Annales Mathematicae et Informaticae
***46** (2016) pp. 143–163*
http://ami.ektf.hu

Methods to analysis of queueing models with state-dependent jump priorities

Agassi Melikov[a], Anar Rustamov[b]
Turan Jafarzade[c], János Sztrik[d]

[a]Institute of Control Science
agassi.melikov@rambler.ru
[b]Qafqaz University
anar.rustemov@gmail.com
[c]National Aviation Academy
turan_jafarzade@hotmail.com
[d]University of Debrecen
sztrik.janos@inf.unideb.hu

Submitted April 20, 2016 — Accepted September 7, 2016

Abstract

In this paper, exact and approximate approaches for studying queuing models with state-dependent jump priorities are developed. Both models with finite separate buffers and finite common buffer for heterogeneous calls are investigated. It is shown that both models might be described by two-dimensional Markov Chains (2-D MC). Exact approach based on solution of appropriate system of balance equations (SBE) for state probabilities faced with big computational challenges for large scale models. To overcome the indicated difficulties an approximate approach based on the state space merging algorithms is developed. This approach allows to construct simple algorithms to calculate the Quality of Service (QoS) metrics of the examined models. The results of numerical experiments are demonstrated.

Keywords: queuing models, jump priority, Markov chains, space merging, numerical analysis

MSC: 60K25, 68M20, 90B22

1. Introduction

Priorities are effective tools to solve the problems of quality of service (QoS) provisioning of heteregeneous calls in queuing systems. By nature the priorities can be broadly divided into two classes: *static* and *dynamic*. Static priorities (relative or preemptive) are defined in advance and they do not change during the whole system operation time [1]. In literature relative static priorities in queuing systems with buffers sometimes are called HOL-priorities (Head-Of-Line), .i.e. in static priorities call for service is chosen from the head of line according to the highest priority. Dynamic priorities in turn are divided into two classes: *dynamical vs time* and *dynamical vs state*. In *dynamical versus time* priorities the priority of the calls can be changed according their waiting times (or sojourn time) [2]. In *dynamical versus state* priorities (they sometimes are called *state-dependent priorities*) calls can change priority according the state of the system where the state is described by vector whose components indicate number of heterogeneous calls in the queue (or in the system) [3].

The drawback of static priorities is that when they are used in real systems the delay of low priority calls is too large espesially for the system with heavy loads of high priority calls. Dynamic priorities allow to avoid the starvation of low priority calls. Detailed review of priority schemas might be found in [4].

As a rule, classical priorities (static or dynamic) are used to determine type of call from the buffer which must be send to channel for servicing. However, some scientific and practical interest represents the priorities which are introduced to change (either increase or decrease) the priorities of calls in buffer. These changes are realized instantaneously so such kind of priorities are called jump priorities (JP). They might be either static or dynamic too. Let us breifly review existing results related to such kind of priorities.

The pioneer work on the analyzing dynamical vs time HOL-priorities with priority jumps (HOL-JP) is [5]. In this paper dynamical vs time HOL-JP was proposed where calls with low priority can jump to another buffer with high priority after waiting some (deterministic) period of time in native buffer; this process goes until a call of any type gets access to a channel or reaches a queue with highest priority. Formulas for calculation of the mean waiting time of the heterogeneous calls were developed in [5].

Dynamical vs state HOL-JP in discrete-time queuing models were proposed in [6-10]. In these models authors included two kinds of calls - high priority calls (H-calls) and low priority calls (L-calls). A scheme of head-of-line merge-by-probability (HOL-MBP) according to which at the end of each time slot all L-calls go to the end of the queue of H-calls with the fixed probability $\beta, 0 < \beta < 1$, was proposed in [6]. A modification of the HOL-MBP scheme was considered in [7]. It was named head-of-line jump-or-serve (HOL-JOS) and, in contrast to the scheme of [6]], in it only one L-call goes from the queue head into the H-queue. Unlike the HOL-JOS scheme, in HOL-JIA$_1$ (Head-Of-Line Jump If-Arrival) scheme [8] transition of the L-call into the H-queue depends not only on the state of the H-queue at

the beginning of the slot, but also on the number of arrivals of L-calls during this slot. The only distinction of the HOL-JIA$_1$ scheme from the HOL-JIA$_2$ scheme [9] lies in that in the latter scheme the L-calls can pass immediately to the H-queue. Formulas for the generating functions of the call queue lengths of both types and the time of H-call waiting on the queue, as well as their moments, were developed in [6-10]. Additionally, the mean time of waiting in the queue of L-calls was determined.

In [5-10] queuing models with infinite buffers are investigated. So, they have little applicability in the real communication networks. In particular, real communication networks have finite buffer capacity. Secondly, investigated JP are defined by state-independent probabilities. Since they cannot be adapted for real situations depending on loads of heterogeneous calls.

Different approach to study queuing models with dynamical vs state HOL-JP can be found in the papers [11–14] and in chapter 5 of the book [15] where new type of randomized state-dependent JP for continuous-time queuing systems with finite buffers was proposed. In papers [11, 12] models with separate buffers for heterogeneous calls have been examined while in paper [13, 14] models with common buffer are investigated. They make it possible pass to from the L-queue into the H-queue only at the instants of arrival of the L-calls, the probability of such transitions depend only on the number of L-calls in the system. In chapter 5 of the book [15] models with separate buffers which jump priorities depending only on the number of H-calls in the system were examined. In the indicated works [11–14] methods of calculation of main QoS metrics of the investigated models are proposed. To the best of our knowledge, models in which JP depends on the number of both types of calls in the system are not examined. In this paper we investigate such kinds of models.

At the end of this section it's worth noting that in [16] queueing models with finite common buffer and two priority classes of calls are investigated, where it is assumed that H-calls can preempt the service of L-calls. Futhermore in [16] various congestion control mechanisms are also proposed. In order to calculate the steady-state probabilities of the investigated models, new calculation approach of the original method based on the theory of generalized invariant subspace is developed. Unlike [16] preemption of H-calls from the services of L-calls is not allowed in our paper. However L-call can jump H-buffer in order to served as H-calls.

The rest of the paper is organized as follows. In section 2 model with separate buffers and state-dependent JP is examined and both exact and approximate methods of calculation its QoS metrics are developed. Similar problems for model with common buffer are investigated in section 3. Section 4 is about numerical results of models with both separate buffers and common buffer. In numerical experiments, we investigate different schemas of changing elements of JP-matrix. Conclusion remarks are given in section 5.

2. Jump priorities in model with separate buffers

The structure scheme of the studied queuing system is depicted in Figure 1. In the single server queueing system two Poisson traffic of heterogeneous calls have different arrival rate $\lambda_i, i = 1, 2$. We determined first type of calls as high priority calls (H-calls) while second type of calls are treated as low priority calls (L-calls). By default H-call from the buffer dominating to be served by the idle server; only in the case of absence of H-call in the buffer, L-calls can be served. If there isn't any call in the buffer, then the channels becomes free. Service intensity of the server is the same for both type of the call where it is determined as μ obeying exponential distribution.

