mathematics

MDPI

Article

Estimation of the Performance Measures of a Group of Servers Taking into Account Blocking and Call Repetition before and after Server Occupation

Sergey Stepanov * and Mikhail Stepanov

Department of communication Networks and Commutation Systems, Moscow Technical University of Communications and Informatics, 8A, Aviamotornaya str., 111024 Moscow, Russia; m.s.stepanov@mtuci.ru

* Correspondence: s.n.stepanov@mtuci.ru or stpnvsrg@gmail.com

Abstract: The model of a fully available group of servers with a Poisson flow of primary calls and the possibility of losses before and after occupying a free server is considered. Additionally, a call can leave the system because of the aging of transmitted information. After each loss, there is some probability that a customer repeats the call. Such models are seen in the modeling of various telecommunication systems such as emergency information services, call and contact centers, access nodes, etc., functioning in overloading situations. The stationary behavior of the system is described by the infinite-state Markov process. It is shown that stationary characteristics of the model can be calculated with the help of an auxiliary model of the same class but without call repetitions due to losses occurring before and after the occupation of a free server and the aging of transmitted information. The performance measurements of the auxiliary model are calculated by solving a system of state equations using a recursive algorithm based on the concept of the truncation of the used state space. This approach allows significant savings of computer resources to be made by ignoring highly unlikely states in the process of calculation. The error caused by truncation is estimated. The presented numerical examples illustrate the use of the model for the elimination of the negative effects of emergency information service overload based on the filtering of the input flow of calls.

Keywords: emergency services; performance evaluation; repeated calls; Markov process; system of state equations; recursive algorithms; truncation of state space

check for updates

Citation: Stepanov, S.; Stepanov, M. Estimation of the Performance Measures of a Group of Servers Taking into Account Blocking and Call Repetition before and after Server Occupation. *Mathematics* **2021**, *9*, 2811. https://doi.org/10.3390/math9212811

Academic Editor: Konstantin Samouylov

Received: 28 September 2021
Accepted: 2 November 2021
Published: 5 November 2021

1. Introduction

Queueing models that take into account customer behavior after being refused service provide a powerful tool for performance evaluation and the planning of many resource-sharing systems functioning in overload conditions. Such tasks often can be witnessed in the telecommunication industry where, by definition, the volume of resource is restricted, and input traffic has random characteristics. The proper solution of such a problem has special significance, especially in emergency situations, which are very common. Modern telecommunication equipment allows people to ask for assistance at any time and from almost anywhere. Technically, such types of connection are organized through public-safety answering points (PSAP) [1,2]. PSAP is the basis of public-safety systems because it allows a citizen to reach primary emergency services such as police, fire- fighting, ambulance, etc. and provide the required intervention resources if necessary. Because of this, the functioning of PSAP should be organized very carefully.

Using PSAP as an example, let us consider the problems related to an overload of calls and possible solutions. Under normal traffic conditions, it is easy to predict the parameters for an emergency service such as the intensity of arriving calls for each period of the day, and estimate the number of operators required to serve the incoming calls. This problem can be solved by available planning tools based on standard queueing models. However, all emergency services are exposed to overloading situations caused by both natural (earthquakes, floods) and human (fires, terrorist attacks, nuclear accidents) factors [1,2]. The vast majority of citizens seeing or suffering the same incident try to reach emergency services almost at the same time, creating conditions for overload. This situation leads to blocking and the subsequent appearance of a large number of repeated attempts, easily created by modern mobile phones equipped with a call-retrial function.

There are two approaches to overcome the negative consequences of overload. The choice depends on the origin of overload and the time needed to solve the problem. If the rising traffic has random characteristics and is caused by some dangerous event, then, to keep the emergency service in normal order, it is necessary to reduce the number of expected requests. This can be done by redirecting some calls to another PSAP with similar functions [1]. After a short period, the intensity of traffic is stabilized. The increase of input flow can be also caused by the involvement of additional customers. In this situation, it is necessary to estimate the intensity of primary requests from results of measurements of total intensity of incoming requests (primary and repeated), and then calculate the required number of operators. The portion of redirected calls and the number of additional operators can be calculated with help of corresponding mathematical models, taking into account the possibility of call repetition due to the nonavailability of required resources.

Moreover, evident interest from researchers dealing with planning and optimizing the infrastructure of such emergency service models has generated a lot of attention from analysts dealing with call and contact centers, and other resource-sharing systems functioning in overload situations. Attracted by the relevance of the issues, specialists of various expertise work in this area. The level of publications varies from practitioner-oriented research (see overview in [3]) to works performed at a high mathematical level (see overview in [4,5]). A survey of the latest research in this field can be found in [6]. The main point of interest for specialists working on queueing models concerning call centers is to estimate the number of agents such that the values of SLA (Service Level Agreement) indicators are met with prescribed values. In the process of model construction, it is commonly assumed that requests for servicing can be well modeled by nonhomogeneous Poisson process [4–8]. Additionally, it is also supposed that arrival rates and the number of agents are stepwise constant functions [6]. It allows the use of stationary queueing models for each time interval for solving planning problems by so-called SIPP procedure [8]. The $M/M/v$ is very often used in practice, especially as a basis for different forms of call-center calculators [5]. This model can be generalized by taking into account the user patience [9,10]. Such models can be written as $M/M/v+G$, where symbol G denotes the generally distributed user patience. Analytical results have been obtained for such type of models that allow us to study the influence of user patience on performance measurements. Another way to increase the accuracy of the call-center model is to take into account different aspects of agent stuffing. Often this feature complicates the model and in this case the only way to calculate the performance measurements is to solve the system of state equation using the Gauss–Seidel iterative algorithm [11].

A family of Erlang models, integrated into a SIPP procedure, forms the basis of numerical methods of estimation the required number of agents. Unfortunately, this approach has several drawbacks. The main among them is the dependence on the intensity of requests that arrive at the beginning of each interval carried over from the previous call-center functioning history, the so-called backlog [6,12]. This property violates the Poisson assumption used for the description of input flow. There are several approaches to dealing with this problem; see for example [12]. A substantial contribution to the formation of a backlog, especially in the case of overload, is repeated attempts that reflect the quality of request servicing has been observed in the past [13].

A major part of the existing research suggests that the reason for call repetition is strictly connected to occupancy of the servers and does not take into account the possibility of blocking and subsequent call repetition before and after occupation of the free server. However, there exist a lot of situations where this peculiarity of input flow formation should be taken into account. For example, in an emergency service, a customer can repeat an attempt after being processed by IVR (Interactive Voice Response) or can be blocked on route to the emergency service. This happens before the occupation of a free operator. The operator can refuse to service a call, and become the reason for retrial because of restricted access to the intervention resources required for proper call servicing. This happens in the presence of a free operator. The listed sequence of bottlenecks occurs in the emergency service chain; see Figure 1 taken from [1]. The same types of blocking should be considered when we model transmission resource occupancy in routes consisting of several links and take into account the possibility of call repetitions because the called party is busy or there is no answer.

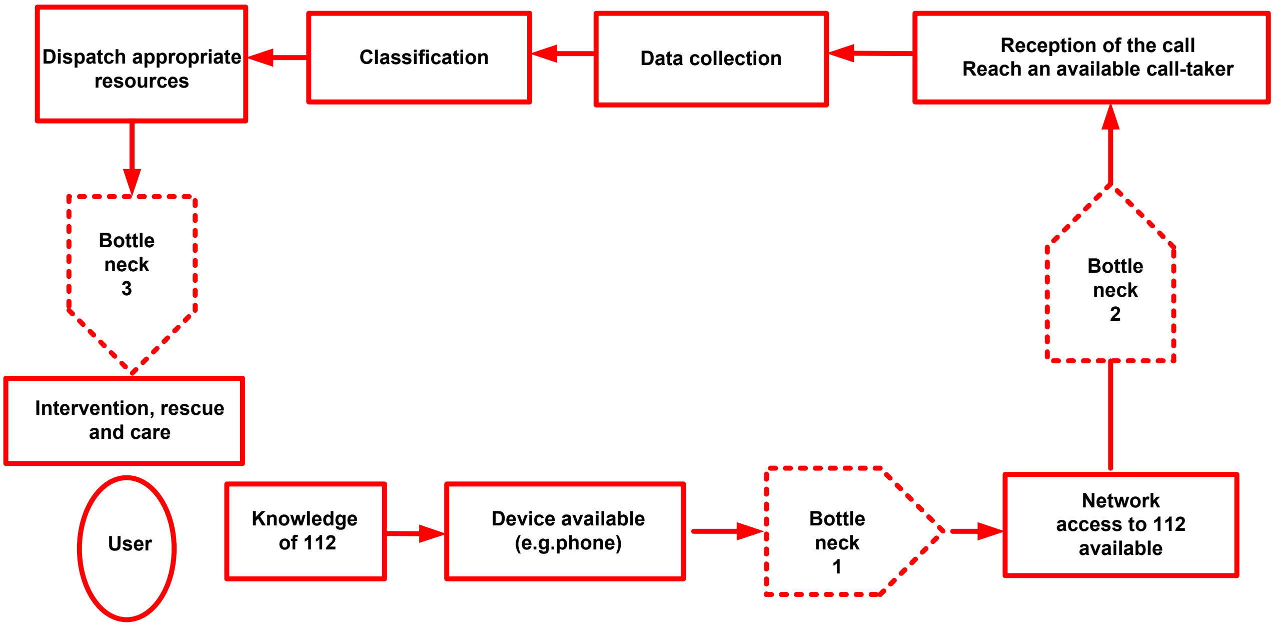

Dispatch appropriate resources
Classification
Data collection
Reception of the call Reach an available call-taker
Bottle neck 3
Bottle neck 2
Intervention, rescue and care
User
Knowledge of 112
Device available (e.g.phone)
Bottle neck 1
Network access to 112 available

Figure 1. Bottlenecks in the emergency service chain [1].

Detailed surveys of publications devoted to queueing systems with retrials have been undertaken, e.g., in [14–17]. By taking into account customer behavior, we can model the input flow in a way that is close to the real processes happening in telecommunication systems that is considered in the overload situation. This aspect creates problems with the theoretical analysis of such models, because time intervals between successive call arrivals are now dependent random variables. We can avoid this problem by including the number of repeated customers in the definition of the model's state. This is done in most publications devoted to this subject. It allows use of Markov processes for model description but seriously complicates the numerical analysis. Performance measures can be found by solving the system of state equations either using the Gauss–Seidel iterative algorithm [2,18–20] or by matrix-geometric methods, when the corresponding Markov process has a suitable structure [17,21,22]. The first approach is time-consuming, and the last is numerically unstable for real values of input parameters because of the necessity to convert a large matrix. Such a situation seriously complicates the use of models with retrials as a component of planning tools, because it requires first numerical stability of the calculation algorithm, and second the possibility to estimate performance measures for any choice of input parameters for a reasonable time.

