THE CONSTRUCTION AND ANALYSIS OF CALL-CENTER MODEL IN OVERLOAD TRAFFIC CONDITION

DOI: 10.36724/2072-8735-2020-14-7-42-50

Sergey N. Stepanov,
MTUCI, Moscow, Russia, ***stpnvsrg@gmail.com***

Maxim O. Shishkin,
MTUCI, Moscow, Russia, ***mackschischkin1@yandex.ru***

Mikhail S. Stepanov,
MTUCI, Moscow, Russia, ***mihstep@yandex.ru***

Hanna M. Zhurko,
MTUCI, Moscow, Russia, ***hazhurko@gmail.com***

Keywords: *call center, Markov process, system of state equations, performance measures, overload.*

The functional and mathematical models of call center working in case of overload are constructed and analyzed. In the model the following features are considered: the possibility of serving of coming request by IVR (Interactive Voice Response); the option of waiting the beginning of service in case of blocking and the opportunity of request repetition in case of occupation of all waiting positions or unsuccessful finishing of waiting time. Markov process that describes model functioning is defined. Main performance measures of requests coming and serving are given with help of values of stationary probabilities of model's states. The values of performance measures are found after solving the system of state equations by Gauss-Zeidel iterative approach. Expressions that relates the model's main performance measures in form of local and global conservation laws are found. The obtained results can be used for indirect measurement of intensity of primary requests and the probability of call repetition. It is shown how to use the model and the derived results for reducing the negative effects of overload by filtering the input flows of primary and repeated attempts. The usage of the model for calculation of the numbers of operators and waiting places required to serve the incoming traffic flows with given value of probability of call losses and mean value of waiting the beginning of service is considered. Numerical results that illustrate the implementation of the derived expressions and algorithms are given.

Information about authors:
Stepanov Sergey, *professor, doctor of science, MTUCI, head of the chair of communication networks and commutation systems, Moscow, Russia*
Maxim O. Shishkin, *graduate student, MTUCI, the chair of multimedia networks and communication services, Moscow, Russia*
Mikhail S. Stepanov, *docent, Cand. Tech. Sciences, MTUCI, the chair of communication networks and commutation systems, Moscow, Russia*
Hanna M. Zhurko, *PhD student, MTUCI, the chair of communication networks and commutation systems, Moscow, Russia*

Для цитирования:
Степанов С.Н., Шишкин М., Степанов М.С., Журко А. Разработка и анализ модели call-центра в условиях перегрузки // T-Comm: Телекоммуникации и транспорт. 2020. Том 14. №7. С. 42-50.

For citation:
Stepanov S.N., Shishkin M.O., Stepanov M.S., Zhurko H.M. (2020) The construction and analysis of call-center model in overload traffic condition. *T-Comm*, vol. 14, no.7, pp. 42-50. (*in Russian*)

1. Introduction

An overload of calls can be a critical issue for some call centers, especially for emergency call centers. As long as in case of large emergencies such as earthquakes, terrorist attacks, multi-vehicle traffic accidents, etc. the risk of a greatly increased call volume can be multiplied and the risk of overload in call centers can also be multiplied.

All emergency services can be affected by an overload of calls and it could be a serious issue when some people in emergency situations cannot reach public-safety answering points (PSAPs) or emergency call centers due to an overload. People who are in life-threatening situations and who need urgent help should gain help or at least be informed about other possibilities of being helped. That is why it is important to prevent the problem of call overload in call centers especially in the emergency services. European emergency services deal with call overload using different approaches. As it will be described later, the goal of dealing with call overloads can be achieved by avoiding bottlenecks in call center's access and implementing measures to minimize them. There are three main bottlenecks will be discussed: network access, call center access, agent access.

In this paper, particularly, the third issue – agent access – will be described and solutions of the problem will be proposed. These solutions are based on results of a mathematical model analysis, which can be used for measurement of primary requests intensity and the probability of call repetition. As a solution to minimize negative impact on call center access and agent access with a help of the model and obtained results filtering of the input flows of primary and repeated attempts can be considered. The usage of the model for calculation of the numbers of operators and waiting places is highly important for creating effective staff planning algorithms and workforce optimization during peak load.

In teletraffic papers and articles theoretical analysis has been done for some of the mentioned above problems [1-14], including the studying of retrials [5-8], overloads in call centers [4, 13-14], the study of possibility of waiting and the investigation of dependence of the request servicing on the operators' skill levels [1-3]. In this paper all mentioned factors are considered together.

As it was mentioned before, three main bottlenecks in call center's access and implementing measures to minimize them will be described in Section 2. Also, in Section 2, the components of the functional model of call center, working in the case of overload, will be introduced, and some approaches to minimize bottlenecks in call centers will be described. Next Section will be devoted to the mathematical description of the process of coming and serving requests.

In Section 4 a Markov process that describes the model functioning will be given together with formal definitions of main performance measures through values of model's stationary probabilities. In Section 5 the system of state equations that relates model's stationary probabilities is introduced. This section also contains the conservation laws that relates the model's main performance measures. In Section 6 the system of state equations is rewritten in the form that is suited for application of Gauss-Zeidel iterative algorithm. The main steps of algorithm are described in Section 6. Numerical results that illustrate the implementation of the derived expressions and algorithms for elimination of negative effects of overload are given in Section 7. Conclusions are given in the last section.

2. Functional Model

As it was mentioned before, the most sensitive to overloads are emergency service call centers, where every customers' call or interaction could be life or death question for someone. That is why it is crucial to determine main bottlenecks which can occur from the time the emergency number was dialed to the time when agent picks up a phone; and finding the ways to minimize these bottlenecks or impact of them to call center are of most importance. The functional model of call center is shown on Fig.1. Proposed model is relevant not only for emergency call centers but for all call centers in general and all bottlenecks might take place not only for emergency call centers as well.

