Model of radio admission control for URLLC and adaptive bit rate eMBB in 5G network

Anna Kushchazli[1], Anastasia Ageeva[1], Irina Kochetkova[1,2], Petr Kharin[1], Alexander Chursin[1] and Sergey Shorgin[2]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2]Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilova St, Moscow, 119333, Russian Federation

Abstract

In today's rapidly developing telecommunications technologies, mobile communication services are widely penetrating into all segments of society. The 5th generation network (5G) will cover a broader range of usage scenarios – enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC). This paper considers a model of joint service of URLLC and eMBB traffic within a single base station. The first type of traffic is supposed to have priority due to its high latency requirements, while the second type of traffic has a high speed. If there are no available resources to serve URLLC traffic, the allocated resources for eMBB traffic will decrease until they are completely withdrawn. We model the above mentioned system as a queuing system with the bit rate degradation and service interruption.

Keywords

5G, URLLC, eMBB, radio admission control, priority, interruption, bit rate degradation, queuing system,

1. Introduction

The life of society has been rapidly developing due to mobile communications, that has become an integral part of the everyday life of each person in society. Due to the rapid increase in the number of users and devices connected to the network, there is the growth of the load on communication networks. At the same time, network delays should be reduced and occur as rarely as possible.

The paper addresses the usage scenarios in 5G networks and joint service of eMBB and URLLC traffic. We apply methods of queuing theory and mathematical teletraffic theory. The research tasks are the following:

1. to analyze works related to joint transmission of eMBB and URLLC traffic;
2. to build a mathematical model with adaptive change in the speed of eMBB traffic;
3. to derive the probabilistic-temporal characteristics of the model.

Workshop on information technology and scientific computing in the framework of the XI International Conference Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems (ITTMM-2021), Moscow, Russian, April 19–23, 2021

aikushch@yandex.ru (A. Kushchazli); anastasia.ageeva.it@gmail.com (A. Ageeva); gudkova-ia@rudn.ru (I. Kochetkova); gruzavjeg@mail.ru (P. Kharin); chursin-aa@rudn.ru (A. Chursin); sshorgin@ipiran.ru (S. Shorgin)

The paper is organized as follows. In Section 2, the characteristics of the coexistence of serving narrow-band URLLC and broadband eMBB traffic are explained. Section 3 presents the system model and probabilistic-temporal attributes for it. Furthermore, Section 4 shows us the numerical analysis and the discussion. As a result, the conclusions are in Section 5.

2. State of The Art

2.1. 5G Usage Scenarios

According to ITU-R recommendation [1], 5G systems must support ultra-low latency and high-reliability communication systems for both users and devices. The quality of service in 5G networks should not degrade under conditions of high system load due to many active users. The 5G usage scenarios include:

- Enhanced Mobile Broadband (eMBB): Mobile broadband addresses the human-centric use cases to access multi-media content, services, and data. This type of scenario come with new application areas and requirements in addition to existing mobile broadband applications for improved performance and an increasingly seamless user experience;
- Ultra-reliable and low latency communications (URLLC): This use case has stringent requirements for capabilities such as throughput, latency, and availability. Some examples include wireless control of industrial manufacturing, medical surgery, distribution automation in a smart grid, transportation, etc.;
- Massive machine-type communications (mMTC): This use case is characterized by a huge number of concurrent users usually transmitting a relatively small amount of data that is not sensitive to latency.

It should be noticed that additional use cases are expected to emerge, which are may currently not foresee. Nevertheless, 5G network will encompass many different features. Figure 1 illustrates some examples of envisioned usage scenarios.