A model with such characteristics can be found in the literature (see, for example [20]). This is a Markov model of a fully available group of servers with the possibility of call repetition only because all servers are busy. Furthermore, we refer to this model as a basic model with repeated calls. Performance measures of this model can be found using a recursive algorithm based on the concept of truncation of the model's state space [2,17,19,20,23]. For this model, the asymptotic formulae for the calculation of characteristics in the case of extreme load can be also found in explicit form [24]. We generalize the basic model in three directions. First, we will take into account the possibility of the aging of transmitted information during the process of repetitions. Second, we will consider the scenario of blocking before occupying a server, and third, the possibility of losses after occupying a server. In the last two cases a customer can repeat an attempt. The aim of this paper is to show that performance measures of a generalized model can be calculated with the help of a basic model with some changes in input parameters.

The closest model with retrials to the model considered in this paper was analyzed in [2,19]. That model consists of a fully available group of servers with the possibility of waiting. For this model, asymptotic expressions of performance measures for large loads are derived. The values of the characteristics are obtained after solving the system of state equations by the Gauss–Seidel iterative algorithm. The distinguishing features of the model and results obtained in this paper, compared to [2,19], are as follows:

- Additional features, such as the possibility of blocking and call repetition after occupying the server, and the possibility of the aging of transmitted information, are considered.
- Performance measures are calculated using a recursive algorithm without construction of the sequence of approximations such as in the realization of the Gauss–Seidel procedure. This significantly reduces the counting time.

A short outline of this paper is as follows. In Section 2, the mathematical description of the basic model with retrials is presented. In Section 3, the same is done for the considered generalized model with retrials. The possibility of using the basic model for the calculation of performance measures of the generalized model is shown by the construction of some number of auxiliary models. That mathematical analysis is presented in Section 4. Section 5 contains the numerical results that show the possibility of using the analyzed model for the elimination of negative effects from PSAP overload based on filtration of the input flow of calls. Section 6 concludes the paper. The Appendix A contains a brief description of the recursive algorithm that allows the estimation of the performance measures of the basic model using an approach based on the concept of truncated state space.

The novelty of the results obtained in this work is as follows:

- A new model for the functioning of a group of serving devices (servers) has been built and investigated, in which the user can additionally receive a refusal and repeat the call before and after seizing the server. It is also possible for transmitted information to age for the user repeating the call. The model can be used for the elimination of negative effects of overload in various telecommunication systems such as emergency information services, call and contact centers, access nodes, etc.
- A procedure for transforming the model under study has been developed, which makes it possible to exclude the reasons for repeating a call before and after occupation of the server and the possibility of the aging of transmitted information. Thus, the calculation of the model is reduced to the calculation of the characteristics of its particular case, with the only reason for the denial of service being the busyness of the servers. This problem is much easier to solve using an approach based on the concept of a truncated state space. This method of performance measurement estimation allows significant savings of computer resources to be made by ignoring highly unlikely states in the process of calculation.

2. Basic Model Description

The basic model consists of a fully available group of v servers with incoming Poisson flow of primary call intensity λ. After being refused servicing, a customer makes another attempt to obtain the free server in a random time with probability H_1 for a primary call and with probability H_2 for a repeated call, exponentially distributed with parameter μ, and, with additional probability, respectively, $1 - H_1$, and $1 - H_2$, the customer stops their attempts to obtain a connection and leaves the system. Without loss of generality, we can suppose that the service time of a call (primary or repeated) is exponentially distributed with parameter equal to one, and it does not depend on the type of a call. It should be noted that the Poisson model is used only for the construction of the flow of primary calls. The total flow of requests for servicing includes repeated calls, and does not follow the Poisson model. This flow is quite complicated. It reflects the quality of application service that has been observed in the past. The possibility of using Poisson model is also discussed in the Introduction section.

Let us denote by $j(t)$ the number of repeating customers and by $i(t)$ the number of busy servers at time t. The model's functioning is described by a two-dimensional Markov process $r(t) = (j(t), i(t))$ with an infinite number of states S, $(j,i) \in S$, $j = 0,1,\ldots,$ $i = 0,1,\ldots,v$. Let us denote by $P(j,i)$ the probability of stationary state (j,i) of $r(t)$. Here, j is the number of repeating customers and i is the number of busy servers. Using the ideas of [25] it can be proved that stationary distribution for $r(t)$ exists if $\mu > 0$ and $H_2 < 1$. If $H_2 = 1$, parameters λ and μ should satisfy the inequalities $\lambda H_1 < v$ and $\mu > 0$. The values of $P(j,i)$ can be found after solving of the following system of state equations:

$$\begin{aligned} P(j,v)\Big(\lambda H_1 + j\mu(1-H_2) + v\Big) &= P(j,v-1)\lambda + P(j-1,v)\lambda H_1 + \\ &P(j+1,v-1)(j+1)\mu + P(j+1,v)(j+1)\mu(1-H_2), \\ &j = 0,1,\ldots, \quad i = v; \\ P(j,i)\Big(\lambda + j\mu + i\Big) &= P(j,i-1)\lambda + P(j+1,i-1)(j+1)\mu + \\ &P(j,i+1)(i+1), \quad j = 0,1,\ldots, \quad i = 0,1,\ldots v-1. \end{aligned} \tag{1}$$

The normalizing condition is held for $P(j,i)$. Here and beyond, we suppose that in the system of state equations, the probability of state with negative-integer components is equal to zero. Let us introduce the following notations:

$$P(i) = \sum_{j=0}^{\infty} P(j,i), \quad J(i) = \sum_{j=0}^{\infty} P(j,i)j, \quad i = 0,1,\ldots,v.$$

In this paper we are using a standard family of performance measures that can be defined for the model with repeated attempts using stationary probabilities of the Markov process that describes model functioning. All such, characteristics can be expressed through $P(i)$ and $J(i)$. For example, the value of I is the mean number of busy servers, the value of J is the mean number of repeating customers, the value of π_c is the ratio of lost calls, the value of τ is the ratio of repeated attempts in the total flow of incoming calls, and the value of M is the mean number of retrials per one primary call can be found from expressions

$$\begin{aligned} &I = \sum_{i=0}^{v} P(i)i, \quad J = \sum_{i=0}^{v} J(i), \quad \pi_c = \frac{\lambda p(v) + J(v)\mu}{\lambda + J\mu}, \\ &\tau = \frac{J\mu}{\lambda + J\mu}, \quad M = \frac{J\mu}{\lambda}. \end{aligned} \tag{2}$$

Considering this further we will work only with $P(i)$ and $J(i)$ and call these characteristics basic. Summing up the system of state Equation (1) over j with i fixes, we obtain the local conservation laws for macrostate (i)

$$P(i)\lambda + J(i)\mu = P(i+1)(i+1), \quad i = 0,1,\ldots,v-1. \tag{3}$$

Summing up the system of state Equation (1) over i with j fixes, we obtain the similar relationships for macrostate (j)

$$P(j,v)\lambda H_1 = P(j+1,v)(j+1)\mu(1-H_2) + \sum_{i=0}^{v-1} P(j+1,i)(j+1)\mu, \qquad j = 0,1,\ldots \tag{4}$$

Summing up (3) and (4) over i and j, we obtain the conservation laws that relate main stationary characteristics of the basic model

$$\lambda(1-P(v)) + (J - J(v))\mu = I, \qquad J\mu = \lambda P(v)H_1 + J(v)\mu H_2. \tag{5}$$

Relationship (5) can be used for the estimation of the error caused by truncation of the used state space (see Appendix A), for the indirect measurement of the intensity of primary calls and for solving a variety of other problems arising in the exact and approximate analysis of the introduced model [2,18–20]. To find $P(i)$ and $J(i)$, it is necessary to solve the system of state Equation (1). The numerical procedure of doing this is briefly described in Appendix A. In more detail, this approach is presented in [20] for the particular case of the studied model, when $H_1 = H_2 = H$. An elaborated calculation method is based on the concept of truncated state space. This approach allows significant savings of computer resources to be made by ignoring highly unlikely states in the process of calculation. The numerical procedure includes two steps: first, it is necessary to decrease the number of unknowns in the system of state equations by not considering the states with negligible probabilities of existence. Second, it is necessary to find the error of estimation of the performance measures caused by truncation.

3. Generalized Model Description

We generalize the introduced model in the following three directions.

1. We take into account the possibility of aging transmitted information during the process of repetitions. This situation can be modeled in the following way. We suppose that starting the process of repetitions also starts the process of aging. The maximum allowed time of aging is exponentially distributed with parameter σ. After this time, the customer stops the process of repetitions and leaves the system, unserved.
2. We also consider the opportunity of losses before occupying the server. This situation will be modeled in the following way. We suppose that incoming call with probability b_p for primary call and with probability b_r for repeated call can be blocked before entering the service system. In this case, with probability H_1 for primary call and with probability H_2 for repeated call, the customer repeats the attempt after a random time, exponentially distributed with parameter μ, and with additional probabilities, respectively, $1-H_1$, and $1-H_2$, the customer stops attempting to obtain a connection and leaves the system, unserved. The probabilities b_p and b_r can be interpreted as losses at the stage of obtaining service between customer and the service system.
3. In the same way, we take into account the possibility of losses after the occupation of the server. The probabilities of this event are a_p for a primary call and a_r for a repeated call. Probabilities of repetitions are, respectively, H_1 and H_2 for primary and repeated attempts. Time between successive repeated attempts in this case is also exponentially distributed with parameter μ. The probabilities of losses a_p and a_r can be interpreted as losses at the stages of servicing between service system and additional resources required for proper call servicing.

The basic model is a particular case of the generalized model. We obtain it by choosing $b_p = b_r = a_p = a_r = \sigma = 0$. To simplify the notation, we will use for performance measures of the generalized model the same symbols that were used for corresponding characteristics of the basic model. Different values of probabilities of losses for primary and repeated calls allow consideration of the following property of functioning of the service systems in overload conditions. After being refused service with some probability in primary attempt, a customer making repeated attempts will obtain a refusal with greater probability. The functional model of the studied service system is presented in Figure 2.

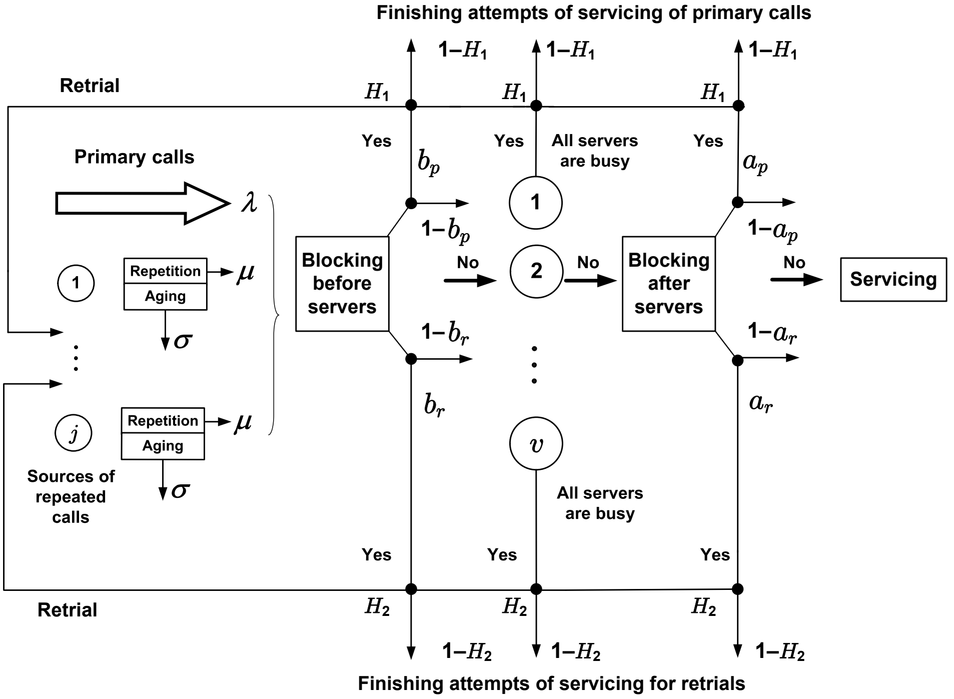


Figure 2. Functional model of the generalized system with blocked call retrials.