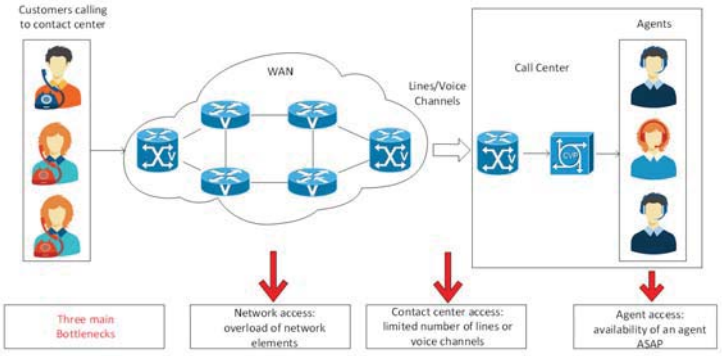

Fig. 1. Functional model of a call center with overloads

There are three main bottlenecks (until the call will be served by an agent) and measures to minimize it can be described:

Network access

In most cases people contacting emergency call centers by phone. Thus, it is a reason why the availability of fixed and mobile telephony networks plays the important role. Emergency calls are routed via one or several public telephone networks, which all are interconnected into Wide Area Network (WAN). An overload can occur on all network elements (routers, switches and interconnection interfaces) and it can lead to blocking of emergency calls. This blocking can not necessarily be caused by high amount of placed calls, but other reasons. And if happens simultaneously with some emergency situation it might cause network overload and unavailability of reaching call center.

As a solution telecommunication operators and public authorities have to prioritize efforts to maintain the public's trouble proof access to emergency services. Both mobile and fixed telephone operators have to prioritize emergency calls in their network and in all interconnections to other operators (e.g. in special trunks). Also, in the case of the phone network congestion, or if the responsible emergency call center is overloaded, emergency calls may be rerouted to another emergency call center or non-emergency call centers [4].

Accessing call center

Public safety answering points or emergency call centers are typically connected to public switched telephone networks (PSTN) by subscriber lines (analog or IP). That is why a number of voice channels is always limited. In case of high calls load these subscriber lines might be busy. As a result, the number of lines or voice channels always should be significantly greater than the number of agents in emergency service call center.

Information about the data and trends in call centers should be captured to monitor situations where call center capacity is compromised which in the future could lead to decisions regarding the correct number of call centers' lines or staffing policy changes.

Inadequate public-warning systems can be a reason for increasing number of calls to emergency call centers which can most likely create an overload situation to happen. In that case additional information regarding an incident should be provided to inform the citizens or customers with the correct steps to do and not to call the dedicated emergency number unless it is absolutely necessary.

In the cases of malfunction (Denial of Service attack to an emergency number) or hoax calls for repeat offenders, those numbers should be detected and 'gray listed' to prevent the system from call avalanches. In most cases it is not intended to block or deny service; however, these calls may be assigned a lower priority than others [4].

Accessing an agent

If a customer can technically reach the call center, emergency services have to ensure that the citizen reaches an available agent as soon as possible. In some situations, when it is possible to predict some catastrophes (forest fires or tsunamis), extra staff or pre-arranged extra shifts can be assigned to the period of time, when it is necessary. On the other hand, there are a lot of unpredictable accidents, we are unable to know about in advance (earthquakes, terror attacks, etc.). These situations appear without any warning but emergency service call centers should be prepared to cope with call avalanches. Previous accidents or major events can be helpful to analyze best strategies that were done before; and based on past statistics some optimization in staffing planning could be performed for the effective handling of future events. As long as workforce planning based on average daily call volumes and activity levels will lead to unpreparedness of an emergency call center for any large incident scenarios and "worst case" events or call avalanches. Workforce have to be optimized to provide a minimum waiting time in all circumstances. In the case of unforeseen events additional numbers of lines or voice channels and number of operators provisioning should be adequate to avoid the overload of emergency call centers. If incoming calls are lost because of busy or engaged lines, it might be very difficult for an emergency call centers to count the number of lost calls and predict future staffing needs.

Repeated calls or calls from the same areas regarding the same incidents can also be a cause of call center overloading. Automatic IVR messages can be used as an announcement to inform customers about already known incidents in order to make lines free as soon as possible. These announcements should be dynamically configured and easily implemented in IVRs.

Non-life/non-property-threatening type of issues can also be a problem for emergency services in case of overload. Therefore, it is possible to be considered the provision of an additional non-emergency number for customers with non-emergency problems. These semi-urgent calls could be handled separately and would not highly impact emergency service call centers. As before, a welcome IVR message may be useful to support this initiative.

The caller geo-location detection also can be used to decrease high call volumes regarding the same situation. When it is possible to determine whether the caller belongs to an area where something has happened then system can decide forward this call to an agent with lower priority or play the announcement about already known incidents; in the case of customer calling from areas where no incidents were detected the probability of some new issue will be reported is extremely high. It is important to differentiate call handling process according to caller location and information about already known incidents to ensure that the emergency service is still available in areas of the country which are not affected by the incident. Callers from the disaster area can not only hear an IVR message about already known situation but be handled by a particular team assigned for this area, providing people with guidelines or additional phone numbers [4].

In the emergency call centers a mechanism of involvement of additional staff members could be a great solution for short periods of time especially during the pick load. An internal procedure which describes the steps to be done in the situation of call avalanche should be prepared, tested and available for emergency call center agents. Emergency call center staff should be also equipped with a remote call handling infrastructure, so where they can quickly go online and start working.

In this article mainly the third bottleneck of agent accessing in call center will be described, as a task which can be optimized by construction of an efficient mathematical model; and some of ideas to minimize overloads or call avalanches will be solved analytically with a help of proposed model

Mathematical Model

Customers' requests for getting service are entering the call center facilities through telephone access lines. After occupying an access line, a customer can be served by IVR (Interactive Voice Response) and if needed by an operator. Let us denote by υ the total number of available operators. It is supposed that number of access lines equals to $\upsilon+\omega$, where ω lines are used for waiting the beginning of service and υ lines are used in the process of call servicing by operator. The flow of primary requests for servicing is described by Poisson model with intensity λ.

