of which is greater than the priority of the given $k+1$st buffer before the jump at point t, and $\{\mathcal{Q}_k(t)=0\}$ is the event, that all buffers, the priority of which is greater than the priority of the given $k+1$st buffer, are empty after the jump at time t.

In the sequel the process $\mathcal{Q}_k(t)$ is called the kth *cumulative* buffer content.

1.4. **Formulation of the problems.** The paper is concerned with the following problems. Let $N^{(1)}$, $N^{(2)}$, $\ldots N^{(\ell)}$ be large positive integer values. Assuming that appropriate limits in probability exist, denote

$$J^{(k)} = \mathbb{P}\text{-}\lim_{t\to\infty}\frac{1}{A_\ell(t)}\sum_{j=1}^{A^{(k)}(t)}\mathbf{I}\Big\{\mathcal{Q}^{(k)}(t_j^{(k)})>N^{(k)}\Big\}, \quad k=1,2,\ldots,\ell. \tag{1.6}$$

$A_\ell(t)$ is the total number of arrivals until time t. Then $J^{(k)}$ is the fraction of arrival instants when the length $N^{(k)}$ of the kth buffer is exceeded. Let $\alpha^{(1)}$, $\alpha^{(2)}$,$\ldots$, $\alpha^{(\ell)}$ be real positive numbers, denoting cost rates, and

$$J=\alpha^{(1)}J^{(1)}+\alpha^{(2)}J^{(2)}+\ldots+\alpha^{(\ell)}J^{(\ell)}. \tag{1.7}$$

Typical questions arising here are the following.

1. Assume that the parameters $N^{(1)}$, $N^{(2)}$, $\ldots$, $N^{(\ell)}$ are given, but the depletion rate C can be controlled. Under what value of the depletion rate C we have $J\leq\varepsilon$, where ε is a given positive small value? This question can be formally written as follows: minimize C subject to $J\leq\varepsilon$.

2. Assume that C is given, but $N^{(1)}$, $N^{(2)}$, $\ldots$, $N^{(\ell)}$ are control variables. Assume additionally that with given $\beta^{(2)}$, $\beta^{(3)}$,$\ldots$,$\beta^{(\ell)}$ the values $N^{(1)}$, $N^{(2)}$, $\ldots$, $N^{(\ell)}$ must satisfy the condition: $N^{(1)}=\lfloor\beta^{(2)}N^{(2)}\rfloor=\lfloor\beta^{(3)}N^{(3)}\rfloor=\ldots=$

$\lfloor \beta^{(\ell)} N^{(\ell)} \rfloor$, where $\lfloor \cdot \rfloor$ is the notation for the integer part of number. The problem is to minimize $N^{(1)}$ subject to $J \leq \varepsilon$.

Remark 1.1. (1.6) applies to the finite and infinite buffers systems. To finite buffers model we prescribe that a complete *arrival group is rejected* when upon arrival the buffer overflows. In the case of the system with infinite buffers, $\mathcal{Q}^{(k)}(t)$, $k = 1, 2, \ldots, \ell$ all are assumed to be right continuous having left limits. In the case of finite buffers model, $\mathcal{Q}^{(k)}(t)$, $k = 1, 2, \ldots, \ell$ are not longer right-continuous. For example, if $N^{(1)}$ is the capacity of the first buffer, and at moment $t_j^{(1)}$ the buffer overflows, then we admit that $\mathcal{Q}^{(1)}\left(t_j^{(1)}\right)$ is greater than $N^{(1)}$ in $t_j^{(1)}$. However in the neighborhood of this point $\mathcal{Q}^{(1)}\left(t_j^{(1)}\right) \leq N^{(1)}$. Then the left and right limits of $\mathcal{Q}^{(1)}(t)$ in point $t_j^{(1)}$ are not greater than $N^{(1)}$, both these limits (with probability 1) are equal and $t_j^{(1)}$ is an isolated point.

1.5. **Brief description of the mathematical ideas, methodology and contribution of the paper.** In this section we describe the mathematical ideas of this paper, as well as the methodology and overall contribution.

We start from the description of the buffer content process. For the highest priority buffer the equation for the buffer content is very simple. It is described by equation (1.1). The equations for the lower priority buffer contents are relatively more complicated and described by equation (1.2). However, the equations for the cumulated buffer contents are simple and described by a difference recurrence equation similar to (1.1)

$$\mathcal{Q}_k(t) = \max\{0, \mathcal{Q}_k(t-) + \triangle\mathcal{A}_k(t) - \triangle\mathcal{D}(t)\}. \tag{1.8}$$

Another form for (1.8) is a stochastic equation

$$\mathcal{Q}_k(t) = \mathcal{A}_k(t) - \sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) \geq j\}\mathrm{d}D(u). \tag{1.9}$$

(In all these two equations $k = 1, 2, \ldots, \ell$.) The stochastic equation (1.9) can be rewritten

$$\mathcal{Q}_k(t) = \mathcal{A}_k(t) - \mathcal{D}(t) + \sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) \leq j-1\}\mathrm{d}D(u),$$

with subsequent reduction to a Skorokhod problem (see [8], [40], [44]). (In the case $C = 1$ such reduction was provided in [33]. For its further application see also [1] and [3].)

It is shown then that representations similar to (1.8) and (1.9) remain valid for finite buffer models. Thus, in all cases the problem reduces to analyzing queueing systems with an autonomous service mechanism.

We use these results for analysis of particular queueing buffer models with priorities. A system with exponentially distributed service times is a special case of a system with an autonomous service mechanism. (A special construction of models with finite and infinite buffers is explained later in Section 5 of the paper.)

In the case of finite buffer models with renewal input and exponentially distributed service times we adapt recent results on asymptotic analysis [2]. As in papers [2], [4] and [6], the analysis is based on reducing the loss probability to a convolution type recurrence relation:

$$f_n = \sum_{j=0}^{n} f_{n-j+1}\pi_j \quad (f_0 > 0),$$

where $\pi_0 > 0$, $\pi_j \geq 0$ for all $j = 1, 2, \ldots$, and $\sum_{j=0}^{\infty} \pi_j = 1$, and applying asymptotic analysis similar to that of the book of Takács [43], p.22-23. Consequently, we provide heavy traffic analysis of these models based on asymptotic expansions of the results obtained under "usual" conditions. The loss probability for the large finite kth cumulative buffer is then not greater than the sum of the loss probabilities in the associated $GI/M^{Y=C}/1/N_i$ queues, $i = 1, 2, \ldots, k$. However, for large values N_i this sum is very small, with the order of this sum being the same as the order of one (maximum) term obtained by asymptotic analysis, and an estimate obtained seems to be better than that estimate obtained by rough methods of large deviation principle and Chernoff's inequality.

Thus, the main mathematical contribution is a general theory of priority buffer models with application to particular priority queueing systems with recurrent input and large buffers.

1.6. **Organization of the paper.** In Section 2, Lemma 2.1 states that the kth cumulative buffer content has the representation (2.1). The intuitive sense of Lemma 2.1 is that the kth cumulative buffer content for the system with infinite buffers is described by the same equation as the queue-length process in the queueing system with autonomous service mechanism, the arrival process of which is $\mathcal{A}_k(t)$ and the departure process $\mathcal{D}(t)$. We further prove a stability theorem. The main condition for stability is (2.4), the proof being based on reduction to the Skorokhod reflection principle and results of Borovkov [12], [13]. In section 3 the model with finite buffers is considered. It is shown that the equations for cumulative buffer contents in this case are similar to the case of a model with infinite buffers. In Section 4 we derive the formula for $J^{(k)}$, $k = 1, 2, \ldots, \ell$, using the level-crossing method based on representation (4.3). In Sections 5 and 6 special models of queueing systems are studied. The results of these sections are illustrative, and we do not discuss general buffer models with batch arrival such as $GI^X/M^{Y=C}/1$ queues, although the asymptotic geometrical bounds for stationary probability to reach high level N in $GI^X/M^Y/1$ queues is known (see [20]). All models considered here are particular cases of the general models discussed in Sections 2 and 3: these models are with independent identically distributed interarrival times. The results of Section 6 are based on an extension of recent results [2]. As in [2] the asymptotic analysis is based on reduction to appropriate representation helping us to use then the Takács theorem on asymptotic behavior of the convolution type recurrence relation [43], p. 22-23. In Section 6.1 the asymptotic behaviour of losses are studied under "usual" conditions, while in Section 6.2 the analysis of losses is done under heavy load conditions. In Section 7 approximation of the initial problem stated in Section 1.4 by another related problem is suggested. In Section 8 algorithms for numerical solution of the problems of Section 7 are proposed. There are concluding remarks in Section 9.

2. The stability theorem for the infinite buffers system

The representation for the buffer content of infinite buffers systems given by (1.1), (1.2) and (1.5) is difficult to analyze. However, for the cumulative buffer contents of infinite buffers systems the representation is simple.