The model's functioning is described by a two-dimensional Markov process $r(t) = (j(t), i(t))$, where $j(t)$ is the number of repeating customers and $i(t)$ is the number of busy servers at time t defined in infinite state space S, $(j,i) \in S, j = 0,1,\dots,\ i = 0,1,\dots,v$. To remove out of consideration the exotic cases, we suppose that $b_p < 1$, $b_r < 1$, $a_p < 1$, $a_r < 1$. Let us denote by $P(j,i)$ the probability of the stationary state (j,i) of $r(t)$. Using the ideas of [25] it can be proved that stationary distribution for $r(t)$ exists if

1. $\mu > 0, \sigma > 0$;
2. $\mu > 0, \sigma = 0$ but $H_2 < 1$;
3. $\mu > 0, \sigma = 0$, $H_2 = 1$ but $\lambda H_1 < v$.

The values of $P(j,i)$ can be found after solving the following system of state equations:

$$\begin{aligned}
&P(j,v)\Big(\lambda H_1 + j(\sigma + \mu(1-H_2)) + v\Big) = P(j,v-1)\lambda(1-b_p)(1-a_p)+\\
&\quad P(j-1,v)\lambda H_1 + P(j+1,v-1)(j+1)\mu(1-b_r)(1-a_r)+\\
&\quad P(j+1,v)(j+1)(\sigma+\mu(1-H_2)),\quad j=0,1,\dots,\quad i=v;\\
&P(j,i)\Big(\lambda(1-(1-H_1)(b_p+(1-b_p)a_p))+\\
&\qquad j(\sigma+\mu(1-H_2(b_r+(1-b_r)a_r)))+i\Big) =\\
&\quad P(j,i-1)\lambda(1-b_p)(1-a_p) + P(j-1,i)\lambda H_1(b_p+(1-b_p)a_p)+\\
&\quad P(j+1,i-1)(j+1)\mu(1-b_r)(1-a_r)+\\
&\quad P(j+1,i)(j+1)(\sigma+\mu(1-H_2)(b_r+(1-b_r)a_r))+\\
&\quad P(j,i+1)(i+1),\quad j=0,1,\dots,\quad i=0,1,\dots v-1.
\end{aligned} \tag{6}$$

The normalizing condition is held for $P(j,i)$. Let us introduce the notations

$$P(i) = \sum_{j=0}^{\infty} P(j,i), \quad J(i) = \sum_{j=0}^{\infty} P(j,i)j, \quad i = 0,1,\ldots,v.$$

In the same way as was done for the basic model, we can prove that local conservation laws for macrostates (i) and (j) are looking as follows:

$$\begin{aligned} &P(i)\lambda(1-b_p)(1-a_p) + J(i)\mu(1-b_r)(1-a_r) = P(i+1)(i+1),\\ &\quad i = 0,1,\ldots,v-1; \end{aligned} \tag{7}$$

$$\begin{aligned} &\sum_{i=0}^{v} P(j,i)\lambda H_1 b_p + P(j,v)\lambda H_1(1-b_p) + \sum_{i=0}^{v-1} P(j,i)\lambda H_1(1-b_p)a_p =\\ &\qquad \sum_{i=0}^{v} P(j+1,i)(j+1)\big(\sigma + \mu b_r(1-H_2)\big) +\\ &\qquad P(j+1,v)(j+1)\mu(1-b_r)(1-H_2) +\\ &\qquad \sum_{i=0}^{v-1} P(j+1,i)(j+1)\mu(1-b_r)(1-a_r H_2), \qquad j = 0,1,\ldots \end{aligned} \tag{8}$$

If values of $P(i)$ and $J(i)$ are known, we can calculate the main stationary characteristics of the model. For example, the corresponding formulae for I, J, π_c, τ and M can be found from expressions (see (2))

$$\begin{aligned} &I = \sum_{i=0}^{v} P(i)i, \quad J = \sum_{i=0}^{v} J(i), \quad \pi_c = \frac{\lambda B_p + B_r\mu + J\sigma}{\lambda + J\mu},\\ &\tau = \frac{J\mu}{\lambda + J\mu}, \quad M = \frac{J\mu}{\lambda}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} B_p &= b_p + (1-b_p)P(v) + (1-b_p)(1-P(v))a_p,\\ B_r &= Jb_r + J(v)(1-b_r) + (J - J(v))(1-b_r)a_r. \end{aligned}$$

Summing up (7) and (8) over i and j, we obtain the conservation laws that relate to the main stationary characteristics of the generalized model

$$\begin{aligned} &\lambda(1-P(v))(1-b_p)(1-a_p) + (J - J(v))(1-b_r)(1-a_r)\mu = I,\\ &J(\mu+\sigma) = \lambda B_p H_1 + B_r\mu H_2. \end{aligned} \tag{10}$$

As we can see from the above expressions, all stationary performance measures of the generalized model can be defined through $P(i)$ and $J(i)$, $i = 0,1,\ldots,v$. To calculate the values of $P(i)$ and $J(i)$, it is necessary to solve the system of state Equation (6). We are not able to find the values of $P(j,i)$ recursively as can be done for the basic model (see Appendix A), because in the generalized model a customer can repeat an attempt for any state of busy servers $i = 0,1,\ldots,v$. In the basic model, this can be done only when $i = v$. What remains is the solving of (6) by standard approaches, either by Gauss–Seidel iterative algorithm or by matrix-geometric methods. The first approach is time-consuming; the last is numerically unstable for real values of input parameters, because of the necessity to convert a large matrix. We overcome these difficulties by showing that values of $P(i)$ and $J(i)$ can be calculated with the help of the corresponding characteristics of the basic model after a suitable choice of values of its input parameters. We prove it by constructing the number of auxiliary models using another probabilistic interpretation of input parameters of the generalized model or by algebraic transformations of the system of state equations.

All four auxiliary models that will appear later (the last is a basic model) will be particular cases of the generalized model, so we can use the same symbols to denote its input parameters and characteristics, differing them only by a superscript. The value of the digit used in the superscript for performance measures $P(i)$ and $J(i)$ means the number of auxiliary model. For input parameters, the digit denotes the number, changing by parameter its value. A parameter without a superscript has the same value as the generalized model.

4. Auxiliary Models

4.1. First Auxiliary Model

The process of call repetition and the aging of transmitted information can be represented in the following way. Let us consider some refusal. If such an event happens, a customer, with probabilities H_x ($x = 1, 2$, depending on the number of attempt), makes another call after a random time T_μ with exponential distribution with parameter μ or, with probabilities $1 - H_x$, stops attempts to obtain service. If refusal in service was obtained for primary calls, then the random time T_σ of aging of the transmitted information for the considered customer starts at the moment of call repetition. The value of T_σ has exponential distribution with parameter σ and does not depend on other exponentially distributed variables realized in the model. If the refusal in servicing was obtained in repeated attempts, then, starting from the moment of call repetition, the remaining time of aging also has exponential distribution with parameter σ and does not depend on other exponentially distributed variables realized in the model. This result follows from the basic property of exponentially distributed random variables.

By considering the above-stated argument, we can remodel the process of call repetition and aging of transmitted information in the following way. After refusal of service, a customer with probability H_x stays in the system for a random time equal to $\min(T_\mu, T_\sigma)$ with exponential distribution with parameter $\mu + \sigma$. Then, with probability of event $T_\mu > T_\sigma$ equal to $\frac{\sigma}{\mu+\sigma}$, at the end of this time a customer leaves the system because of the aging of transmitted information and with the probability of event $T_\mu < T_\sigma$ equal to $\frac{\mu}{\mu+\sigma}$ a customer makes a repeated attempt. The fact that with some known probability the customer stays in the system without subsequent call repetition and then leaves the system does not influence the process of server occupation.

Let us construct the first auxiliary model by making the following changes in the generalized model. We suppose that after being refused, a customer with probability $\frac{H_x\mu}{\mu+\sigma}$ makes another attempt in a time interval with exponential distribution with parameter $\mu + \sigma$, and with additional probability $1 - \frac{H_x\mu}{\mu+\sigma}$ the customer immediately leaves the system. It is easy to notice that the first auxiliary model is a particular case of the generalized model. We obtain it by choosing in the generalized model the following values of input parameters:

$$H_1^{(1)} = \frac{H_1\mu}{\mu+\sigma}, \quad H_2^{(1)} = \frac{H_2\mu}{\mu+\sigma}, \quad \mu^{(1)} = \mu + \sigma, \quad \sigma^{(1)} = 0. \tag{11}$$

Other parameters remain the same as in the generalized model. From the above considerations (7) and (12), we can conclude that the following relations are true:

$$P(i) = P^{(1)}(i), \quad J(i) = J^{(1)}(i)\,\frac{\mu+\sigma}{\mu}, \quad ,i = 0, 1, \ldots, v, \tag{12}$$

where

$$P^{(1)}(i) = \sum_{j=0}^{\infty} P^{(1)}(j,i), \quad J^{(1)}(i) = \sum_{j=0}^{\infty} P^{(1)}(j,i)j, \quad i = 0, 1, \ldots, v$$

and $P^{(1)}(j, i)$ are the stationary probabilities of state (j, i) of a Markov process that describes the functioning of the first auxiliary model. Relationship (12) for particular cases of the basic model was also proven in [26,27] by algebraic transformations of the system of state equations. Let us consider the first auxiliary model and continue the process of its simplification.

4.2. Second Auxiliary Model

All input parameters of this model have the same values as the corresponding parameters of the first auxiliary model. We only divide the repeating customers into two groups. In the first group, the number of repeating customers increases by one with probability $H_1^{(1)}(b_p + (1 - b_p)a_p)$ each time when the primary call appears in the first auxiliary model. The repeating customer of this group leaves the group with probability $1 - H_2^{(1)}(b_r + (1 - b_r)a_r)$ after a random time with exponential distribution with parameter μ and with additional probability of continuing to stay in the first group.

Please note that increasing and decreasing of the number of repeating customers in the first group happens with the same probabilities for all possible values of busy servers. We put all other repeating subscribers of first auxiliary model in the second group. After leaving the first group, a customer can occupy the server, move to the second group of repeating customers, or stop attempts for service. Otherwise, the functioning of a new auxiliary model does not change compared to the first auxiliary model. A more detailed description of the changing of the states of the second auxiliary model can be seen from the system of state Equation (13) that will appear below.