2.2. Joint Scheduling of URLLC and eMBB

Since both eMBB and URLLC are essential components of communication traffic in 5G networks, various studies have looked at the coexistence of these services. In terms of network bandwidth, eMBB generates a massive amount of data traffic. Unlike eMBB, URLLC produces fewer data because of its stringent latency and reliability requirements. Consequently, the coexistence of these two services is associated with achieving the sufficient eMBB throughput while meeting the URLLC requirements. Due to the time sensitivity of critical applications such as UAV automation, autonomous vehicle control, and critical medical equipment management, URLLC takes precedence over eMBB for scheduling. Typically, eMBB scheduling involves increasing the network capacity to improve the spectral efficiency, while the packet delivery reliability is ensured through re-transmissions. However, eMBB scheduling approaches may not guarantee the reliability and latency thresholds required for URLLC and thus cannot be applied in URLLC scheduling. In contrast, URLLC involves the transmission of short packets with specified latency and reliability margins.

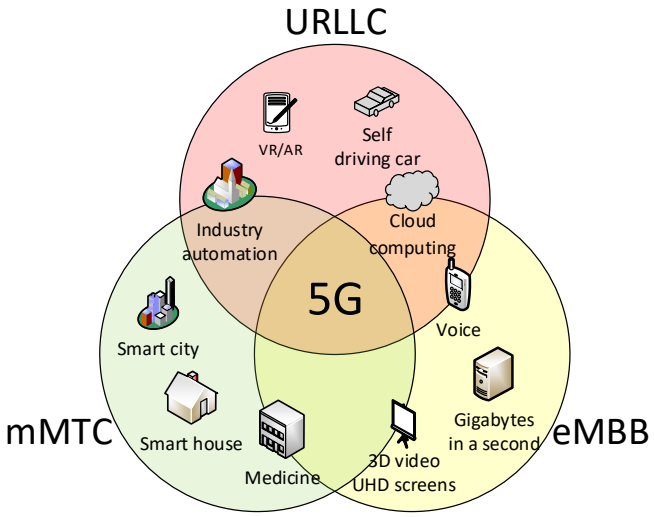


Figure 1: 5G usage scenarios

By the way, 3GPP has proposed a short and long transmission within the time interval (TTI) frame allocation for these coexistence scenarios. In this frame allocation, eMBB traffic is scheduled for a long TTI, and URLLC is automatically scheduled for a short TTI over existing eMBB traffic by adopting a puncture or overlay scheme [2].

Exceptionally, the authors of [3] propose an efficient scheduling scheme for the coexistence of eMBB and URLLC by dynamically applying to puncture or overlay schemes. The base station performs eMBB scheduling at the start of a long TTI, and URLLC scheduling is performed for a short TTI using a puncture or overlay scheme.

Regarding that paper, we consider that an incoming eMBB session arrives and takes the whole slot while URLLC traffic takes a mini-slot. Indeed, one PRB equals one slot, and the duration of it is 5 ms. We will describe the model with an adaptive change of eMBB traffic rate in more detail in the next section.

2.3. Related Works

There are several approaches to URLLC and eMBB coexistence. So [4] proposes the resource reservation for URLLC traffic. In [5], network slicing is used for both heterogeneous orthogonal multiple access (H-OMA) and heterogeneous non-orthogonal multiple access (H-NOMA), which was also studied in [6] and [7]. In addition to URLLC priority access , the authors of [8] and [9] emphasize the eMBB quality of service, they use the methods of stochastic geometry and queuing theory. A feature of paper [10] is a queuing system with random requirements.

In [11, 12], the authors consider the URLLC priority access with eMBB session interruption, and in [13] we analyzed eMBB session delay. This paper comparing to [11, 12] also proposes a preliminary transmission speed reduction of the eMBB session before its interruption.

