The Study of Wireless Network Resources while Transmitting Heterogeneous Traffic

Veronika Antonova
Bauman Moscow State Technical University, Moscow Technical University of Communications and Informatics
Moscow, Russia
v.m.antonova@mtuci.ru

***Abstract*—Today's network traffic can be broadly classified into two categories: real-time traffic and elastic one. Real-time traffic is delay sensitive and is to be conveyed at a fixed bit rate. This type of traffic comprises voice, videoconferencing of different quality, etc. Elastic traffic is generally used for data transmission; therefore its transfer rate can vary in proportion to the residual cell capacity. It might happen, for instance, while downloading files, transmitting machine-to-machine data for the Internet of things, etc. The process of conveying elastic traffic allows increasing the efficiency of network resources considerably. According to teletraffic theory, requests for file transmission at the expense of spontaneous resource seizing force requests for traffic transmission of real-time services out of processing. This is essential for mobile (wireless) communications networks of any standard where subscribers use devices with intelligent modems that increase the elastic traffic portion significantly. Simultaneous transmission of heterogeneous traffic over a network obviously needs some control tools that will ensure the desired quality of service (QoS) for arriving requests. Thereby when the transmission rate increases or decreases, the mean value of residual time for data file transfer decreases or increases proportionally. One of the ways to control the cellular network throughput is to set limits to the speed of elastic data traffic transmission. The given study considers the problem of planning the network capacity and acceptable amount of traffic transmitted through the network cell while transferring it with the desired QoS.**

I. Introduction

The analysis of trends in mobile network development made in [1] shows that real-time and elastic types of traffic increase. One of the major problems in mobile networks is to improve the efficiency of the resource for data transmission. This can be done by means of managing the rate of elastic traffic transfer that allows increasing the network throughput as well as improving the quality of service. This raises the challenge of developing a new method of traffic management in mobile networks. The aim of the given study is to build a mathematical model for simultaneous transfer of two types of traffic: the real-time one and the traffic the rate of which changes depending on the cell resource loading. At present, this model can be widely used for controlling the speed of heterogeneous traffic transfer to minimize radio resource.

It is the properties of elastic traffic that have the potential to significantly improve efficient use of data transmission resources. This is of particular importance for prospective mobile telecommunications networks that will utilize user devices with intelligent modems raising the portion of elastic traffic [2]. It is clear that simultaneous transmission of heterogeneous traffic needs efficient tools for managing that will ensure the given quality of service indicators for arriving requests. The simplest way of managing is to set some threshold limits for elastic data traffic bit rate.

First models of wireless network resource assessment included only a homogeneous type of traffic; analytical solutions were found and recurrent algorithms were developed for them. Later, solution algorithms were developed for heterogeneous traffic with constant rates. Due to the emergence of applications that do not require a constant data rate, models of non-real time traffic began to be calculated. However, those studies were conducted without taking into account access schemes that implement priority processing in multiservice wireless networks.

A similar study was carried out in [3]. It proposes an efficient call admission control algorithm that relies on an adaptive multi-level bandwidth allocation scheme for non-real time calls. The scheme allows reducing the call dropping probability along with increasing bandwidth utilization. However, the article is narrowed down to one-dimensional Markov process that does not take into account changes in the number of calls by decreasing its speed that is one of the most important properties of non-real time traffic.

In [4] there is a description of routers that use virtual queues to mark packets depending on the level of congestion. But the algorithm itself is based on the fact that non-real time traffic is simply delayed in the queue during congestions. Moreover, as in the majority of articles on such subjects, modeling is not confirmed by the corresponding mathematical calculations.

This study is the practical one; nevertheless, the numerical experiment helps better understand theoretical properties of the mathematical model for simultaneous transmission of real-time and elastic traffics through wireless networks.