Let us denote by $j_1(t)$ and $j_2(t)$ the number of repeating customers at time t in the first and second groups, respectively, and by $i(t)$ we denote the number of busy servers at time t. The dynamic of changing the model states is described by a three-dimensional Markov process of the type $r(t) = (j_1(t), j_2(t), i(t))$ with infinite number of states S, $(j_1, j_2, i) \in S$, $j_1 = 0, 1, \ldots,$ $j_2 = 0, 1, \ldots,$ $i = 0, 1, \ldots, v$. Let us denote by $P^{(2)}(j_1, j_2, i)$ the probability of stationary state (j_1, j_2, i) of $r(t)$. The system of state equations can be written in the following way

$$\begin{aligned} P^{(2)}(j_1, j_2, v)\Big(\lambda H_1^{(1)} + j_1\mu^{(1)}(1 - H_2^{(1)}(b_r + (1 - b_r)a_r)) + j_2\mu^{(1)}(1 - H_2^{(1)}) + v\Big) = \\ P^{(2)}(j_1, j_2, v - 1)\lambda(1 - b_p)(1 - a_p)+ \\ P^{(2)}(j_1 - 1, j_2, v)\lambda H_1^{(1)}(b_p + (1 - b_p)a_p)+ \\ P^{(2)}(j_1, j_2 - 1, v)\lambda H_1^{(1)}(1 - b_p)(1 - a_p)+ \\ P^{(2)}(j_1 + 1, j_2, v - 1)(j_1 + 1)\mu^{(1)}(1 - b_r)(1 - a_r)+ \\ P^{(2)}(j_1 + 1, j_2, v)(j_1 + 1)\mu^{(1)}(1 - H_2^{(1)})+ \\ P^{(2)}(j_1 + 1, j_2 - 1, v)(j_1 + 1)\mu^{(1)}H_2^{(1)}(1 - b_r)(1 - a_r)+ \\ P^{(2)}(j_1, j_2 + 1, v - 1)(j_2 + 1)\mu^{(1)}(1 - b_r)(1 - a_r)+ \\ P^{(2)}(j_1, j_2 + 1, v)(j_2 + 1)\mu^{(1)}(1 - H_2^{(1)}), \\ j_1 = 0, 1, \ldots, \quad j_2 = 0, 1, \ldots, \quad i = v; \end{aligned} \tag{13}$$

$$\begin{aligned}
P^{(2)}(j_1,j_2,i)\Big(&\lambda(1-(1-H_1^{(1)})(b_p+(1-b_p)a_p))+j_1\mu^{(1)}(1-H_2^{(1)}(b_r+(1-b_r)a_r))+\\
&j_2\mu^{(1)}(1-H_2^{(1)}(b_r+(1-b_r)a_r))+i\Big)=\\
&P^{(2)}(j_1,j_2,i-1)\lambda(1-b_p)(1-a_p)+\\
&P^{(2)}(j_1-1,j_2,i)\lambda H_1^{(1)}(b_p+(1-b_p)a_p)+\\
&P^{(2)}(j_1+1,j_2,i-1)(j_1+1)\mu^{(1)}(1-b_r)(1-a_r)+\\
&P^{(2)}(j_1+1,j_2,i)(j_1+1)\mu^{(1)}(1-H_2^{(1)})(b_r+(1-b_r)a_r)+\\
&P^{(2)}(j_1,j_2+1,i-1)(j_2+1)\mu^{(1)}(1-b_r)(1-a_r)+\\
&P^{(2)}(j_1,j_2+1,i)(j_2+1)\mu^{(1)}(1-H_2^{(1)})(b_r+(1-b_r)a_r)+\\
&P^{(2)}(j_1,j_2,i+1)(i+1),\\
&\quad j_1=0,1,\dots,\quad j_2=0,1,\dots,\quad i=0,1,\dots,v-1.
\end{aligned}$$

The normalization condition is held for $P^{(2)}(j_1,j_2,i)$. The input parameters of the first and the second auxiliary models are the same. The process of call repetition and server occupation is the same. We only introduce the rule of separating the repeating customers into two groups. As result, between the corresponding characteristics of the first and the second auxiliary models the following relations are valid:

$$P^{(1)}(i)=P^{(2)}(i),\quad J^{(1)}(i)=J_1^{(2)}(i)+J_2^{(2)}(i)=J^{(2)}(i),\quad i=0,1,\dots,v, \tag{14}$$

where

$$\begin{aligned}
&P^{(2)}(i)=\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}P^{(2)}(j_1,j_2,i),\quad J_1^{(2)}(i)=\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}P^{(2)}(j_1,j_2,i)j_1,\\
&J_2^{(2)}(i)=\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}P^{(2)}(j_1,j_2,i)j_2,\quad i=0,1,\dots,v.
\end{aligned}$$

4.3. Third Auxiliary Model

Let us show that the solution of (13) can be represented in the form

$$\begin{gathered}
P^{(2)}(j_1,j_2,i)=P^{(2)}(0,j_2,i)\,\frac{D^{j_1}}{j_1!},\quad j_1=0,1,\dots,\\
(j_1,j_2,i)\in S,
\end{gathered} \tag{15}$$

where

$$D=\frac{\lambda H_1^{(1)}(b_p+(1-b_p)a_p)}{\mu^{(1)}\big(1-H_2^{(1)}(b_r+(1-b_r)a_r)\big)}.$$

After the substitution of (15) into (13) with changing $P^{(2)}(0,j_2,i)$ into $P^{(3)}(j,i)$, we obtain the following system of linear equations

$$\begin{aligned}
P^{(3)}(j,v)\Big(\lambda^{(1)}H_1^{(2)}+j(\mu^{(2)}(1-H_2^{(1)})+\sigma^{(1)})+v\Big)&=P^{(3)}(j,v-1)\lambda^{(1)}+\\
P^{(3)}(j-1,v)\lambda^{(1)}H_1^{(2)}&+P^{(3)}(j+1,v-1)(j+1)\mu^{(2)}+\\
+P^{(3)}(j+1,v)(j+1)&(\mu^{(2)}(1-H_2^{(1)})+\sigma^{(1)}),\\
j=0,1,\dots,&\quad i=v;
\end{aligned} \tag{16}$$

$$P^{(3)}(j,i)\Big(\lambda^{(1)}+j(\mu^{(2)}+\sigma^{(1)})+i\Big)=P^{(3)}(j,i-1)\lambda^{(1)}+P^{(3)}(j+1,i-1)(j+1)\mu^{(2)}+$$
$$P^{(3)}(j+1,i)(j+1)\sigma^{(1)}+P^{(3)}(j,i+1)(i+1),$$
$$j=0,1,\dots,\quad i=0,1,\dots v-1,$$

where

$$\begin{aligned}
&\lambda^{(1)}=\lambda\big((1-b_p)(1-a_p)+(1-b_r)(1-a_r)A\big),\\
&A=\frac{H_1^{(1)}(b_p+(1-b_p)a_p)}{1-H_2^{(1)}(b_r+(1-b_r)a_r)},\\
&\mu^{(2)}=\mu^{(1)}(1-b_r)(1-a_r),\\
&\sigma^{(1)}=\mu^{(1)}(1-H_2^{(1)})(b_r+(1-b_r)a_r),\\
&H_1^{(2)}=\frac{(1-b_p)(1-a_p)H_1^{(1)}+(1-b_r)(1-a_r)H_2^{(1)}A}{(1-b_p)(1-a_p)+(1-b_r)(1-a_r)A}.
\end{aligned}\tag{17}$$

Because $H_1^{(2)}\le 1$ and $H_2^{(1)}\le 1$ then we can conclude that the system of linear Equation (16) will be the system of state Equation (6) of a particular case of generalized model with the choice of blocking probabilities before and after the group of v servers in the form $b_p^{(1)}=b_r^{(1)}=a_p^{(1)}=a_r^{(1)}=0$ and other parameters defined by (11) and (17). Provided that the ergodicity properties hold for Markov process $r(t)$ describing the generalized model, the system of state Equation (16) has a unique solution and the solution of (13) has the form (15). Let us call the constructed particular case of generalized model as the third auxiliary model. From (15) we follow the relationship between corresponding characteristics of the second and the third auxiliary models

$$P^{(2)}(i)=P^{(3)}(i),\quad J^{(2)}(i)=J^{(3)}(i)+D\,P^{(3)}(i),\quad i=0,1,\dots,v,\tag{18}$$

where

$$P^{(3)}(i)=\sum_{j=0}^{\infty}P^{(3)}(j,i),\quad J^{(3)}(i)=\sum_{j=0}^{\infty}P^{(3)}(j,i)j,\quad i=0,1,\dots,v.$$

4.4. Fourth Auxiliary Model

The last step in constructing a basic model is to eliminate the event of aging of the transmitted information from the third auxiliary model. The eliminated event is described by a random time with an exponential distribution with parameter $\sigma^{(1)}$. We have already discovered how to solve this problem when analyzing the first auxiliary model (see, Section 4.1). According to these results, the fourth auxiliary model is a particular case of the generalized model with the following choice of input parameters

$$\begin{aligned}
&\lambda^{(1)}=\lambda\big((1-b_p)(1-a_p)+(1-b_r)(1-a_r)A\big),\\
&\mu^{(3)}=\mu^{(2)}+\sigma^{(1)},\\
&\sigma^{(2)}=b_p^{(1)}=b_r^{(1)}=a_p^{(1)}=a_r^{(1)}=0,\\
&H_1^{(3)}=H_1^{(2)}\frac{\mu^{(2)}}{\mu^{(2)}+\sigma^{(1)}},\\
&H_2^{(2)}=H_2^{(1)}\frac{\mu^{(2)}}{\mu^{(2)}+\sigma^{(1)}}.
\end{aligned}\tag{19}$$

This choice of parameters tells us that the fourth auxiliary model is equivalent to the basic model. Similar to the (12), we obtain the relationship between basic characteristics of the third and fourth auxiliary models

$$P^{(3)}(i) = P^{(4)}(i), \quad J^{(3)}(i) = J^{(4)}(i)\,\frac{\mu^{(2)} + \sigma^{(1)}}{\mu^{(2)}}, \quad i = 0, 1, \dots, v, \tag{20}$$

where

$$P^{(4)}(i) = \sum_{j=0}^{\infty} P^{(4)}(j,i), \quad J^{(4)}(i) = \sum_{j=0}^{\infty} P^{(4)}(j,i)j, \quad i = 0, 1, \dots, v$$

and $P^{(4)}(j,i)$ are stationary probabilities of states (j,i) of Markov process that describes the functioning of the fourth auxiliary model that is equivalent to the basic model.