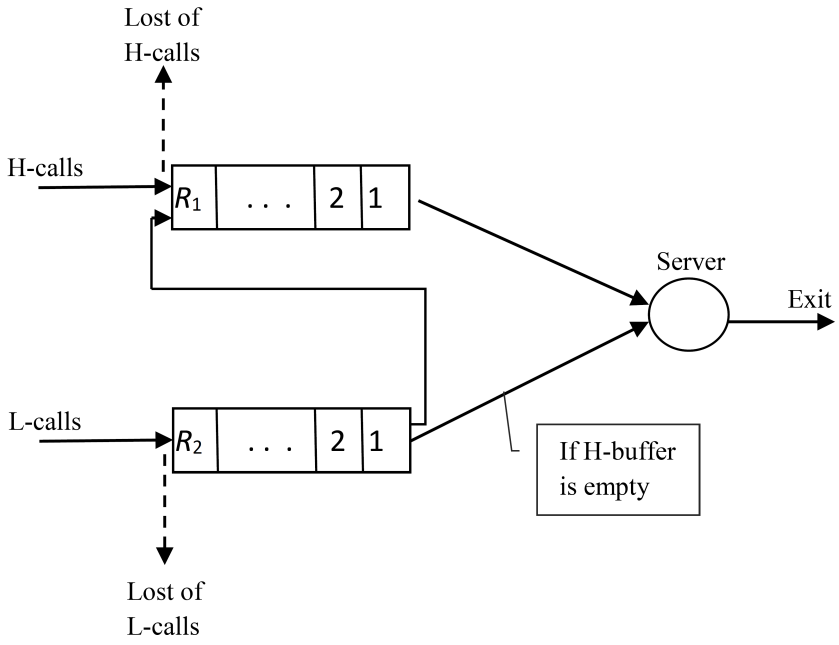

Figure 1: Structure of the queuing system with jump priorities

First let us consider the model with separate buffers, i.e. it is assumed that there are two isolated buffers – H-buffer (for waiting H-calls) and L-buffer (for waiting L-calls) with size of R_1 and$R_2(0 < R_i < \infty, i = 1, 2)$ respectively.

Decision epochs (i.e. jumping moments of L-calls to H-buffer) coincide with the arrival moments of L-calls. In this model state-dependent HOL-JP is defined as follows:

- High priority calls are always accepted to the H-buffer with the probability 1 if there is a free place in this buffer. If the H-buffer is full then arrived H-call will be dropped with probability 1.

- If upon arrival L-call number of calls of this type equals $i, i < R_2$, and number of H-call equals $j, j < R_1$, then L-call joins the H-buffer with probability $\alpha_i(j)$ and in future it will be served as H-call; and arrived L-call joins the L-buffer with probability $1 - \alpha_i(j)$.

- If upon arrival L-call number of H-call equals R_1, then L-call joins the L-buffer if there is free place in this buffer; otherwise, arrived L-call will be dropped with probability 1.

- If upon arrival L-call L-buffer is full and number of H-call equals $j, j < R_1$, then L-call joins the H-buffer with probability $\alpha_{R_2}(j)$; and arrived L-call will be dropped with probability $1 - \alpha_{R_2}(j)$.

In other words, to define JP matrix with dimension $(R_2 + 1)$ x R_1 is introduced. Entities of this matrix (JP-matrix) are $\alpha_i(j), i = 0, 1, \ldots, R_2, j = 0, 1, \ldots, R_1 - 1$. Thus in this scheme entities of JP-matrix depends on both number of heterogeneous calls in the appropriate buffers.

Let us note some important special schemes regarding the jump priorities mentioned above.

1) *The uniform schemas.* In this schemas, the elements of JP-matrix does not depend on the number of heteregeneous calls in the buffers. So, if the elements of JP-matrix does not depend on the number of H-calls in the buffer, i.e., $\alpha_i(j) = \alpha_i$ for any $j = 0, 1, \ldots, R_1 - 1$, then we obtain JP-schema which was proposed in [11, 12]. Here in the special case $\alpha_i = 0$, we obtain the classical HOL-priorities. Alternative case is that the elements of JP-matrix do not depend on the number of L-calls in the buffer, i.e., $\alpha_i(j) = \alpha(j)$ for any $i = 0, 1, \ldots, R_2$. Such kind of JP has been investigated in [15]. In last scheme in the special case $\alpha(j) = 1$ for any $i = 0, 1, \ldots, R_2$, we obtain the classical queuing system with single traffic.

2) *The threshold-based schemas.* In these schemas, the threshold parameters $T_i, 0 \le T_i \le R_1 - 1, i = 0, 1, \ldots, R_2$, are introduced, and elements of JP-matrix are defined as follows:

$$\alpha_i(j) = \begin{cases} \alpha_{i1}, & \text{if } 0 \le j \le T_i, \\ \alpha_{i2}, & \text{if } j > T_i. \end{cases}$$

The probabilities $\alpha_{ij},\ i = 0, 1, \ldots, R_2,\ j = 1, 2$, can be defined in various ways. In special cases, i.e. when these probabilities equal either 0 or 1 we obtain different non-randomized threshold-based JP-schemas.

The problem is finding the QoS metrics for this model. The main QoS metrics are the following: the stationary probability of lossing the calls of the ith type (CLP_i), the mean number of the ith type calls in the buffers (L_i) and the mean call transmission delay of the ith type calls (CTD_i), i=1, 2.

2.1. Exact method for model with separate buffers

The state of the system is defined by two dimensional vectors $\boldsymbol{n} = (n_1, n_2)$ where the first component indicates the number of H-calls and the second one the number

of L-calls respectively. In other words, operation of this system is described by the two-dimensional Markov Chain (2-D MC) with the following state space:

$$S = \{\boldsymbol{n} : \ n_i = 0, 1, \ldots, R_i, \ i = 1, 2\}. \tag{2.1}$$

Transition intensity from state $\boldsymbol{n} \in S$ to state $\boldsymbol{n'} \in S$ are denoted by $q(\boldsymbol{n}, \boldsymbol{n'})$. Then nonnegative elements of the generating matrix (Q-matrix) of the given 2-D MC can be calculated as below (see Figure 2):