The call center functioning is considered in the conditions of overload. From this assumption follows that besides of primary requests, the call center serves repeated requests. In the model we study two reasons of call repetition. The first reason is related to insufficient amount of available operators and access lines where customer can wait the beginning of servicing.

If coming request finds that all υ operators and ω access lines are occupied then customer can repeat the call. Another reason of repetition is related with procedure of waiting. It is supposed that maximum allowed time of waiting is restricted and if this time is finishing without success then customer can repeat the request. In both considered cases, a customer with some probability H repeats the request for servicing after random exponentially distributed time having parameter ν and with additional probability $1-H$ the blocked customer stops his attempts to seize a operator and leaves the system. In the model it is supposed that maximum allowed time of waiting the starting of operator service is exponentially distributed with parameter σ.

The process of servicing of coming request may have two phases. The first one is getting a recorded message from the IVR and second – receiving the information from an operator. In the constructed model we suppose that durations of operator's service have exponential distribution with parameter μ.

The transition to operator' service is described by two probabilities depending on the type of the request: primary or repeated. With probability q_p for primary request and with probability q_r for repeated attempt after getting service from IVR a customer is trying to get servicing from operator. With additional probability $1-q_p$ for primary request and with additional probability of $1-q_r$ for repeated attempt a customer leaves the system receiving the servicing at IVR.

In the constructed model the process of call servicing at IVR may have two interpretations. In the first scenario the IVR-message contains the service information asked by customer. In this case IVR can be considered as robotized operator that give the administration of call center a cheap way to decrease the required number of costly operators in case of overload. In the second scenario the IVR-message can keep the directives to subscriber concerning his behavior in case of overload. In particular it can contain the urgent advice do not repeat the call in case of overload. In doing this we can differ between primary and repeated attempts. This approach also allows the administration of call center to filter the input flow and as result to decrease the required number of operators in case of overload.

Model's main parameters and the process of coming and serving requests are shown on Fig. 2.

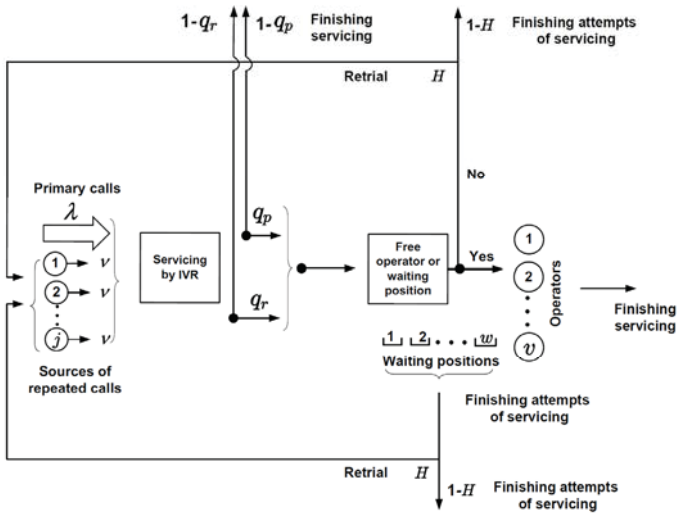

Fig. 2. The structure of call center mathematical model

3. Markov process and main performance measures

Let us denote the state of the model by vector (j,i) where j is the number of customers repeating a request for servicing and i is the number of occupied operators and waiting positions. The values of j,i varies as follows:

$$j=0,1,...,;\ i=0,1,...,\upsilon+\omega; \tag{1}$$

Let us denote by $j(t)$ the number of customers repeating attempt at time t and by $i(t)$ we denote the number of operators and waiting positions occupied at time t. The changing of model states is described by Markov process of the type $r(t)=(j(t),i(t))$ defined on the infinite space of states S given by (1). The Markov character of the constructed process $r(t)$ follows from the fact that durations of model staying at each state (1) have exponential distribution and are independent from each other just like transition probabilities from state to state. The model functioning is considered in stationary mode. This regime always exists if the following inequalities are satisfied $H<1$ and $\nu>0$. In the considered case a customer leaves the system with probability tending to one after some time having bounded average.

Let us denote by $p(j,i)$ the probability of stationary state $(j,i)\in S$ of the considered model. Due to the ergodicity property of the $r(t)$ the values of $p(j,i)$ can be interpreted as the portion of time the model stands in the state (j,i). This interpretation allows us to introduce the definitions for main performance measures that characterize the process of requests coming and serving in the analyzed model of call center.

We start from definitions of mean values of components of the model state. Let us denote by M_r, M_i, M_ω mean numbers of, respectively, customers repeating a request, occupied operators, occupied waiting positions for operators. Significant role in analyzing the model functioning are playing the intensities of coming and blocking requests. We denote by I_b, I_o and I_t correspondingly the intensity of requests lost in attempt to get service from operators, intensity of primary and repeated requests coming to get service from operators and intensity of primary and repeated requests coming to get service from call center. Key performance measures of call center are defined by ratios of lost requests. Let us denote by π_0 the ratio of lost primary and repeated requests coming to get service from operators, by v the ratio of lost primary and repeated requests coming to get service from call center, by t_ω we denote the mean time for request to be on waiting or servicing, by M we denote the mean number of retrials per one primary attempt, by τ we denote the portion of repeated calls in the total flow coming requests.