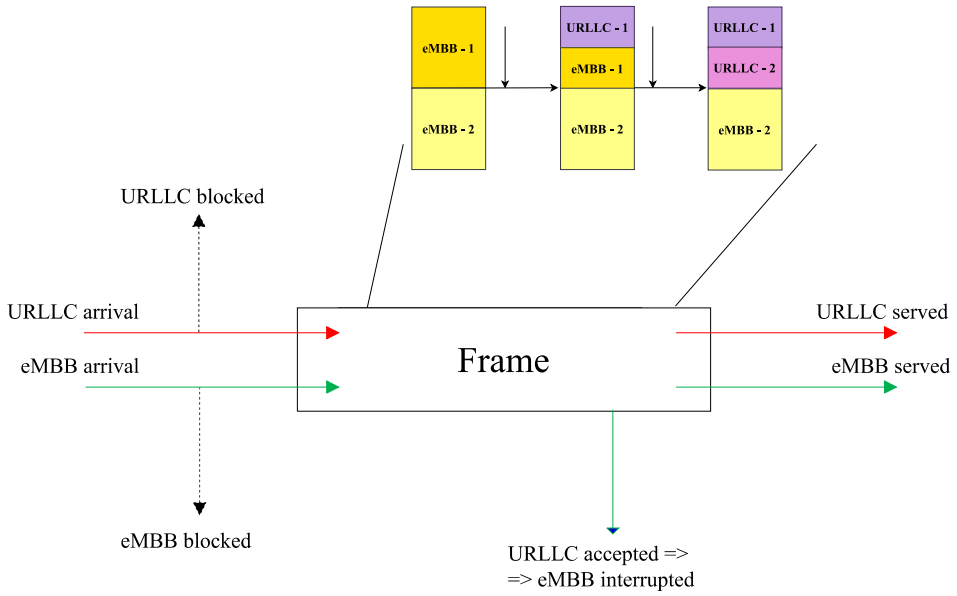


Figure 2: eMBB bit rate degradation and service interruption.

3. Queuing Model

3.1. Assumptions and Parameters

Let us consider a model where resources are allocated within one base station to service URLLC and eMBB traffic. eMBB traffic requests are distributed by the scheduler to the resource units of each slot. When a URLLC request is received, it can be assigned to any free resource unit of any subsequent mini-slot due to the delay requirements. Furthermore, the structure of the frame is presented in Figure 2.

Sessions of eMBB traffic occupy one resource block or b_1 resource units (RU), while URLLC sessions occupy one resource unit. There are N resource blocks in the system, i.e. the maximum number of eMBB sessions. Then the maximum number of URLLC sessions in the system is $C = b_1 \cdot N$.

Sessions of both traffic arrive according to the Poisson process, and the arrival rates are equal to λ_m and λ_u respectively. An eMBB session occupies the maximum speed. However, if all the resources of the system are busy when a URLLC session arrives, the rate of a eMBB session can be reduced since the second traffic is in priority. At the same time, the maximum speed will first decrease, and then the lower ones in descending order. If several applications are served at the same speed, then the choice is made randomly.

Let us denote n the total number of URLLC sessions active at a certain moment of time. To designate the number of eMBB sessions, we introduce the vector $\vec{m} = (m_1, .., m_K)$, where K the number of speeds at which eMBB sessions can be serviced, equal to the number of resource units b_1. Also, we will use a service rate vector $\vec{b} = (b_1, .., b_K)$ such that $b_1 > b_2 > ... > b_K$. Then the state of the system will take the form $(m_1, ..., m_K, n) = (\vec{m}, n) = \vec{x}$. Since the requests are

Table 1
System parameters

Parameter	Description
λ_m / λ_u	Session arrival rate of eMBB / URLLC
μ_m^{-1} / μ_u^{-1}	Average eMBB / URLLC session duration
b_1	Number of resource units
N	Total number of resource block
$C = b_1 N$	Total number of resource units
$\vec{m} = (m_1, .., m_K)$	Number of eMBB active sessions on K speed
n	Number of URLLC active sessions
$\vec{b} = (b_1, .., b_K)$	eMBB session service speed

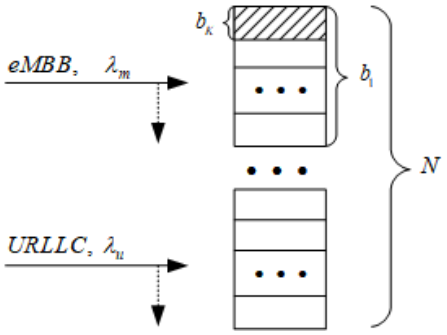


Figure 3: Model scheme

served according to an exponential distribution, the eMBB service rate will be μ_m, and URLLC – μ_u. All the main parameters of the system used in this work are presented in Table 1.