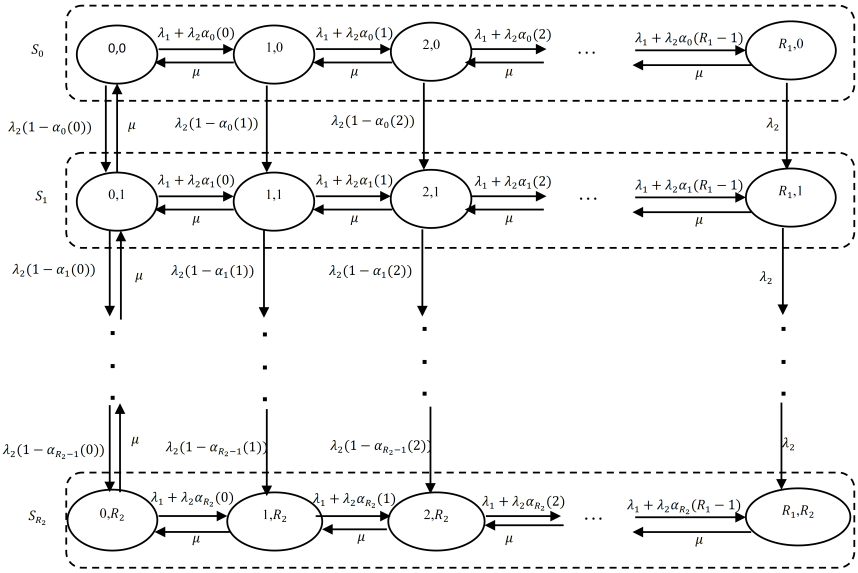

Figure 2: State diagram of the model with separate buffers

$$q(\boldsymbol{n}, \boldsymbol{n'}) = \begin{cases} \lambda_1 + \lambda_2\alpha_{n_2}(n_1), & \text{if } n_1 < R_1, \ \boldsymbol{n'} = \boldsymbol{n} + \boldsymbol{e}_1, \\ \lambda_2(1 - \alpha_{n_2}(n_1)), & \text{if } n_1 < R_1, \ n_2 < R_2, \boldsymbol{n'} = \boldsymbol{n} + e_2, \\ \lambda_2, & \text{if } n_1 = R_1, \ \boldsymbol{n'} = \boldsymbol{n} + e_2, \\ \mu, & \text{if } n_1 > 0, \ \boldsymbol{n'} = \boldsymbol{n} - \boldsymbol{e}_1 \text{ or} \\ & \quad n_1 = 0, \ \boldsymbol{n'} = \boldsymbol{n} - \boldsymbol{e}_2, \\ 0 & \text{in other cases,} \end{cases} \tag{2.2}$$

where $\boldsymbol{e}_1 = (1, 0), \ \boldsymbol{e_2} = (0, 1)$.

For any positive values of the parameters of incoming traffics, all states of the investigated finite MC are communicating (see Figure 2), and consequently, stationary distribution of the investigated 2-D MC exists.

The stationary probability of state $\boldsymbol{n} \in S$ is denoted by $p(\boldsymbol{n})$. Construction and solution of the corresponding system of balance equations (SBE) for the given

2-D MC is the standard way for determining the stationary state probabilities. It is constructed with regard to (2.2) and has the following form:

Case $n_1 < R_1,\ n_2 < R_2$:

$$\begin{aligned} &p\left(n_1, n_2\right)\left(\lambda_1 + \lambda_2 + \mu I\left(n_1 + n_2 > 0\right)\right) = \\ &p\left(n_1 - 1, n_2\right) I\left(n_1 > 0\right)\left(\lambda_1 + \lambda_2 \alpha_{n_2}\left(n_1 - 1\right)\right) + \\ &p\left(n_1, n_2 - 1\right) I\left(n_2 > 0\right) \lambda_2 \left(1 - \alpha_{n_2 - 1}\left(n_1\right)\right) + \\ &p\left(n_1 + 1, n_2\right) \mu + p\left(n_1, n_2 + 1\right) \mu I\left(n_1 = 0\right) \end{aligned} \tag{2.3}$$

Case $n_1 = R_1,\ n_2 < R_2$:

$$\begin{aligned} &p\left(R_1, n_2\right)\left(\mu + \lambda_2\right) = \\ &p\left(R_1 - 1, n_2\right)\left(\lambda_1 + \lambda_2 \alpha_{n_2}\left(R_1 - 1\right)\right) + p\left(R_1, n_2 - 1\right) I\left(n_2 > 0\right) \lambda_2; \end{aligned} \tag{2.4}$$

Case $n_2 = R_2$:

$$\begin{aligned} &p\left(n_1, R_2\right)\left(\mu + \lambda_1 + \lambda_2 \alpha_{R_2}\left(n_1\right) I\left(n_1 < R_1\right)\right) = \\ &p\left(n_1 - 1, R_2\right) \mu I\left(n_1 > 0\right) + p\left(n_1, R_2 - 1\right)_2 \left(1 - \alpha_{R_2 - 1}\left(n_1\right)\right)\ . \end{aligned} \tag{2.5}$$

The SBE (2.3)–(2.5) should be completed with normalizing condition over state space (2.1):

$$\sum_{\boldsymbol{n} \in S} p\left(n\right) = 1. \tag{2.6}$$

After determining the state probabilities from SBE (2.3)–(2.6), one can establish its QoS metrics. As indicated above, H-calls are lost if upon their arrivals H-buffer is full. Hence, the loss probability for H-calls (CLP_1) can be determined as follows:

$$CLP_1 = \sum_{i=0}^{R_2} p\left(R_1, i\right). \tag{2.7}$$

Similarly, we conclude that L-calls are lost in the following cases: (2.1) at the time an L-call arrives, both buffers are full (in such case L-call is lost with probability 1); (2.2) at the time an L-call arrives, L-buffer is full but there is free place in H-buffer (in such case L-call is lost with probability$1 - \alpha_{R_2}(i)$). Thus the loss probability of L-calls (CLP_2) is given by

$$CLP_2 = p\left(R_1, R_2\right) + \sum_{i=0}^{R_1 - 1} p\left(i, R_2\right)\left(1 - \alpha_{R_2}\left(i\right)\right). \tag{2.8}$$

The first and second terms of the sum in the formula (2.8) denote the probability of events (2.1) and (2.2), respectively.

The mean numbers of the H-calls (L_1) and L-calls (L_2) in the queue are determined as the expected values of appropriate discrete random variables:

$$L_1 = \sum_{i=1}^{R_1} i \sum_{j=0}^{R_2} p\,(i,j); \tag{2.9}$$

$$L_2 = \sum_{i=1}^{R_2} i \sum_{j=0}^{R_1} p\,(j,i). \tag{2.10}$$

Further, formulas (2.3)–(2.6) and modified Little's formula can be used to evaluate the mean times of transmission delay for the heterogeneous:

$$CTD_1 = \frac{L_1}{\lambda_1^{(c)}}\ ; \tag{2.11}$$

$$CTD_2 = \frac{L_2}{\lambda_1^{(c)} + \lambda_2^{(c)}}, \tag{2.12}$$

where $\lambda_1^{(c)}$ and $\lambda_2^{(c)}$ are carried loads of H-calls and L-calls, respectively. These parameters are calculated as follows:

$$\lambda_1^{(c)} = \lambda_1 \left(1 - \sum_{j=0}^{R_2} p\,(R_1,j)\right) + \lambda_2 \sum_{i=0}^{R_1-1} \sum_{j=0}^{R_2} p\,(i,j)\,\alpha_j\,(i);$$

$$\lambda_2^{(c)} = \lambda_2 \sum_{i=0}^{R_1} \sum_{j=0}^{R_2-1} p\,(i,j)\,(1-\alpha_j\,(i)).$$

Thus, to find QoS metrics (2.7)–(2.12), it is necessary to determine the steady-state probabilities of the model from the corresponding SBE (2.3)–(2.6). By implementation of programming languages it is possible to solve the SBE (2.3)–(2.6) for the steady-state probabilities $p\,(\boldsymbol{n})$, $\boldsymbol{n} \in S$ with a help of numerical methods of the linear algebra. This method of calculation of QoS metrics is called the exact (precise) method. In cases of moderate capacity of state space (2.1) this methods is reasonable to calculate QoS metrics of the system. But for large scale system (i.e. when system has large buffers) it isn't suitable. Therefore, we need to find out a more efficient method to calculate the QoS metrics of the models with large buffers.