II. Analytical Modeling and Simulating of Admission Control Scheme for Simultaneous Traffic Transfer

A. Description of the mathematical model for simultaneous traffic transmission

In this study we analyze the process of resource allocation in a network cell. First, denote by (i_r, i_d) the vector that indicates the mode for the number of requests serviced by the cell, where

- i_r – is the number of requests for real-time traffic transmission;
- i_d – is the number of requests for elastic (or data file) traffic transmission.

We assume that the service time of every request i_r for transferring real-time traffic has an exponential distribution with the parameter μ_r. While supporting real-time traffic in the mode (i_r, i_d), cell resources are allocated in the quantity of $i_r \cdot c_r$ bps, where c_r is the required resource for real-time traffic transmission. Further, suppose that the service time i_d of each request for elastic traffic transferring in the mode (i_r, i_d) is exponentially distributed. By μ_d denote the parameter of this distribution. The value μ_d depends on the cell capacity. Assume that the capacity of the cell is equal to C. In this case if the ratio

$$i_r \cdot c_r + i_d \cdot c_2 \leq C$$

is performed, each of the accepted requests i_d is transferred at the maximum rate c_2. It is readily seen that for the elastic traffic service in the mode (i_r, i_d), resources are allocated in the quantity of $i_d \cdot c_2$ bps, wherein the part $C - i_r \cdot c_r - i_d \cdot c_2$ of the cell resources remains untapped due to the restrictions on the maximum possible data transfer rate. This assures a constant bit rate to the end users. And vice versa when the ratio equals

$$i_r \cdot c_r + i_d \cdot c_2 > C,$$

each of the served requests i_d in the mode (i_r, i_d) is transmitted at the

$$c_d = (C - i_r \cdot c_r) / i_d$$

bit rate. Obviously, all residue cell resources in the quantity of $C - i_r \cdot c_r$ bps are allocated for elastic traffic service.

It follows that the average value of the residual service time for every request when elastic traffic is transmitted increases or decreases in proportion to increasing or decreasing transfer rates [5], [6]. The data transfer rate changes dynamically subjected to the cell loading. If the cell traffic is low, data is transmitted at the maximum bit rate c_1 that is supported by technical facilities of the wireless network; otherwise the minimum bit rate c_2 is applied. It should be noted here that in this case the resource being used and, consequently, the traffic speed of real-time services are constant [7].

B. The example of the process of request arriving

In order to provide an example of resource allocation for data transmission rates, the process of request arrival and service is considered in the situation when the cell capacity C is 10 Mbps, the real-time traffic rate c_r equals 1 Mbps, and the transmission rate of elastic traffic requests varies from c_1 = 1 Mbps to c_2 = 2 Mbps (see Fig.1: hatched cells show real-time traffic transfer, squared cells mean elastic traffic transmission). While the cell is in the mode (2, 2), i.e. there are two real-time requests and two elastic traffic requests to be served, the cell resource $c_{r,d} = 2 + 2 \cdot 2 = 6$ Mbps is occupied at the moment of time t_0 and the residue cell capacity is 4 Mbps.

Suppose that at the moment of time t_1 we have a new real-time request to transmit. This request is accepted to be transferred and the system switches to the mode (3, 2). In this mode the cell resource $c_{r,d}$ equaled to $3 + 2 \cdot 2 = 7$ Mbps is occupied and the residue cell capacity is 3 Mbps.

Then, suppose that at the moment of time t_2 we receive a new elastic traffic request to transmit. This request is accepted to be transferred with the maximum rate c_2 = 2 Mbps and the system switches to the mode (3, 3). In this mode the cell resource $c_{r,d} = 3 + 3 \cdot 2 = 9$ Mbps is occupied and only 1 Mbps is left.