Lemma 2.1. *For all $k = 1, 2, \ldots, \ell$ the following equation for the kth cumulative buffer content $\mathcal{Q}_k(t)$ holds:*

$$\mathcal{Q}_k(t) = \mathcal{A}_k(t) - \sum_{j=1}^{C} \int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) \geq j\} \mathrm{d}D(u), \tag{2.1}$$

where $D(t) = \frac{\mathcal{D}(t)}{C}$.

The proof of this lemma is given in Appendix A.

The statement of Lemma 2.1 has a simple intuitive explanation. For example, in the case $\ell = 2$ we have two classes of units, and clearly the cumulative buffer content process $\mathcal{Q}_2(t) = \mathcal{Q}^{(1)}(t) + \mathcal{Q}^{(2)}(t)$ contains two unit classes together, and therefore must behave as a usual (i.e. without priorities) queue-length process with an autonomous service mechanism, the arrival process of which is $\mathcal{A}_2(t) = \mathcal{A}^{(1)}(t) + \mathcal{A}^{(2)}(t)$, and the departure process is $\mathcal{D}(t)$. This intuitive explanation is easily extended to the case of arbitrary $k = 1, 2, \ldots, \ell$ number of classes.

The right-hand side of this equation contains the sum

$$\sum_{j=1}^{C} \int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) \geq j\} \mathrm{d}D(u).$$

Nevertheless, the problem can be reduced to the Skorokhod reflection principle.

Denote $S_k(t) = \mathcal{A}_k(t) - \mathcal{D}(t)$, $k = 1, 2, \ldots, \ell$. Then,

$$\mathcal{Q}_k(t) = S_k(t) + \sum_{j=1}^{C} \int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) \leq j-1\} \mathrm{d}D(u). \tag{2.2}$$

Equation (2.2) implies that $\mathcal{Q}_k(t)$ is the normal reflection of the process $S_k(t)$ ($S_k(0) = 0$) at zero. More accurately, $\mathcal{Q}_k(t)$ is the nonnegative solution of the Skorokhod problem of the normal reflection of the process $S_k(t)$ at zero (see Skorokhod [40] as well as Tanaka [44] and Anulova and Liptser [8], Ramanan [38]). This is because the function

$$\phi_k(t) = \sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_k(u-)\leq j-1\}\mathrm{d}D(u)$$

satisfies the following two properties:

(a) $\int_0^t h[\mathcal{Q}_k(u)]\mathrm{d}\phi_k(t) = 0$ for any continuous nonnegative function $h(x)$ with $h(0)=0$;

(b) the function $\int_0^t [Y(u)-\mathcal{Q}_k(u)]\mathrm{d}\phi_k(u)$ is not decreasing for any non-negative right-continuous function $Y(u)$ having the left limits.

Let us show (a). We have

$$\begin{aligned}\int_0^t h[\mathcal{Q}_k(u)]\mathrm{d}\phi_k(u) &= \int_0^t h[\triangle\mathcal{Q}_k(u)]\mathrm{d}\phi_k(u)\\ &= \int_0^t h[\triangle\mathcal{Q}_k(u)]\mathrm{d}\left(\sum_{j=1}^{C}\int_0^u \mathbf{I}\{\mathcal{Q}_k(v-)\leq j-1\}\mathrm{d}D(v)\right).\end{aligned}$$

Let u_i denote the points of jump of the process $\mathcal{Q}_k(u)$ in the interval $[0,t]$. For the last integral we have the following representation:

$$\begin{aligned}&\int_0^t h[\triangle\mathcal{Q}_k(u)]\mathrm{d}\left(\sum_{j=1}^{C}\int_0^u \mathbf{I}\{\mathcal{Q}_k(v-)\leq j-1\}\mathrm{d}D(v)\right)\\ &= \sum_{0\leq u_i\leq t}\sum_{j=1}^{C} h[\triangle\mathcal{Q}_k(u_i)]\triangle\mathcal{A}_k(u_i)\mathbf{I}\{\mathcal{Q}_k(u_i-)\leq j-1\}\triangle D(u_i).\end{aligned}$$

The last sum is a finite sum: the number of points u_i is finite in any finite interval $[0,t]$ with probability 1. Any value of jump $\triangle\mathcal{Q}_k(u_i)$ is bounded with probability 1, and the nonnegative continuous function $h[\triangle\mathcal{Q}_k(u)]$, satisfying the property $h(0)=0$ is therefore bounded for all $0\leq u\leq t$. In addition, taking into account that the jumps of the processes $\mathcal{A}_k(u)$ and $D(u)$ are disjoint, i.e. either $\triangle\mathcal{A}_k(u_i)=0$ or $\triangle D(u_i)=0$ with probability 1, we arrive at the conclusion that $\int_0^t h[\mathcal{Q}_k(u)]\mathrm{d}\phi_k(u)=0$. (a) follows.

(b) is implied by (a).

It follows from the Skorokhod reflection principle that the function $\phi_k(t)$ has the following representation:

$$\phi_k(t) = -\inf_{u\leq t} S_k(u).$$

Therefore $\mathcal{Q}_k(t)$ has the following representation

$$\begin{gathered}\mathcal{Q}_k(t) = S_k(t) - \inf_{u\leq t} S_k(u),\\ k=1,2,\ldots,\ell.\end{gathered}\tag{2.3}$$

Equation (2.3) is well-known in queueing theory. Following Borovkov [12], we have the following statement of the stability.

Theorem 2.2. *Assume*

$$\mathbb{P}\left\{\lim_{t\to\infty}\frac{\mathcal{A}_\ell(t)-\mathcal{D}(t)}{t}=r<0\right\}, \tag{2.4}$$

and $\widetilde{S}_k(t)$, $k=1,2,\ldots,\ell$, *are stationary point processes, the increments of which coincide in distribution with the corresponding increments of the processes* $\mathcal{A}_k(t)-\mathcal{D}(t)$, $k=1,2,\ldots,\ell$.

Then there exist stationary processes $\mathcal{Q}^{(k)}(T)$, $k=1,2,\ldots,\ell$, *such that*

$$\mathcal{Q}^{(1)}(T)\overset{'d'}{=}\sup_{u\le T}\left[\widetilde{S}_1(T)-\widetilde{S}_1(u)\right], \tag{2.5}$$

and

$$\begin{aligned}&\mathcal{Q}^{(k)}(T)\overset{'d'}{=}\sup_{u\le T}\left[\widetilde{S}_k(T)-\widetilde{S}_k(u)\right]-\sup_{u\le T}\left[\widetilde{S}_{k-1}(T)-\widetilde{S}_{k-1}(u)\right]\\ &k=2,3,\ldots,\ell.\end{aligned} \tag{2.6}$$

Proof. The proof is based on representation (2.3) and can be found in Borovkov [12]. Specifically, it follows from that proof that there are stationary processes $\mathcal{Q}_k(T)$, $k=1,2,\ldots,\ell$ such that

$$\mathcal{Q}_k(T)\overset{'d'}{=}\sup_{u\le T}\left[\widetilde{S}_k(T)-\widetilde{S}_k(u)\right]. \tag{2.7}$$

Therefore, keeping in mind that $\mathcal{Q}^{(k)}(t)=\mathcal{Q}_k(t)-\mathcal{Q}_{k-1}(t)$, $k=2,3,\ldots,\ell$ and $\mathcal{Q}^{(1)}(t)=\mathcal{Q}_1(t)$, from (2.7) we have (2.5) and (2.6). □

3. The finite buffers model

Equation (2.1) and other related equations for infinite buffers content can be easily extended for the model with finite buffers. It is assumed that if upon arrival of a batch the buffer of a given class overflows, then the complete arrival batch is rejected, see Remark 1.1.

For the analysis of the finite buffers case we introduce new arrival processes $\overline{\mathcal{A}}^{(k)}(t)$, which are derived from the initial processes $\mathcal{A}^{(k)}(t)$ as follows. We set

$$\triangle\overline{\mathcal{A}}^{(k)}(t)=\triangle\mathcal{A}^{(k)}(t)\mathbf{I}\left\{\mathcal{Q}^{(k)}(t)\le N^{(k)}\right\}. \tag{3.1}$$

The arrival processes $\overline{\mathcal{A}}^{(k)}(t)$ take into account only jumps of real buffer content process. Thus $A(t)-\overline{A}(t)$ is the number of lost units during time t, and $\mathcal{A}(t)-\overline{\mathcal{A}}(t)$ is their total length during that time t.