2.2. Approximate method for model with separate buffers

Below we consider asymptotic analysis of the QoS metrics for large scale models, i.e. when R_1 and R_2 take large values. The developed approximate method has high accuracy for heavy traffic regime of H-calls. In other words, below we consider asymptotic analysis of the large scale model with heavy loads of H-calls, i.e. it is assumed that $\nu_1 \gg \nu_2$,where $\nu_i = \lambda_i/\mu, i = 1,2$. Note that this assumption is not something extraordinary because introduction of the jump priorities for the L-calls makes sense, namely in the systems with heavy loads of H-calls.

Consider the following splitting of the state space (2.1):

$$S = \bigcup_{i=0}^{R_2} S_i, S_i \bigcap S_j = \emptyset, i \neq j, \tag{2.13}$$

where $S_i = \{\boldsymbol{n} \in S : n_2 = i\}$, $i = 0, 1, 2, \ldots, R_2$.

We notice that the assumption made about the relation of the loads of the heterogeneous calls enables one to satisfy the condition for correct use of the algorithms of state space merging of the 2-D MC (see [3, Appendix]): transition intensities within classes S_i, $i = 0, 1, \ldots, R_2$, are essentially higher than those between states of different classes.

The classes of microstates S_i are united into individual merged states $< i >$, and in the original state space S the following merge function is defined:

$$U(\boldsymbol{n}) =< i >, \text{ if } \boldsymbol{n} \in S_i \ . \tag{2.14}$$

The function (2.14) defines a merged model with the state space

$$\Omega = \{< i >: i = 0, 1, 2, \ldots, R_2\}.$$

Let us consider the problem of calculation of state probabilities inside the splitting models. The stationary probability of the state (k, i) in the split model with the state space S_i is denoted by $\rho_i(k)$, $i = 0, 1, 2, \ldots, R_2$, $k = 0, 1, 2, \ldots, R_1$.

Each split model with state space S_i is a one-dimensional birth and death process with the parameters that are calculated as follows (see Figure 2):

$$q_i(k_1, k_2) = \begin{cases} \lambda_1 + \lambda_2 \alpha_i(k_1), & \text{if } k_2 = k_1 + 1 \\ \mu, & \text{if } k_2 = k_1 - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.15}$$

Consequently, we have

$$\rho_i(k) = \prod_{j=0}^{k-1} (\nu_1 + \nu_2 \alpha_i(j)) \rho_i(0), \ k = 1, \ldots, R_1, \tag{2.16}$$

where $\rho_i(0) = \left(1 + \sum_{k=1}^{R_1} \prod_{j=0}^{k-1} (\nu_1 + \nu_2 \alpha_i(j))\right)^{-1}$.

The elements of the Q-matrix of the merged model are denoted by

$$q(< k >, < k' >), < k >, < k' >\in \Omega.$$

According to the algorithm of state space merging of the 2-D MC (see [3, Appendix]) these elements are given by

$$q(< k >, < k' >) = \sum_{\substack{\boldsymbol{n} \in \boldsymbol{S_k} \\ \boldsymbol{n'} \in \boldsymbol{S_{k'}}}} q(\boldsymbol{n}, \boldsymbol{n'}) \rho_{n_1}(n_2). \tag{2.17}$$

So, by using (2.2), (2.16) and (2.17) after some mathematical transformations the following formula are obtained (see Figure 2):

$$q\left(< k >, < k' >\right) = \begin{cases} \lambda_2 \left(\rho_k\left(R_1\right) + \sum_{i=0}^{R_1-1}\left(1-\alpha_k\left(i\right)\right)\rho_k\left(i\right)\right), & \text{if } k' = k+1, \\ \mu\rho_k\left(0\right), & \text{if } k' = k-1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

From (2.18) we can calculate the probabilities of the merged states

$$\pi\left(< k >\right), < k > \in \Omega$$

as follows:

$$\pi\left(< k >\right) = {\nu_2}^k \prod_{j=0}^{k-1} \Lambda_j \pi\left(< 0 >\right), \; k = 1, \ldots, R_2, \tag{2.19}$$

where

$$\pi\left(< 0 >\right) = \left(1 + \sum_{k=1}^{R_2} {\nu_2}^k \prod_{j=0}^{k-1} \Lambda_j\right)^{-1},$$

$$\Lambda_j = \frac{\rho_j\left(R_1\right) + \sum_{i=0}^{R_1-1}\left(1-\alpha_j\left(i\right)\right)\rho_j\left(i\right)}{\rho_{j+1}\left(0\right)}, \quad j = 0, \ldots, R_2 - 1.$$

The state probabilities of the initial 2-D MC are determined approximately as follows (see [3, Appendix]):

$$p\left(i,j\right) \approx \rho_j\left(i\right)\pi\left(< j >\right). \tag{2.20}$$

By taking into account (2.16), (2.19) and (2.20) we can calculate approximate values of state probabilities of initial 2-D MC, and omitting the intermediate mathematical calculations the following approximate formulae to calculate the QoS metrics (2.7)-(2.10) are obtained:

$$CLP_1 \approx \sum_{i=0}^{R_2} \rho_i\left(R_1\right)\pi\left(< i >\right); \tag{2.21}$$

$$CLP_2 \approx \pi\left(< R_2 >\right)\left(\rho_{R_2}\left(R_1\right) + \sum_{i=0}^{R_1-1} \rho_{R_2}\left(i\right)\left(1-\alpha_{R_2}\left(i\right)\right)\right); \tag{2.22}$$

$$L_1 \approx \sum_{i=1}^{R_1} i \sum_{k=0}^{R_2} \rho_k\left(i\right)\pi\left(< k >\right); \tag{2.23}$$

$$L_2 \approx \sum_{k=1}^{R_2} k\ \pi\left(< k >\right). \tag{2.24}$$

The QoS metrics CTD_k are determined from (2.11) and (2.12) after the calculation of the parameters L_k and $\lambda_k^{(c)},\ k = 1, 2$.

3. Jump priorities in models with common buffer

Now let us consider the model with common buffer. In this case it is assumed that there is single buffer with size $R,\ 0 < R < \infty$, for both types of calls. In this model the state-dependent HOL-JP is defined as follows.

- High priority calls are always accepted to the common buffer with the probability 1 if there is a free place in the buffer. If the common buffer is full then arriving H-call will be dropped with probability 1.
- If upon arrival of L-call the number of H-calls equals n_1 and number of L-calls equals n_2, where $n_1 + n_2 < R$, then arriving L-call becomes H-call with probability $\beta_{n_2}(n_1)$ or it will be accepted to the buffer as L-call with probability $1 - \beta_{n_2}(n_1)$.
- If upon arrival of L-call common buffer is full (i.e. in case $n_1 + n_2 = R$) it will be dropped with probability 1.

The main QoS metrics of the model are the same with the model with separate buffers.

3.1. Exact method for model with common buffer

As it is mentioned in section 2, the 2-D vector $\boldsymbol{n} = (n_1, n_2)$ is used to describe the state of the system and state space for this model is determined as follows (see Figure 3):

$$S = \{\boldsymbol{n} :\ n_i = 0, 1, \ldots, R,\ i = 1, 2;\ n_1 + n_2 = R\}\ . \tag{3.1}$$

Note 1. Hereinafter, for simplicity, we use the same notations for the state spaces, state probabilities etc. in different models. This will not lead to misunderstanding since from the context it will be clear which model is being considered.