3.2. Admission Control and State Space

The scheme of the described model is presented in Figure 3. The system state space is:

$$\mathcal{X} = \left\{ \begin{array}{l} (m_1, \dots, m_K, n): \quad n \geq 0, \quad m_k \geq 0, \quad k = 1, \dots, K, \\ \sum\limits_{k=1}^{K} m_k \leq N, \quad n + \sum\limits_{k=1}^{K} b_k m_k \leq C \end{array} \right\} \tag{1}$$

If the system is in a state $(\vec{m}, n)$, then various events are possible to occur with different intensities. All possible situations are presented in Table 2. For the convenience of describing transitions, we introduce the unit vector $\vec{e_k} = (0, ..., 0, 1, 0, ..., 0)$, where 1 is at the k^{th} place.

First, consider the possible situations occurring with the intensity λ_m, i.e. when a new request eMBB arrives:

- If the system has free resources, namely a free resource block, the request will be accepted for service. In Table 2, this transition from the central state is presented as number 1.
- If there are no free resources in the system, then the session will be blocked.

In the case when an eMBB session service ends, i.e. an event with intensity μ_m which is presented as under number 2 from Table 2 occurs, the request leaves the system, and the resources are released.

We turn to events that occur with intensity λ_u, i.e. upon an receipt of the URLCC application:

- If the system has at least one free resource unit, the session is accepted for service. In Table 2 it is presented as number 3.
- If there are no free resources, but at least one eMBB session served not at the minimum speed, then the service speed of the eMBB session decreases, and the URLLC session is accepted for service. In Table 2 it is displayed as number 4.
- If there are no free resources, but at least one eMBB session served at the minimum speed, then an eMBB session service is interrupted, and the URLLC session is accepted for service. In Table 2 this is confirmed as number 5.
- If there are no free resources and no eMBB sessions in the system, the session will be blocked.

When a URLLC session leaves the system, events occur with an intensity μ_u:

- If there are no eMBB sessions in the system or they are served at the maximum speed, then the URLLC session leaves the system and frees up resources. In Table 2 it is presented as number 6.
- If the system has at least one eMBB session that is not served at full speed, then after the termination of the URLLC session service, the speed of the active eMBB session is restored. In Table 2 it is shown as number 7.

3.3. Performance Measures

Now let us talk about performance indicators of priority service of URLLC traffic. Having formed an infinitesimal generator matrix and solving the resulting system of equilibrium equations, we can find the stationary probability distribution $p(\vec{m}, n), (\vec{m}, n) \in \mathcal{X}$. Based on it, we obtain the following probabilistic-temporal characteristics of the model:

1. Average number of eMBB sessions $\overline{m_k}$ and URLLC sessions $\overline{n}$:

$$\overline{m_k} = \sum_{\vec{x} \in \mathcal{X}} m_k \cdot p(m_1, \dots, m_K, n), \quad k = 1, \dots K; \tag{2}$$

$$\overline{n} = \sum_{\vec{x} \in \mathcal{X}} n \cdot p(m_1, \dots, m_K, n); \tag{3}$$

2. Blocking probability of eMBB sessions B_m and URLLC sessions B_u:

$$B_m = \sum_{\vec{x} \in \mathcal{B}_2} p(m_1, \dots, m_L, n), \quad \mathcal{B}_2 = \left\{ \vec{x} \in \mathcal{X} \ : \ n + \sum_{k=1}^{K} b_k m_k + b_1 > C \right\}; \tag{4}$$