Then the buffer content process $\mathcal{Q}^{(1)}(t)$ is defined by the pair of equations

$$\mathcal{Q}^{(1)}(t)=\max\left\{0,\mathcal{Q}^{(1)}(t-)+\triangle\mathcal{A}^{(1)}(t)-\triangle\mathcal{D}(t)\right\}, \tag{3.2}$$

$$\mathcal{Q}^{(1)}(t+) = \max\left\{0, \mathcal{Q}^{(1)}(t-) + \triangle\overline{\mathcal{A}}^{(1)}(t) - \triangle\mathcal{D}(t)\right\}. \tag{3.3}$$

Thus in the case $\triangle\mathcal{A}^{(1)}(t) = \triangle\overline{\mathcal{A}}^{(1)}(t)$ the buffer contents $\mathcal{Q}^{(1)}(t)$ and $\mathcal{Q}^{(1)}(t+)$ are equal and there is no loss at time t. Otherwise, if $\triangle\mathcal{A}^{(1)}(t) \neq \triangle\overline{\mathcal{A}}^{(1)}(t)$, i.e. $\triangle\overline{\mathcal{A}}^{(1)}(t) = 0$ and $\triangle\mathcal{A}^{(1)}(t) > 0$, then there is a loss of a unit in time t.

Next, similarly to (1.5) for $k = 1, 2, \ldots, \ell - 1$ we have another pair of equations:

$$\begin{aligned}\mathcal{Q}^{(k+1)}(t) = \max\Big\{0, \mathcal{Q}^{(k+1)}(t-) + \triangle\mathcal{A}^{(k+1)}(t) \\ - \Big[\triangle\mathcal{D}(t) - \mathcal{Q}_k(t-) - \triangle\mathcal{A}_k(t)\Big]\mathbf{I}\{\mathcal{Q}_k(t) = 0\}\Big\},\end{aligned} \tag{3.4}$$

$$\begin{aligned}\mathcal{Q}^{(k+1)}(t+) = \max\Big\{0, \mathcal{Q}^{(k+1)}(t-) + \triangle\overline{\mathcal{A}}^{(k+1)}(t) \\ - \Big[\triangle\mathcal{D}(t) - \mathcal{Q}_k(t-) - \triangle\overline{\mathcal{A}}_k(t)\Big]\mathbf{I}\{\mathcal{Q}_k(t) = 0\}\Big\}.\end{aligned} \tag{3.5}$$

Similarly to Lemma 2.1, for the finite buffers model we have the following lemma.

Lemma 3.1. *For all continuity points of the kth cumulative buffer content process $\mathcal{Q}_k(t)$, $k = 1, 2, \ldots, \ell$, we have:*

$$\mathcal{Q}_k(t) = \overline{\mathcal{A}}_k(t) - \sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) \geq j\}\mathrm{d}D(u). \tag{3.6}$$

4. The formula for $J^{(k)}$

In this section we study the dynamics of the buffer lengths by level-crossings analysis for the infinite buffers model. It is assumed throughout that condition (2.4) for the stability is fulfilled.

In addition to the stability condition assume:

$$\mathbb{P}\left\{\lim_{t\to\infty}\frac{A^{(k)}(t)}{t} = \lambda^{(k)}\right\} = 1, \quad k = 1, 2, \ldots, \ell, \tag{4.1}$$

and

$$\mathbb{P}\left\{\lim_{t\to\infty}\frac{D(t)}{t} = \mu\right\} = 1. \tag{4.2}$$

Then according to (2.4) the sequences $\frac{1}{n}\sum_{i=1}^{n}\vartheta_i^{(k)}$, $k = 1, 2, \ldots, \ell$, as $n \to \infty$, also converges with probability 1.

Recall that $t_1^{(k)} = \tau_1^{(k)}$, $t_2^{(k)} = \tau_1^{(k)} + \tau_2^{(k)}$, ..., $(k = 1, 2, \ldots, \ell)$ denote the sequence of points (arrival moments) of the process $\mathcal{A}^{(k)}(t)$, and $x_1 = \chi_1$,

$x_2 = \chi_1 + \chi_2, \ldots$ denote the sequence of points (the moments of possible departure jumps) of $\mathcal{D}(t)$.

Then, for the number of up- and down-crossings for $m \geq 1$ we have the following equation:

$$
\begin{aligned}
&\sum_{i=1}^{A^{(k)}(t)} \mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right) \geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right) < m\right\} \\
&= \sum_{j=1}^{D(t)} \mathbf{I}\left\{m \leq \mathcal{Q}^{(k)}\left(x_j-\right) \leq m-1+C\right\} \\
&\quad + \mathbf{I}\left\{\mathcal{Q}^{(k)}(t) \geq m\right\} \\
&= \sum_{l=1}^{C}\sum_{j=1}^{D(t)} \mathbf{I}\left\{\mathcal{Q}^{(k)}\left(x_j-\right) = m-1+l\right\} \\
&\quad + \mathbf{I}\left\{\mathcal{Q}^{(k)}(t) \geq m\right\},
\end{aligned}
\tag{4.3}
$$

where $\mathcal{Q}^{(k)}(0) = 0$. Equation (4.3) can be explained as follows. The left-hand side of the equation is the number of arrivals until time t, seeing before arrival the buffer content less than m and at the moment of arrival not smaller than m. This constitutes the number of up-crossings of the level m until time t, i.e. the number of instants where arrivals jump over the level $m-1$. The first term of the right-hand side describes the number of departure moments when immediately before departure the buffer content is between m and $m+C-1$. (Then after the departure the buffer content is between $\max\{0, m-C\}$ and $m-1$, and this constitutes the number of down-crossings of the level m). The difference between the number of up-crossings and down-crossings of level m can be either 1 or 0, and the second term of the right-hand side compensates for this difference.

Dividing the both sides of (4.3) by t, and letting t increase unboundedly, we obtain:

$$
\begin{aligned}
&\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{i=1}^{A^{(k)}(t)} \mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right) \geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right) < m\right\} \\
&= \lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{l=1}^{C}\sum_{j=1}^{D(t)} \mathbf{I}\left\{\mathcal{Q}^{(k)}(x_j-) = m-1+l\right\},
\end{aligned}
\tag{4.4}
$$

and after elementary transformations (see Appendix B) we arrive at

$$\begin{aligned}
J^{(k)} &= \frac{\lambda^{(k)}}{\lambda^{(1)}+\lambda^{(2)}+\ldots+\lambda^{(\ell)}}\cdot\frac{1}{\lambda^{(k)}}\\
&\quad\times \lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t\sum_{l=1}^{C}\mathbf{I}\left\{\mathcal{Q}^{(k)}(u-)\geq N^{(k)}+l\right\}\mathrm{d}D(u)\\
&= \frac{1}{\lambda_\ell}\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{l=1}^{C}\int_0^t\mathbf{I}\left\{\mathcal{Q}^{(k)}(u-)\geq N^{(k)}+l\right\}\mathrm{d}D(u),\\
&\quad k=1,2,\ldots,\ell,
\end{aligned}\tag{4.5}$$

where $\lambda_\ell=\lambda^{(1)}+\lambda^{(2)}+\ldots+\lambda^{(\ell)}$.

5. The buffers content distribution of $GI/M^{Y=C}/1$ queues

5.1. **Main result.** We start this section with a representation for the buffer content processes in the case where the arrival processes $A^{(k)}(t)$, $k=1,2,\ldots,\ell$ all satisfy (4.1), and the process $D(t)$ is Poisson. Assume also that $\vartheta_n=1$ for all n.

We have:

$$\begin{aligned}
&\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t\mathbf{I}\{\mathcal{Q}_k(u-)=m-1\}\mathrm{d}A_k(u)\\
&=\mu\lim_{t\to\infty}\frac{1}{t}\sum_{l=1}^{C}\int_0^t\mathbb{P}\{\mathcal{Q}_k(u)=m-1+l\}\mathrm{d}u,\\
&\quad k=1,2,\ldots,\ell.
\end{aligned}\tag{5.1}$$

However (5.1) does not permit us to obtain explicit results for the stationary probabilities even in the case where the processes $A^{(k)}(t)$, $k=1,2,\ldots,\ell$ all are renewal processes. Moreover, in the case where all processes $A^{(k)}(t)$, $k=1,2,\ldots,\ell$ are renewal, Lemma 2.1 is no longer useful in general, because stationary interarrival times to cumulative buffers are dependent in general, and the corresponding stationary arrival processes are not longer renewal.