In this case the nonnegative elements of the Q-matrix of the given 2-D MC can be calculated as follows (see Figure 3):

$$q\left(\boldsymbol{n}, \boldsymbol{n'}\right) = \begin{cases} \lambda_1 + \lambda_2 \alpha_{n_2}(n_1), & \text{if } n_1 + n_2 < R,\ \boldsymbol{n'} = \boldsymbol{n} + \boldsymbol{e}_1, \\ \lambda_2 \left(1 - \alpha_{n_2}(n_1)\right), & \text{if } n_1 + n_2 < R,\ \boldsymbol{n'} = \boldsymbol{n} + e_2, \\ \mu, & \text{if } n_1 > 0,\ \boldsymbol{n'} = \boldsymbol{n} - \boldsymbol{e}_1 \text{ or} \\ & \quad n_1 = 0,\ \boldsymbol{n'} = \boldsymbol{n} - \boldsymbol{e}_2, \\ 0 & \text{in other cases.} \end{cases} \tag{3.2}$$

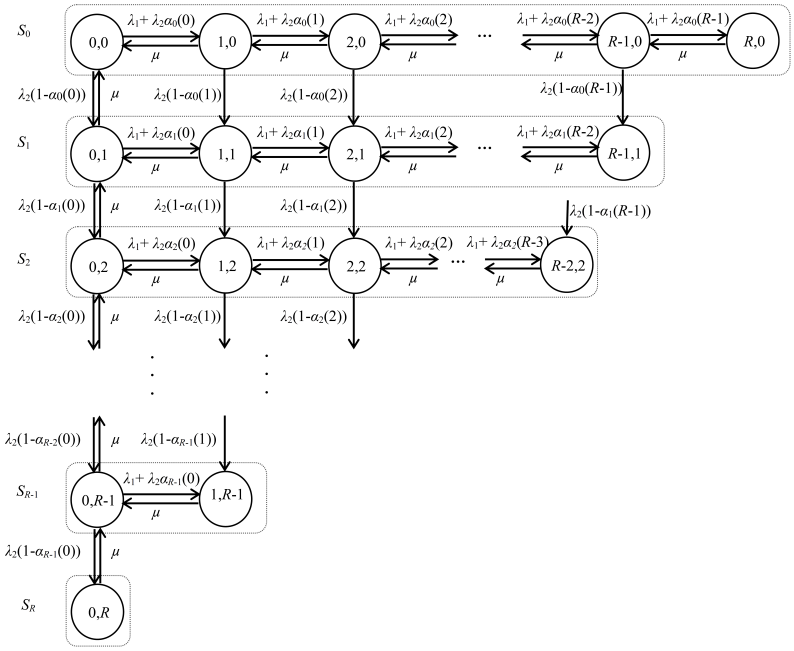

Figure 3: State diagram of the model with common buffer

Loss probabilities of heterogeneous calls in this model equal each other and are given by

$$CLP_1 = CLP_2 = \sum_{\boldsymbol{n} \in S_d} p\left(\boldsymbol{n}\right), \tag{3.3}$$

where $S_d = \{\boldsymbol{n} \in S : \ n_1 + n_2 = R\}$ is set of diagonal states (see Figure 3).

In this model the mean number of the H-calls and L-calls in the buffer are determined as follows:

$$L_k = \sum_{i=1}^{R} i \xi_k \left(i\right) \tag{3.4}$$

where $\xi_k = \sum_{\boldsymbol{n} \in S} p\left(\boldsymbol{n}\right) \delta\left(n_k = i\right)$, $k = 1, 2$; $\delta\left(i, j\right)$ are Kronecker's symbols.

The mean transmission delays of the heterogeneous calls are determined by Eqs. (2.11) and (2.12).

Computational difficulties of the exact approach to solve appropriate SBE (it is constructed with regard to (3.2) and here it is omitted) are observed in the models with large scale too. In order to overcome the above-mentioned computational difficulties to calculate the QoS metrics (3.3), (3.4) an approximate approach to asymptotic analysis can be applied as well.

3.2. Approximate method for model with common buffer

Similar to (2.13) splitting of the state space (3.1) is considered:

$$S = \bigcup_{i=0}^{R} S_i,\ S_i \bigcap S_j = \emptyset,\ i \neq j, \tag{3.5}$$

where $S_i = \{\boldsymbol{n} \in S : n_2 = i\}$, $i = 0, 1, \ldots, R$.

Not repeating the stages of the approximate approach, final formula to calculate QoS metrics are given below. In this case stationary distribution of the split model with state space $S_i,\ i = 0, 1, \ldots, R-1,$ is defined as follows:

$$\rho_i(k) = \prod_{j=0}^{k-1} (\nu_1 + \nu_2 \beta_i(j)) \rho_i(0), \quad k = 1, \ldots, R-i, \tag{3.6}$$

where $\rho_i(0) = \left(1 + \sum_{k=1}^{R-i} \prod_{j=0}^{k-1} (\nu_1 + \nu_2 \beta_i(j))\right)^{-1}$.

Note 2. Split model with state space S_R contains only one micro-state $(0, R) \in S$ so we set $\rho_R(0) = 1$.

According to the algorithm of state space merging of the 2-D MC, the elements of the Q-matrix of the merged model in this case (see formulae (2.17), (3.2) and (3.6)) are given by

$$q(< k >, < k' >) = \begin{cases} \lambda_2 \sum_{i=0}^{R-k-1} \rho_k(i)(1 - \beta_k(i)), & \text{if } k' = k+1, \\ \mu \rho_k(0), & \text{if } k' = k-1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

So, the probabilities of the merged states in this model are calculated as follows:

$$\pi(< k >) = {\nu_2}^k \prod_{j=0}^{k-1} M_j \pi(< 0 >), \ k = 1, \ldots, R, \tag{3.8}$$

where $\pi(< 0 >) = \left(1 + \sum_{k=1}^{R} {\nu_2}^k \prod_{j=0}^{k-1} M_j\right)^{-1}$, $M_j = \frac{\sum_{i=0}^{R-j-1} \rho_j(i)(1-\beta_j(i))}{\rho_{j+1}(0)}$, $j = 0, \ldots, R-1$.

So, by using (3.6) and (3.8) from (2.20) we can calculate approximate values of steady-state probabilities of initial 2-D MC for the model with common buffer. Consequently, for asymptotic analysis of QoS metrics of the investigated model the following approximate formula are obtained:

$$CLP_1 \approx CLP_2 \approx \sum_{i=0}^{R} \rho_i(R-i) \pi(< i >); \tag{3.9}$$

$$L_1 \approx \sum_{k=1}^{R} k \sum_{i=0}^{R-k} \rho_i(k) \pi(< i >); \tag{3.10}$$

$$L_2 \approx \sum_{k=1}^{R} k \; \pi\left(< k >\right). \tag{3.11}$$

Based on (3.10) and (3.11) approximate values of transmission delays of the heterogeneous calls are calculated by (2.11) and (2.12).

4. Numerical results

The developed approximate formula allow one to carry out an authentic analysis of QoS metrics over any range of change of values of loading parameters of the heterogeneous traffic, satisfying assumption concerning their ration (i.e. when$\nu_1 \gg \nu_2$) and also at any buffers sizes.