Final expressions for input parameters of the equivalent basic model through parameters of the generalized model are

$$\begin{aligned}
&\lambda^{(1)} = \lambda\big((1-b_p)(1-a_p) + (1-b_r)(1-a_r)A\big),\\
&A = \frac{H_1\mu\big(b_p + (1-b_p)a_p\big)}{\sigma + \mu\big(1 - H_2(b_r + (1-b_r)a_r)\big)},\\
&\mu^{(3)} = \sigma + \mu\big(1 - H_2(b_r + (1-b_r)a_r)\big),\\
&H_1^{(3)} = \frac{(1-b_p)(1-a_p)H_1 + (1-b_r)(1-a_r)H_2A}{(1-b_p)(1-a_p) + (1-b_r)(1-a_r)A} \times\\
&\qquad \frac{\mu(1 - (b_r + (1-b_r)a_r))}{\sigma + \mu(1 - H_2(b_r + (1-b_r)a_r))},\\
&H_2^{(2)} = \frac{H_2\mu\big(1 - (b_r + (1-b_r)a_r)\big)}{\sigma + \mu(1 - H_2(b_r + (1-b_r)a_r))}.
\end{aligned} \tag{21}$$

Using (21) as input parameters of the basic model, we can calculate the characteristics of the generalized model $P(i), J(i)$, $i = 0, 1, \dots, v$ for a given set of input parameters $\lambda, b_p, b_r, H_1, H_2, a_p, a_r, \mu, \sigma$. The final results are as follows:

$$\begin{aligned}
&P(i) = P^{(4)}(i),\\
&J(i) = \frac{\sigma + \mu\big(1 - H_2(b_r + (1-b_r)a_r)\big)}{\mu(1-b_r)(1-a_r)}\,J^{(4)}(i) +\\
&\qquad \frac{\lambda H_1(b_p + (1-b_p)a_p)}{\sigma + \mu\big(1 - H_2(b_r + (1-b_r)a_r)\big)}\,P^{(4)}(i),\\
&\qquad\qquad i = 0, 1, \dots, v.
\end{aligned} \tag{22}$$

5. Numerical Examples

Let us consider a few numerical examples that illustrate the usage of the constructed model for practical purposes. We start by showing the influence of blocking before and after the servers on values of main performance measurements: π_c—the ratio of lost calls—and $\delta = \frac{I}{v}$—the mean value of one server usage. Figure 3 illustrates the dependence of π_c and δ on the probability of primary and repeated calls blocking before servers.

The values of b_p are varying in interval from 0 to 0.6. For fixed b_p the value of b_r is calculated from relationship $b_r = b_p \times 1.3$. Other fixed parameters used for Figure 3 are as follows: $\lambda = 22$; $v = 30$; $\mu = 5$; $H_1 = H_2 = 0.8$; $\sigma = 0.2$; $a_p = a_r = 0$.

When $b_p = b_r = 0$ the blocking before servers does not affect the model performance measurements and π_c, δ have acceptable for practical implementation levels $\pi_c = 0.0429$, $\delta = 0.7261$. With the increasing of b_p and b_r the ratio of lost calls strongly increases. In contrast, the mean value of one server usage decreases. The presented numerical results show the negative aspects of call losses before servers on values of main performance measures. The situation continues to deteriorate if losses also happen after servers. Additional reasons for call repetition increase the ratio of lost calls and decrease the mean value of server use compared to Figure 3. This result is illustrated in Figure 4 for the same values of input parameters that was used in Figure 3 only with $a_p = b_p$ and $a_r = b_r$.

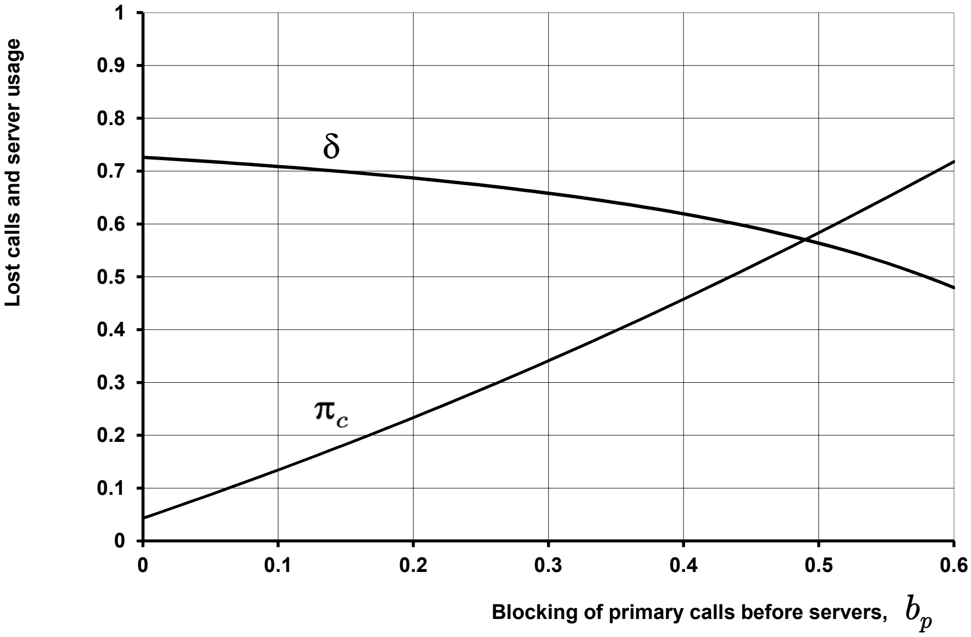


Figure 3. Dependence of π_c and δ on b_p and b_r.

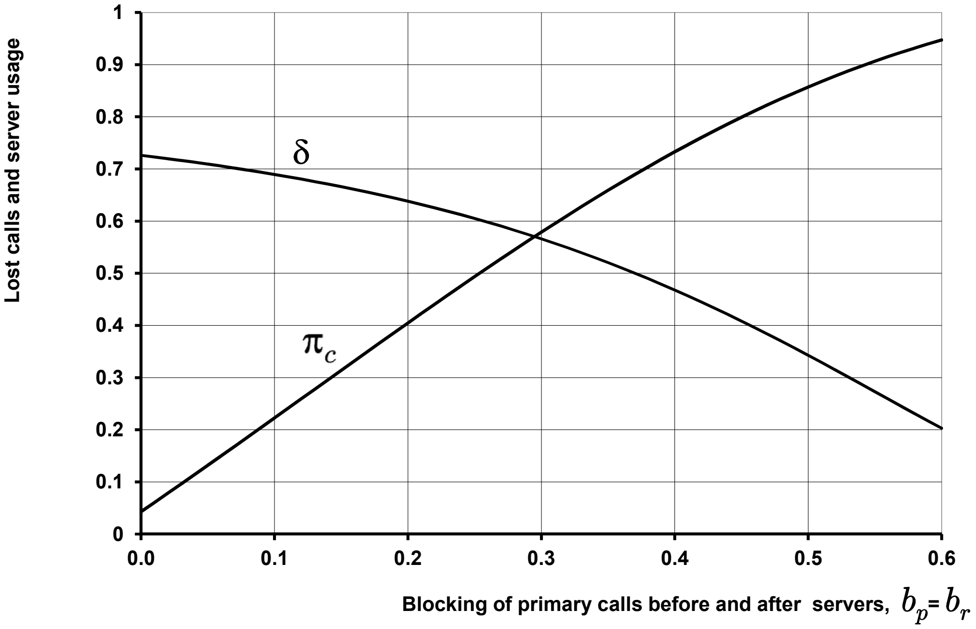


Figure 4. Dependence of π_c and δ on b_p, b_r, a_p and a_r.

It is well known that emergency services are an object of overload. The overload can be caused by many reasons that are discussed in Section 1. To decrease the negative consequences of overload we can redirect part of the input flow of primary calls to other emergency services with similar functions. The proportion of redirected calls can be found with the help of the constructed model. Let us consider a numerical example. The model's input parameters are as follows: $\lambda = 45$; $v = 30$; $\mu = 5$; $H_1 = H_2$ and takes values 0.7; 0.8; 0.9; 0.99; $\sigma = 0.2$; $b_p = b_r = a_p = a_r = 0$.

Figure 5 illustrates the dependence of π_c on the portion of redirected primary calls r that varies from 0 (no redirection) to 0.6 for different values of persistence function $H_1 = H_2$

that takes values 0.7; 0.8; 0.9; 0.99. The results of calculations show that by redirecting some part of primary calls we can decrease the value of losses to the prescribed level (here the restriction on π_c is chosen as $\pi_c < 0.05$). It is worth mentioning that the required value of redirected primary calls is weakly dependent on the value of probability of repetitions.

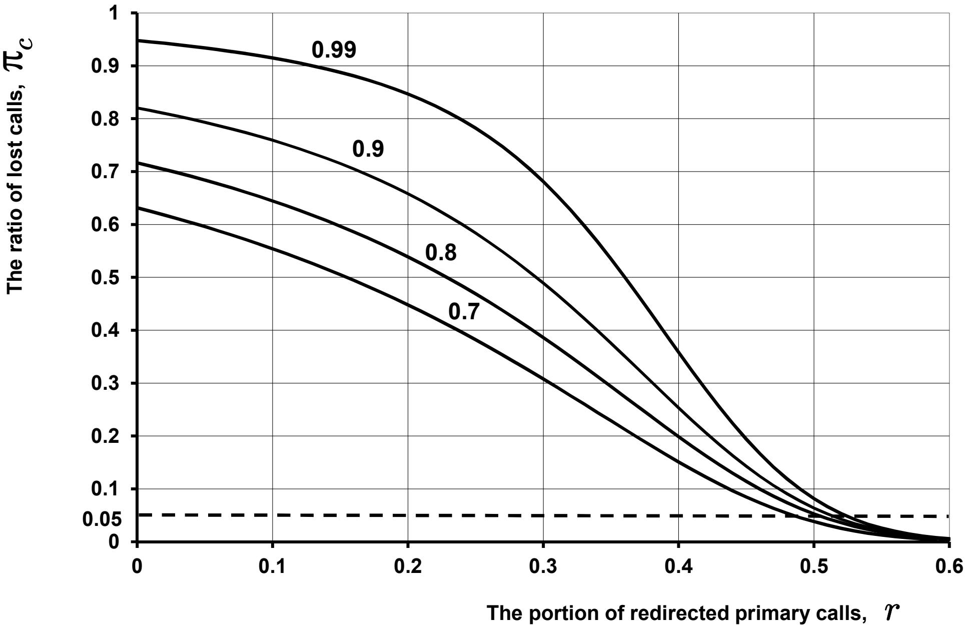


Figure 5. Dependence of π_c on the portion of redirected primary calls r for different values of $H_1 = H_2$.

6. Discussion

The model of a fully available group of servers with Poisson flow of primary calls and possibility of losses before and after occupying the free server is constructed and analyzed. It is supposed that a call can leave the system because of the aging of transmitted information. All random variables realized in the model have exponential distribution and do not depend on each other. After every loss, a customer with some probability depending on the number of unsuccessful attempt repeats the call. Such models are seen in the modeling of various telecommunication systems such as emergency information services, call and contact centers, access nodes, etc., functioning in an overload situation. The stationary behavior of the system is described by the infinite state Markov process.

A procedure for transforming the model under study has been developed, which makes it possible to exclude the reasons for repeating a call before and after occupation of the server and the possibility of the aging of transmitted information. The procedure consists of four consecutive steps, where each step includes the construction and analysis of an auxiliary model that simplifies the functioning of the generalized model (see Sections 4.1–4.4). Thus, the calculation of the model's performance measures is reduced to the calculation of the characteristics of its particular case with the only reason for the denial of service being the busyness of the servers. The process of calculation is based on constructing and solving the system of state equations. This problem is much easier to solve using an approach based on the concept of a truncated state space. The error caused by truncation is estimated.