Therefore we consider the following special case of the general buffers model. Let $A(t)$ be a point process of arrivals satisfying the condition $\mathbb{P}\{\lim_{t\to\infty}\frac{A(t)}{t}=\lambda\}=1$. Let $\pi^{(1)}$, $\pi^{(2)}$,..., $\pi^{(\ell)}$ be positive probabilities, $\sum_{k=1}^{\ell}\pi^{(k)}=1$, where $\pi^{(k)}$ is a probability that an arriving customer belongs to the class k. Then the points processes $A^{(k)}(t)$, $k=1,2,\ldots,\ell$, are all thinnings of the original process $A(t)$, and in the case where $A(t)$ is a renewal process all the processes $A^{(k)}(t)$, $k=1,2,\ldots,\ell$ are renewal processes as well with intensities $\lambda^{(k)}=\lambda\pi^{(k)}$ correspondingly. Consequently, the processes $A_k(t)$, $k=1,2,\ldots,\ell$, are renewal processes with intensities λ_k $=\lambda^{(1)}+\lambda^{(2)}+\ldots+\lambda^{(k)}$, and one can apply the theory to the $GI/M^{Y=C}/1$ and $GI/M^{Y=C}/1/N_k$ queues with large buffers N_k.

By $GI/M^{Y=C}/1$ queue we mean a single-server queueing system with recurrent input and exponentially distributed service time of the constant size batch C. In the sequel we use the notation $GI/M^C/1$ for these queueing systems. $GI/M^C/1$ queueing systems are particular systems with an autonomous service mechanism, and they are therefore described by buffer type stochastic differential equations or by one of the above equivalent forms of these equations. For these queueing systems therefore Lemma 2.1 remains true. Specifically, from this lemma one can conclude that the cumulative buffer content processes are described by the steady-state distributions of the usual queue-length processes of the $GI/M^C/1$ queues. The stability condition for these queues is $\rho_\ell = \frac{\lambda_\ell}{\mu C} < 1$.

Using a standard method, the limiting and stationary probabilities for the cumulative buffer contents of $GI/M^C/1$ queues are calculated as follows. Let $t_{k,j}$ denote the jth arrival moment to one of the first k buffers. Then for limiting and stationary probability we have the following.

Theorem 5.1. *For cumulative buffer contents $\mathcal{Q}_k(t)$, $k = 1, 2, \ldots, \ell$.*

$$P_{k,m} = \lim_{j\to\infty} \mathbb{P}\{\mathcal{Q}_k(t_{k,j}-) = m\} = \varsigma_k^m(1-\varsigma_k), \tag{5.2}$$

where ς_k is the (unique) root of the functional equation

$$z = \widehat{B}_k(\mu - \mu z^C) \tag{5.3}$$

in the interval (0,1), and $\widehat{B}_k(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}B_k(x)$ is the Laplace-Stieltjes transform of the stationary distribution of interarrival time $B_k(x)$ to the first k buffers.

Proof. The stationary probabilities of $GI^X/M^Y/1$ queues can be found in Economou and Fakinos [20], and the statement for $GI/M^C/1$ queues can be deduced from their result.[1] However, the direct proof of the result for the $GI/M^C/1$ queue is much simpler than that reduction from the aforementioned general result. Therefore below the direct proof of this theorem is provided.

First of all notice, that according to (5.1) the state probabilities immediately before arrival, $P_{k,m}$, are

$$P_{k,m} = \lim_{t\to\infty} \frac{1}{\lambda_k t}\mathbb{E}\int_0^t \mathbf{I}\{\mathcal{Q}_k(u-) = m\}\mathrm{d}A_k(u),$$

and $P_{k,m} = (1-z)z^m$ for some $z < 1$. Let f_m denote the number of up-(down-) crossing of level m during a busy period of $GI/M^C/1$ queue (the number of cases where immediately before arrival there are m customers in the system). Then, by renewal arguments $\mathbb{E}f_m = z^m$, and according to the

[1] The following additional condition is missed in the main statement of [20]: the common divisor of possible values X and Y must be equal to 1.

total expectation formula for any $m \geq 1$ we have the following equation:

$$z^m = \sum_{i=0}^{\infty} \int_0^{\infty} \mathrm{e}^{-\mu x} \frac{\left(\mu x z^C\right)^i}{i!} \; z^{m-1} \mathrm{d}B_k(x), \tag{5.4}$$

where $B_k(x)$ is the probability distribution function of interarrival time. Therefore, from (5.4) we obtain

$$\begin{aligned} z &= \sum_{i=0}^{\infty} \int_0^{\infty} \mathrm{e}^{-\mu x} \frac{\left(\mu x z^C\right)^i}{i!} \; \mathrm{d}B_k(x) \\ &= \widehat{B}_k(\mu - \mu z^C), \end{aligned}$$

and the statement of Theorem 5.1 follows. By standard method (see e.g. [42], [28]) one can prove that under the assumption $\rho_\ell < 1$ there exists a unique root of equation $z = \widehat{B}_k(\mu - \mu z^C)$ in the interval (0,1). □

5.2. **Particular case.** We consider an $M/M^C/1$ queueing system with infinite buffers. This particular case is easily deduced from the statement of Theorem 5.1. Specifically, in the case of Poisson arrivals from (5.3) we obtain the equation:

$$z = \frac{\lambda_k}{\lambda_k + \mu - \mu z^C}.$$

Then, the constant ς_k must be the solution of equation

$$\frac{\lambda_k}{\mu} = \sum_{i=1}^{C} z^i, \tag{5.5}$$

belonging to the interval (0,1). A similar result can be also found in [15] for nodes of a network, the customers of which are served by random batches.

6. Loss probabilities for cumulative buffers

In this section we discuss loss probabilities assuming that the kth cumulative buffer content has large capacity N_k. We study buffer loss probabilities under "usual" and heavy load conditions. By "usual" conditions we mean the case when the load parameter of the queueing system is fixed, while in the case of heavy load conditions the sequence of load parameters, associated with series of queueing systems, approaches 1.

6.1. **Loss probabilities under "usual" conditions.** We use the notation $GI/M^C/1/N_k$ for the queueing systems with finite capacity N_k, similar to the notation used for the queueing systems with infinite capacity in the previous section. According to Lemma 3.1 the cumulative buffer contents in continuity points of the process $\mathcal{Q}_k(t)$ behave as usual $GI/M^C/1/N_k$ queues. However, the behavior of the number of losses, the main characteristic of interest, is essentially different, that is the losses in $GI/M^C/1/N_k$ queues are not equal to the losses in the corresponding cumulative buffers $\mathcal{Q}_k(t)$. Specifically, the losses in $GI/M^C/1/N_k$ queues occur only in the case in which the buffer overflowed when the arriving customer met all waiting places busy. The losses in the cumulative buffers $\mathcal{Q}_k(t)$ can occur in many cases when one of specific buffers, say jth buffer, $1 \leq j \leq k$, has overflowed.

However, in some cases when the values $N_1 < N_2 < \ldots < N_\ell$ all are large, a correspondence between $GI/M^C/1/N_k$ queues and finite buffers models, may give useful asymptotic results.

Specifically, the loss probability of a customer arriving at one of the first k buffers is not greater than $p_1 + p_2 + \ldots + p_k$, where p_i denotes the loss probability in the corresponding $GI/M^C/1/N_i$ queueing system, the probability distribution of interarrival time of which is $B_i(x)$. All probabilities p_k, $1 \leq k \leq \ell$ are very small as is N_k large. They decrease geometrically fast (see Theorem 6.1 below), and the finite sum of these probabilities seems to remain a good upper bound for the buffers loss probability.

Theorem 6.1. *The buffer contents loss probability is not greater than* $p_1 + p_2 + \ldots + p_k$, $k = 1, 2, \ldots, \ell$, *where*

$$p_k = \frac{(1-\rho_k)[1 + C\mu\widehat{B}'_k(\mu - \mu\varsigma_k^C)]\varsigma_k^{N_k}}{(1-\rho_k)(1 + \varsigma_k + \ldots + \varsigma_k^{C-1}) - \rho_k[1 + C\mu\widehat{B}'_k(\mu - \mu\varsigma_k^C)]\varsigma_k^{N_k}} + o\left(\varsigma_k^{2N_k}\right), \tag{6.1}$$

$$\rho_k = \frac{\lambda_k}{C\mu},$$

and ς_k *is the (least) root of the functional equation*

$$z = \widehat{B}_k(\mu - \mu z^C)$$

in the interval (0,1).

Proof. We consider the $GI/M^C/1/N_k$ queueing system. Following Miyazawa [36], the loss probability for the $GI/M^Y/1/N_k$ queueing system is determined by the formula

$$p_k = \frac{1}{\sum_{j=0}^{N_k} \pi_{k,j}},$$

where the generating function of $\pi_{k,j}$, $j = 1, 2, \ldots$ is

$$\Pi_k(z) = \sum_{j=0}^{\infty} \pi_{k,j} z^j = \frac{(1 - Y(z))\widehat{B}_k(\mu - \mu Y(z))}{\widehat{B}_k(\mu - \mu Y(z)) - z}, \tag{6.2}$$

and $Y(z)$ is the generating function of complete service batch. In the case of the $GI/M^C/1/N_k$ queueing system $Y(z) = z^C$, and (6.2) can be then rewritten as

$$\Pi_k(z) = \frac{(1 - z^C)\widehat{B}_k(\mu - \mu z^C)}{\widehat{B}_k(\mu - \mu z^C) - z}. \tag{6.3}$$

In the particular case of $C = 1$ the asymptotic behaviour of the loss probability has been studied in [2] and [17]. In the case of $\rho_k < 1$ it was based on an application of the Takács theorem [43], p. 22-23.