Let us first examine the results of the numerical experiments for the model with separate buffers. The following initial data for hypothetical model was selected: $R_1 + R_2 = 110,\ \lambda_1 = 2,\ \lambda_2 = 0.5,\ \mu = 3,$ i.e. $\nu_1 = 2/3,\ \ \nu_2 = 1/6$. The numerical results are analyzed based on the two schemas for changing the elements of JP-matrix. In schema 1 it is assumed that they are changed with respect to both parameters (state-dependent JP) and defined as $\alpha_i\left(j\right) = \frac{i+1}{i+j+2}$ while in second one we assume that they are constant (state-independent JP), i.e. $\alpha_i\left(j\right) = 0.5$ for any i and j. In other words, in schema 1 probabilities of jumping to H-buffer are decreasing function with respect to number of H-calls in buffer at fixed values of L-calls but in schema 2 they do not depend on number of heterogeneous calls in buffers.

Figures 4–6 show dependences of QoS metrics on R_1. As it was expected the loss probability of H-calls is positively related to the buffer size of H-buffer (Figures 4) while loss probability of L-calls is increasing function versus R_1. As we see from Figures 4, rate of change of the indicated functions are high enough. Also from this figure we conclude that schema 1 is favorable for the loss probabilities of H-calls while for loss probabilities of L-calls schema 2 is favorable. Moreover, differences between values of loss probabilities of H-calls are essential in different schemas especially at large buffer sizes but values of loss probabilities of L-calls are very close to each other in different schemas.

Let us note that from this graph for both schemas we may find such values of buffer sizes for which difference between loss probabilities of heterogeneous calls is less than given ϵ>0 (such kind of problems are called ϵ-fair servicing policy).

Dependency of length of heterogeneous calls on the H-buffer size is shown in Figures 5. In both schemas the mean queue length of the H-calls positively related to the H-buffer size but length of the L-calls is negatively related to the H-buffer size. From this figure we conclude that schema 1 again is favorable for length of the H-calls while length of the L-calls is invariant to different schemas. Behavior of the indicated QoS metrics is interesting. So, length of the H-calls in both schemas increases with low rates for small values of R_1, and for about $R_1 > 20$ they are almost constant; alternative situation occurs for length of the L-calls in both

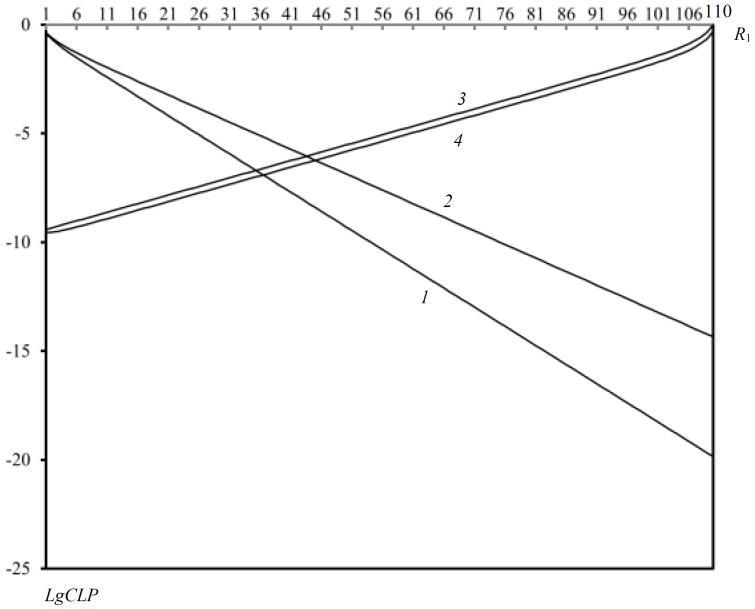

Figure 4: Dependence of loss probabilities versus R_1 in model with separate buffers: 1- CLP_1 in *schema 1*; 2- CLP_1 in *schema 2*; 3- CLP_2 in *schema 1*; 4- CLP_2 in *schema 2*

schemas, i.e. they are almost constant for $R_1 < 100$, and after that they decrease with low rates.

Behavior of both functions CTD_1 and CTD_2 are very similar to behavior of functions L_1 and L_2 respectively (see Figures 6). In other words, schema 1 again is favorable for CTD_1 while CTD_2 is almost constant in different schemas; in both schemes CTD_1 increases with low rates for small values of R_1, and for about $R_1 > 20$ they are almost constant and CTD_2 in both schemas is almost constant for $R_1 < 100$, and for $R_1 > 100$ it decreases with low rates.

Let us now consider the results for model with common buffer based on above indicated different schemas of changing of parameters $\alpha_i(j)$, $i = 0, 1, 2, \ldots, R-1$, $j = 0, 1, \ldots, R-i-1$. Loads of this model are unchanged, i.e. we select$\nu_1 = 2/3$, $\nu_2 = 1/6$.

Dependency of function CLP (as it was mentioned above loss probabilities of heterogeneous calls in this model equal each other, see (3.2)) on the buffer size is shown in Figures 7. It is seen from this figure that in both schemas the loss probability of calls strictly decreases (with high rate) versus the common buffer size and as it was expected schema 1 is favorable for the loss probabilities. Note that differences between values of loss probabilities in different schemas are increased versus buffer size.

In Figures 8 the dependency of functions L_1 and L_2 on the buffer size is shown.

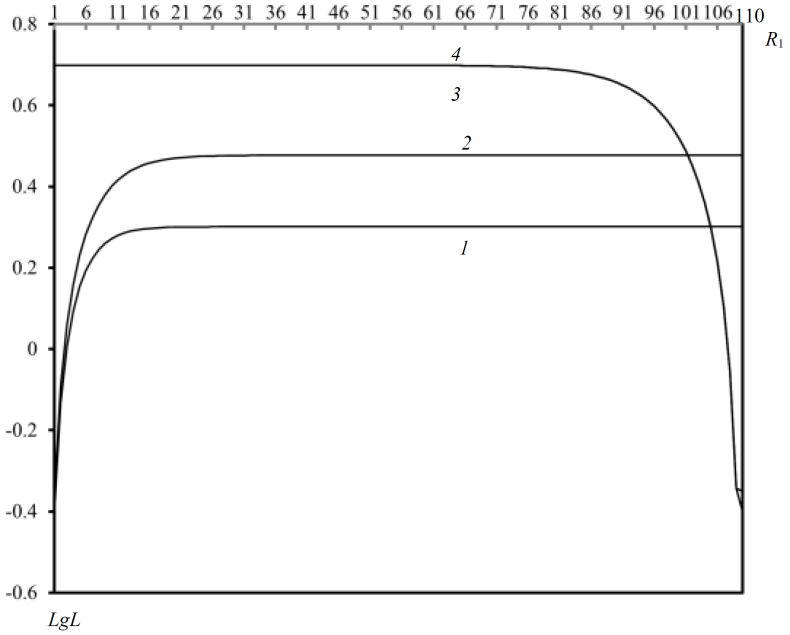

Figure 5: Dependence of mean length of queues versus R_1 in model with separate buffers: 1- L_1 in *schema 1*; 2- L_1 in *schema 2*; 3- L_2 in *schema 1*; 4- L_2 in *schema 2*

In both schemas these functions increase versus the common buffer size. From this figure we conclude that schema 1 is favorable for length of the H-calls while schema 2 is favorable for length of the L-calls. As in case of the model with separate buffers, in this model the rate of change (increasing) of these functions are very small too, i.e. about $R > 15$ they are almost constant.