The positive features of estimation of the performance measurements based on the concept of a truncated state space lies in the following characteristics:

- The realization of such and approach makes it possible to evaluate characteristics faster, due to the reduction of the used state space and the ability to evaluate characteristics with a predetermined accuracy (see, Table A2). The reduction of counting time, especially noticeable in the analysis of asymptotic cases when intensity of input flow, is increasing (overload) [20].

- A suggested approach allows the avoidance of overflow/underflow problems in computer calculation by taking out of consideration highly unlikely states.
- The values of probabilities are determined from simple recursive relations and do not require matrix inversion. All these properties are of great importance for the implementation of the algorithm in planning tools such as network calculators.

The presented numerical examples illustrate the usage of the model for showing the negative effect of losses before and after occupation of the servers on the model's performance measures and for the elimination of PSAP overload based on the filtering of the input flow of calls.

The limitations of the obtained results can be summed up as follows:

- The usage of exponential distribution with the same parameter for modeling the service time for all types of incoming requests. This property is necessary for the estimation of the error caused by truncation of the used state space. It does not mean that without this specific characteristic of the model we are not able to implement truncation concept, but in such scenarios finding errors is quite difficult. In most cases it can be done only empirically [20].
- The absence of waiting room for blocked requests. Because of this property, we are not able to estimate the distribution function of waiting time and mean value of waiting.
- The usage of Poisson assumption as a model for primary requests. It should be noted that the Poisson model is used only for constructing the flow of primary calls. The total flow of requests for servicing includes repeated calls and does not follow the Poisson model. By doing this, we are reconstructing the input flow closer to reality, especially when service system is functioning in overload conditions. Retrials are forming a backlog that reflects the quality of request servicing that has been observed in the past. Nevertheless,the acceptability of using Poisson assumption for primary incoming requests needs additional study.

Directions for further research are formulated by taking into consideration above given limitations:

- Find the error of performance measurement estimation caused by the truncation in the general case when service times depend on the type of a call and has exponential distribution with different mean values.
- Generalize the model considered in the paper by taking into account the possibility of waiting and construct an effective algorithm of performance measurement estimation including the mean waiting time and waiting time distribution. Another direction of generalization considers the dependance of probability of call repetition on the type of refusal.
- Study the acceptability of using Poisson assumption for primary requests.

Author Contributions: Conceptualization, S.S.; methodology, S.S. and M.S.; software, M.S.; validation, S.S. and M.S.; formal analysis, S.S. and M.S.; investigation, M.S.; writing, original draft preparation, S.S. and M.S.; writing, review and editing, S.S. and M.S.; supervision, S.S.; project administration, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

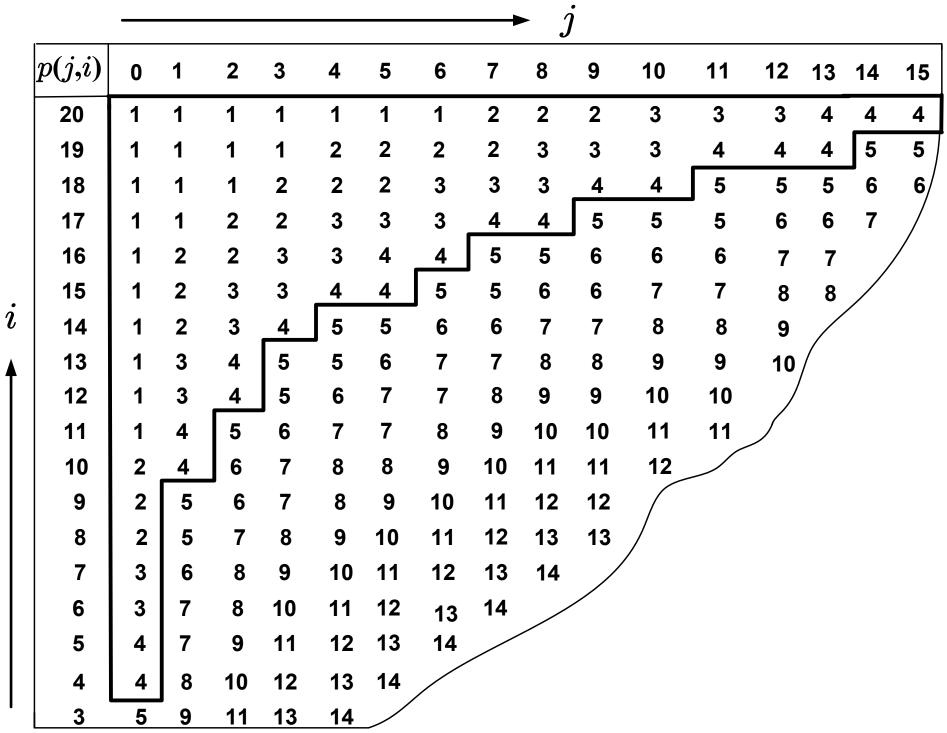
Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Basic Model Analysis Based on the Concept of Truncation

Appendix A.1. Truncated State Space

Here in the example of the basic model taking into account retrials, introduced in Section 2, we realize the ideas of estimating the performance measures of models with retrials based on the concept of truncation of the used state space. In more detail the used approach is presented in [20] for a particular case of the studied model when $H_1 = H_2 = H$. The necessity to truncate the state space follows from the fact that in the model description the number j of repeated customers is unlimited. For simplicity we suppose here that $\max(H_1, H_2) < 1$. The general case is analyzed in the same way but needs to consider more particular cases.

To have the possibility to use standard procedures of linear algebra for solving the system of state equations it is necessary to restrict j by some integer number N chosen sufficiently large. The possibility of doing this is based on the fact that stationary probabilities $P(j,i)$ of the states (j,i), where i is the number of busy servers, strongly decrease when j increases for i fixed. In the majority of cases, we can also suspect that $P(j,i)$ strongly decrease when i decreases for j fixed. This property based on the obvious characteristic of the models with taking into account retrials: the probability of having many repeated customers and a lot of free servers should be very small. Let us consider a numerical example that illustrates this property. Figure A1 shows the dependence of $-\lfloor \log_{10} P(j,i) \rfloor$ on j and i. Model parameters are as follows: $v = 20$; $\lambda = 20$; $\mu = 10$; $H_1 = 0.7$; $H_2 = 0.9$.



$p(j,i)$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
20	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4
19	1	1	1	1	2	2	2	2	3	3	3	4	4	4	5	5
18	1	1	1	2	2	2	3	3	3	4	4	5	5	5	6	6
17	1	1	2	2	3	3	3	4	4	5	5	5	6	6	7	
16	1	2	2	3	3	4	4	5	5	6	6	6	7	7		
15	1	2	3	3	4	4	5	5	6	6	7	7	8	8		
14	1	2	3	4	5	5	6	6	7	7	8	8	9			
13	1	3	4	5	5	6	7	7	8	8	9	9	10			
12	1	3	4	5	6	7	7	8	9	9	10	10				
11	1	4	5	6	7	7	8	9	10	10	11	11				
10	2	4	6	7	8	8	9	10	11	11	12					
9	2	5	6	7	8	9	10	11	12	12						
8	2	5	7	8	9	10	11	12	13	13						
7	3	6	8	9	10	11	12	13	14							
6	3	7	8	10	11	12	13	14								
5	4	7	9	11	12	13	14									
4	4	8	10	12	13	14										
3	5	9	11	13	14											

Figure A1. Dependence of $-\lfloor \log_{10} P(j,i) \rfloor$ on j and i for $v = 20$; $\lambda = 20$; $\mu = 10$; $H_1 = 0.7$; $H_2 = 0.9$.

The bold line shows the borders of the set of states (j,i), where $P(j,i) \geq 10^{-4}$. We define the truncated state space R as an arbitrary subset of S with exception of S itself. The border B of R will be defined as subset of R from which $r(t)$ may leave R in one transition. Numerical experiments point out that set of states (j,i), where $P(j,i) \geq \varepsilon$ has the same geometrical properties as the truncated state space providing the given relative error ε of performance measures estimation.

In the majority of situations, the truncated state space that can be used for the estimation of performance measurements with given relative error of the form presented in Figure A2.

Arrows indicate the directions in which the initial process $r(t)$ can leave the considered truncated state space in one transition. The border states are shaded. Exceptions concern asymptotic cases that have other forms of truncated state space but can be studied in the same way [20]. Formal definitions of the accepted further truncated state space R are as follows: $(j,i) \in R, \;\; j = 0,1,\ldots,N; \;\; L(j) \le i \le v$. Here, $L(j)$ is a nondecreasing integer function defined on $0,1,\ldots,N$.

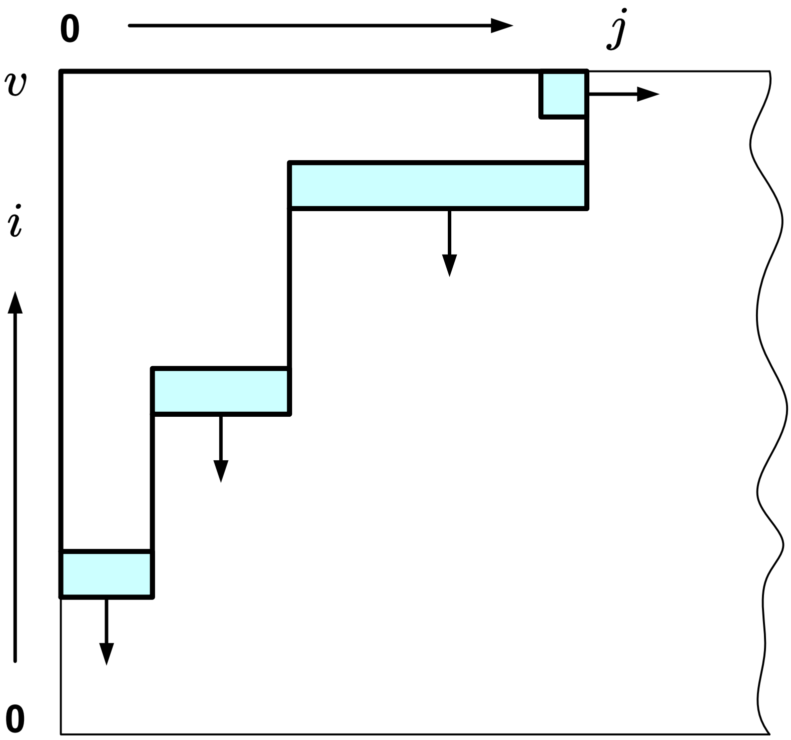


Figure A2. Accepted form of truncated state space. Arrows indicate the directions in which the initial process $r(t)$ can leave the considered truncated state space. The border states are shaded.

In the process of construction of the calculation procedure based on the concept of truncation state space, we need to solve two problems: first, it is necessary to decrease the number of unknowns in the system of state equations by not considering the states with negligible probabilities of existence, and second, it is necessary to find the error of estimation of the performance measures caused by truncation. Let us consider two realizations of the formulated approach: primary and advanced. In the primary version we truncate only the number of repeating customers (see Appendix A.2). The advanced version additionally includes the truncation of occupied servers (see Appendix A.3). Both realizations have the possibility of estimation the relative error caused by truncation in terms of characteristics of auxiliary process defined only on the truncated state space. The choice of the borders of truncated state space for proposed relative error of performance measures estimation will be considered in Appendix A.5.