In the case of $C>1$ the scheme of the proof is similar. Expanding (1-z^C) in the numerator of (6.3) as $1-z^C=(1-z)(1+z+\ldots+z^{C-1})$ we have

$$\Pi_k(z)=\frac{(1-z)(1+z+z^2+\ldots+z^{C-1})\widehat{B}_k(\mu-\mu z^C)}{\widehat{B}_k(\mu-\mu z^C)-z}. \tag{6.4}$$

Therefore, the other generating function $\widetilde{\Pi}_k(z)=\frac{1}{1-z}\Pi_k(z)$ is

$$\begin{aligned}\widetilde{\Pi}_k(z)&=\sum_{i=0}^{\infty}\widetilde{\pi}_{k,i}z^i=\sum_{i=0}^{\infty}\left(\sum_{j=0}^{i}\pi_{k,j}\right)z^i\\&=\frac{(1+z+z^2+\ldots+z^{C-1})\widehat{B}_k(\mu-\mu z^C)}{\widehat{B}_k(\mu-\mu z^C)-z},\end{aligned} \tag{6.5}$$

and the loss probability is

$$p_k=\frac{1}{\widetilde{\pi}_{k,N_k}}. \tag{6.6}$$

Our goal is therefore to find the asymptotic behaviour of $\widetilde{\pi}_{k,N_k}$ as $N_k\to\infty$.

The equation $z=\widehat{B}_k(\mu-\mu z^C)$ has exactly one solution ς_k in the interval (0,1). Furthermore, $\widehat{B}_k(\mu-\mu z^C)$ is the probability generating function of some integer random variable, i.e.

$$\widehat{B}_k(\mu-\mu z^C)=R(z)=\sum_{j=0}^{\infty}r_jz^j,$$

and

$$\widetilde{\Pi}_k(z)=F(z)\sum_{i=0}^{C-1}z^i,$$

where

$$F(z)=\sum_{n=0}^{\infty}f_nz^n=\frac{R(z)}{R(z)-z}.$$

Therefore (see Takács [43]), the sequence $\{f_n\}$ satisfies the recurrence relation

$$f_n=\sum_{j=0}^{n}f_{n-j+1}r_j,\quad f_0>0.$$

Since $\varsigma_k<1$, we correspondingly obtain $\gamma=\sum_{n=1}^{\infty}nr_n>1$. According to formula (35) of [43], p. 23

$$\lim_{n\to\infty}\left[f_n-\frac{f_0\delta^{-n}}{[1-F'(\delta)]}\right]=\frac{f_0}{1-\gamma}, \tag{6.7}$$

where δ is the least root of equation $z=R(z)$ in the interval (0,1).

In our case $\rho_k = \frac{1}{\gamma}$, and $\varsigma_k = \delta$, and we have:

$$\begin{aligned}\widetilde{\pi}_{k,N_k} =& \frac{1+\varsigma_k+\ldots+\varsigma_k^{C-1}}{\varsigma_k^{N_k}} \cdot \frac{1}{1+C\mu\widehat{B}_k'(\mu-\mu\varsigma_k^C)} \\ &+ \frac{(1+\varsigma_k+\ldots+\varsigma_k^{C-1})\rho_k}{\rho_k-1} + o(1),\end{aligned}$$

and the statement of the proposition follows from (6.6) after some algebraic transformations. □

6.2. **Loss probabilities under heavy load conditions.** The loss probabilities under heavy load conditions for $GI/M/1/n$ queues have been recently studied in [2] and [46]. For the further development of these results see also [6] and [47]. In this specific case, the behaviour of the system under heavy load condition differs from the classic cases considered in these papers.

We consider the case of heavy load conditions and assume that the load parameter $\rho_\ell = \frac{\lambda_\ell}{C\mu}$ is close to 1. More specifically, we assume that $\rho_\ell = \rho_\ell(\delta) < 1$ (δ is a small parameter), and $\rho_\ell(\delta)$ approaches 1 from the left as δ vanishes. Denote $\rho_{\ell,j} = \mu^j \int_0^\infty x^j \mathrm{d}B_\ell(x)$ $(\rho_{\ell,1} = \frac{1}{C\rho_\ell})$. We have the following result.

Theorem 6.2. *Assume that $\rho_\ell(\delta) < 1$, $\rho_\ell(\delta)$ approaches 1 from the left, and $\delta N_\ell(\delta) \to \Delta > 0$ as δ vanishes. Assume also that $\widetilde{\rho}_{\ell,2} = \lim_{\delta\to 0}\rho_{\ell,2}(\delta)$, and $\rho_{\ell,3}(\delta)$ remains bounded as δ vanishes. Then*

$$p_\ell = \frac{\delta \exp\left(-\frac{\Delta}{\binom{C}{2}\widetilde{\rho}_{\ell,2}}\right)}{C - \exp\left(-\frac{\Delta}{\binom{C}{2}\widetilde{\rho}_{\ell,2}}\right)} \cdot [1+o(1)]. \tag{6.8}$$

Proof. Let us first derive the expansion for the least root of equation $z = \widehat{B}_\ell(\mu - \mu z^C)$ under the assumption of the theorem. Clearly, the root of this equation approaches 1 as δ vanishes. Therefore, using the Taylor expansion of $\widehat{B}_\ell(\mu-\mu z^C)$ as δ vanishes, we obtain the following equation for ς_ℓ

$$z = 1 - (1-\delta)(1-z) + \binom{C}{2}\widetilde{\rho}_{\ell,2}(1-z)^2 + o(1-z)^3.$$

Ignoring the last term $o(1-z)^3$ we have the quadratic equation, the solutions of which are $z=1$ and $z = 1-\delta/\left[\binom{C}{2}\widetilde{\rho}_{\ell,2}\right]$. Therefore we obtain

$$\varsigma_\ell = 1 - \frac{\delta}{\binom{C}{2}\widetilde{\rho}_{\ell,2}} + o(\delta). \tag{6.9}$$

Notice, that representation similar to (6.9) for the root of equation $z = \widehat{B}_\ell(\mu - \mu z)$ (particular case where C=1) has been obtained in Subhankulov [41], p. 326.

Next, the asymptotic representation for p_k, $k = 1, 2, \dots, \ell$ is given by (6.1). For $k = \ell$ the main term of asymptotic expansion of

$$[1 + C\mu\widehat{B}'_\ell(\mu - \mu\varsigma_\ell^C)]\varsigma_\ell^{N_\ell}$$

is given by

$$\delta \exp\left(-\frac{\Delta}{\binom{C}{2}\widetilde{\rho}_{\ell,2}}\right)$$

(see [2], [6] for details of the proof), and according to (6.9) the main term of the asymptotic expansion of

$$1 + \varsigma_\ell + \dots + \varsigma_\ell^{C-1}$$

is given by

$$C - \frac{\delta}{\widetilde{\rho}_{\ell,2}}.$$

Next notice, that the expansion for the term

$$1 - \rho_\ell - \rho_\ell[1 + C\mu\widehat{B}'_k(\mu - \mu\varsigma_\ell^C)]\varsigma_\ell^{N_\ell}$$

is

$$\delta\left[1 - \exp\left(-\frac{\Delta}{\binom{C}{2}\widetilde{\rho}_{\ell,2}}\right)\right][1 + o(1)].$$

Therefore, asymptotic relation (6.8) follows. □

7. Approximation of the solution in particular cases

In this section we discuss the approximation of the solution for the problem stated in Section 1.4 that is to minimize functional (1.7) containing the terms $\alpha^{(k)}J^{(k)}$ associated with buffer contents $\mathcal{Q}^{(k)}(t)$, $k = 1, 2, \dots, \ell$.

However, in all particular cases above the explicit solutions were obtained for the cumulative buffer contents $\mathcal{Q}_k(t)$, $k = 1, 2, \dots, \ell$, and the solution of the problem in the initial terms seems to be hard. Therefore we formulate and solve the problem in new terms. This solution of the new problem is then used to approximate the desired solution of the initial problem.

Let us first introduce new functionals J_k instead of the $J^{(k)}$ ($k = 1, 2, \dots, \ell$), which were introduced in Section 1.4.