Introduced characteristics can be easily defined by means of stationary probabilities of model states

$$\begin{aligned}
M_r &= \sum_{\{(j,i)\in S|j>0\}} p(j,i)\,j;\\
M_i &= \sum_{\{(j,i)\in S|0<i<\upsilon\}} p(j,i)i + \sum_{\{(j,i)\in S|i>\upsilon\}} p(j,i)\upsilon;\\
M_\omega &= \sum_{\{(j,i)\in S|i>\upsilon\}} p(j,i)(i-\upsilon);\\
I_b &= \sum_{\{(j,i)\in S|i=\upsilon+\omega\}} p(j,i)(\lambda q_p + j\nu q_r);\\
I_o &= \lambda q_p + \sum_{\{(j,i)\in S|j>0\}} p(j,i)\,j\nu q_r;\\
I_t &= \lambda + \sum_{\{(j,i)\in S|j>0\}} p(j,i)\,j\nu;\\
\pi_o &= \frac{I_b + M_\omega\sigma}{I_o};\quad \pi_c = \frac{I_b + M_\omega\sigma}{I_t};\quad t_\omega = \frac{M_i + M_\omega}{I_o - I_b};
\end{aligned} \tag{2}$$

$$M = \frac{M_r \nu}{\lambda}; \quad \tau = \frac{M_r \nu}{I_t}.$$

4. System of state equations

The model's performance measures that are introduced in the previous chapter are expressed through values of $p(j,i)$. Unnormalized values of $p(j,i)$ we denote by $P(j,i)$.

The values of $P(j,i)$ can be found from the solution of the system of state equations. To obtain this system it is necessary to equate the intensity of leaving the arbitrary model state $(j,i) \in S$ to the intensity of entering the state (j,i). By using the indicator function we write the system of state equations in one relation. This form of representing the system of state equations will be very convenient for solving it by Gauss-Zeidel iterative algorithm. We show it later. The system of state equations is looking as follows:

$$\begin{aligned}
&P(j,i)\{j\nu(1-q_r)+(\lambda q_p + j\nu q_r)I(i<\upsilon+\omega)+ \quad (3)\\
&+(\lambda q_p H + j\nu q_r(1-H))I(i=\upsilon+\omega)+\\
&+i\mu I(i\le\upsilon)+\upsilon\mu I(i>\upsilon)+(i-\upsilon)\sigma I(i>\upsilon)\}=\\
&= P(j,i-1)\lambda q_p I(i>0)+\\
&+P(j+1,i-1)(j+1)\nu q_r I(i>0)+\\
&+P(j-1,i)\lambda q_p H I(j>0,i=\upsilon+\omega)+\\
&+P(j+1,i)(j+1)\nu q_r(1-H)I(i=\upsilon+\omega)+\\
&+P(j+1,i)(j+1)\nu(1-q_r)+P(j,i+1)(i+1)\mu I(i+1\le\upsilon)+\\
&+P(j,i+1)\upsilon\mu I(\upsilon<i+1\le\upsilon+\omega)+\\
&+P(j-1,i+1)(i+1-\upsilon)\sigma H I(j>0,\upsilon<i+1\le\upsilon+\omega)+\\
&+P(j,i+1)(i+1-\upsilon)\sigma(1-H)I(\upsilon<i+1\le\upsilon+\omega).
\end{aligned}$$

By $I(\cdot)$ the indicator function is defined;

$I(\cdot)=1$, if condition formulated in brackets is fulfilled;

$I(\cdot)=0$, if condition formulated in brackets is not fulfilled.

Values of $P(j,i)$ should be normalized.

The system of state equations can be used to derive the conservation laws that relates the model's main performance measures. Let us sum up (3) over i from 0 to $\upsilon+\omega$ with j fixed. The summation gives the following relations

$$\begin{aligned}
&p(j,\upsilon+\omega)\lambda q_r H + \sum_{i=\upsilon+1}^{\upsilon+\omega} p(j,i)(i-\upsilon)\sigma H =\\
&= p(j+1,\upsilon+\omega)(j+1)\nu(1-q_r H)+\\
&+\sum_{i=0}^{\upsilon+\omega-1} p(j+1,i)(j+1)\nu, \quad j=0,1,... \qquad (4)
\end{aligned}$$

Now let us sum up (3) over j from 0 to ∞ with i fixed. The summation gives the following relations

$$\begin{aligned}
&\sum_{j=0}^{\infty} p(j,i)(\lambda q_p + j\nu q_r) =\\
&= \sum_{j=0}^{\infty} p(j,i+1)((i+1)\mu I(i+1\le\upsilon)+\\
&+(\upsilon\mu+(i+1-\upsilon)\sigma)I(i+1>\upsilon)), \qquad (5)
\end{aligned}$$

$i = 0,1,...,\upsilon+\omega-1.$

Now let us sum up (4) over j from 0 to ∞ and (5) over i from 0 to $\upsilon+\omega$. The summation gives the following two relations

$$M_r \nu = (I_b + M_\omega \sigma)H; \qquad (6)$$

$$I_t = \lambda(1-q_p) + M_r \nu(1-q_r) + I_b + M_\omega \qquad i$$

The relation (7) can rewritten in the form

$$I_o = \lambda q_p + M_r \nu q_r = I_b + M_\omega \sigma + M_i \mu. \qquad (8)$$

Relations (6) – (8) can be also proved with help of Little's formula. Let us prove (6). From the model description follows that the mean time for a customer to repeat the request equals $\frac{1}{\nu}$. The usage of the Little formula gives that $\frac{1}{\nu}$ equals to the mean number of customers M_r that are in the state of request repetition divided to the total intensity of requests that are going to repeat an attempt $(I_b + M_\omega \sigma)H$. As result we have the relation

$$\frac{1}{\nu} = \frac{M_r}{(I_b + M_\omega \sigma)H}$$

that is equivalent to the (6). Now let us prove (7). The mean time of request servicing at operator equals to $\frac{1}{\mu}$. From the Little formula follows that this time equals to the ratio of the mean number of requests being on servicing M_i to the intensity of requests accepted by operator

$$(I_t - \lambda(1-q_p) - M_r \nu(1-q_r) - I_b - M_\omega \sigma).$$

As result we have the relation

$$\frac{1}{\mu} = \frac{M_i}{I_t - \lambda(1-q_p) - M_r \nu(1-q_r) - I_b - M_\omega \sigma}$$

that is equivalent to the (7).