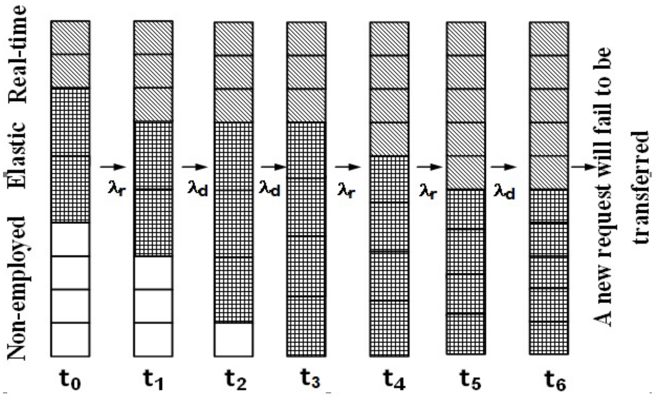

Fig. 1. The example of resource allocation for the researched network cell model

Further, assume that at the moment of time t_3 we have a new elastic traffic request to transmit. This request is accepted to be transferred and the system switches to the mode (3, 4). But the system can't transmit all elastic traffic requests at the maximum rate. Thus elastic traffic request rates decrease and are calculated as (10 – 3) / 4 = 1,75 Mbps.

Now assume that at the moment of time t_4 a new real-time request is received to be transmitted. The system accepts it to transfer and switches to the mode (4, 4). All elastic traffic request rates are reduced again and determined as (10 – 4) / /4 = 1,5 Mbps.

At the next moment of time t_5 we have a new real-time request to transmit. This request is accepted to be transferred and the system switches to the mode (5, 4). All elastic traffic request rates decrease and are calculated as (10 – 5) / 4 = =1,25 Mbps.

Further, suppose that at the moment of time t_6 a new elastic traffic request arrives at the system. This request is accepted to be transferred and the system switches to the mode (5, 5). All elastic traffic rates are reduced once more and can be determined as (10 – 5) / 5 = 1 Mbps; this means that all elastic traffic requests are transmitted at the minimum rate.

If the system hasn't completed any of the requests being processed at the moment, it won't accept a new one of any kind, therefore this new request will fail to be transferred.

Quality of service for real-time requests is assessed by a portion of lost requests and by the average value of used cell resources expressed by bits per second [8], [9].

C. Representation of the results of the data transfer modeling

Let us define quality of service for elastic traffic requests by the portion of requests that have been denied admission to service, and by average delivery time of the corresponding message [10].

For assessing the portion of real-time requests that haven't been processed (or the lost ones) as well as employed resource volume and transmission time for elastic traffic it's sufficient to know how much time the cell is in the modes where the numbers of real-time and elastic traffics don't vary [11-14]. At the same time it is necessary to determine the stochastic process modes and components for which the assessment of the given indicators is carried out. Denote by:

- $i_r(t)$ the number of requests for real-time traffic transmission that are being serviced at a time t;
- $i_d(t)$ the number of requests for elastic traffic transmission that are being serviced at a time t.

Two-dimensional Markov process is thereby introduced:

$$r(t) = (i_r(t), i_d(t)).$$

This process is defined on the finite mode space S. The space includes vectors (i_r, i_d) with components i_r, i_d and can take values

$$i_r = 0,1,...,\left\lfloor \frac{C}{c_r} \right\rfloor$$

$$i_d = 0,1,...,\left\lfloor \frac{C - i_r c_r}{c_1} \right\rfloor.$$

By $p(i_r, i_d)$ denote values of stationary probabilities for modes $(i_r, i_d) \in S$. The probabilities represent the time when the cell is in the mode (i_r, i_d); they can be used for core indicators assessment while servicing simultaneous arriving requests. In the model, uppercase letters will be utilized for denoting unnormalized values of probabilities for the model modes and lowercase letters will be used for normalized values.

The servicing procedure of requests for real-time traffic is evaluated by the portion of lost requests and by the average value of employed cell resources measured in bits per second. QoS for elastic flow is assessed by the portion of each type of requests that have been denied admission to service as well as by the average time of corresponding message delivery.

These characteristics as well as a number of other characteristics that are considered in the Markov model, which is analyzed in this paper, can be calculated by summing up the stationary probabilities $p(i_r, i_d)$ of the Markov process $r(t)$ for certain subsets of the state modes S.