Again behavior of both functions CTD_1 and CTD_2 is very similar to behavior of functions L_1 and L_2 respectively (see Figures 9). It is interesting that in this case about $R > 15$ values of the function CTD_1 in schema 1 are almost same with values of the function CTD_2 in schema 2.

Presented numerical results allow to take some comparisons proposed in two buffer management mechanisms. So, for instance, values of both functions CLP_1 and CLP_2 in model with separate buffers equal (approximately) $10^{-6.5}$ and this value corresponds to buffers size $R_1 = 35$, $R_2 = 75$ ($R_1 + R_2 = 110$). However, the indicated value for both kinds of calls might be provided in model with common buffer at size $R = 36$. In other words, common buffer is essentially effective buffer management mechanisms for call loss probabilities. Other interesting conclusions with respect to the rest QoS metrics in different buffer management mechanisms might be carried out.

Another goal of performing numerical experiments was the estimation of the

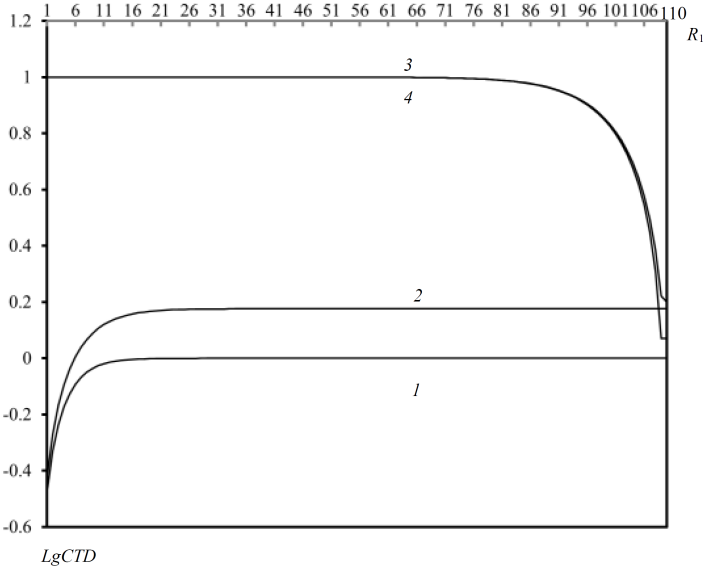

Figure 6: Dependence mean transmission delays versus R_1 in model with separate buffers: 1- CTD_1 in *schema 1*; 2- CTD_1 in *schema 2*; 3- CTD_2 in *schema 1*; 4- CTD_2 in *schema 2*

proposed approximate formulae accuracy. As it was indicated in section 2, the exact values (EV) of the QoS metrics are determined by the appropriate SBE (such an approach allows studying QoS metrics of the model only for small buffer stores).

In order to be short, here in Table 1 and the results only for the model with separate buffers are demonstrated only for the schema 1 (similar results are obtained for the model with common buffer in both schemas as well). Here initial data was selected as above, i.e. $R_1 + R_2 = 110,\ \lambda_1 = 2,\ \lambda_2 = 0.5,\ \mu = 3$.

As it is given in the tables accuracy of the proposed approximate formulae are acceptable for engineering practice. The bigger the ratio v_1/v_2, the higher accuracy of approximate value (AV).

5. Conclusion

This paper proposed a new class of state-dependent JP in queueing systems with finite separate buffers and finite common buffer for heterogeneous calls. An exact and effective approximate approaches for calculating the QoS metrics of heterogeneous calls in such systems are developed. The important advantage of approximate approach lies in the use of explicit formulae to calculate the QoS metrics, which enables our approach to be used for models of any dimension. In addition, it is possible to use the proposed formulae to find the optimal (in given sense) values of

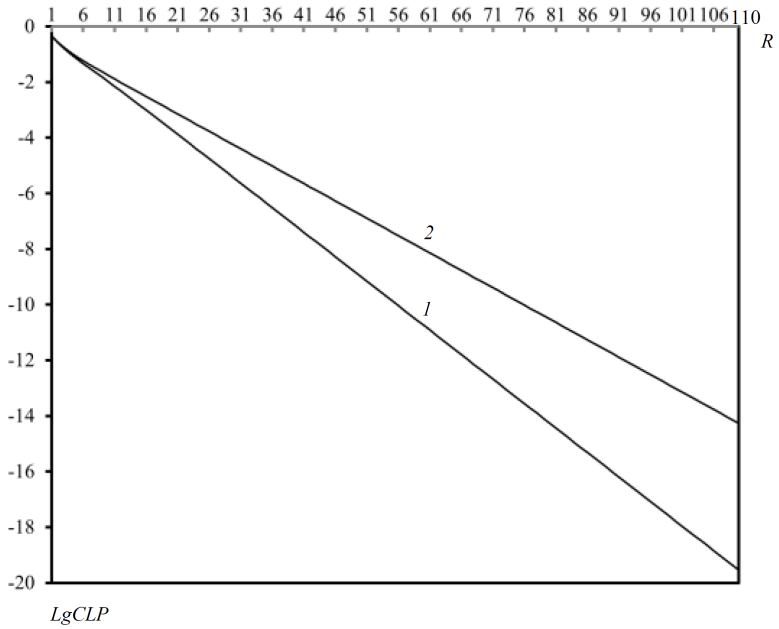

Figure 7: Dependence of loss probabilities versus R in model with common buffer: 1-*schema 1*, 2-*schema 2*

JP-matrix. Latest problems are important especially for the threshold-based non-randomized JP-schemas (see end of section 2) and they are a subject for further study.

Acknowledgements. The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

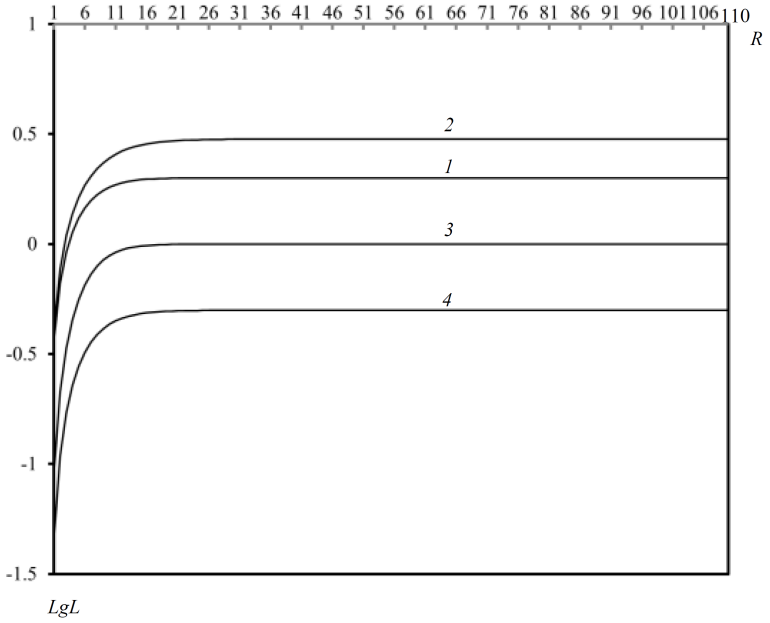

Figure 8: Dependence of mean length of queues versus R in model with common buffer: 1- L_1 in *schema 1*; 2- L_1 in *schema 2*; 3- L_2 in *schema 1*; 4- L_2 in *schema 2*