$$B_u = p(0, \dots 0, C). \tag{5}$$

Table 2
Transitions

No.	Output state	Intensity	Conditions
1	$(\vec{m}+\vec{e}_1, n)$	λ_m	$\sum_{k=1}^{K} m_k \cdot b_k + n + b_1 \le C$;
2	$(\vec{m}-\vec{e}_k, n)$	$m_k \cdot \mu_m$	$m_k > 0$;
3	$(\vec{m}, n+1)$	λ_u	$\sum_{k=1}^{K} m_k \cdot b_k + n + 1 \le C$;
4	$(\vec{m}-\vec{e}_k+\vec{e}_{k+1}, n+1)$	λ_u	$\begin{cases} \sum_{k=1}^{K} m_k \cdot b_k + n + 1 > C, \\ \sum_{i=1}^{k-1} m_i = 0, m_k > 0, k = \overline{1, K-1}; \end{cases}$
5	$(\vec{m}-\vec{e}_K, n+1)$	λ_u	$\begin{cases} \sum_{k=1}^{K} m_k \cdot b_k + n + 1 > C, \\ \sum_{k=1}^{K-1} m_k = 0, m_K > 0; \end{cases}$
6	$(\vec{m}, n-1)$	$n \cdot \mu_u$	$\begin{cases} n > 0, \\ \sum_{k=2}^{K} m_k = 0; \end{cases}$
7	$(\vec{m}-\vec{e}_k+\vec{e}_{k-1}, n-1)$	$n \cdot \mu_u$	$\begin{cases} n > 0, \sum_{i=k+1}^{K} m_i = 0, m_k > 0, \\ k = \overline{2, K}; \end{cases}$

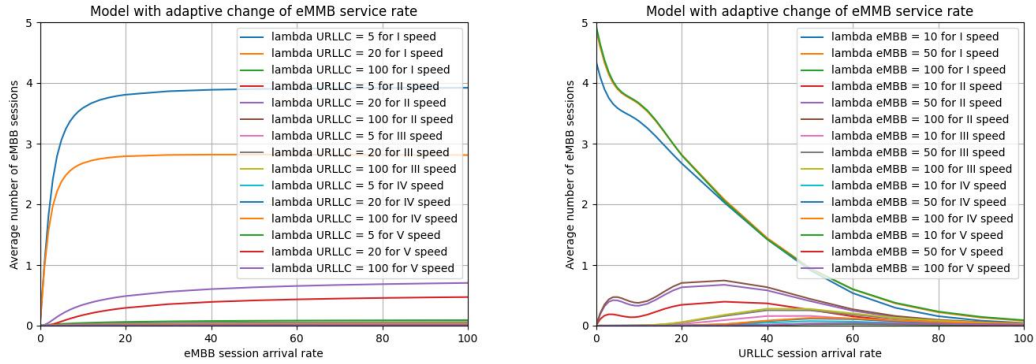
Table 3
System parameters for numerical analysis

Parameter	Scenario 1	Scenario 2
N	5	5
K	5	5
λ_m	0 - 100	10, 50, 100
λ_u	5, 20, 100	0 - 100
${\mu_m}^{-1}$	1^{-1}	1^{-1}
${\mu_u}^{-1}$	3^{-1}	3^{-1}

4. Numerical Results

4.1. Input Data

In this section, we present the results of a numerical analysis, namely the average number of eMBB and URLLC sessions and the blocking probability of eMBB/URLLC sessions. We will use two scenarios, which are presented in Table 3. The idea is that firstly λ_m is gradually increased for various constant values of λ_u, and vice versa. So we have that $N = 5$ which is the total number of the resource blocks. We have $K = 5$, which is the number of speeds, so we will have $b_1 = 5, b_2 = 4, b_3 = 3, b_4 = 2, b_5 = 1$.



(a) Scenario 1

(b) Scenario 2

Figure 4: Average number of eMBB sessions

4.2. Average Number of Sessions

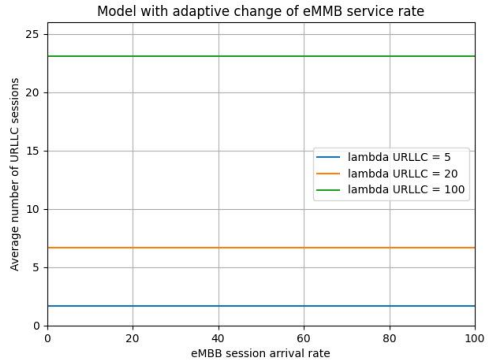
Figures 4.a-4.b show graphs of changes in the average number of eMBB requests in the system, which are served at various speeds: the first one is the highest, while the fifth one is the lowest. With a fixed arrival rate of URLLC traffic, an increase in the number of eMBB requests in the system is observed with an increase in their arrival rate. At a low arrival rate of URLLC sessions, almost all eMBB applications are served at maximum speed. When there are not enough resources to serve eMBB at maximum speed, the speed of servicing eMBB sessions begins to decrease, so the average number of sessions served at the maximum speed decreases, and the average number of sessions served at the remaining four speeds increases. This happens until there are so many URLLC requests that the system does not begin to block the connection of eMBB sessions, and those that still enter the system are served at the minimum speed.