The average number of requests for real-time traffic transmitting m_r is defined as:

$$m_r = \sum_{(i_r, i_d) \in S} p(i_r, i_d) i_r .$$

The average number of requests for elastic traffic transmitting m_d is determined as:

$$m_d = \sum_{(i_r, i_d) \in S} p(i_r, i_d) i_d .$$

The average value of resources for transmitting information in the cell s_r that are employed for elastic traffic transfer is expressed by the ratio

$$s_r = \sum_{(i_r, i_d) \in S} p(i_r, i_d) i_r c_r .$$

The average value of resources for transmitting information in the cell s_d that are employed for elastic traffic transfer is expressed by the ratio

$$s_d = \sum_{(i_r, i_d) \in S} p(i_r, i_d) i_d \min\left\{ c_2, \frac{C - i_r c_r}{i_d} \right\} .$$

To select the numerical values of the minimum and maximum allowed bit rates for elastic traffic transmission it is necessary to develop a set of mathematical models that would increase the efficiency of simultaneous receiving and processing requests from modern communication applications for heterogeneous traffic transfer and to study the constructed model in order to formulate recommendations on applying the obtained results.

D. Analyzing the dependence of basic probabilistic characteristics on the traffic increase

Let us analyze the dependence of basic probabilistic characteristics for the given model on the traffic increase. For example, we assume that input parameter points are equal to C = 100 Mbps, c_r = 3 Mbps, c_1 = 1 Mbps, c_2 = 5 Mbps and the average size of a request being transmitted is F = 16 Mbit.

Further assume that the average time for real-time traffic transfer T is equal to 300 s, and the average minimum and maximum file transfer times are, respectively:

$$T_{d,1} = F / c_1 = 16 \text{ s and}$$

$$T_{d,2} = F / c_2 = 3{,}2 \text{ s.}$$

Then assume that in case of an emergency situation, requests for elastic traffic transfer are received a hundred times more frequently than requests for real-time traffic transfer, i.e. the ratio $\lambda_r = 100\lambda_d$ is true.

By assumption that requests for elastic traffic transfer arrive a hundred times more frequently than requests for real-time traffic transmission, denote by ρ the minimum potential loading of a cell transfer resource unit.

Fig. 2 presents the results of the calculation of average values m_r and m_d for the requests that are being serviced while cell resource loading ρ is increasing. The graph shows that when the value ρ is rising, the average number of real-time requests taken for servicing is increasing at first and then it is decreasing. This is due to the fact that the minimal transfer resources necessary for processing one elastic traffic request are less than the resources needed for processing one real-time request. For this reason, elastic traffic requests force real-time traffic requests out of servicing. This concept is usual for processing multiservice real-time traffic.

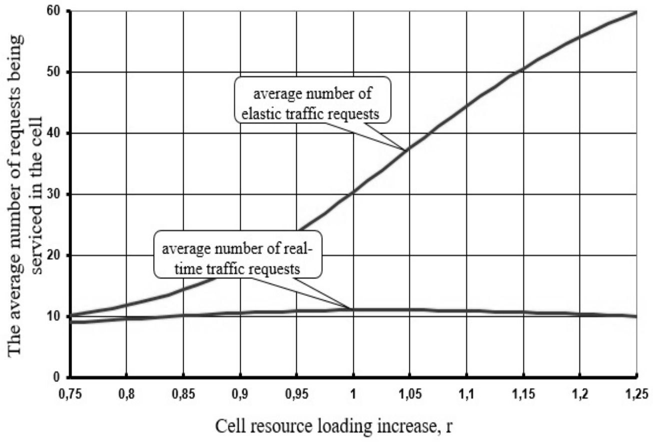

Fig. 2. The average values m_r and m_d for the requests that are being serviced while cell resource loading ρ is increasing

Fig. 3 shows the results of the calculated average values of rates s_r and s_d used in the cell that are used for traffic transmission of real-time requests and elastic traffic correspondingly.