Appendix A.2. Primary Truncation

Let us denote by R^N the truncated state space of rectangular form $(j,i) \in R^N$, $j = 0,1,\ldots,N, \;\; i = 0,1,\ldots,v$. The state space R^N with respect to $r(t)$ has only one border state (N,v). The process $r(t)$ moves out of R^N when primary request comes in the state (N,v) and customer with probability H_1 decides to repeat an attempt. We prevent this transition if at this moment we take out of the model one repeating customer. Let us denote obtained in this way, auxiliary process by symbol $r^b(t) = (j^b(t), i^b(t))$. Figure A3 illustrates the results of primary truncation of the used state space S and the definition of $r^b(t)$.

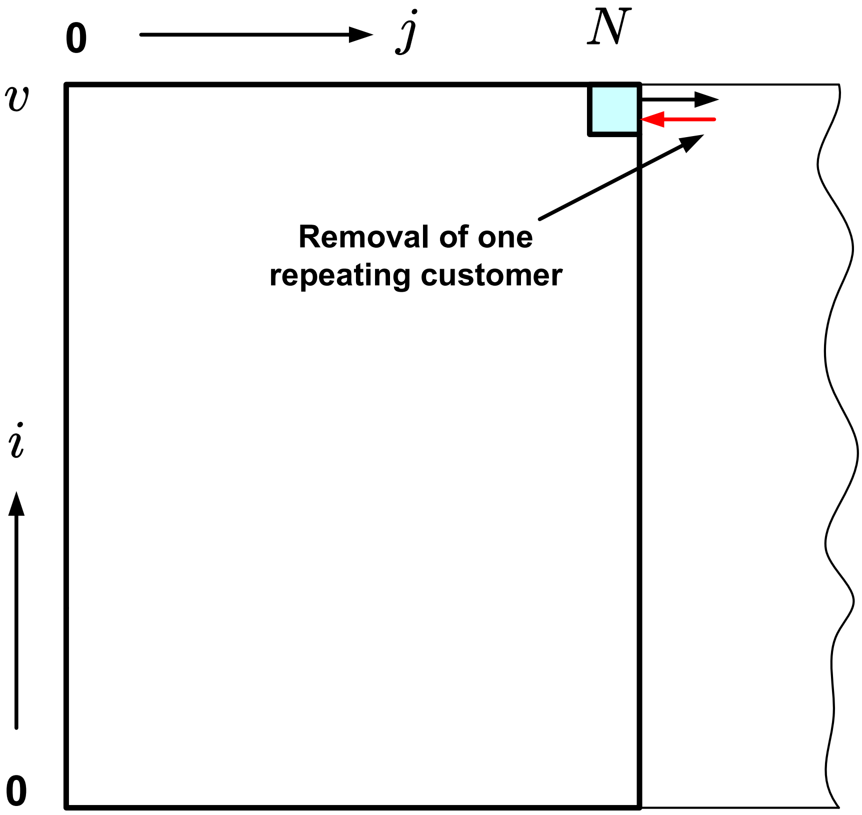


Figure A3. Results of primary truncation of the used state space S.

Let us denote the performance measures of truncated model by the same symbols that used for initial model only with superscript b and find the error caused by truncation. The analog of (5) for truncated model is looking as follows

$$\lambda(1 - P^b(v)) + (J^b - J^b(v))\mu = I^b; \qquad J^b\mu = \lambda P^b(v)H_1 + J^b(v)\mu H_2 - \alpha_b, \tag{A1}$$

where

$$\alpha_b = \lambda P^b(N, v)H_1. \tag{A2}$$

Using the basic property of exponentially distributed variables and ideas used in [20] it can be proved that the following inequalities are true

$$P(v) - P^b(v) \geq 0; \quad I - I^b \geq 0; \quad J(v) - J^b(v) \geq 0; \quad J - J^b \geq 0. \tag{A3}$$

For main performance measures from (A1)–(A3) follow upper estimates of absolute error caused by truncation as function of α_b

$$\begin{aligned} &0 \leq P(v) - P^b(v) \leq \frac{\alpha_b}{\lambda(1 - \max(H_1, H_2))}; \\ &0 \leq I - I^b \leq \alpha_b; \\ &0 \leq J(v) - J^b(v) \leq \frac{\alpha_b}{\mu(1 - H_2)}; \\ &0 \leq J - J^b \leq \frac{\alpha_b}{\mu(1 - \max(H_1, H_2))}. \end{aligned} \tag{A4}$$

For other model characteristics that can be expressed as a function of $P(v)$, I, $J(v)$, I and model's input parameters the estimation of relative error caused by truncation can be obtained with help of (A4). For example, for π_c the following inequality is true

$$\delta(\pi_c) = \left|\frac{\pi_c - \pi_c^b}{\pi_c^b}\right| \le \frac{\alpha_b}{1 - max(H_1, H_2)}\left(\frac{1}{\lambda P^b(v) + J^b(v)\mu} + \frac{1}{\lambda + J^b\mu}\right). \tag{A5}$$

Appendix A.3. Advanced Truncation

At this step we pass from state space R^N to R defined as follows: $(j,i) \in R,\ \ j = 0,1,\dots,N;$ $L(j) \le i \le v$. Here, $L(j)$ is a nondecreasing integer function defined by $0,1,\dots,N$. The state space R with respect to $r^b(t)$ has the set of border states B, defined as $(j,i) \in B$, $j = 0,1,\dots,N;\ \ i = L(j)$. The process $r^b(t)$ moves out of R when one of i occupied servers in state $(j,i) \in B$ completes servicing. We prevent the transition of $r^b(t)$ out of R by causing at this moment the servicing of some additional fictitious call. Let us denote, obtained in this way, auxiliary process by symbol $r^{ba}(t) = (j^{ba}(t), i^{ba}(t))$. Figure A4 illustrates the procedure of construction of $r^{ba}(t)$.

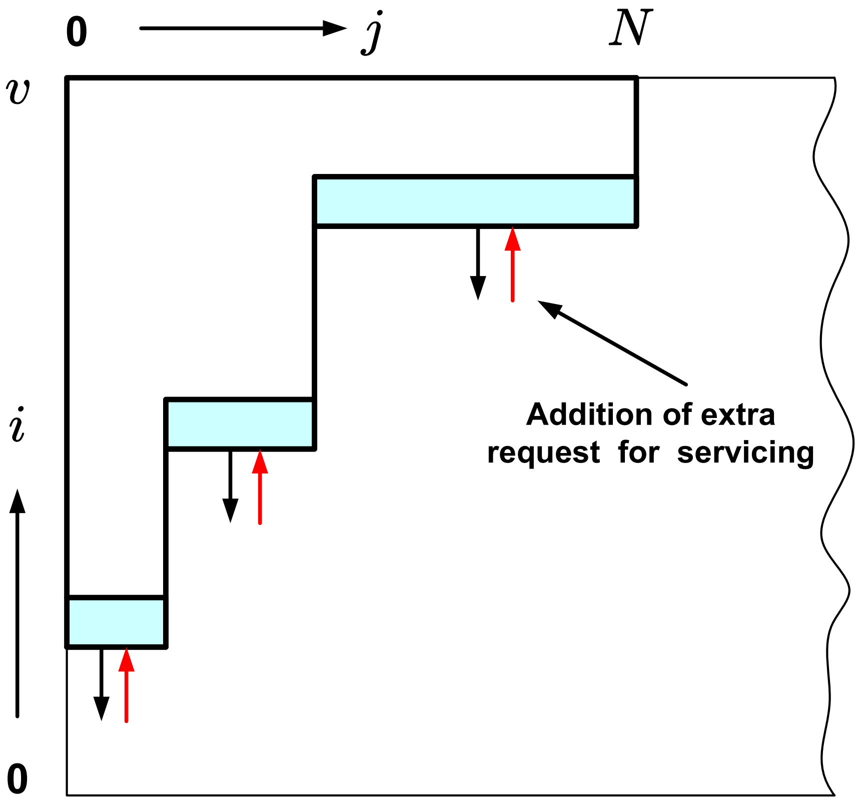


Figure A4. The procedure of construction of $r^{ba}(t)$ on R.

Let us denote performance measures of $r^{ba}(t)$ by the same symbols that are used for $r^b(t)$ only with superscript ba and find the error of estimation of characteristics $r^b(t)$ by corresponding characteristics of $r^{ba}(t)$ caused by truncation. The analog of (A1) for $r^{ba}(t)$ is as follows

$$\begin{gathered}\lambda(1 - P^{ba}(v)) + (J^{ba} - J^{ba}(v))\mu = I^{ba} - \beta_{ba}; \\ J^{ba}\mu = \lambda P^{ba}(v)H_1 + J^{ba}(v)\mu H_2 - \alpha_{ba},\end{gathered} \tag{A6}$$

where

$$\beta^{ba} = \sum_{(j,i)\in B} P^{ba}(j,i)i; \quad \alpha_{ba} = \lambda P^{ba}(N,v)H_1. \tag{A7}$$

Using the basic property of exponentially distributed variables and ideas used in [20] it can be proved that the following inequalities are true:

$$\begin{aligned} &P^{ba}(v) - P^b(v) \geq 0; \quad I^{ba} - I^b \geq 0; \\ &J^{ba}(v) - J^b(v) \geq 0; \quad J^{ba} - J^b \geq 0; \\ &P^{ba}(N,v) - P^b(N,v) \geq 0. \end{aligned} \tag{A8}$$

For main performance measures from (A6)–(A8) follow upper estimates of error caused by truncation as function of β_{ba}

$$\begin{aligned} &0 \leq P^{ba}(v) - P^b(v) \leq \frac{\beta_{ba}}{\lambda(1-\max(H_1,H_2))}; \\ &0 \leq I^{ba} - I^b \leq \beta_{ba}; \\ &0 \leq J^{ba}(v) - J^b(v) \leq \frac{\beta_{ba}}{\mu(1-H_2)}; \\ &0 \leq J^{ba} - J^b \leq \frac{\beta_{ba}}{\mu(1-\max(H_1,H_2))}. \end{aligned} \tag{A9}$$

Combining (A4) and (A9) we obtain upper bounds for absolute error of estimation $P(v), I, J(v), J$ with help of $P^{ba}(v), I^{ba}, J^{ba}(v), J^{ba}$

$$\begin{aligned} &|P^{ba}(v) - P(v)| \leq \frac{\max(\beta_{ba},\alpha_{ba})}{\lambda(1-\max(H_1,H_2))}; \\ &|I^{ba} - I| \leq \max(\beta_{ba},\alpha_{ba}); \\ &|J^{ba}(v) - J(v)| \leq \frac{\max(\beta_{ba},\alpha_{ba})}{\mu(1-H_2)}; \\ &|J^{ba} - J| \leq \frac{\max(\beta_{ba},\alpha_{ba})\max(H_1,H_2)}{\mu(1-\max(H_1,H_2))}. \end{aligned} \tag{A10}$$

For other model characteristics that can be expressed as a function of $P(v)$, I, $J(v)$, I and model input parameters the estimation of relative error caused by truncation can be obtained with help of (A10) in the same way as it was done for π_c by relation (A5).