The conservation laws (6), (7) have the following interpretation. The relation (6) means that intensity of retrials (left part of (6)) equals to the intensity of events that caused the repeated attempt (right part of (6)). The relation (7) means that the total intensity of requests coming to call center (left part of (7) equals to the sum of intensities of requests that are leaving call center after servicing at IVR, blocked because of full occupancy of operators and waiting positions, leaving waiting positions without servicing and leaving the call center after servicing at operator (right part of (6).

The relations (6) – (8) can be used for indirect measurements of performance measures that are hard to estimate by direct methods because of difficulties in separation of primary and repeated requests. Let us demonstrate it. Let us sum up (6) with λ. Using definitions of I_t and π_c we obtain the relation

$$\lambda = I_t(1-\pi_c H) \qquad (9)$$

If we know the value of H and results of measurement of I_t and π_c (these characteristics can be estimated without separating primary and repeated requests) then (9) gives the value of intensity of primary requests.

In opposite, if know the value of λ then we can estimate the value of H. From (9), definitions of I_t and π_c follows

$$H = (I_t - \lambda)\frac{1}{\pi_c I_t} = (1 - \frac{\lambda}{I_t})\frac{1}{\pi_c} \qquad (10)$$

Let us find alternative formulae for estimation of M and τ. From definition of π_c and (6) follows

$$M_r \nu = \pi_c I_t H .$$

Using this relation, (9) and definition of M we obtain

$$M = \frac{\pi_c H}{1 - \pi_c H}$$

In a similar way we obtain the relation for estimation of τ

$$\tau = \frac{M_r \nu}{I_t} = \frac{M_r \nu}{\lambda + M_r \nu} = \frac{1}{\frac{1}{M} + 1} = \pi_c H .$$

5. Solution of the system of state equations

The model's performance measures are expressed through values of $p(j,i)$, that are related by system of state equations (3). In order to solve (3) by standard algorithms of linear algebra it is necessarily to restrict the number of unknowns in (3). Let us change the model functioning by limiting the number of customers repeating the attempt by some integer j_m. The value of j_m can be found through numerical experiments (see details in [4,10]). When calculating the performance measures introduced in Section 4 we suppose that the value of j varies in interval $j = 0,1,..., j_m$. In this case model's space of states S will be finite $(j,i) \in \mathrm{S},\ j = 0,1,..., j_m; i = 0,1,..., \upsilon + \omega.$. System of state equation can be easily rewritten for the case $j_m < \infty$ and looks in the following way

$$\begin{aligned}
&\mathrm{P}(j,i)\{j\nu(1-q_r) + (\lambda q_p + j\nu q_r) I(i < \upsilon + \omega) + \qquad (11)\\
&+(\lambda q_p H I(j < j_m) + j\nu q_r (1-H)) I(i = \upsilon + \omega) +\\
&+i\mu I(i \le \upsilon) + \upsilon\mu I(i > \upsilon) + (i-\upsilon)\sigma I(i > \upsilon)\} =\\
&= P(j,i-1)\lambda q_p I(i > 0) + \mathrm{P}(j+1,i-1)(j+1)\nu q_r I(j+1 \le j_m, i > 0) +\\
&+P(j-1,i)\lambda q_p H I(j > 0, i = \upsilon + \omega) +\\
&+P(j+1,i)(j+1)\nu q_r (1-H) I(j+1 \le j_m, i = \upsilon + \omega) +\\
&+P(j+1,i)(j+1)\nu(1-q_r)\,\mathrm{I}(j+1 \le j_m) + P(j,i+1)(i+1)\mu\,\mathrm{I}(i+1 \le \upsilon) +\\
&+P(j,i+1)\upsilon\nu\,\mathrm{I}(\upsilon < i+1 \le \upsilon + \omega) +\\
&+P(j,i+1)(i+1-\upsilon)\sigma(1-H) I(j < j_m, \upsilon < i+1 \le \upsilon + \omega) +\\
&+P(j,i+1)(i+1-\upsilon)\sigma I(j = j_m, \upsilon < i+1 \le \upsilon + \omega) +\\
&+P(j-1,i+1)(i+1-\upsilon)\sigma H I(j > 0, \upsilon < i+1 \le \upsilon + \omega),\\
&j = 0,1,..., j_m; i = 0,1,..., \upsilon + \omega.
\end{aligned}$$

The values $P(j,i)$ should be normalized

$$\sum_{j=0}^{j_m} \sum_{i=0}^{\upsilon+\omega} p(j,i) = 1.$$

Almost all elements of the matrix of the (11) are zeros. In this case the effective approach to solve (11) consist in using Gauss-Seidel iterative algorithm [4, 10]. It is known that convergence of Gauss-Zeidel algorithm for singular system (11) is not guaranteed. In order to achieve the convergence, it is sufficient to put one of unknown probabilities $P(j,i)$ in (11) to one or other positive value, remove from (11) equation corresponding to the chosen probability and after apply Gauss-Seidel algorithm to obtained nonsingular system of linear equations. For transformed in such a way system the convergence of Gauss-Seidel algorithm is guaranteed because the matrix of the constructed system of linear equations belongs to the class of irreducibly diagonally dominant matrices [4,10]. At this point it is necessary to say that convergence of Gauss-Seidel algorithm for singular system (11) is much faster than for considered above nonsingular case.