This graph as well as the graph in Fig.2 show that when the load ρ grows, the cell resource used for traffic transmission of real-time requests increases at first but then it decreases. The reason for this is that the minimal resource needed for servicing an elastic traffic request is less than the resource necessary for transmitting a real-time request. As a result, the elastic traffic requests force the real-time ones out. Fig. 3 also illustrates summarized resource use in the cell. When the load ρ increases, it approximates to the maximum cell capacity C = 100 Mbps.

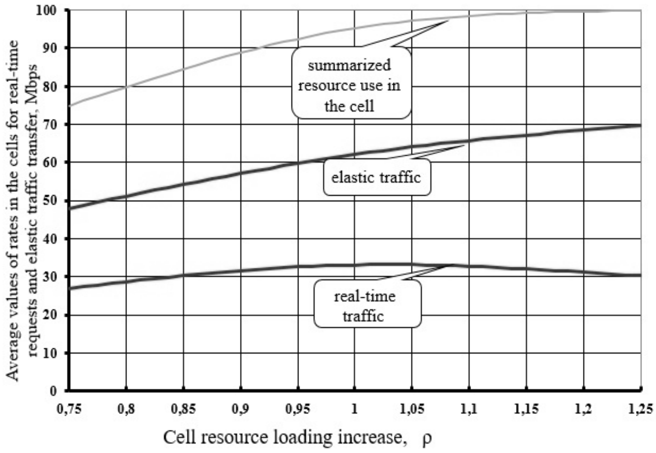

Fig. 3. The average cell resource usage for traffic transmission of real-time requests and elastic traffic while increasing the minimal potential load of a transmission resource unit in a cell with traffic continuing to arrive

III. THE PROBLEM OF MANAGING THE RATIO OF THE MINIMAL AND THE MAXIMUM BIT RATES FOR THE ELASTIC FLOW

The model construction necessitates that when the cell load is low, the elastic traffic transmission will have the maximum allowed bitrate. In the considered example the current elastic traffic request bitrate c_d tends to 5 Mbps. Fig. 4 shows the assessment of c_d. In this situation when the load ρ increases, the elastic traffic grows gradually and starts to be served at the minimum allowed bitrate c_1 = 1 Mbps, therefore overloads occur.

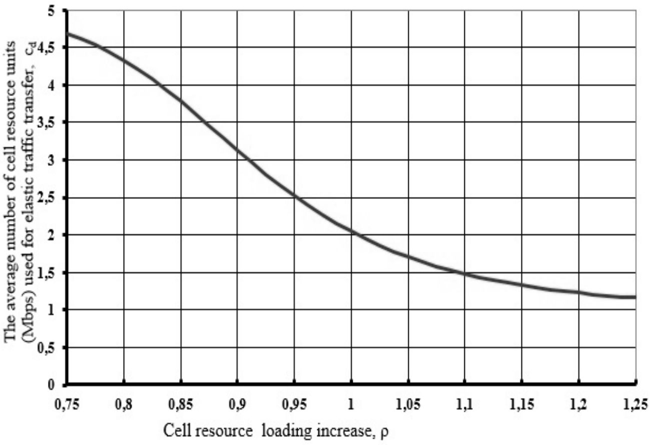

Fig. 4. The average rate of elastic traffic in the cell while increasing the minimal potential load of the cell transfer resource with traffic continuing to arrive

Further, consider the problem of managing the ratio of the minimal c_1 and the maximum c_2 bit rates for the elastic flow. It is clear that if $c_1 = c_2$, Erlang model is applied to the servicing procedure. When the upper bit rate limit c_2 is increasing, the data (elastic) transmission rate is rising at the expense of the cell throughput unemployed by real-time traffic processing. The numerical example given below will help illustrate this.

Consider an example of a cell with the following values of numerical parameters:

C = 100 Mbps, c_r = 3 Mbps, c_1 = 1 Mbps, c_2 = 1 Mbps, F = 16 Mb, $\mu_r = 1/300$, $\lambda_r = 0{,}04$, $\lambda_d = 4$ requests per second. As it was mentioned above, the given set of initial parameters will allow using Erlang model for the servicing procedure in the cell. Then if c_2 starts rising while other parameters are constant, the cell throughput capacity will increase due to faster file transfer. The numerical examples below illustrate this.

