UDC 621.391

Three algorithms for traffic limitation in emergencies

Andrey K. Levakov[*], **Nikolai A. Sokolov**[†]

[*] *Department of Technical Director*
Center Macro-regional Branch (Center MRF) of PJSC Rostelecom
"Comcity", household 6, build. 1, Kievskoe shosse 22km, Moscow, 108811
[†] *Department of Network Planning*
Saint-Petersburg branch of Central science research telecommunication institute
Warshavskaya 11, St. Petersburg, 196128, Russia

Email: levakov1966@list.ru, sokolov@niits.ru

Three algorithms for traffic limitation arising in case of significant congestion which are caused by emergencies of different nature are considered. The first algorithm is based on the introduction of a pause between the moments of call arrival from each source. Each call represents the attempt to establish a connection from a fixed or mobile phone. The second algorithm is based on restriction of the communication session duration. With regard to telephone communication, session is considered as the conversation time. The third algorithm provides for the sorting of calls and the possibility of separating the number of servers into two groups taking into account the nature of specific emergencies. In this case, servers are terminals in the emergency call center. The effectiveness of the proposed algorithms in emergencies is analyzed. Some examples of practical implementation of the proposed solutions are given. In conclusion, the directions of further work are formulated. Mainly, these directions are based on interdisciplinary research in relation to the operation of a multiservice network in emergencies.

Key words and phrases: teletraffic system, call, pauses between demands, limitation of holding time, calls sorting, effectiveness of the algorithm.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the IX Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 19-Apr-2019, published at http://ceur-ws.org

1. Introduction

The teletraffic theory studies telecommunications networks and their elements as stochastic systems [1, 2]. The available research papers and articles on teletraffic theory rarely consider whether it is possible to change the nature of the incoming request flow and their service time deliberately and divide the service units into several groups. The studies are mainly focused on the request service processes and choice of the appropriate algorithms. This approach is reasonable when the majority of requests are generated to satisfy the communication and information needs [3] of the telecommunication networks' users. However, it is an entirely different case when the requests come largely in response to extraordinary events, such as, for example, emergency situations [4].

In emergencies, certain elements of telecommunications networks often experience a considerable overload, which can be decreased in different ways. Among them, three algorithms are of ultimate importance because they allow increasing the portion of requests processed successfully.

This article describes a teletraffic system that technically allows implementing three changes. First, we can modify the nature of the incoming flow by introducing pauses between requests created by each load source. Second, the request service time can be restricted. Third, the requests that arrive at first responders can be sorted [5] by the probable cause, which also helps to increase the portion of requests processed successfully.

The proposed algorithms have been successfully tested for processing of requests created by fixed or mobile networks' subscribers. The corresponding examples are given at the end hereof.

2. Mathematical model of a teletraffic system

The model under study can be represented as a black box [6] with the defined processes $A(t)$, $B(t)$, $C(t)$, $D(t)$ and $P(t)$ (see Fig. 1). The black box notion is typically used to simplify the research of complex systems. Representing a complex system as a black box does not require a profound understanding of its operating principles. Normally, it is sufficient to study the system input and output process and controls.

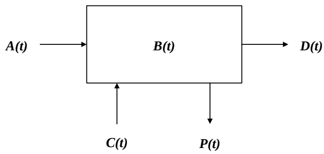

Figure 1. Teletraffic system modeled as a black box

Process $A(t)$ at the queuing system input relates to the request flow, which is one of the most important notions in the teletraffic theory. A request refers to any element that should be processed. In this article, a request means an attempt to establish a connection from a fixed or mobile phone. The word "service" is used as a generic term for a set of actions to establish a connection between the voice communication terminals. Process $B(t)$ is assumed to describe servicing. In some cases, a request cannot be processed; then it leaves the teletraffic system. Process $P(t)$ reflects such events. The successfully processed requests form a flow, which process $D(t)$ represents. This request flow is commonly referred to as an incoming flow. Process $C(t)$ represents the black box control functions. Together with other management operations, this process provides for the functions implemented with the three algorithms mentioned above.