Appendix A.4. Calculation of Estimates

Both auxiliary processes $r^b(t)$ and $r^{ba}(t)$ are defined in finite state space (see Figures A3 and A4). The stationary probabilities are found after solving the system of state equations. A formal description of algorithm for solving the system of state equations for both auxiliary processes $r^b(t)$ and $r^{ba}(t)$ is quite simple and follows in main steps to algorithm described in [20]. In both cases it can be done recursively. The order of doing recursions can be easily seen from Table A1 where for the case $v = 5$, $N = 2$, $L(0) = 1$, $L(1) = 2$, $L(2) = 4$ is shown as the block structure of nonzero elements (marked by $*$) of the system of state equations for $r^{ba}(t)$. It is easily found that for the ordering of unknown probabilities used in Table A1, the solution of system of state equations split to the solution of $(N+1)$ tridiagonal subsystems of size $(v+1-L(j))(v+1-L(j))$, $j = N, N-1, \ldots, 0$ that are solved in the following order: $j = N, N-1, \ldots, 0$. By doing this we can easily express all probabilities $P^{ba}(j,i)$ through $P^{ba}(N, L(N))$ and afterwards find true values of $P^{ba}(j,i)$ from normalizing condition.

Table A1. Block structure of nonzero elements (marked by $*$) of the system of state equations of $r^{ba}(t)$ for the case $v=5,\ N=2,\ L(0)=1,\ L(1)=2,\ L(2)=4$.

(j,i)	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**1,2**	**1,3**	**1,4**	**1,5**	**2,4**	**2,5**
0,1	*	*									
0,2	*	*	*								
0,3		*	*	*		*					
0,4			*	*	*		*				
0,5				*	*			*	*		
1,2						*	*				
1,3						*	*	*			
1,4							*	*	*	*	
1,5					*			*	*	*	*
2,4										*	*
2,5									*	*	*

Appendix A.5. Borders of Truncated State Space

Formulae (A4) and (A10) allow us to find the error caused using the truncated state space through values of model's input parameters and probabilities of the auxiliary process to be in the border states. These relationships can be also used for solving the reverse problem: determination of the borders of the truncated state space providing a given relative error of performance measures estimation. To realize this idea, it is necessary to find a simple procedure for estimation of stationary probabilities $P(j,i)$ of random process $r(t)$ describing the functioning of basic model with retrials. Let us suggest approach based on simplified equations of states [20]. They are easily derived if we suppose that local conservation laws (3) and (4) that are valid for macrostates (i) and (j) consequently are also valid in each microstate (j,i). Let us denote by $P_{j,i}$ the estimation of $P(j,i)$ obtained in this way. The formulated assumptions give us recurrence for estimation of $P_{j,i}$. They are looking in the following way

$$\begin{aligned} &P_{j,i}(\lambda+j\mu)=P_{j,i+1}(i+1); \quad j=0,1,\dots, \quad i=0,1,\dots,v-1; \\ &P_{j,v}\lambda H_1+P_{j+1,v}(j+1)\mu H_2=\sum_{i=0}^{v}P_{j+1,i}(j+1)\mu; \quad j=0,1,\dots \end{aligned} \tag{A11}$$

The last relation in (A11) can be easily transformed in recurrence

$$P_{j+1,v}=P_{j,v}\frac{\lambda H_1 E(v,\lambda+(j+1)\mu)}{(j+1)\mu(1-H_2E(v,\lambda+(j+1)\mu))}, \quad j=0,1,\dots, \tag{A12}$$

where $E(v,a)$ is an Erlang formula

$$E(v,a)=\frac{\frac{a^v}{v!}}{1+a+\frac{a^2}{2!}+\dots+\frac{a^v}{v!}}.$$

Let us consider a numerical example that illustrates the advantages of suggested approach. Table A2 shows values of J^{ba}, found for proposed relative error ε by advanced truncation of state space (see, Appendix A.3); exact values of the relative error $|J^{ba}-J|/J$; the upper estimates of $|J^{ba}-J|/J$ calculated from inequalities (A10). Model input parameters are as follows: $v=50$; $\lambda=50$; $H_1=0.7$; $H_2=0.9$; $\mu=10$. Proposed relative error varies from 10^{-1} to 10^{-10}. The borders of R are obtained with help of relations (A10)–(A12). The effect of truncation is estimated by the ratio of the number of states in the system of state equations obtained after advanced truncation to the number of states in the state space used in realization of the traditional approach. For definiteness, the last number is the number of states obtained by primary truncation for relative error at the level 10^{-12}.

Table A2. Numerical analysis of estimation the performance measures of basic model with retrials based on the truncation of state space.

ε	J^{ba}	$\frac{\lvert J^{ba}-J\rvert}{J}$	**Estimation** $\frac{\lvert J^{ba}-J\rvert}{J}$	**Effect of Truncation**
10^{-1}	1.4166969791	1.5×10^{-2}	2.4×10^{-2}	0.085
10^{-2}	1.4350225264	1.8×10^{-3}	2.7×10^{-3}	0.113
10^{-3}	1.4371620706	3.0×10^{-4}	4.5×10^{-4}	0.138
10^{-4}	1.4375365971	4.4×10^{-5}	6.2×10^{-5}	0.166
10^{-5}	1.4375920789	5.5×10^{-6}	7.5×10^{-6}	0.196
10^{-6}	1.4375991299	6.0×10^{-7}	8.0×10^{-7}	0.224
10^{-7}	1.4375999046	5.7×10^{-8}	7.4×10^{-8}	0.253
10^{-8}	1.4375999792	4.8×10^{-9}	6.1×10^{-9}	0.282
10^{-9}	1.4375999848	8.6×10^{-10}	1.1×10^{-9}	0.311
10^{-10}	1.4375999860	5.9×10^{-11}	7.4×10^{-11}	0.342

Numerical results presented in Table A2 show the correctness of the suggested approach of choosing the borders of truncated state space providing the given a priory relative error of performance measurement estimation. These properties are very important in the design of different kinds of calculators used as network planning tools. Calculation based on truncation of state space allows the performance of correct mathematical analysis of asymptotic cases when some of the input parameters tend to their limit values [20]. For such cases, traditional calculation approaches meet difficulties due to underflow/overflow problems. It should be noted that ideas based on the concept of truncated state space and illustrated here on the example of the basic model with retrials are very clear, and can be generalized to other types of models with heterogeneous properties in the distribution of stationary probabilities [20].

References

1. The European Emergency Number Association Website. Overload of Calls. Available online: https://eena.org/document/overload-of-calls/ (accessed on 28 September 2021).
2. Stepanov, S.N.; Stepanov, M.S.; Shishkin, M.O. Performance Measures of Emergency Services in Case of Overload. In *Lecture Notes in Computer Science (LNCS)*; Vishnevskiy, V., Samouylov, K., Dmitry, V., Kozyrev, D., Eds.; Springer: Cham, Switzerland, 2020; Volume 12563, pp. 436–449.
3. Koole, G.M. *A Deep Dive into Call Center Workforce Management*; MG Books: Amsterdam, The Netherlands, 2020.
4. Aksin, O.Z.; Armony, M.; Mehrotra, V. The modern call-center: A multidisciplinary perspective on operations management research. *Prod. Oper. Manag.* **2007**, *16*, 665–688. [CrossRef]
5. Gans, N.; Koole, G.M.; Mandelbaum, A. Telephone call centers: Tutorial, review, and research prospects. *Manuf. Serv. Oper. Manag.* **2003**, *5*, 79–141. [CrossRef]
6. Koole, G.; Li, S. A practice-oriented overview of call center workforce planning. *arXiv* **2021**, arXiv:2101.10122.
7. Kim, S.; Whitt, W. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manuf. Serv. Oper. Manag.* **2014**, *16*, 464–480. [CrossRef]
8. Green L.; Kolesar P. The pointwise stationary approximation for queues with nonstationary arrivals. *Manag. Sci.* **1991**, *37*, 84–97. [CrossRef]
9. Mandelbaum, A.; Zeltyn, S. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* **2009**, *57*, 1189–1205. [CrossRef]
10. Koole, G.M.; Jouini, O.; Roubos, A. Performance indicators for call centers with impatience. *IIE Trans.* **2013**, *45*, 341–354.
11. Stolletz, R.; Helber, S. Performance Analysis of an Inbound Call Center with Skills-Based Routing: A Priority Queueing System with Two Classes of Impatient Customers and Heterogeneous Agents. *OR Spectrum.* **2004**, *26*, 331–352. [CrossRef]
12. Stolletz, R. Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models. *Eur. J. Oper. Res.* **2008**, *190*, 478–493. [CrossRef]
13. Ding, S.; Koole, G.; van der Mei, R.D. On the estimation of the true demand in call centers with redials and reconnects. *Eur. J. Oper. Res.* **2015**, *246*, 250–262. [CrossRef]
14. Artalejo, R. Accessible bibliography on retrial queues. *Math. Comput. Model.* **1999**, *30*, 1–6. [CrossRef]
15. Artalejo, J.R. Accessible bibliography on retrial queues: Progress in 2000–2009. *Math. Comput. Model.* **2010**, *51*, 1071–1081. [CrossRef]
16. Falin, G.I.; Templeton, J.G.C. *Retrial Queues*; Chapman and Hall: London, UK, 1997.

17. Artalejo R.; Gomez-Corral, A. *Retrial Queueing Systems: A Computational Approach*; Springer: Berlin, Germany, 2008.
18. Stepanov, S.N.; Stepanov, M.S. Algorithms for Estimating Throughput Characteristics in a Generalized Call Center Model. *Autom. Remote Control.* **2016**, *77*, 1195–1207. [CrossRef]
19. Stepanov, S.N.; Shishkin, M.O.; Stepanov, M.S.; Zhurko, H.M. The construction and analysis of call-center model in overload traffic condition. *T-Comm* **2020**, *14*, 42–50. [CrossRef]
20. Stepanov, S.N. Markov models with retrials: The calculation of stationary performance measures based on the concept of truncation. *Math. Comput. Model.* **1999**, *30*, 207–228. [CrossRef]
21. Kim, C.S.; Klimenok, V.I.; Dudin, A.N. Priority tandem queueing system with retrials and reservation of channels as a model of call center. *Comput. Ind. Eng.* **2016**, *96*, 61–71. [CrossRef]
22. Dudin, S.; Dudina, O. Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. *Appl. Math. Model.* **2019**, *65*, 676–695. [CrossRef]
23. Anisimov, V.; Artalejo, J. Approximation of multiserver retrial queues by means of generalized truncated models. *TOP Off. J. Span. Soc. Stat. Oper. Res.* **2002**, *10*, 51–66. [CrossRef]
24. Stepanov, S.N. Generalized model with retrials in case of extreme load. *Queueing Syst.* **1998**, *27*, 131–151. [CrossRef]
25. Deul, N. Stationary conditions for multi-server queueing systems with repeated calls. *J. Inf. Process. Cybern.* **1980**, *16*, 607–613.
26. Ionin, G.L.; Sedol, I.I. Telephone systems with repeated calls. In Proceedings of the 6th International Teletraffic Congress, Munich, Germany, 9–15 September 1970; pp. 435.1–435.5.
27. Kornyshev, Y.N. A single-line system with repeated orders and preliminary servicing. *Eng. Cybern.* **1977**, *15*, 63–68.