Fig. 5 shows the results of calculating the proportion of the time of the full load of the cell resource with an increase of the rate c_2. As expected, when the rate c_2 increases, the load on the cell increases. The unevenness of the curve is related to a variety of possible combinations of request distribution in a cell with a fixed value of the overall bit rate of the transmission channel C and simultaneous transfer of heterogeneous traffic.

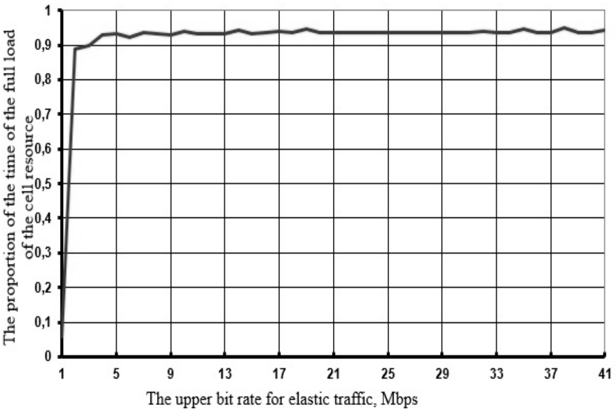

Fig. 5. The results of calculating the proportion of the time of the full load of the cell resource while increasing the maximum availability of the resource for transmitting elastic traffic

Fig. 6 shows the results of calculating the average number of requests m_d and m_r that are being serviced in the cell while increasing c_2. As it may be seen from the graph, the growth of c_2 leads to a drastic decrease in the average number of accepted requests for data transfer. This is due to the fact that they leave the system faster having additional opportunities for accelerated processing.

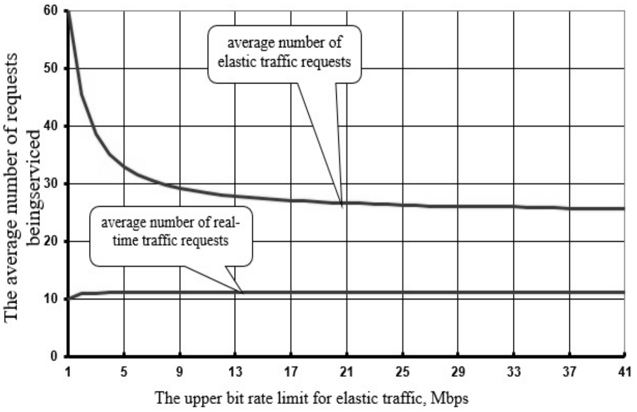

Fig. 6. The average number of requests that are being serviced in the cell while increasing the maximum availability of the resource for elastic traffic transfer

Fig. 7 shows the results of the calculated average values of rates s_r and s_d used for cell traffic transmission of real-time requests and elastic traffic correspondingly as well as their cumulative use of cell.

When the upper limit bit rate c_2 is increasing, a significant change of the resource use occurs only at the initial moment and then it becomes less relevant.

When the probability of losing requests for data transfer decreases slightly, request losses increase due to uncontrolled network resource seizing. This can be avoided by introducing the guaranteed threshold for the maximum bit rate of elastic traffic; for existing networks this threshold is to be calculated on the basis of ongoing monitoring.

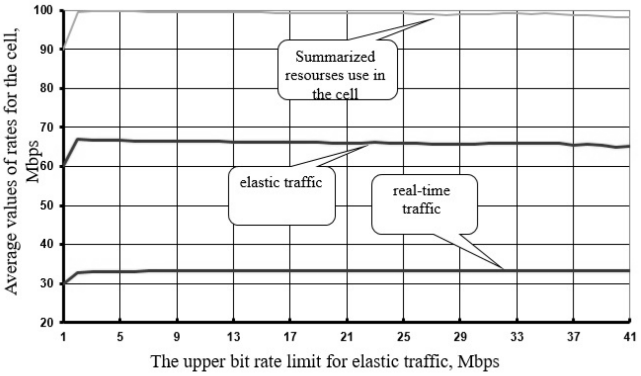

Fig.7. The average use of cell resources for real-time and elastic traffics while maximizing access to the resources for elastic traffic transfer

Fig. 8 shows the results of calculating the average use of the cell resource for file transfer c_d. It can be seen that this parameter increases with the growth of c_2, but this change occurs only at the initial moment.