The mathematical model is also defined using the modified Kendall's notation [7]. Generally, the model under study can be represented as follows:

$$GI/GI/V/r/f_0 \tag{1}$$

Symbol GI indicates that two distributions (duration of intervals between the request arrival time and their service time) are subject to any probability law; the number of service units is equal to V; the number of waiting places in a queue amounts to r; the requests are processed on a first-served basis without priorities.

With the following simplifications, we observed a number of regularities: the request flow follows a Poisson law and the service time distribution follows an exponential law. In some applications, it is appropriate to consider that there are no waiting places in the queue ($r = 0$). In the modified Kendall's notation, such an idealized (simplified) model is denoted as follows:

$$M/M/V/r/f_0 \tag{2}$$

To analyze both models, it is generally appropriate to use the simulation method. In most cases, getting the required dependencies for model (2) is based on the known analytical relations [1, 2].

3. Introducing pauses between call attempts in telecommunication networks

In the teletraffic theory, load intensity Y [2, 8] is defined as the product of n (total number of the connected terminals), λ (average intensity of requests from one terminal per unit of time) and h (mathematical expectation of the time the network resources are occupied):

$$Y = n\lambda h \tag{3}$$

The loss probability defined by the Erlang formula is often denoted as follows: $E(Y, V)$. It is convenient to represent the value of the pause duration τ as a fraction of the mean value of occupation time h. This fraction is share is denoted by letter ϑ. It is appropriate to denote the corresponding request loss probability as follows: $E(Y, V, \vartheta)$.

If the network is overloaded during an emergency situation, value Y often exceeds value V by several times. To get the quantitative estimations, we first consider the case when $V = 30$ for three levels of Y that define the following probability values $E(Y, V)$:

- $E(Y = 3V, V = 30) \approx 0.67$;
- $E(Y = 6V, V = 30) \approx 0.83$;
- $E(Y = 9V, V = 30) \approx 0.89$.

Fig. 2 shows functions $E(Y, V, \vartheta)$ for different duration values of the pause between the call attempts. We chose to use the logarithmic scale for the ordinate axis. It is evident that introducing comparatively short pauses between the connection attempts reduces the number of lost calls considerably. For example, to maintain the loss probability level below 0.01 even when $Y = 9V$, it is sufficient to introduce pauses that last about 3% of an average conversation time.

Unfortunately, we have to introduce longer pauses for small values V. In particular, simulation results for $V = 3$ showed that if $Y = 9V$, then the loss probability level stays below 0.01 when the pause lasts at least 50% of an average conversation time.

The model analysis (1) revealed that the behavior of the functions shown in Fig. 2 does not change [9, 10] when the nature of distributions $A(t)$ and $B(t)$ changes. However, choosing a proper value τ becomes a more complicated task. To solve it, we should follow a procedure that involves periodic traffic monitoring of value Y, awareness of the maximum allowed traffic intensity Y_{MAX}, preventive selection of the pause duration change limits – τ_{min} and τ_{max}, respectively. Modern switching systems and call centers allow estimating value Y in real time. We can assume that value τ_{min} is equal to zero. Threshold τ_{max} is estimated giving due consideration to psychological factors of users'

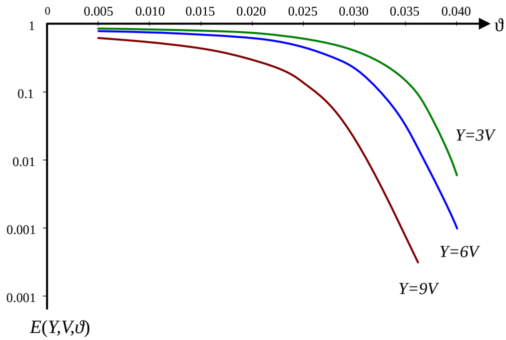

Figure 2. Changes in the request loss probability with the introduced pauses

behavior in emergency situations [11, 12]. The preliminary studies indicated that most users perceive value $\tau_{max} = 30s$ as an acceptable waiting time.

Fig. 3 shows how value τ (with some simplifications) is selected based on a set of comparatively simple calculations. This procedure does not seem difficult to implement in modern telecommunication equipment.