$\boldsymbol{R}_1$	$\boldsymbol{CLP}_1$		$\boldsymbol{L}_1$	
	EV	***AV***	***EV***	***AV***
15	6.76E-03	7.62E-03	1.30929	1.97560
22	1.52E-05	4.46E-05	1.64089	1.99795
29	3.63E-05	2.61E-06	1.75494	1.99984
36	8.87E-06	1.53E-07	1.79172	1.99998
43	2.21E-08	8.93E-09	1.80311	1.99999
50	5.53E-09	5.23E-10	1.80654	2
57	1.43E-10	3.06E-11	1.80756	2
64	3.58E-11	1.79E-12	1.80787	2
71	9.22E-12	1.05E-13	1.80797	2
78	2.39E-13	6.13E-15	1.80808	2
85	6.21E-14	3.59E-16	1.80844	2
92	1.63E-16	2.12E-17	1.80998	2
99	4.46E-17	1.23E-18	1.81776	2
106	1.67E-18	7.24E-20	1.87771	2

Table 1: Comparison for H-calls in model with separate buffers in schema 1

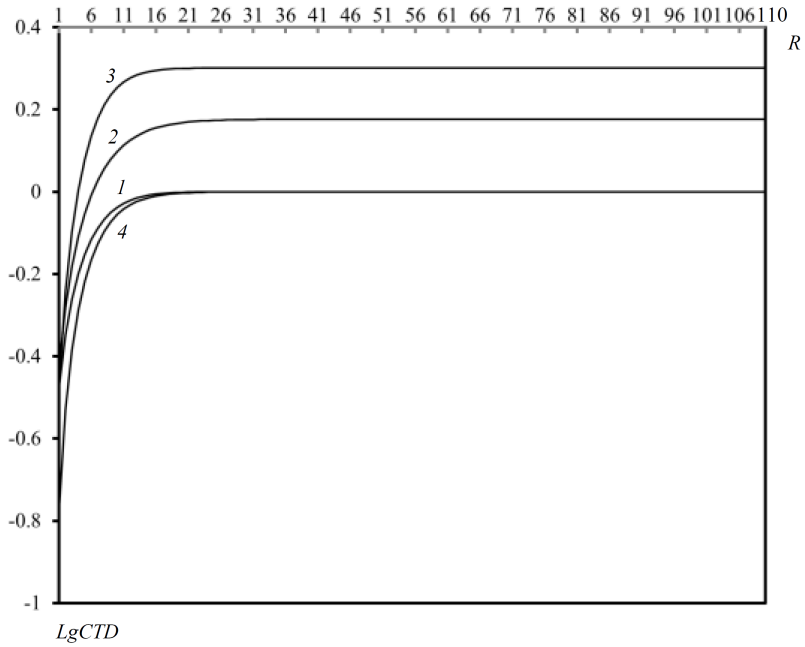

Figure 9: Dependence mean transmission delays versus R in model with common buffer: 1- CTD_1 in *schema 1*; 2- CTD_1 in *schema 2*; 3- CTD_2 in *schema 1*; 4- CTD_2 in *schema 2*

R_1	CLP_2		L_2	
	EV	***AV***	***EV***	***AV***
15	3.21E-10	5.01E-09	4.193002	4.999998
22	6.25E-09	1.79E-08	4.201397	4.999992
29	4.13E-08	6.43E-08	4.196922	4.999974
36	1.93E-07	2.33E-07	4.193955	4.999914
43	7.84E-07	8.25E-07	4.192694	4.999719
50	9.02E-07	2.96E-06	4.192222	4.999098
57	1.15E-06	1.06E-05	4.192017	4.997138
64	4.36E-06	3.89E-05	4.191798	4.991074
71	1.68E-05	1.36E-04	4.191215	4.972766
78	6.61E-05	4.89E-04	4.189308	4.919352
85	2.67E-04	1.76E-03	4.182937	4.770876
92	1.13E-03	6.462E-03	4.161524	4.386067
99	5.26E-03	2.53E-02	3.08775	3.484102
106	1.20E-02	1.34E-01	1.795528	1.640507

Table 2: Comparison for L-calls in model with separate buffers in schema 1

References

[1] Jaiswal HK. Priority queues. New York, *Academic Press* (1968).

[2] Kleinrock L. A delay dependent queue discipline.*Naval Research Logistics Quarterly Journal*, Vol. 11 (1964), 329–341.

[3] Melikov A, Ponomarenko L, Kim CS., Performance Analysis and Optimization of Multi-Traffic on Communication Networks. Heidelberg: Springer (2010).

[4] Wittevrongel S, De Vuyst S, Sys C, Bruneel H., A reservation-based scheduling mechanism for fair QoS provisioning in packet-based networks. *In: Proceeding of the 26th IEEE International Teletraffic Congress*, Karlskrona (2014), 55–62.

[5] Lim Y, Kobza JE., Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch, *IEEE Transactions on Communication*, Vol. 38(5)(1990),659–665.

[6] Maertens T, Walraevens J, Bruneel H., On priority queues with priority jumps, *Performance Evaluation*, Vol. 63(12)(2006): 1235–1252.

[7] Maertens T, Walraevens J, Bruneel H., A modified HOL priority scheduling discipline: performance analysis, *European Journal of Operation Research*, Vol. 180(3)(2007): 1168–1185.

[8] Maertens T, Walraevens J, Moeneclaey M, Bruneel H., A new dynamic priority scheme: performance analysis, *In: Proceeding of the 13th International Conference on Analytical and Stochastic Modeling Techniques and Applications*, Bonn (2006), 74–84.

[9] Maertens T, Walraevens J, Bruneel H., Performance comparison of several priority schemes with priority jumps, *Annals of Operation Research*, Vol.! 162 (2008), 109–125.

[10] Walraevens J, Steyaert B, Bruneel H., Performance analysis of single-server ATM queue with priority scheduling, *Computers and Operation Research*, Vol. 30(12)(2003), 1807–1829.

[11] Melikov AZ, Ponomarenko LA., Kim CS., Algorithmic approach to analysis of queuing system with finite buffers and jump priorities, *Journal of Automation and Information Sciences*, Vol. 44(12)(2012),43–54.

[12] Kim CS, Oh Y, Melikov AZ., A space merging approach to the analysis of the performance of queueing models with buffers and priority jumps, *Industrial Engineering and Management Systems*, Vol. 12(3)(2013),274–280.

[13] Melikov AZ, Ponomarenko LA, Kim CS., Approximate method for analysis of queuing models with jump priorities, *Automation and Remote Control*, Vol. 74(1)(2013), 62–75.

[14] Melikov AZ, Ponomarenko LA, Kim CS., Numerical method for analysis of queuing models with priority jumps, *Cybernetics and System Analysis*, Vol. 49(1)(2013), 55–61.

[15] Melikov A., Ponomarenko L., Multidimensional queueing models in telecommunication networks, *Heidelberg: Springer*, (2014).

[16] Tran HT., Do TV., Pap L., Analysis of a queue with two priority classes and feedback control, *Vietnam Journal of Computer Science*, Vol. 1 (2014), 71–78.