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 raising the portion of elastic traffic: the real-time one and the traffic the rate of which changes depending on the cell resource loading. At present, this phenomenon is not confirmed by the corresponding mathematical modeling. Moreover, as in the majority of articles on such subjects, traffic is simply delayed in the queue during congestions. But the algorithm itself is based on the fact that non-real time traffic is not delay sensitive and can be managed using the QoS properties of 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.

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;
- \( i_{id} \) – 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 \( \mu_r \). While supporting real-time traffic in the mode \((i_r, i_{id})\), 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 \( i_{id} \) of each request for elastic traffic transferring in the mode \((i_r, i_{id})\) is exponentially distributed. By \( \mu_{id} \) denote the parameter of this distribution. The value \( \mu_{id} \) depends on the cell capacity. Assume that the capacity of the cell is equal to \( C \). In this case if the ratio

\[
i_r \cdot c_r + i_{id} \cdot c_{id} \leq C
\]

is performed, each of the accepted requests \( i_{id} \) is transferred at the maximum rate \( c_{id} \). It is readily seen that for the elastic traffic service in the mode \((i_r, i_{id})\), resources are allocated in the quantity of \( i_{id} \cdot c_{id} \) bps, wherein the part \( C - i_r \cdot c_r - i_{id} \cdot c_{id} \) 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

\[
i_r \cdot c_r + i_{id} \cdot c_{id} > C
\]

each of the served requests \( i_{id} \) in the mode \((i_r, i_{id})\) is transmitted at the

\[c_{id} = (C - i_r \cdot c_r) / i_{id}\]

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 \( c_{id} \) that is supported by technical facilities of the wireless network; otherwise the minimum bit rate \( c_{id} \) 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 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 \( c_{id} = 1 \) Mbps to \( c_{id} = 2 \) 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 \( c_{id} = 2 + 2 - 2 = 2 \) 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 \( c_{id} \) 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 \( c_{id} = 2 \) Mbps and the system switches to the mode \((3, 3)\). In this mode the cell resource \( c_{id} = 3 + 3 \cdot 2 = 9 \) Mbps is occupied and only 1 Mbps is left.

Further, assume that at the moment of time \( t_3 \) we have a new real-time 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 and needs to be transmitted. The system accepts it to transfer and switches to the mode \((4, 4)\). All elastic traffic request rates are reduced once more 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

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

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 express is by the ratio

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

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:

\[ T_{d,1} = F / c_1 = 16 \text{ s} \text{ and} \]
\[ T_{d,2} = F / c_2 = 3.2 \text{ s}. \]

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 \( \lambda_e = 100 \lambda_d \) 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 \( \rho \) 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 \( \rho \) is increasing. The graph shows that when the value \( \rho \) 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.

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 \( \rho \) increases, the elastic traffic grows gradually and starts to be served at the minimum allowed bitrate \( c_1 = 1 \text{ Mbps} \), therefore overloads occur.

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:

\[
C = 100 \text{ Mbps}, \quad c_r = 3 \text{ Mbps}, \quad c_1 = 1 \text{ Mbps}, \quad c_2 = 1 \text{ Mbps}, \quad F = 16 \text{ Mb}, \quad \mu_r = 1/300, \quad \lambda_r = 0.04, \quad \lambda_d = 4 \text{ requests per second}. \]

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 $m_t$ 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. 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. 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. 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.
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