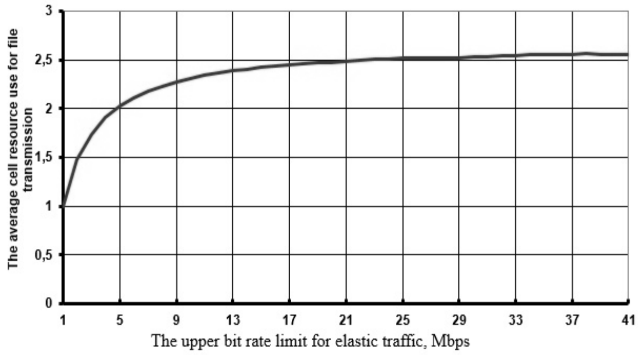

Fig.8. The average usage of cell resources for file transmission while increasing the maximum rate of the resource for transmitting traffic data

Fig. 9 illustrates the results of calculating the average file transfer time with increasing c_2. We can see that when c_2 rises, the average file transfer time decreases starting from the value of 16 seconds when a single cell resource is used for file transfer, and tends to a constant in accordance with the cell load.

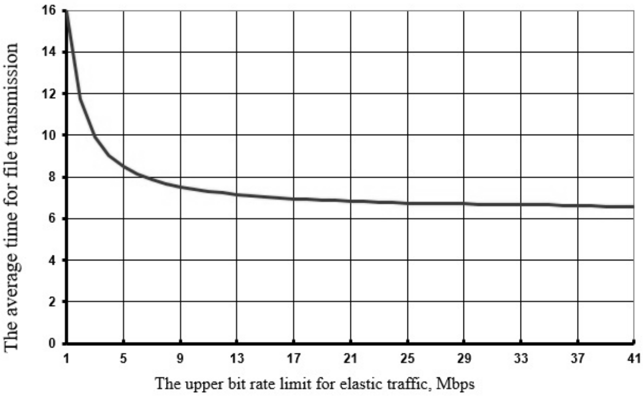

Fig.9. The average utilization of the cell resource for file transmission while maximizing access to the resources for elastic traffic transmitting

Signaling overhead applied to the model will make two-dimensional Markov process a three-dimensional one so these calculations will become more complicated.

The theoretical relevance of the paper consists in the construction and study of a model that takes into account the dependence of receiving and maintaining requests of real-time and elastic traffic. This model can be used for analyzing procedures based on the threshold limits of improvement assessment of effective resource use in a network segment. This mathematical model makes it possible to numerically evaluate the advantages of the simultaneous transfer of heterogeneous traffic.

The general nature of the assumptions made in the paper allows applying the mathematical model and the calculation algorithms created on its basis for addressing practical challenges in the radio interface of mobile networks.

IV. Conclusion

The parameter characteristics of simultaneous request servicing in the wireless network are studied, i.e. real-time and elastic traffic are considered to increase the load of the cell resource. This model makes it possible to numerically evaluate the advantages of simultaneous transmission of speech and data traffic. The solution for the planning problem of the cell capacity and the allowable traffic that can be transmitted with the given quality of service indicators is proposed.

The general nature of the assumptions accepted in this article allows applying the mathematical model and the calculation algorithms created on its basis for the majority of practical problems arising on the radio interface of mobile networks. The developed tools are recommended for using in the design and operation of mobile networks of the fifth generation networks.

The given paper has considered one of the solutions for the problem of determining the ratio between the bit rate limits for data transmission in order to improve the efficiency of cell resource use.