The checking of the convergence can be done with help of conservation laws (6) – (7). The analog of (6) – (7) for truncated model is looking as follows

$$M_r \nu = (I_b + M_\omega \sigma) H; \qquad (12)$$

$$I_t = \lambda(1-q_p) + M_r \nu (1-q_r) + I_b + M_\omega \sigma + M_i \mu - \delta, \qquad (13)$$

where δ defined as

$$\delta = p(j_m, \upsilon + \omega)\lambda H q_p + \sum_{i=\upsilon+1}^{\upsilon+\omega} p(j_m, i)(i-\upsilon)\sigma H.$$

Relations (12) – (13) can be also used for choosing the truncation level j_m. It can be proved [4, 10] that relative error of performance measures estimation caused by truncation can be estimated by value of δ.

6. Implementation of the model

6.1. The filtering of retrials

The negative consequences of overload caused by random increasing of the intensity of coming requests for servicing can be removed by filtering the input flow. We can do it by using IVR. We suppose that the IVR-message has directives to subscriber concerning his behavior in case of overload. In particular it can contain the urgent advice do not try to get service at call center for primary calls and do not repeat the call in case of overload. In doing this we can differ between primary and repeated attempts. Let us consider for the beginning the procedure of filtering only retrials.

We illustrate the procedure of filtering by numerical example. Model's input parameters are as follows: $\lambda = 24$; $\upsilon = 10$; $\omega = 5$; $q_p = 0,5$; H =0,9; ν =5; j_m =50; μ =1; σ =0,1. As a time, unit was chosen the mean time of servicing a request by operator. The probability of filtering of retrials is $1 - q_r$ and varies from 0 to 0,95. The required level of service should satisfy the inequality $\pi_c < 0,1$.

The Figure 3 shows the dependence of π_c on $1-q_r$.

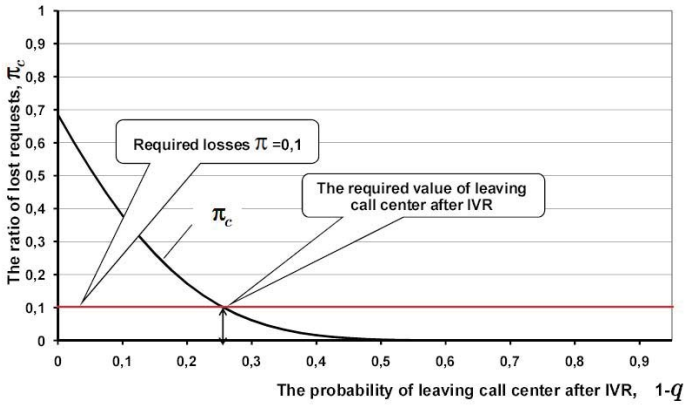

Fig. 3. The dependence of π_c on the coefficient of $1-q_r$ of retrials filtering

The results of calculations show that by filtering the retrials we can decrease the value of losses but we can do it up to certain level determined by intensity of primary calls λ. If value of λ is big it is necessary to filter both primary and repeated attempts.

6.2. The usage of ω for decreasing the required number of υ

The possibility of waiting the beginning of servicing in case of call blocking can decrease the number of operators required to serve input flow of calls with given value of call blocking. It is clear that time of waiting is restricted by subscriber impatience. In the model the mean time of being in the system on waiting or servicing is estimated by value of t_ω and this characteristic should be used as additional measure of sufficiency of the found number of operators.

Let us suggest the procedure of estimation the number of operators υ and waiting places ω required to serve coming requests with given values of call losses $\pi_c = \pi$ and the mean time of a request being in the system on waiting or servicing $t_\omega = W$. At the first step of the approach we put the number of waiting position ω to zero fix all other model's parameters and find the number of operators υ_0 satisfying the inequality $\pi_c < \pi_0$. On the second step we simultaneously decrease the value of υ_0 by one and increase the number of waiting positions by one. Next, we calculate the current value of π_c and check the inequality $\pi_c < \pi$. If $\pi_c < \pi$ then we decrease the current value of υ and increase the current value of ω in opposite case we fix the current value of υ and increase the current value of ω. After doing this we pass to calculation of π_c and t_ω and checking inequalities $\pi_c < \pi$ and $t_\omega < W$.

At the third and other steps we consequently decrease the current value of υ and increase the current value of ω in the same way as it was done at second step. The process of estimation of υ and ω stops when one or both inequalities $\pi_c < \pi$, $t_\omega < W$ are violated.

The found values of υ and ω will be the solution of formulated problem. Let us consider a numerical example that illustrates the solution.

Model's fixed input parameters are as follows:

λ =60; q_p =0,5; q_r =0,9; H =0,9; ν =5; j_m =50; μ =1; σ =0,05. The required number of operators and waiting positions should satisfy the inequality $\pi_0 < 0,03$ and $t_\omega < 1,5$. As a time unit was chosen the mean time of servicing a request by operator. The initial value of operators according to suggested algorithm is $\upsilon_0 = 38$. For this choice $\pi_c = 0,0297$ and $t_\omega = 0$.

The Figure 4 shows the dependence of π_c both on υ and ω and Figure 5 shows the dependence of t_ω both on υ and ω in the process of realizing the formulated above algorithm of estimation υ and ω. The answer is: $\upsilon = 31; \omega = 32$. The value of control performance measures: $\pi_c = 0,0299$; $t_\omega = 1,3136$.