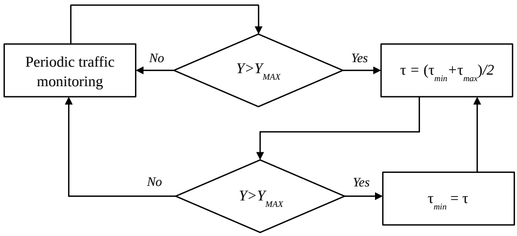

Figure 3. Pause duration τ selection procedure

To improve the public perception of the proposed suggested algorithm, telecommunications networks' users should be provided with explanations. This task has two essential aspects. First, the public should be informed about possible changes in call processing via various mass media. Second, so-called voice prompts during emergency situations can help users gain a better understanding of the connection establishment mechanism. Namely, the following phrase can be used as a prompt: "You can make another call attempt in 30 seconds."

4. Restriction of the communication session duration

According to formula (3), the load intensity can be decreased by restricting the maximum conversation time h_{max} by a predefined threshold X. Value X and threshold τ_{max} are selected giving due consideration to psychological factors of users' behavior in emergency situations [11, 12]. According to a number of experts, the initial value X can be set to 2 minutes. In addition, a voice prompt should be used, for example: "The conversation time is limited to 2 minutes."

The conversation time restriction should be the first measure implemented before pauses between the call attempts are introduced. Thus, both algorithms can be reasonably considered to form a joint solution for restricting the load intensity in emergency situations (see Fig. 4). To simplify the example, we assume that the duration of the pause between the call attempts increases only once, which makes it longer by summand $\delta\tau$.

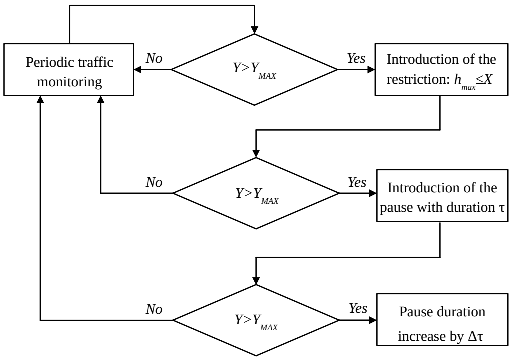

Figure 4. Algorithm to limit the load intensity in emergency situations

Formula (3) gives grounds for the following statement: the load decrease efficiency is defined by the relation of values h and X. On the other hand, we should bear in mind that if the maximum conversation time is decreased noticeably, the number of the repeated call attempts will grow [1].

5. Sorting the calls in the emergency call center

At present and in the near future, telecommunications are the main information exchange channel between users and first responders. The first responders' workplaces are located in a specialized center, which is normally connected to the 112 emergency telephone system [5, 14]. In a large-scale emergency situation, the number of incoming calls to first responders increases dramatically. Most of the calls are related to the same event and contain almost identical information.

To minimize the number of such calls, voice prompts should inform the users that the first responders are aware of the emergency situation and taking measures. The conducted studies proved that such a solution could reduce the number of calls induced by the emergency situation by at least 40% [15]. To solve this task, we need to define a certain area S_{ES}, which has the lasting effects of the emergency situation. Fig. 5 pictures this area as an ellipsis within the first responders' service area S_0. It is evident that $S_0 > S_{ES}$ in most cases; however, the following two inequalities can be true for the load intensity values Y_0 and Y_{ES}: $Y_0 \geq Y_{ES}$ and $Y_{ES} \geq Y_0$. The right part of Fig. 5 shows the queue of r calls that arrived from two areas, S_{ES} and $(S_0 - S_{ES})$. The software in the emergency call center is represented as a queue management unit, which sorts the calls in such a way so that the service procedure is close to the so-called

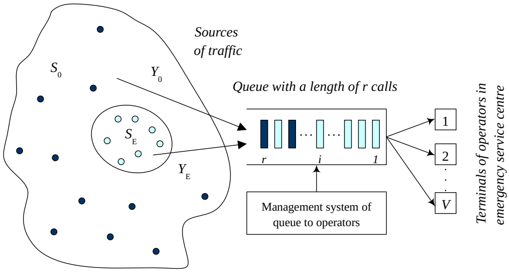

Figure 5. Sorting the calls in the emergency call center during an emergency situation

egalitarian algorithms [16,17]. The "egalitarity" (equality, derived from French "égalité") is determined by the equivalence of the call loss probability or waiting time values.