Now consider the average number of eMBB sessions in the system for the second scenario presented. In Figure 4.b, almost the same behavior of the graph is observed, because lines overlap almost perfectly. The slight difference is explained by the fact that with a low intensity of eMBB arrival, there will be fewer requests in the system. All the graphs show a strong dependence on the increase in the intensity of incoming URLLC sessions. First, the average number of sessions decreases at maximum speed, while at others, it increases.

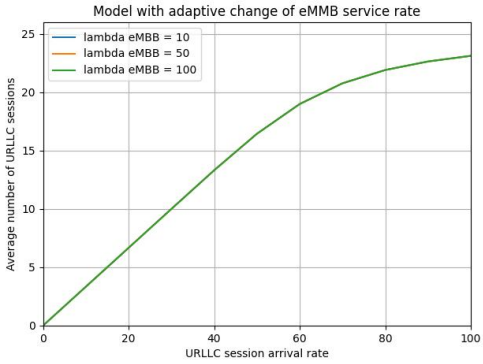
The average number of URLLC sessions in the system is shown in Figures 5.a and 5.b. Based on the presented graphs, we can conclude that this value depends only on the intensity of URLLC. This behavior is due to the priority of the URLLC traffic.

4.3. Blocking Probability

The following is the blocking probability of eMBB sessions. Obviously, the higher the intensity of incoming eMBB sessions, the greater the blocking probability of them. Figure 6.a clearly shows that with a very high rate of arrival of URLLC sessions, the rate of arrival of eMBB practically does not affect the blocking probability of eMBB sessions, which is very large and tends to be 1, since all resources are occupied by URLLC traffic. At the same time, when the

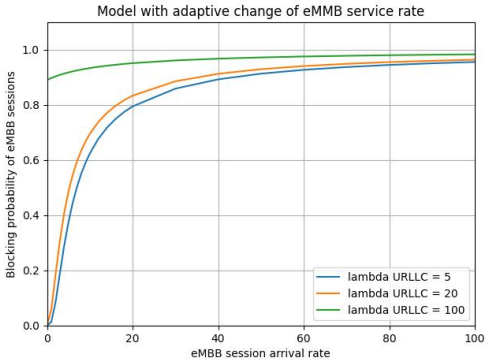


(a) Scenario 1

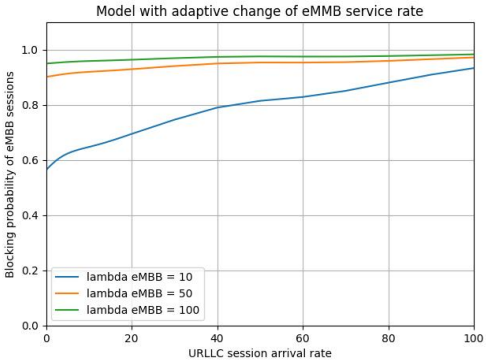


(b) Scenario 2

Figure 5: Average number of URLLC sessions.



(a) Scenario 1



(b) Scenario 2

Figure 6: Blocking probability of eMBB sessions.

intensity of URLLC sessions is low, the blocking probability of eMBB will be small since the system will cope with the load from the eMBB stream. The graph in Figure 6.b confirms the assumptions made. It can also be concluded that eMBB blocking is more influenced by the intensity of incoming URLLC sessions since the schedule behaves almost the same for the selected constant intensities of incoming eMBB traffic.