The work is financially supported by the Russian Foundation for Basic Research (project No. 19-07-00525 A – Developing flow-based models of routing problems in telecommunications networks)

Acknowledgment

I would like to pay special thankfulness, warmth and appreciation to my Scientific Adviser, Academician of the RAS, Dr. Kuznetsov and to Dr. Stepanov, Head of the MTUCI's Department of Communications Networks and Switching Systems, for their vital support and assistance. They helped me at every bit of my work, encouraged me to meet the deadlines and made it possible to achieve the goal.

References

[1] V.G. Skrynnikov, *UMTS/LTE Radio Subsystems. Theory and Practice*. Moscow: Sport and Culture – 2000 Publishing House, 2012.

[2] S.N. Stepanov, "The Model for Servicing the Real Time Traffic and Data with a Dynamically Changeable Speed of Transmission", *Automation and Remote Control*, vol.71, issue 1, 2010, pp.18-33.

[3] M.Z. Chowdhury, Y.M. Jang, and Z.J. Haas, "Call admission control based on adaptive bandwidth allocation for wireless networks", *Journal of Communications and Networks*, vol. 15, issue 1, Feb. 2013, pp. 15 – 24.

[4] T. Tung, J. Walrand, "Providing QoS for Real-time Applications", *in Proc. Communications, Internet, and Information Technology (CIIT 2003)*, USA, Nov. 2003, pp. 121-130.

[5] S.N. Stepanov, "Model of joint servicing of real-time service traffic and data traffic. I", *Automation and Remote Control*, vol.72, issue 4, 2011, pp. 121-132.

[6] S.N. Stepanov, "Model of joint servicing of real-time service traffic and data traffic. II", *Automation and Remote Control*, vol.72, issue 5, 2011, pp. 139-147.

[7] V.M. Antonova, "Channel Resource Assessment for Multi-Mode Links on an LTE Network Segment", *Natural and Technical Sciences*, no.10, 2014, pp. 356-358.

[8] V.M. Antonova and E.E. Malikova, "The Research of the Influence of the Service and the Information Traffic on Each Other in LTE Networks", *T-Comm – Telecommunications and Transport*, no.9, 2014, pp. 17-19.

[9] V.M. Antonova and I.A. Tsirik, "Access Control of New Requests in an LTE Network Segment", *Fundamental Problems of Radioengineering and Device Construction*, vol.15, no.5, 2015, pp. 226-228.

[10] V.M. Antonova, I.A. Gudkova, E.V. Markova, and P.O. Abaev, "Analytical Modeling and simulation of admission control scheme for non-real time services in LTE networks", *in Proc. the 29th European Conference on Modelling and Simulation, ECMS,* Varna, Bulgaria, ISBN: 978-0-9932440-0-1 / ISBN: 978-0-9932440-1-8 (CD), 2015, pp. 1156-1163.

[11] N.A. Kuznetsov, D.V. Myasnikov, and K.V. Semenikhin, "Optimization of two-phase queuing system and its application to the control of data transmission between two robotic agents", *Journal of Communications Technology and Electronics*, vol.62, no.12, 2017, pp.1484-1498.

[12] N.A. Kuznetsov, D.V. Myasnikov, and K.V. Semenikhin, "Two-phase queuing system optimization in applications to data transmission control", *Procedia Engineering*, vol.201, 2017, pp.567-577.

[13] N.A. Kuznetsov, I.K. Minashina, N.G. Ryabykh, E.M. Zakharova, and F.F. Pashchenko, "Design and Comparison of Freight Scheduling Algorithms for Intelligent Control Systems", *Procedia Computer Science*, vol. 98, 2016, ISSN 1877-0509, pp. 56–63.

[14] N.A. Kuznetsov, D.V. Myasnikov, and K.V. Semenikhin, "Optimal control of data transmission in a mobile two-agent robotic system", *Journal of Communications Technology and Electronics*, vol.61, no.12, 2016, pp.1456-1465