The right part of Fig. 5 shows the queue of r calls that arrived from two areas, S_{ES} and $(S_0 - S_{ES})$. The software in the emergency call center is represented as a queue management unit, which sorts the calls in such a way so that the service procedure is close to the so-called egalitarian algorithms [16, 17]. The egalitarity (equality, derived from French “égalité”) is determined by the equivalence of the call loss probability or waiting time values.

The key research findings that allow us to define the call sorting rules are given in [13]. This article examines another call sorting aspect based on dividing V first responders' workplaces (service units in terms of the teletraffic theory) into two groups, V_0 and V_{ES}. According to the teletraffic theory [2], when the load characteristics Y_0 and Y_{ES} are identical, dividing the service units into two or more groups results in $V_0 + V_{ES} > V$.

Let us assume that a group of first responders has been specially trained to process the calls typical for a certain emergency situation. When the situation occurs, the first responders are capable to process calls from area S_{ES} faster. This is possible because value Y_{ES} decreases with the decrease of factor h in formula (3).

To illustrate this statement, let us consider a numeric example with the following initial conditions:

- The emergency call center receives over 500 requests per hour.
- Half of the calls arrive from area S_{ES}.
- The average request processing time is 3 minutes.
- The allowed call loss probability is set to 0.01.
- The average request processing time for this emergency situation can be reduced to 1.5 minutes if a trained group of first responders is involved.

To simplify the calculations, we assume that $r = 0$. Then, we use the first Erlang formula to estimate the loss probability [1, 2]. To comply with the allowed call loss probability standard without the preventive training of first responders on how to address specific issues of a particular emergency situation, 36 workplaces should be created. With the assumptions made, when two groups of first responders are created, $V_0 = 21$ and $V_{ES} = 13$. It means that the total number of workplaces can be decreased to 34.

Thus, call sorting helps in addressing a range of topical issues related to call processing in emergency situations. It should be emphasized that training of first responders should become an ongoing process considering the changes in the nature of certain emergency situations and the new information technology developments aimed at increasing security.

6. Examples of practical applications of the proposed algorithms

The proposed algorithms have been tested during the development and implementation of the 112 emergency telephone system equipment in the Russian Federation by PROTEI company. Based on the conducted research, the “Traffic Limitation Mechanism in Emergencies” study was contributed to the 13th Study Group of the Standardization Sector of the International Telecommunication Union [18].

The emergency situation effects can be efficiently minimized based on the interdisciplinary research results [19]. For this reason, several stakeholders have approved the practical guidelines on the implementation of the proposed solutions, which gives hope of getting a synergistic effect [4]. In particular, the rational use of the emergency notification system resources reduces the volume of traffic that arrives at the emergency call center. However, the opposite process can also take place: less informative messages in the emergency notification systems can lead to mass panic, which, in turn, stimulates the snowballing growth of traffic in the telecommunications networks.

In emergency situations, traffic can be successfully limited with the efficient feedback communication between all the participants in the information and communication sector, such as research centers, equipment designers, design institutes, construction companies, maintenance companies, and information resource providers. The feedback communication can be arranged using the key principles of dia$par [20], which creates a cybernetic model of real-life objects and processes. This model is called a digital twin and allows analyzing the results of the emergency situation consequence liquidation (in terms of the telecommunications networks’ stability) and reproducing various scenarios of changes in the multiservice traffic.

7. Conclusions

The three algorithms that limit traffic in emergency situations, which we examined herein, have proven to be efficient. It is important that the proposed solutions can be implemented with comparatively simple modifications in the telecommunication equipment software.