Namely, let N_1, $N_2, \dots, N_\ell$ denote large integer numbers, $N_1 < N_2 < \dots < N_\ell$. We set

$$J_k = \mathbb{P}\text{-}\lim_{t\to\infty} \frac{1}{A_\ell(t)} \sum_{j=1}^{A_k(t)} \mathbf{I}\left\{\mathcal{Q}_k(t_{k,j}) > N_k\right\}, \quad k = 1, 2, \dots, \ell.$$

Note first (see (4.5)) that for $J_k,\ k=1,2,\dots,\ell$ we have the following representation:

$$J_k=\frac{1}{\lambda_\ell}\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{l=1}^{C}\int_0^t \mathbf{I}\left\{\mathcal{Q}_k(u-)\geq N_k+l\right\}\mathrm{d}D(u),\qquad k=1,2,\dots,\ell, \tag{7.1}$$

where $\lambda_k=\lambda^{(1)}+\lambda^{(2)}+\ldots+\lambda^{(k)}$, $k=1,2,\dots,\ell$. The proof of representation (7.1) is similar to the proof of (4.5) with minor difference in the notation.

Replacing functional (1.7) by

$$\overline{J}=\alpha_1J_1+\alpha_2J_2+\ldots+\alpha_\ell J_\ell, \tag{7.2}$$

we have then the following problems similar to the problems formulated above in Section 1.4.

1. Assuming that $N_1,\ N_2,\dots,N_\ell$ are known, minimize C subject to $\overline{J}\leq\varepsilon$.

2. Assume that C is known, but $N_1,\ N_2,\dots,N_\ell$ are unknown. Assume additionally that with given $\beta_1,\ \beta_2,\dots,\beta_{\ell-1}$ the values $N_1,\ N_2,\dots,\ N_\ell$ must satisfy the condition: $N_1=\lfloor\beta_1N_2\rfloor=\lfloor\beta_2N_3\rfloor=\ldots=\lfloor\beta_{\ell-1}N_\ell\rfloor$. The problem is to minimize N_1 subject to $\overline{J}\leq\varepsilon$.

The values $\alpha_1,\ \alpha_2,\ \dots,\ \alpha_\ell;\ N_1,\ N_2,\dots,\ N_\ell$ are unknown, and by approximation of the solution of the problem we hope to find a correspondence between the vectors $(\alpha_1,\ \alpha_2,\ \dots,\ \alpha_\ell)$ and $(\alpha^{(1)},\ \alpha^{(2)},\ \dots,\ \alpha^{(\ell)})$ and between the vectors $(N_1,\ N_2,\dots,\ N_\ell)$ and $(N^{(1)},\ N^{(2)},\dots,\ N^{(\ell)})$ such that the solution of the initial problems formulated in Section 1.4 and the problems formulated in this section would be approximately the same.

Consider first the queueing systems with infinite number of waiting places, say $GI/M^C/1$ queues.

Notice, that $\alpha_1=\alpha^{(1)}$ and $N_1=N^{(1)}$. According to Theorem 5.1 the expected queue-length of the kth cumulative buffer content immediately before arrival of a unit is

$$\lim_{t\to\infty}\frac{1}{t}\sum_{m=1}^{\infty}m\int_0^t\mathbf{I}\{\mathcal{Q}_k(u)=m\}\mathrm{d}u=\frac{\varsigma_k}{1-\varsigma_k}. \tag{7.3}$$

From (7.3) we have the following. Put

$$p_{2,1}=\frac{\dfrac{\varsigma_1}{1-\varsigma_1}}{\dfrac{\varsigma_2}{1-\varsigma_2}}=\frac{\varsigma_1(1-\varsigma_2)}{\varsigma_2(1-\varsigma_1)},\qquad p_{2,2}=1-p_{2,1}, \tag{7.4}$$

and then

$$\alpha_2=\alpha_1p_{2,1}+\alpha^{(2)}p_{2,2}. \tag{7.5}$$

Similarly to (7.4) and (7.5) for $k=1,2,\ldots,\ell-1$ we set

$$p_{k+1,1}=\frac{\varsigma_k(1-\varsigma_{k+1})}{\varsigma_{k+1}(1-\varsigma_k)},\qquad (7.6)$$
$$p_{k+1,2}=1-p_{k+1,1},$$

and

$$\alpha_{k+1}=\alpha_k p_{k+1,1}+\alpha^{(k+1)}p_{k+1,2}. \qquad (7.7)$$

Let us now express the correspondence between the vectors ($N^{(1)}$, $N^{(2)}$,..., $N^{(\ell)}$) and (N_1, N_2,..., N_ℓ). Let $\beta^{(2)}$, $\beta^{(3)}$,..., $\beta^{(\ell)}$ be such the real numbers that $N^{(1)}=\lfloor\beta^{(2)}N^{(2)}\rfloor=\lfloor\beta^{(3)}N^{(3)}\rfloor=\ldots=\lfloor\beta^{(\ell)}N^{(\ell)}\rfloor$.

Then for the purpose of approximation the values β_2, β_3,...,β_ℓ are taken as

$$\beta_k=\frac{\beta^{(k)}}{1+\beta^{(k)}},\ k=2,3\ldots,\ell, \qquad (7.8)$$

and $N_1=\lfloor\beta_2N_2\rfloor=\lfloor\beta_3N_3\rfloor=\ldots=\lfloor\beta_\ell N_\ell\rfloor$.

For the queueing model with large finite buffers, say $GI/M^C/1/N_k$, $k=1,2,\ldots,\ell$, the approximation is similar. Specifically, since the buffers are large, approximation for (α_1, α_2,..., α_ℓ) can be given by (7.3)-(7.6). The values β_2, β_3,...,β_ℓ are assumed to be taken by the same relation (7.8).

8. Minimization algorithms for the functional $\overline{J}$

In this section we discuss the problem of minimization of the functional $\overline{J}$ defined by (7.2). J_1, J_2,..., J_ℓ depends on parameters C, N_1, N_2,..., N_ℓ.

8.1. C is known while N_1 is unknown.

Assume first that C is known, N_1 is unknown, and the problem is to find the value N_1 minimizing the functional $\overline{J}$ in the buffer models, where explicit representation for the state probabilities as well as for J_k, $k=1,2,\ldots,\ell$ are known. These models are considered in Sections 5 and 6.

To be specific we refer to the models of infinite buffers of $GI/M^C/1$ queues. The algorithm has the following steps.

- *Step 1.* Calculate ς_k, $k=1,2,\ldots,\ell$. Recall that ς_k is the root of the functional equation (5.3) in the interval (0,1). For each k it can be calculated by the fixed point method or by one of other well-known methods, say direct search method or gold section method (e.g. see [29]).

- *Step 2.* We have ℓ geometric distributions obtained in Step 1, and therefore one can compute the corresponding values N_1, N_2,...,N_ℓ at which each of the tails of the geometric distributions multiplied to the corresponding coefficient α_k, $k=1,2,\ldots,\ell$ will be less than ε (i.e. $\alpha_kJ_k<\varepsilon$).

- *Step 3.* By using the known coefficients β_2, β_3, ..., β_ℓ one can find the value N_1^{lower}. N_1^{lower} is the maximum amongst all minimal values of

N_1 under which $\alpha_k J_k < \varepsilon$ for all $k = 1, 2, \ldots, \ell$. Specifically, we have the system:

$$\min N_1 : \ \alpha_1 J_1(N_1) < \varepsilon,$$
$$\min N_1 : \ \alpha_2 J_2(N_1) < \varepsilon,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\min N_1 : \ \alpha_\ell J_\ell(N_1) < \varepsilon,$$

and N_1^{lower} is the maximum amongst ℓ obtained values of N_1.

• *Step 4.* By using the same known coefficients β_2, β_3, $\ldots$, β_ℓ one can find the value N_1^{upper}. N_1^{upper} is the minimum amongst all maximal values of N_1 under which $\alpha_k J_k < \frac{\varepsilon}{\ell}$ for all $k = 1, 2, \ldots, \ell$. Specifically, we have the system:

$$\min N_1 : \ \alpha_1 J_1(N_1) < \frac{\varepsilon}{\ell},$$
$$\min N_1 : \ \alpha_2 J_2(N_1) < \frac{\varepsilon}{\ell},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\min N_1 : \ \alpha_\ell J_\ell(N_1) < \frac{\varepsilon}{\ell},$$

and N_1^{upper} is the maximum amongst ℓ obtained values of N_1.

• *Step 5.* We solve the following integer programming problem:

$$\text{minimize } N_1 : N_1^{lower} \leq N_1 \leq N_1^{upper}$$

$$\text{subject to } \alpha_1 J_1 + \alpha_2 J_2 + \ldots + \alpha_\ell J_\ell \leq \varepsilon.$$

8.2. **N_1, N_2,…,N_ℓ are known while C is unknown.** In the case where N_1, N_2,…, N_ℓ all are known but C is unknown the algorithm of the problem solution is the following.

• *Step 1.* From the stability condition find the lower (integer) bound for C:

$$C^{lower} = \min\left\{C : \frac{\lambda_\ell}{C\mu} < 1\right\}.$$

• *Step 2.* Find ς_k, $k = 1, 2, \ldots, \ell$.