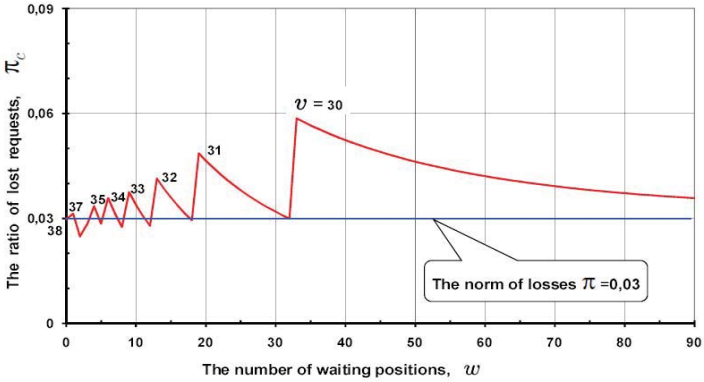

Fig. 4. The dependence of π_c on υ and ω in the process of realizing the formulated algorithm of estimation υ and ω

It is clear that the process of simultaneous decreasing of υ and increasing of ω for achieving the restriction on the values of π_c and t_ω has only few steps. This is clearly seen on Figures 4,5. For large values of ω the rate of decreasing π_c is stabilized and π_c tends to some limit as $\omega \to \infty$ depending on the value of σ. This limit can be more then π.

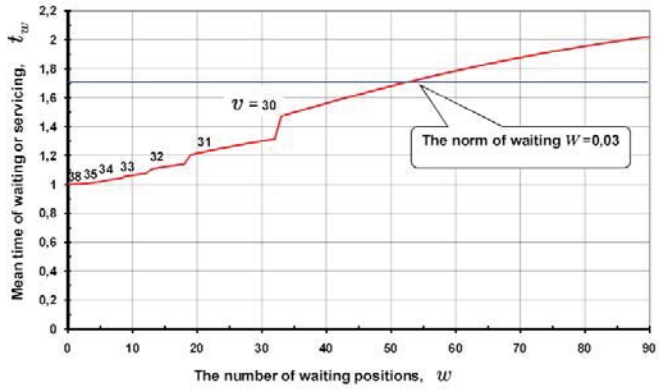

Fig. 5. The dependence of t_ω on υ and ω in the process of realizing the formulated algorithm of estimation υ and ω

Conclusion

In the paper the functional and mathematical models of call center working in case of overload are constructed and analyzed. In the model the following features are considered: the possibility of serving of coming request by IVR (Interactive Voice Response); the option of waiting the beginning of service in case of blocking and the opportunity of request repetition in case of occupation of all waiting places or unsuccessful finishing of waiting time. Primary and repeated requests for servicing are coming after exponentially distributed time intervals. It is supposed that service times of operators and maximum allowed time of waiting the beginning of service also have exponential distribution with corresponding parameters. The components of

Markov process that describes model functioning are defined. Main performance measures of requests coming and serving are given with help of values of stationary probabilities of model's states. The values of performance measures are found after solving the system of state equations by Gauss-Zeidel iterative approach. Expressions that relates the model's main performance measures in form of local and global conservation laws are found. The obtained results can be used for indirect measurement of intensity of primary requests, the probability of call repetition and other performance measures that difficult to measure because of necessity to separate primary and repeated attempts. It is shown how to use the model and the derived results for reducing the negative effects of overload by filtering the input flows of primary and repeated attempts. The usage of the model for calculation of the numbers of operators and waiting places required to serve the incoming traffic flows with given value of probability of call losses and mean value of waiting the beginning of service is considered. Numerical results that illustrate the implementation of the derived expressions and algorithms are given.

The model and derived algorithms of performance measures estimation can be used to produce the quantitative and qualitative analysis of the dependence of model's performance measures on the values of input parameters with considering the customer behavior in case of overload. Proposed model can be further developed to include the possibility of non-exponential distribution of time interval between retrials, service times and maximum allowed waiting times.

References

1. Gans N., ole M., Mandelbaum A. (2003) Telephone Call-Centers: Tutorial, Review and Research Prospects. *Manuf. Service Manage*, no. 5, pp. 79-141.

2. Stolletz R., Helber S. (2004) Perfomance Analysis of an Inbound Call-Center with Skills-Based Routing. *Springer-Vellag, Hannover*.

3. Mandelba A., Zeltyn S. (2009) Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Operations Research. Vol. 7, no 5, pp. 1189-1205.

4. Ruiz Mart ez D., Mclaren P., Norman J., Ronkko M. (2012) Overload of calls. EENA Operations Document.

5. Stepanov S.N. (1983) Chisle ye metody rascheta sistem s povtornymi vyzovami (Numerical Methods for Analysis of Systems with Repeated Calls). *Nauka*, Moscow. (in Russian)

6. Stepanov S.N., Stepan , M.S. (2014) Construction and Analysis of a Generalized Contact Center Model. *Automation and Remote Control*. Vol 75, no 11, pp. 1936-1947.

7. Stepanov S ., Stepanov M.S. (2016) Algorithms for Estimating Throughput Characteristics in a Generalized Call Center Model. *Automation and Remote Control*. Vol. 77, no 7, pp. 1195-1207.

8. Aguir S., K aesmen F., Aksin O.Z., Chauvet F. (2004) The impact of retrials on call center performance. *OR Spectrum*. Vol. 26, no 3, pp. 353-376.

9. Stepanov S.N. (2010) Osn y teletraffika multiservisnykh setei (Fundamentals of Multiservice Networks). *Eqo-Trends*, Moscow. (in Russian)

10. S anov S.N. Teoriya teletraffika: kontseptsii, modeli, prilozheniya (Theory of Teletraffic: Concepts, Models, Applications). (2015) *Goryachaya Liniya-Telekom*, Moscow. (in Russian)

11. Stepanov S.N. (1999) Markov Models with Retrials: Th alculation of Stationary Performance Measures Based on the Concept of Truncation. *Mathematical and Computer Modelling*. Vol. 30, pp. 207-228.

12. Stepa S.N. (1998) Generalized model with retrials in case of extreme load. *Queueing Systems*. Vol. 27, pp. 131-151.

13. S anov S.N., Stepanov M.S., Zhurko H. The Modeling of Call Center Functioning in Case of Overload. In: Vishnevskiy V., Samouylov K. (eds) DCCN 2019. *Lecture Notes in Computer Science (LNCS). Springer, Cham*. Vol 11965, pp. 391-406.