The blocking probability of URLLC sessions is affected by the rate of only this type of traffic since the system prioritizes it.

5. Conclusion

This paper is focused on a mathematical model for eMBB and URLLC coexistence in the form of a queuing system with priority service for URLLC traffic – reducing and interrupting the transmission rate of eMBB. In particular, we formulated indicators of priority access efficiency

such as the probability of reducing the transmission rate, the probability of service interruption, the average transmission rate of eMBB traffic. Moreover, the numerical results show that URLLC has a significant impact on eMBB. In the future, we will consider the model that considers the spatial location of devices generating URLLC and eMBB traffic.

Acknowledgments

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipients Alexander Chursin, Petr Kharin, and Anna Kushchazli). The work was supported by the RFBR, project 20-37-70079 (recipients Irina Kochetkova and Petr Kharin).

References

[1] I. Vision, Framework and overall objectives of the future development of imt for 2020 and beyond, International Telecommunication Union (ITU), Document, Radiocommunication Study Groups (2015).

[2] G. T. 38.913, Technical specification group radio access network; study on scenarios and requirements for next generation access technologies;(release 14), Tech. Rep. (2016).

[3] A. Manzoor, S. A. Kazmi, S. R. Pandey, C. S. Hong, Contract-based scheduling of urllc packets in incumbent embb traffic, IEEE Access 8 (2020) 167516–167526.

[4] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, C. S. Hong, embb-urllc resource slicing: A risk-sensitive approach, IEEE Communications Letters 23 (2019) 740–743.

[5] E. J. dos Santos, R. D. Souza, J. L. Rebelatto, H. Alves, Network slicing for urllc and embb with max-matching diversity channel allocation, IEEE Communications Letters 24 (2019) 658–661.

[6] T. De Cola, I. Bisio, Qos optimisation of embb services in converged 5g-satellite networks, IEEE Transactions on Vehicular Technology 69 (2020) 12098–12110.

[7] I. Gerasin, A. Krasilov, E. Khorov, Dynamic multiplexing of urllc traffic and embb traffic in an uplink using nonorthogonal multiple access, Journal of Communications Technology and Electronics 65 (2020) 750–755.

[8] E. Markova, D. Moltchanov, R. Pirmagomedov, D. Ivanova, Y. Koucheryavy, K. Samouylov, Prioritized service of urllc traffic in industrial deployments of 5g nr systems, in: International Conference on Distributed Computer and Communication Networks, Springer, 2020, pp. 497–509.

[9] E. Markoval, D. Moltchanov, R. Pirmagomedov, D. Ivanova, Y. Koucheryavy, K. Samouylov, Priority-based coexistence of embb and urllc traffic in industrial 5g nr deployments, in: 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, 2020, pp. 1–6.

[10] E. Sopin, V. Begishev, D. Moltchanov, A. Samuylov, Resource queuing system with preemptive priority for performance analysis of 5g nr systems, in: International Conference on Distributed Computer and Communication Networks, Springer, 2020, pp. 87–99.

[11] E. Makeeva, N. Polyakov, P. Kharin, I. Gudkova, Probability model for performance analysis

of joint urllc and embb transmission in 5g networks, in: Internet of things, smart spaces, and next generation networks and systems, Springer, 2019, pp. 635–648.

[12] E. Makeeva, N. Polyakov, P. Kharin, I. Gudkova, Veroyatnostnaya model' dlya analiza harakteristik sovmestnoj peredachi trafika urllc i embb v besprovodnyh setyah [probability model for perfomance analysis of joint urllc and embb transmission in 5g networks] (2020) 33–42.

[13] P. Kharin, E. Makeeva, I. Kochetkova, D. Efrosinin, S. Shorgin, Sistema massovogo obsluzhivaniya s orbitami dlya analiza sovmestnogo obsluzhivaniya trafika s malymi zaderzhkami urllc i shirokopolosnogo dostupa embb v besprovodnyh setyah pyatogo pokoleniya [retrial queuing model for analyzing joint urllc and embb transmission in 5g networks] 14 (2020) 17–24.