• *Step 3.* Compute the functional $\overline{J}$.

• If $\overline{J} > \varepsilon$, then find a new value C and repeat steps 1-3. These procedure should be repeated more and more while $\overline{J} > \varepsilon$. Since the upper bound of C is unknown, the value C should be found according to the special search procedure offered by Rubalskii [39].

Rubalskii [39] proposed the minimization algorithm for a unimodal function on an unbounded set. The optimal algorithm is an extension of the standard Fibonacci procedure.

9. Concluding remarks

In this paper we studied queueing systems with priority classes and infinite and finite buffers. We derived general type equations for buffer content processes assuming that service mechanism is autonomous. The results of general theory were then applied to special queueing models with exponentially distributed service times. These queueing systems are a particular case of systems with an autonomous service mechanism. For the model having large buffers we derived an asymptotic result for the loss probability. We developed an algorithm for a solution of the problem numerically.

Acknowledgements. The author expresses his gratitude to Prof. Robert Liptser for providing the author relevant information about this problem. The author indebts to Prof. Boris Miller for his advice and some comments. Careful reading and comments of Prof. Aidan Sudbury are appreciated. The author also thanks the anonymous referee for careful reading and comments, which substantially improved the presentation. The research was supported by the Australian Research Council, grant # DP0771338.

APPENDIX A: Proof of Lemma 2.1

We start from equation (1.1). In order to write this equation in the customary form of a stochastic equation, we use the process $D(t)=\sum_{n=1}^{\infty}\mathbf{I}\{x_n\leq t\}$. The jumps of the process $D(t)$ are equal to 1, and according to the definition, we have $CD(t)=\mathcal{D}(t)$ for all $t\geq 0$. Then due to the assumption that arrival and departure jumps are disjoint, (1.1) can be rewritten

$$\mathcal{Q}^{(1)}(t)=\mathcal{A}^{(1)}(t)-\sum_{j=1}^{C}\int_0^t\mathbf{I}\Big\{\mathcal{Q}^{(1)}(u-)\geq j\Big\}\mathrm{d}D(u). \tag{A.1}$$

The equivalence of representations of (1.1) and (A.1) can be easily checked by considering a small time interval $(t-\delta,t]$ containing exactly one event as either arrival or departure of a unit. Then the term

$$\mathbf{I}\Big\{\mathcal{Q}^{(1)}(u-)\geq j\Big\} \tag{A.2}$$

of the integrand shows that if u is the point of jump of the process $\mathcal{D}(t)$, and $\mathcal{Q}^{(1)}(u-)=n\leq C$, then $\mathcal{Q}^{(1)}(u)=0$.

Similarly to (A.1), equation (1.2) can be rewritten as follows:

$$\mathcal{Q}^{(2)}(t)=\mathcal{A}^{(2)}(t)-\sum_{j=1}^{C}\int_0^t\mathbf{I}\left\{\mathcal{Q}^{(1)}(u)=0\right\} \times\mathbf{I}\left\{\mathcal{Q}^{(2)}(u-)\geq j-\mathcal{Q}^{(1)}(u-)\right\}\mathrm{d}D(u). \tag{A.3}$$

The explanation of the equivalence of (A.1) and (A.3) is similar to the above case, but slightly more complicated in details. Specifically, the presence of the term $\mathbf{I}\left\{\mathcal{Q}^{(1)}(u)=0\right\}$ in the integrand is obvious, and the validation of the term

$$\mathbf{I}\left\{\mathcal{Q}^{(2)}(u-)\geq j-\mathcal{Q}^{(1)}(u-)\right\}$$

is explained similarly to that of (A.2).

Let us now find the representation for $\mathcal{Q}_2(t)=\mathcal{Q}^{(1)}(t)+\mathcal{Q}^{(2)}(t)$. Keeping in mind that $\mathcal{Q}_1(t)=\mathcal{Q}^{(1)}(t)$ and $\mathcal{A}_1(t)=\mathcal{A}^{(1)}(t)$ from (A.1) and (A.3) we obtain:

$$\begin{aligned}\mathcal{Q}_2(t) &= \mathcal{A}_2(t)-\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_1(u-)\geq j\}\mathrm{d}D(u)\\ &\quad -\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_1(u-)<j\}\mathbf{I}\{\mathcal{Q}_2(u-)\geq j\}\mathrm{d}D(u)\\ &= \mathcal{A}_2(t)-\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_2(u-)\geq j\}\mathrm{d}D(u).\end{aligned}$$

The term

$$\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_1(u-)\geq j\}\mathrm{d}D(u)$$

characterizes departure lengths from the highest priority buffer, while the term

$$\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_1(u-)<j\}\mathbf{I}\{\mathcal{Q}_2(u-)\geq j\}\mathrm{d}D(u)$$

characterizes that from the second buffer of the lower priority.

Thus for $k=1,2$ we have already shown

$$\mathcal{Q}_k(t)=\mathcal{A}_k(t)-\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_k(u-)\geq j\}\mathrm{d}D(u).$$

Let us prove (2.1) by using induction.

For this purpose let us write first a representation for $\mathcal{Q}^{(k+1)}(t)$, $k=1,2,\ldots,\ell-1$. Similarly to (A.3) we have:

$$\begin{aligned}\mathcal{Q}^{(k+1)}(t)=\mathcal{A}^{(k+1)}(t)-\sum_{j=1}^{C}\int_0^t \mathbf{I}\left\{\mathcal{Q}_k(u)=0\right\}\\ \times\mathbf{I}\left\{\mathcal{Q}^{(k+1)}(u-)\geq j-\mathcal{Q}_k(u-)\right\}\mathrm{d}D(u).\end{aligned} \tag{A.4}$$

Equation (A.4) is a straightforward extension of (A.3). Therefore, assuming that (2.1) is valid for some k and adding $\mathcal{Q}_k(t)$ and $\mathcal{Q}^{(k+1)}(t)$, and similarly to the above for the $k+1$st cumulative buffer content we obtain:

$$\mathcal{Q}_{k+1}(t) = \mathcal{A}_{k+1}(t)-\sum_{j=1}^{C}\int_0^t \mathbf{I}\{\mathcal{Q}_{k+1}(u-)\geq j\}\mathrm{d}D(u).$$

Representation (2.1) is proved.

APPENDIX B: Deriving (4.5)

Equation (4.4) is a basic equation for our analysis. With $\infty \cdot 0 = 0$ for the left-hand side of (4.4) we have:

$$
\begin{aligned}
&\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{i=1}^{A^{(k)}(t)}\mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right)\geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right)<m\right\}\\
&=\lambda^{(k)}\lim_{t\to\infty}\mathbb{E}\frac{1}{A^{(k)}(t)}\sum_{i=1}^{A^{(k)}(t)}\mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right)\geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right)<m\right\}\\
&=\lambda^{(k)}\lim_{t\to\infty}\sum_{l=0}^{\infty}\frac{1}{l}\sum_{i=1}^{l}\mathbb{P}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right)\geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right)<m\right\}\\
&\quad\times\mathbb{P}\left\{A^{(k)}(t)=l\right\}\\
&=\lambda^{(k)}\lim_{t\to\infty}\mathbb{E}\sum_{l=0}^{\infty}\frac{1}{l}\sum_{i=1}^{l}\mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right)\geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right)<m\right\}\\
&\quad\times\mathbb{P}\left\{A^{(k)}(t)=l\right\}\\
&=\lambda^{(k)}\mathbb{P}\text{-}\lim_{t\to\infty}\frac{1}{A^{(k)}(t)}\sum_{i=1}^{A^{(k)}(t)}\mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right)\geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right)<m\right\},\\
&\quad k=1,2,\ldots,\ell.
\end{aligned}
$$

The right-hand side of (4.4) can be rewritten as follows:

$$
\begin{aligned}
&\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{l=1}^{C}\sum_{j=1}^{D(t)}\mathbf{I}\left\{\mathcal{Q}^{(k)}(x_j-)=m-1+l\right\}\\
&=\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{l=1}^{C}\int_0^t\mathbf{I}\left\{\mathcal{Q}^{(k)}(u-)=m-1+l\right\}\mathrm{d}D(u),\\
&\quad k=1,2,\ldots,\ell.
\end{aligned}
$$

From these last two equations we obtain:

$$
\begin{aligned}
&\mathbb{P}\text{-}\lim_{t\to\infty}\frac{1}{A^{(k)}(t)}\sum_{i=1}^{A^{(k)}(t)}\mathbf{I}\left\{\mathcal{Q}^{(k)}\left(t_i^{(k)}\right)\geq m,\ \mathcal{Q}^{(k)}\left(t_i^{(k)}-\right)<m\right\}\\
&=\frac{1}{\lambda^{(k)}}\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{l=1}^{C}\int_0^t\mathbf{I}\left\{\mathcal{Q}^{(k)}(u-)=m-1+l\right\}\mathrm{d}D(u),\\
&\quad k=1,2,\ldots,\ell,
\end{aligned}
$$

and taking into account (1.6) we finally obtain (4.5).