14. S anov S.N., Shishkin M.O., Sosnovikov G.K, Stepanov M.S., Vorobeychikov L.A., Zhurko H.M. (2019) The Analysis of Call Center Model in Case of Overload. *T-Comm*. Vol. 13, no.11, pp. 68-76.

РАЗРАБОТКА И АНАЛИЗ МОДЕЛИ CALL-ЦЕНТРА В УСЛОВИЯХ ПЕРЕГРУЗКИ

Степанов Сергей Николаевич, *Московский Технический Университет Связи и Информатики (МТУСИ), Москва, Россия*
Шишкин Максим, *Московский Технический Университет Связи и Информатики (МТУСИ), Москва, Россия*
Степанов Михаил Сергеевич, *Московский Технический Университет Связи и Информатики (МТУСИ), Москва, Россия*
Журко Анна, *Московский Технический Университет Связи и Информатики (МТУСИ), Москва, Россия*

Аннотация
Построены и проанализированы функциональные и математические модели call-центра, работающего в случае перегрузки. В модели рассматриваются следующие особенности: возможность обслуживания поступающего запроса посредством IVR (Interactive Voice Response); возможность ожидания начала обслуживания в случае блокировки и возможность повторения запроса в случае занятия всех позиций ожидания или неудачного завершения времени ожидания. Определен марковский процесс, описывающий функционирование модели. Основные показатели эффективности поступления и обслуживания запросов приведены с помощью значений стационарных вероятностей состояний модели. Значения показателей эффективности находят после решения системы уравнений состояния с помощью итерационного подхода Гаусса-Зейделя. Найдены выражения, которые связывают основные показатели эффективности модели в форме локальных и глобальных законов сохранения. Полученные результаты могут быть использованы для косвенного измерения интенсивности первичных запросов и вероятности повторения вызова. Показано, как использовать модель и полученные результаты для снижения негативных последствий перегрузки путем фильтрации входных потоков первичных и повторных попыток. Рассмотрено использование модели для расчета количества операторов и мест ожидания, необходимых для обслуживания входящих потоков трафика, с заданным значением вероятности потерь при вызове и средней величиной ожидания начала обслуживания. Приведены численные результаты, иллюстрирующие реализацию полученных выражений и алгоритмов.

Ключевые слова: *call центр, система уравнений равновесия, оценка производительности, многопрофильная маршрутизация, повторные вызовы..*

Литература

1. *Gans N., Koole M., Mandelbaum A.* (2003) Telephone Call-Centers: Tutorial, Review and Research Prospects. Manuf. Service Manage, no. 5, pp. 79-141.
2. *Stolletz R., Helber S.* (2004). Perfomance Analysis of an Inbound Call-Center with Skills-Based Routing. Springer-Vellag, Hannover.
3. *Mandelbaum A., Zeltyn S.* Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Operations Research. 2009. Vol. 7, no 5, pp. 1189-1205.
4. *Ruiz Martinez D., Mclaren P., Norman J., Ronkko M.* Overload of calls. EENA Operations Document, 2012.
5. *Степанов С.Н.* Численные методы расчета систем с повторными вызовами. М.: Наука. 1983.
6. *Степанов С.Н., Степанов М.С.* Построение и анализ обобщенной модели контакт-центра // Автоматика и телемеханика. 2014. № 11. С. 55-69.
7. *Stepanov S.N., Stepanov M.S.* Algorithms for Estimating Throughput Characteristics in a Generalized Call Center Model. Automation and Remote Control. 2016. Vol. 77, no 7, pp. 1195-1207.
8. *Aguir S., Karaesmen F., Aksin O.Z., Chauvet F.* The impact of retrials on call center performance. OR Spectrum. 2004. Vol. 26, no 3, pp. 353-376.
9. *Степанов С.Н.* Основы телетрафика мультисервисных сетей. М.: Эко-Трендз. 2010.
10. *Степанов С.Н.* Теория телетрафика: концепции, модели, приложения // Серия "Теория и практика инфокоммуникаций". М.: Горячая линия - Телеком, 2015.
11. *Stepanov S.N.* Markov Models with Retrials: The Calculation of Stationary Performance Measures Based on the Concept of Truncation. Mathematical and Computer Modelling. 1999. Vol. 30, pp. 207-228.
12. *Stepanov S.N.* Generalized model with retrials in case of extreme load. Queueing Systems. 1998. Vol. 27, pp. 131-151.
13. *Stepanov S.N., Stepanov M.S., Zhurko H.* The Modeling of Call Center Functioning in Case of Overload. In: Vishnevskiy V., Samouylov K. (eds) DCCN 2019. Lecture Notes in Computer Science (LNCS). Springer, Cham. Vol 11965, pp. 391-406.
14. *Stepanov S.N., Shishkin M.O., Sosnovikov G.K, Stepanov M.S., Vorobeychikov L.A., Zhurko H.M.* The Analysis of Call Center Model in Case of Overload. T-Comm. 2019. Vol. 13, no.11, pp. 68-76.

Информация об авторах:
Степанов Сергей Николаевич, *Московский Технический Университет Связи и Информатики (МТУСИ), заведущий кафедрой сети связи и системы коммутации, д.т.н., Москва, Россия*
Шишкин Максим, *Московский Технический Университет Связи и Информатики (МТУСИ), кафедра мультимедийных сетей и услуг связи, магистрант, Москва, Россия*
Степанов Михаил Сергеевич, *Московский Технический Университет Связи и Информатики (МТУСИ), доцент кафедры сетей связи и систем коммутации, к.т.н., Москва, Россия*
Журко Анна, *Московский Технический Университет Связи и Информатики (МТУСИ), кафедра сети связи и системы коммутации, аспирант, Москва, Россия*