The mentioned algorithms have been adapted to process the voice traffic. In the future, the share of other types of traffic sent as SMS, MMS and otherwise will grow. It means that the traffic will become more and more multiservice, which will require the new algorithms to be developed or the existing ones to be modified.

The suggestions on how to minimize the redundant information on the Internet nowadays are given in [21], for example. According to the authors of the article, such an approach is extremely important in emergency situations when decision-makers should never be distracted from the essence of the task they are to solve. The use of neural networks [22] and other artificial intelligence techniques [23] is the principal direction of future developments in the traffic limitation algorithms.

References

1. S. N. Stepanov, Teletraffic theory: concepts, models, applications, Goryachaya Liniya – Telecom, 2015, 867 p.
2. Y. N. Kornyshev, A. P. Pshenichnikov, A. D. Kharkevich, Teletraffic theory, Radio i Svyaz’, 1996, 272 p.
3. Maslow A. G, Motivation and personality, Eurasia, 2001, 478 p.
4. A. K. Levakov, Features of the next generation network in emergency situations, IRIAS, 2012, 108 p.
5. B. S. Goldstein, A. K. Levakov, N. A. Sokolov, Access to the call center "112", Vestnik Svyazi, 2012, no.1, p. 5-8.
6. M. Bunge, A General Black Box Theory, Philosophy of Science, Vol. 30, No. 4, 1963, pp. 346-358.

7. N. A. Sokolov, The tasks of telecommunication network planning, Tehnika svyazi, 2012, 432 p.
8. ITU-D. Teletraffic Engineering Handbook (edited by V. B. Iversen), Geneva, 2003, 321 p.
9. A. K. Levakov, The results of the simulation of the NGN network with a significant increase in traffic. Part I, Telecommunications, 2012, no.7, pp. 32-34.
10. A. K. Levakov, The results of the simulation of the NGN network with a significant increase in traffic. Part II, Telecommunications, 2012, no.8, pp. 24-25.
11. T. N. Gurenkova, I. N. Eliseeva, T. Yu. Kuznetsova, O. L. Makarova, T. Y. Matafonova, M. V. Pavlova, Y. S. Shoigu, Psychology of extreme situations, Akademiya, 2009, 320 p.
12. B. V. Boev, V. S. Yastrebov, Prediction of mass panic processes in man-made accidents and disasters, Journal of Neurology and Psychiatry, 2009, no.11, p. 81-88.
13. A. K. Levakov, Sorting calls with increasing load in the "System-112", Vestnik Svyazi, 2013, no.1, p. 26-29.
14. COCOM 18-03. Working Document "Implementation of the European emergency number 112 - Results of the eleventh data-gathering round", Brussels, 2018, 11 p.
15. A. K. Levakov, M. V. Kabanov, N. V. Pinchuk, N. A. Sokolov, Evaluation of methods to reduce telephone traffic generated by the reaction of subscribers to the event, Vestnik Svyazi, 2015, no.2, p. 12-15.
16. A. Demers, S. Keshav, S. Shenker, Analysis and simulation of a fair queuing algorithm, ACM SIGCOMM Computer Communication Review, 1989, Volume 19 (4), pp. 1-12.
17. S. F. Yashkov, A. S. Yashkova, Egalitarian division of the processor, Information processes, 2006, Volume 6, No. 4, p. 396-444.
18. `http://www.itu.int/md/meetingdoc.asp?lang=en&parent=T13-SG13-140707-C&PageLB=100` (accessed on 11 February 2019).
19. N. N. Moiseev, Selected Works. In 2 volumes. Volume 2. Interdisciplinary research of global problems. Publicism and social issues, Taydeks Ko, 2003, 264 p.
20. `https://diasparbusiness.com/cis-ru/what-is-diaspar` (accessed on 11 February 2019).
21. A. K. Levakov, N. A. Sokolov, The concept of "modified reality", Vestnik Svyazi, 2018, no.11, p. 3-6.
22. C. C. Aggarwal, Neural Networks and Deep Learning, Springer, 2018, 497 p.
23. D. L. Poole, A. K. Mackworth, Artificial Intelligence: Foundations of Computational Agents, Cambridge University Press, 2018, 820 p.