References

[1] Abramov, V.M. (2000). A large closed queueing network with autonomous service and bottleneck. *Queueing Systems*, 35: 23-54.

[2] Abramov, V.M. (2002). Asymptotic analysis of the $GI/M/1/n$ loss system as n increases to infinity. *Annals of Operations Research*, 112: 35-41.

[3] Abramov, V.M. (2004). Large closed queueing networks containing two types of node and multiple customer classes: One bottleneck station. *Queueing Systems*, 48: 45-73.

[4] Abramov, V.M. (2004). Asymptotic behaviour of the number of lost messages. *SIAM Journal on Applied Mathematics*, 64: 746-761.

[5] Abramov, V.M. (2005). The stability of join-the-shortest-queue models with general input and output processes. arXiv : math/PR 0505040.

[6] Abramov, V.M. (2007). Asymptotic analysis of loss probabilities in $GI/M/m/n$ queueing systems as n increases to infinity. *Quality Technology and Quantitative Management*, 4: 379-393.

[7] Abramov, V.M. (2008). Large closed queueing networks in semi-Markov environment and their applications. *Acta Applicandae Mathematicae*, 100: 201-226.

[8] Anulova, S.V. and Liptser, R. Sh. (1990). Diffusion approximation for the processes with normal reflection. *Theory of Probability and Its Application*, 35: 413-423.

[9] Berger, A.W. and Whitt, W. (1998). Effective bandwidths with priorities. *IEEE/ACM Transaction on Networking*, 6 (4): 447-460.

[10] Berger, A.W. and Whitt, W. (1998). Extending the effective bandwidth concept to networks with priority classes. *IEEE Communication Magazine*, 36 (8): 78-83.

[11] Bertsimas, D., Paschalidis, I.C. and Tsitsiklis, J.N. (1998). Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Transactions on Automatic Control* 43: 315-335.

[12] Borovkov, A.A. (1976). *Stochastic Processes in Queueing Theory*. Springer, Berlin.

[13] Borovkov, A.A. (1984). *Asymptotic Methods in Queueing Theory*. John Wiley, New York.

[14] Botvich, D.D. and Duffield, N.G. (1995). Large deviations, the shape of loss curve, and economics of scale in large multiplexers. *Queueing Systems*: 20, 293-320.

[15] Chao, X., Pinedo, M. and Shaw, D. (1996). Network of queues with batch services and customer coalescence. *Journal of Applied Probability* 33: 858-869.

[16] Choi, B.D. and Kim, B. (2000). Sharp results on convergence rates for the distribution of the $GI/M/1/K$ queues as K tends to infinity. *Journal of Applied Probability* 37: 1010-1019.

[17] Choi, B.D., Kim, B. and Wee, I.-S. (2000). Asymptotic behavior of loss probability in $GI/M/1/K$ queue as K tends to infinity. *Queueing Systems*, 36: 437-442.

[18] Choudhury, G.L., Lucantoni, D.M. and Whitt, W. (1996). Squeezing the most out of ATM. *IEEE Transactions in Communications* 44: 203-217.

[19] Courcoubetis, C., Siris, V.A. and Stamoulis, G. (1999). Application of many sources asymptotic and effective bandwidth for traffic engineering. *Telecommunication Systems*, 12: 167-191.

[20] Economou, A. and Fakinos, D. (2003). On the stationary distribution of the $GI^X/M^Y/1$ queueing system. *Stochastic Analysis and Applications*, 21: 559-565.

[21] Elwalid, A.I. and Mitra, D. (1995). Analysis, approximations and admission control of a multi-service multiplexing system with priorities. *Proc. IEEE INFOCOM'95*, 463-472.

[22] Elwalid, A.I. and Mitra, D. (1999). Design of generalized processor sharing schedulers with statistically multiplex heterogeneous QoS classes. *Proc. IEEE INFOCOM'99*, 1220-1230.

[23] Evans, J.S. and Everitt, D. (1999). Effective bandwidth-based admission control for multiservice CDMA cellular networks. *IEEE Transactions on Vehicular Technology*, 48: 36-46.

[24] Filippov, A.F. (1988). *Differential Equations with Discontinuous Right-Hand Side*. Kluwer, Dordrecht.

[25] Fricker, C. (1986). Etude d'une file GI/G/1 á service autonome (avec vacances du serveur). *Advances in Applied Probability*, 18: 283-286.

[26] Fricker, C. (1987). Note sur un modele de file GI/G/1 á service autonomé (avec vacances du serveur). *Advances in Applied Probability*, 19: 289-291.

[27] Gelenbe, E. and Iasnogorodski, R. (1980). A queue with server of walking type (autonomous service). *Ann. Inst. H. Poincare*, 16, 63-73.

[28] Gnedenko, B.V. and Kovalenko, I.N. (1968). *Introduction to the Theory of Queues*. Israel Program for Scientific Translations, Jerusalem.

[29] Himmelblau, D.M. (1972). *Applied Non-Linear Programming*. McGraw-Hill, New York.

[30] Kella, O. (1993). Parallel and tandem fluid networks with dependent Lévy inputs. *The Annals of Applied Probability*, 3: 682-695.

[31] Kella, O. and Whitt, W. (1992). A tandem fluid network with Lévy input. In *Queues and Related Models* (I. Basawa and U. Bhat eds) 112-128, Oxford University Press.

[32] Kelly, F.P. (1996). Notes on effective bandwidth. In *Stochastic Networks: Theory and Applications* (F.P.Kelly, S.Zachary and I.B.Ziedins eds). Oxford University Press, Oxford, 1996, 141-168.

[33] Kogan, Ya. and Liptser, R. Sh. (1993). Limit non-stationary behaviour of large closed queueing network with bottlenecks. *Queueing Systems*, 14: 33-55.

[34] Kumaran, K., Margrave, G.E., Mitra, D. and Stanley, K.R. (2000). Novel techniques for the design and control of generalized processor-sharing schedulers for multiple QoS classes. *Proc. INFOCOM'00*, 2: 932-941.

[35] Lee, J.Y., Kim, S., Kim, D. and Sung, D.K.(2005). Bandwidth optimization for internet traffic in generalized processor-sharing servers. *IEEE Transactions on Parallel Distributed Systems*, 16: 324-334.

[36] Miyazawa, M. (1990). Complementary generating functions for the $M^X/GI/1/k$ and $GI/M^Y/1/k$ queues and their application to the comparison for loss probabilities. *Journal of Applied Probability*, 27: 684-692.

[37] Paschalidis, I.C. (1996). Large deviations in high-speed communications networks. PhD thesis. MIT Laboratory for Information and Decision Systems, Cambridge, MA, USA.

[38] Ramanan, K. (2006). Reflected diffusions defined via extended Skorokhod map. *Electronic Journal of Probability*, 11: 934-992.

[39] Rubalskii, G.B. (1982). The search of an extremum of unimodal function of one variable in an unbounded set. *U.S.S.R. Comput. Maths. Math. Phys.*, 22 (1): 8-15. Transl. from Russian: *Zhurnal Vychislitelnoi Matematiki i Matematicheskoi Fiziki*, 22 (1): 10-16, 251.

[40] Skorokhod, A.V. (1961). Stochastic equations for difusion processes in a bounded region. *Theory Probabilities and Its Application* 6: 264-274.

[41] Subhankulov, M.A. (1976). *Tauberian Theorems with Remainder*. Nauka, Moscow. (In Russian.)

[42] Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press, New York/London.

[43] Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes*. John Wiley, New York.

[44] Tanaka, H. (1979). Stochastic differential equations with reflected boundary condition in convex regions. *Hiroshima Mathematical Journal*, 9: 163-177.

[45] Whitt, W. (1993). Tail probabilities with statistical multiplexing and effective bandwidth in multiclass queues. *Telecommunication Systems*, 2: 71-107.

[46] Whitt, W. (2004). Heavy-traffic limits for loss proportions in single-server queues. *Queueing Systems*, 46: 507-536.

[47] Whitt, W. (2005). Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research*, 30: 1-27.

[48] Wischik, D. (1999). The output of switch, or, effective bandwidth for networks. *Queueing Systems*, 32: 383-396.

[49] Wischik, D. (2001). Sample path large deviations for queues with many outputs. *The Annals of Applied Probability*, 11: 379-404.

School of Mathematical Sciences, Monash University, Building 28M, Clayton Campus, Clayton, VIC 3800, Australia

E-mail address: vyacheslav.abramov@sci.monash.edu.au